CENGAGE Learning

11a. ed.



Anderson
Sweeney
Williams

Estadística para negocios y economía

ESTADÍSTICA PARA NEGOCIOS Y ECONOMÍA

11a. ed.

ESTADÍSTICA PARA NEGOCIOS Y ECONOMÍA
11a. ed.

David R. Anderson
University of Cincinnati

Dennis J. Sweeney
University of Cincinnati

Thomas A. Williams
Rochester Institute of Technology

Lorena Peralta Rosales
María Elsa Ocampo Malagamba
Traductoras profesionales

Revisión técnica

María de Guadalupe Arroyo Satisteban
Academia de Matemáticas ECEE
Universidad Panamericana

Iren Castillo Saldaña
Academia de Matemáticas ECEE
Universidad Panamericana

Ignacio García Juárez
Academia de Matemáticas ECEE
Universidad Panamericana

Vinicio Pérez Fonseca
Academia de Matemáticas ECEE
Universidad Panamericana

José Cruz Ramos Báez
Academia de Matemáticas ECEE
Universidad Panamericana

Roberto Palma Pacheco
Facultad de Economía y Negocios
Universidad Anáhuac-México Norte

CENGAGE Learning™

Australia • Brasil • Corea • España • Estados Unidos • Japón • México • Reino Unido • Singapur

CENGAGE Learning™

Estadística para negocios y economía, 11a. ed.
David R. Anderson
Dennis J. Sweeney
Thomas A. Williams

Presidente de Cengage Learning Latinoamérica
Fernando Valenzuela Migoya

Director de producto y desarrollo Latinoamérica
Daniel Oti Yvonnet

Director editorial y de producción Latinoamérica
Raúl D. Zendejas Espejel

Editor senior
Javier Reyes Martínez

Coordinadora de producción editorial
Abril Vega Orozco

Editora de producción
Gloria Luz Olguín Sarmiento

Coordinador de manufactura
Rafael Pérez González

Diseño de portada
Craig Ramsdell

Imagen de portada
Getty Images/GlowImages

Composición tipográfica
Heriberto Gachúz Chávez

© D.R. 2012 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.
Corporativo Santa Fe
Av. Santa Fe, núm. 505, piso 12
Col. Cruz Manca, Santa Fe
C.P. 05349, México, D.F.
Cengage Learning™ es una marca registrada usada bajo permiso.

DERECHOS RESERVADOS. Ninguna parte de este trabajo amparado por la Ley Federal del Derecho de Autor podrá ser reproducida, transmitida, almacenada o utilizada, en cualquier forma o por cualquier medio, ya sea gráfico, electrónico o mecánico, incluyendo, pero sin limitarse a lo siguiente: fotocopiado, reproducción, escaneo, digitalización, grabación en audio, distribución en Internet, distribución en redes de información o almacenamiento y recopilación en sistemas de información, a excepción de lo permitido en el capítulo III, artículo 27 de la Ley Federal del Derecho de Autor, sin el consentimiento por escrito de la editorial.

Traducido del libro:
Statistics for Business and Economics, 11a. ed.
Publicado en inglés por
South-Western Cengage Learning
ISBN 13: 978-0-324-78324-7
ISBN 10: 0-324-78324-8

Datos para catalogación bibliográfica:
Anderson, David R., Dennis J. Sweeney, Thomas A. Williams,
Estadística para negocios y economía, 11a. ed.
ISBN-13: 978-607-481-750-8
ISBN-10: 607-481-750-2

Visite nuestro sitio en:
http://latinoamerica.cengage.com

Impreso en México
1 2 3 4 5 6 7 8 9 12 11

Dedicado a
Marcia, Cherri y Robbie

Contenido breve

Prefacio xxv

Acerca de los autores xxix

Capítulo 1 Datos y estadística 1

Capítulo 2 Estadística descriptiva: presentaciones tabulares y gráficas 31

Capítulo 3 Estadística descriptiva: medidas numéricas 85

Capítulo 4 Introducción a la probabilidad 148

Capítulo 5 Distribuciones de probabilidad discreta 193

Capítulo 6 Distribuciones de probabilidad continua 232

Capítulo 7 Muestreo y distribuciones de muestreo 265

Capítulo 8 Estimación por intervalo 308

Capítulo 9 Pruebas de hipótesis 348

Capítulo 10 Inferencia estadística acerca de medias y proporciones con dos poblaciones 406

Capítulo 11 Inferencias acerca de varianzas poblacionales 448

Capítulo 12 Pruebas de bondad de ajuste e independencia 472

Capítulo 13 Diseño de experimentos y análisis de varianza 506

Capítulo 14 Regresión lineal simple 560

Capítulo 15 Regresión múltiple 642

Capítulo 16 Análisis de regresión: construcción de modelos 712

Capítulo 17 Números índice 763

Capítulo 18 Análisis de series de tiempo y elaboración de pronósticos 784

Capítulo 19 Métodos no paramétricos 855

Capítulo 20 Métodos estadísticos para el control de la calidad 903

Capítulo 21 Análisis de decisiones 937

Capítulo 22 Sample Survey On Website

Apéndice A Referencias y bibliografía 976

Apéndice B Tablas 978

Apéndice C Notación de suma 1005

Apéndice D Soluciones a las autoevaluaciones y respuestas a los ejercicios de números pares 1007

Apéndice E Uso de las funciones de Excel 1062

Apéndice F Cálculo de los valores-*p* utilizando Minitab y Excel 1067

Índice analítico 1071

Contenido

Prefacio xxv
Acerca de los autores xxix

Capítulo 1 Datos y estadística 1

***Estadística en la práctica*. BusinessWeek 2**
1.1 Aplicaciones en negocios y economía 3
Contabilidad 3
Finanzas 4
Marketing 4
Producción 4
Economía 4
1.2 Datos 5
Elementos, variables y observaciones 5
Escalas de medición 6
Datos categóricos y cuantitativos 7
Datos de corte transversal y de series de tiempo 7
1.3 Fuentes de datos 10
Fuentes existentes 10
Estudios estadísticos 11
Errores en la adquisición de los datos 13
1.4 Estadística descriptiva 13
1.5 Inferencia estadística 15
1.6 Computadoras y análisis estadístico 17
1.7 Minería de datos 17
1.8 Lineamientos éticos para la práctica estadística 18
Resumen 20
Glosario 20
Ejercicios complementarios 21
Apéndice Una introducción a StatTools 28

Capítulo 2 Estadística descriptiva: presentaciones tabulares y gráficas 31

***Estadística en la práctica*. Colgate-Palmolive Company 32**
2.1 Resumen de datos cualitativos 33
Distribución de frecuencia 33
Distribuciones de frecuencia relativa y frecuencia porcentual 34
Gráficas de barras y circulares 34

2.2 Resumen de datos cuantitativos 39
Distribución de frecuencia 39
Distribuciones de frecuencia relativa y frecuencia porcentual 41
Diagrama de puntos 41
Histograma 41
Distribuciones acumuladas 43
Ojiva 44
2.3 Análisis de datos exploratorios: el diagrama de tallo y hoja 48
2.4 Tabulaciones cruzadas y diagramas de dispersión 53
Tabulación cruzada 53
La paradoja de Simpson 56
Diagrama de dispersión y línea de tendencia 57
Resumen 63
Glosario 64
Fórmulas clave 65
Ejercicios complementarios 65
***Caso a resolver* 1 Pelican Stores 71**
***Caso a resolver* 2 Industria del cine 72**
Apéndice 2.1 Uso de Minitab para presentaciones tabulares y gráficas 73
Apéndice 2.2 Uso de Excel para presentaciones tabulares y gráficas 75
Apéndice 2.3 Uso de StatTools para presentaciones tabulares y gráficas 84

Capítulo 3 Estadística descriptiva: medidas numéricas 85

***Estadística en la práctica*. Small Fry Design 86**
3.1 Medidas de posición o localización 87
Media 87
Mediana 88
Moda 89
Percentiles 90
Cuartiles 91
3.2 Medidas de variabilidad 95
Rango 96
Rango intercuartílico 96
Varianza 97
Desviación estándar 99
Coeficiente de variación 99
3.3 Medidas de la forma de la distribución, posición relativa y detección de observaciones atípicas 102
Forma de la distribución 102
Valor *z* 103
Teorema de Chebyshev 104
Regla empírica 105
Detección de observaciones atípicas 106

3.4 Análisis exploratorio de datos 109
Resumen de cinco números 109
Diagrama de caja 110
3.5 Medidas de asociación entre dos variables 115
Covarianza 115
Interpretación de la covarianza 117
Coeficiente de correlación 119
Interpretación del coeficiente de correlación 120
3.6 Media ponderada y trabajo con datos agrupados 124
Media ponderada 124
Datos agrupados 125
Resumen 129
Glosario 130
Fórmulas clave 131
Ejercicios complementarios 133
***Caso a resolver* 1 Pelican Stores 137**
***Caso a resolver* 2 Industria del cine 138**
***Caso a resolver* 3 Escuelas de negocios de Asia-Pacífico 139**
***Caso a resolver* 4 Transacciones del sitio web de Heavenly Chocolates 139**
Apéndice 3.1 Estadística descriptiva usando Minitab 142
Apéndice 3.2 Estadística descriptiva usando Excel 143
Apéndice 3.3 Estadística descriptiva usando StatTools 146

Capítulo 4 Introducción a la probabilidad 148

***Estadística en la práctica*. Oceanwide Seafood 149**
4.1 Experimentos, reglas de conteo y asignación de probabilidades 150
Reglas de conteo, combinaciones y permutaciones 151
Asignación de probabilidades 155
Probabilidades para el proyecto de KP&L 157
4.2 Eventos y sus probabilidades 160
4.3 Algunas relaciones básicas de probabilidad 164
Complemento de un evento 164
Ley de la adición 165
4.4 Probabilidad condicional 171
Eventos independientes 174
Ley de la multiplicación 174
4.5 Teorema de Bayes 178
Método tabular 182
Resumen 184
Glosario 184

Fórmulas clave 185
Ejercicios complementarios 186
***Caso a resolver* Jueces del condado de Hamilton 190**

Capítulo 5 Distribuciones de probabilidad discreta 193

***Estadística en la práctica*. Citibank 194**
5.1 Variables aleatorias 194
Variables aleatorias discretas 195
Variables aleatorias continuas 196
5.2 Distribuciones de probabilidad discreta 197
5.3 Valor esperado y varianza 202
Valor esperado 202
Varianza 203
5.4 Distribución de probabilidad binomial 207
Un experimento binomial 208
El problema de Martin Clothing Store 209
Uso de tablas de probabilidades binomiales 213
Valor esperado y varianza de la distribución binomial 214
5.5 Distribución de probabilidad de Poisson 218
Un ejemplo con intervalos de tiempo 218
Un ejemplo con intervalos de longitud o de distancia 220
5.6 Distribución de probabilidad hipergeométrica 221
Resumen 225
Glosario 225
Fórmulas clave 226
Ejercicios complementarios 227
Apéndice 5.1 Distribuciones de probabilidad discretas con Minitab 230
Apéndice 5.2 Distribuciones de probabilidad discretas con Excel 230

Capítulo 6 Distribuciones de probabilidad continua 232

***Estadística en la práctica*. Procter & Gamble 233**
6.1 Distribución de probabilidad uniforme 234
El área como medida de la probabilidad 235
6.2 Distribución de probabilidad normal 238
Curva normal 238
Distribución de probabilidad normal estándar 240
Cálculo de probabilidades para cualquier distribución de probabilidad normal 245
El problema de Grear Tire Company 246
6.3 Aproximación normal de las probabilidades binomiales 250
6.4 Distribución de probabilidad exponencial 253
Cálculo de probabilidades para la distribución exponencial 254
Relación entre las distribuciones de Poisson y exponencial 255

Resumen 257
Glosario 258
Fórmulas clave 258
Ejercicios complementarios 258
***Caso a resolver* Specialty Toys 261**
Apéndice 6.1 Distribuciones de probabilidad continua con Minitab 262
Apéndice 6.2 Distribuciones de probabilidad continua con Excel 263

Capítulo 7 Muestreo y distribuciones de muestreo 265

***Estadística en la práctica*. MeadWestvaco Corporation 266**
7.1 El problema de muestreo de Electronics Associates 267
7.2 Selección de una muestra 268
Muestreo de una población finita 268
Muestreo de una población infinita 270
7.3 Estimación puntual 273
Consejo práctico 275
7.4 Introducción a las distribuciones muestrales o de muestreo 276
7.5 Distribución de muestreo de $\bar{x}$ 278
Valor esperado de $\bar{x}$ 279
Desviación estándar de $\bar{x}$ 280
Forma de la distribución de muestreo de $\bar{x}$ 281
Distribución de muestreo de $\bar{x}$ en el problema de EAI 283
Valor práctico de la distribución de muestreo de $\bar{x}$ 283
Relación entre el tamaño de la muestra y la distribución de muestreo de $\bar{x}$ 285
7.6 Distribución de muestreo de $\bar{p}$ 289
Valor esperado de $\bar{p}$ 289
Desviación estándar de $\bar{p}$ 290
Forma de la distribución de muestreo de $\bar{p}$ 291
Valor práctico de la distribución de muestreo de $\bar{p}$ 291
7.7 Propiedades de los estimadores puntuales 295
Insesgadez 295
Eficiencia 296
Consistencia 297
7.8 Otros métodos de muestreo 297
Muestreo aleatorio estratificado 297
Muestreo por conglomerados 298
Muestreo sistemático 298
Muestreo de conveniencia 299
Muestreo subjetivo 299
Resumen 300
Glosario 300
Fórmulas clave 301

Ejercicios complementarios 302
Apéndice 7.1 Valor esperado y desviación estándar de $\bar{x}$ 304
Apéndice 7.2 Muestreo aleatorio con Minitab 306
Apéndice 7.3 Muestreo aleatorio con Excel 306
Apéndice 7.4 Muestreo aleatorio con StatTools 307

Capítulo 8 Estimación por intervalo 308

***Estadística en la práctica*. Food Lion 309**
8.1 Media poblacional: σ conocida 310
Margen de error y estimación por intervalo 310
Consejo práctico 314
8.2 Media poblacional: σ desconocida 316
Margen de error y estimación por intervalo 317
Consejo práctico 320
Uso de una muestra pequeña 320
Resumen de los procedimientos de estimación por intervalo 322
8.3 Determinación del tamaño de la muestra 325
8.4 Proporción poblacional 328
Determinación del tamaño de la muestra 330
Resumen 333
Glosario 334
Fórmulas clave 335
Ejercicios complementarios 335
***Caso a resolver* 1 Revista *Young Professional* 338**
***Caso a resolver* 2 Gulf Real Estate Properties 339**
***Caso a resolver* 3 Metropolitan Research, Inc. 341**
Apéndice 8.1 Estimación por intervalo con Minitab 341
Apéndice 8.2 Estimación por intervalo usando Excel 343
Apéndice 8.3 Estimación por intervalo con StatTools 346

Capítulo 9 Pruebas de hipótesis 348

***Estadística en la práctica*. John Morrell & Company 349**
9.1 Formulación de las hipótesis nula y alternativa 350
La hipótesis alternativa como hipótesis de investigación 350
La hipótesis nula como un supuesto para ser rebatido 351
Resumen de las formas para las hipótesis nula y alternativa 352
9.2 Errores tipo I y tipo II 353
9.3 Media poblacional: σ conocida 356
Prueba de una cola 356
Prueba de dos colas 362
Resumen y consejo práctico 365

Relación entre estimación por intervalo y prueba de hipótesis 366
9.4 Media poblacional: σ desconocida 370
Prueba de una cola 371
Prueba de dos colas 372
Resumen y consejo práctico 373
9.5 Proporción poblacional 376
Resumen 379
9.6 Prueba de hipótesis y toma de decisiones 381
9.7 Cálculo de la probabilidad de los errores tipo II 382
9.8 Determinación del tamaño de la muestra en una prueba de hipótesis para la media poblacional 387
Resumen 391
Glosario 392
Fórmulas clave 392
Ejercicios complementarios 393
***Caso a resolver* 1 Quality Associates, Inc. 396**
***Caso a resolver* 2 Comportamiento ético de los estudiantes de negocios en la Universidad de Bayview 397**
Apéndice 9.1 Pruebas de hipótesis con Minitab 398
Apéndice 9.2 Pruebas de hipótesis con Excel 400
Apéndice 9.3 Pruebas de hipótesis con StatTools 404

Capítulo 10 Inferencia estadística acerca de medias y proporciones con dos poblaciones 406

***Estadística en la práctica*. U.S. Food and Drug Administration 407**
10.1 Inferencias acerca de la diferencia entre dos medias poblacionales: σ_1 y σ_2 conocidas 408
Estimación por intervalo para $\mu_1 - \mu_2$ 408
Pruebas de hipótesis acerca de $\mu_1 - \mu_2$ 410
Consejo práctico 412
10.2 Inferencias acerca de la diferencia entre dos medias poblacionales: σ_1 y σ_2 desconocidas 415
Estimación por intervalo para $\mu_1 - \mu_2$ 415
Pruebas de hipótesis acerca de $\mu_1 - \mu_2$ 417
Consejo práctico 419
10.3 Inferencias acerca de la diferencia entre dos medias poblacionales: muestras pareadas 423
10.4 Inferencias acerca de la diferencia entre dos proporciones poblacionales 429
Estimación por intervalo para $p_1 - p_2$ 429
Prueba de hipótesis acerca de $p_1 - p_2$ 431
Resumen 436
Glosario 436

Fórmulas clave 437
Ejercicios complementarios 438
***Caso a resolver* Par, Inc. 441**
Apéndice 10.1 Inferencias acerca de dos poblaciones usando Minitab 442
Apéndice 10.2 Inferencias acerca de dos poblaciones usando Excel 444
Apéndice 10.3 Inferencias acerca de dos poblaciones usando StatTools 446

Capítulo 11 Inferencias acerca de varianzas poblacionales 448

***Estadística en la práctica*. U.S. Government Accountability Office 449**
11.1 Inferencias acerca de una varianza poblacional 450
Estimación por intervalo 450
Pruebas de hipótesis 454
11.2 Inferencias acerca de dos varianzas poblacionales 460
Resumen 466
Fórmulas clave 467
Ejercicios complementarios 467
***Caso a resolver* Programa de capacitación de la Fuerza Aérea 469**
Apéndice 11.1 Varianzas poblacionales con Minitab 470
Apéndice 11.2 Varianzas poblacionales con Excel 470
Apéndice 11.3 Desviación estándar poblacional simple con StatTools 471

Capítulo 12 Pruebas de bondad de ajuste e independencia 472

***Estadística en la práctica*. United Way 473**
12.1 Prueba de bondad de ajuste: una población multinomial 474
12.2 Prueba de independencia 479
12.3 Prueba de bondad de ajuste: distribuciones de Poisson y normal 487
Distribución de Poisson 487
Distribución normal 491
Resumen 496
Glosario 497
Fórmulas clave 497
Ejercicios complementarios 497
***Caso a resolver* Una agenda bipartidista para el cambio 501**
Apéndice 12.1 Pruebas de bondad de ajuste e independencia con Minitab 502
Apéndice 12.2 Pruebas de bondad de ajuste e independencia con Excel 503

Capítulo 13 Diseño de experimentos y análisis de varianza 506

***Estadística en la práctica*. Burke Marketing Services, Inc. 507**
13.1 Introducción al diseño de experimentos y al análisis de varianza 508

Recolección de datos 509
Supuestos para el análisis de varianza 510
Análisis de varianza: una perspectiva conceptual 510
13.2 Análisis de varianza y el diseño completamente aleatorizado 513
Estimación de la varianza poblacional entre tratamientos 514
Estimación de la varianza poblacional dentro de los tratamientos 515
Comparación de las estimaciones de las varianzas: la prueba *F* 516
Tabla de ANOVA 518
Resultados de computadora para el análisis de varianza 519
Prueba para la igualdad de *k* medias poblacionales: un estudio observacional 520
13.3 Procedimientos de comparación múltiple 524
LSD de Fisher 524
Tasas de error tipo I 527
13.4 Diseño de bloques aleatorizado 530
Prueba de estrés para controladores de tráfico aéreo 531
Procedimiento ANOVA 532
Cálculos y conclusiones 533
13.5 Experimento factorial 537
Procedimiento ANOVA 539
Cálculos y conclusiones 539
Resumen 544
Glosario 545
Fórmulas clave 545
Ejercicios complementarios 547
***Caso a resolver* 1 Wentworth Medical Center 552**
***Caso a resolver* 2 Compensación para profesionales de ventas 553**
Apéndice 13.1 Análisis de varianza con Minitab 554
Apéndice 13.2 Análisis de varianza con Excel 555
Apéndice 13.3 Análisis de un diseño completamente aleatorizado usando StatTools 557

Capítulo 14 Regresión lineal simple 560

***Estadística en la práctica*. Alliance Data Systems 561**
14.1 Modelo de regresión lineal simple 562
Modelo de regresión y ecuación de regresión 562
Ecuación de regresión estimada 563
14.2 Método de mínimos cuadrados 565
14.3 Coeficiente de determinación 576
Coeficiente de correlación 579
14.4 Supuestos del modelo 583
14.5 Prueba de significancia 585
Estimación de σ^2 585
Prueba *t* 586

Intervalo de confianza para β_1 587
Prueba F 588
Algunas advertencias acerca de la interpretación de las pruebas de significancia 590

14.6 Uso de la ecuación de regresión estimada para estimación y predicción 594
Estimación puntual 594
Estimación por intervalo 594
Intervalo de confianza para el valor medio de y 595
Intervalo de predicción para un solo valor de y 596

14.7 Solución por computadora 600

14.8 Análisis de residuales: confirmación de los supuestos del modelo 605
Gráfica de residuales contra x 606
Gráfica de residuales contra $\hat{y}$ 607
Residuales estandarizados 607
Gráfica de probabilidad normal 610

14.9 Análisis de residuales: observaciones atípicas y observaciones influyentes 614
Detección de observaciones atípicas 614
Detección de observaciones influyentes 616

Resumen 621

Glosario 622

Fórmulas clave 623

Ejercicios complementarios 625

***Caso a resolver* 1 Medición del riesgo en el mercado bursátil 631**

***Caso a resolver* 2 Departamento de Transporte de Estados Unidos 632**

***Caso a resolver* 3 Donaciones de exalumnos 633**

***Caso a resolver* 4 Estadísticas del PGA Tour 633**

Apéndice 14.1 Deducción de la fórmula de mínimos cuadrados basada en el cálculo 635

Apéndice 14.2 Prueba de significancia usando correlación 636

Apéndice 14.3 Análisis de regresión con Minitab 637

Apéndice 14.4 Análisis de regresión con Excel 638

Apéndice 14.5 Análisis de regresión con StatTools 640

Capítulo 15 Regresión múltiple 642

***Estadística en la práctica*. dunnhumby 643**

15.1 Modelo de regresión múltiple 644
Modelo de regresión y ecuación de regresión 644
Ecuación de regresión múltiple estimada 644

15.2 Método de mínimos cuadrados 645
Un ejemplo: Butler Trucking Company 646
Nota sobre la interpretación de los coeficientes 648

15.3 Coeficiente de determinación múltiple 654

15.4 Supuestos del modelo 657

15.5 Prueba de significancia 658
Prueba *F* 658
Prueba *t* 661
Multicolinealidad 662
15.6 Uso de la ecuación de regresión estimada para estimaciones y predicciones 665
15.7 Variables independientes cualitativas 668
Un ejemplo: Johnson Filtration, Inc. 668
Interpretación de los parámetros 670
Variables cualitativas más complejas 672
15.8 Análisis residual 676
Detección de observaciones atípicas 678
Residuales eliminados estudentizados y observaciones atípicas 678
Observaciones influyentes 679
Uso de la medida de la distancia de Cook para identificar observaciones influyentes 679
15.9 Regresión logística 683
Ecuación de regresión logística 684
Estimación de la ecuación de regresión logística 685
Prueba de significancia 687
Uso en la administración 688
Interpretación de la ecuación de regresión logística 688
Transformación logit 691
Resumen 694
Glosario 695
Fórmulas clave 696
Ejercicios complementarios 698
***Caso a resolver* 1 Consumer Research, Inc. 704**
***Caso a resolver* 2 Aportaciones de exalumnos 705**
***Caso a resolver* 3 Estadísticas del PGA Tour 705**
***Caso a resolver* 4 Predicción del porcentaje de triunfos de la NFL 708**
Apéndice 15.1 Regresión múltiple con Minitab 708
Apéndice 15.2 Regresión múltiple con Excel 709
Apéndice 15.3 Regresión logística con Minitab 710
Apéndice 15.4 Análisis de regresión múltiple con StatTools 711

Capítulo 16 Análisis de regresión: construcción de modelos 712

***Estadística en la práctica*. Monsanto Company 713**
16.1 Modelo lineal general 714
Modelado de relaciones curvilíneas 714
Interacción 718

Transformaciones que involucran la variable dependiente 720
Modelos no lineales que son intrínsecamente lineales 724

16.2 Determinación de cuándo agregar o eliminar variables 729
Caso general 730
Uso de los valores-*p* 732

16.3 Análisis de un problema mayor 735

16.4 Procedimientos de selección de variables 739
Regresión por pasos 739
Selección hacia adelante 740
Eliminación hacia atrás 741
Regresión de los mejores subconjuntos 741
La elección final 742

16.5 Método de regresión múltiple para el diseño de experimentos 745

16.6 Autocorrelación y la prueba de Durbin-Watson 750

Resumen 754

Glosario 754

Fórmulas clave 754

Ejercicios complementarios 755

***Caso a resolver* 1 Análisis de las estadísticas de la PGA Tour 758**

***Caso a resolver* 2 Rendimiento de combustible en los automóviles 759**

Apéndice 16.1 Procedimientos de selección de variables con Minitab 760

Apéndice 16.2 Procedimientos de selección de variables con StatTools 761

Capítulo 17 Números índice 763

***Estadística en la práctica*. Oficina de Estadísticas Laborales, Departamento del Trabajo de Estados Unidos 764**

17.1 Precios relativos 765

17.2 Índices de precios agregados 765

17.3 Cálculo del índice de precios agregado a partir de los precios relativos 769

17.4 Algunos índices de precios importantes 771
Índice de precios al consumidor 771
Índice de precios al productor 771
Promedios Dow Jones 772

17.5 Deflactación de una serie mediante índices de precios 773

17.6 Índices de precios: otras consideraciones 777
Selección de artículos 777
Selección de un periodo base 777
Variaciones en la calidad 777

17.7 Índices de cantidad 778

Resumen 780
Glosario 780
Fórmulas clave 780
Ejercicios complementarios 781

Capítulo 18 Análisis de series de tiempo y elaboración de pronósticos 784

***Estadística en la práctica*. Nevada Occupational Health Clinic 785**

18.1 Patrones de una serie de tiempo 786
Patrón horizontal 786
Patrón de tendencia 788
Patrón estacional 788
Patrones de tendencia y estacional 789
Patrón cíclico 789
Selección de un método de elaboración de pronósticos 791

18.2 Exactitud del pronóstico 792

18.3 Promedios móviles y suavizamiento exponencial 797
Promedios móviles 797
Promedios móviles ponderados 800
Suavizamiento exponencial 800

18.4 Proyección de la tendencia 807
Regresión de tendencia lineal 807
Suavizamiento exponencial lineal de Holt 812
Regresión de tendencia no lineal 814

18.5 Estacionalidad y tendencia 820
Estacionalidad sin tendencia 820
Estacionalidad y tendencia 823
Modelos basados en datos mensuales 825

18.6 Descomposición de series de tiempo 829
Cálculo de los índices estacionales 830
Desestacionalización de una serie de tiempo 834
Uso de una serie de tiempo desestacionalizada para identificar tendencias 834
Ajustes estacionales 836
Modelos basados en datos mensuales 837
Patrón cíclico 837

Resumen 839
Glosario 840
Fórmulas clave 841
Ejercicios complementarios 842
***Caso a resolver* 1 Pronóstico de ventas de alimentos y bebidas 846**
***Caso a resolver* 2 Elaboración del pronóstico de pérdidas de ventas 847**
Apéndice 18.1 Elaboración de pronósticos con Minitab 848
Apéndice 18.2 Elaboración de pronósticos con Excel 851
Apéndice 18.3 Elaboración de pronósticos con StatTools 852

Capítulo 19 Métodos no paramétricos 855

***Estadística en la práctica*. West Shell Realtors 856**

19.1 La prueba de signos 857
Prueba de hipótesis acerca de una mediana poblacional 857
Prueba de hipótesis con muestras pareadas 862

19.2 Prueba de rangos con signo de Wilcoxon 865

19.3 Prueba de Mann-Whitney-Wilcoxon 871

19.4 Prueba de Kruskal-Wallis 882

19.5 Correlación de rangos 887

Resumen 891

Glosario 892

Fórmulas clave 893

Ejercicios complementarios 893

Apéndice 19.1 Métodos no paramétricos con Minitab 896

Apéndice 19.2 Métodos no paramétricos con Excel 899

Apéndice 19.3 Métodos no paramétricos con StatTools 901

Capítulo 20 Métodos estadísticos para el control de la calidad 903

***Estadística en la práctica*. Dow Chemical Company 904**

20.1 Filosofías y marcos de referencia 905
El Malcolm Baldrige National Quality Award 906
ISO 9000 906
Six Sigma 906

20.2 Control estadístico de procesos 908
Gráficas de control 909
Carta $\bar{x}$: media y desviación estándar del proceso conocidas 910
Carta $\bar{x}$: media y desviación estándar del proceso desconocidas 912
Gráfica *R* 915
Gráfica *p* 917
Gráfica *np* 919
Interpretación de las gráficas de control 920

20.3 Muestreo de aceptación 922
KALI, Inc.: Uun ejemplo de muestreo de aceptación 924
Cálculo de la probabilidad de aceptación de un lote 924
Selección de un plan de muestreo de aceptación 928
Planes de muestreo múltiple 930

Resumen 931

Glosario 931

Fórmulas clave 932

Ejercicios complementarios 933

Apéndice 20.1 Gráficas de control con Minitab 935

Apéndice 20.2 Gráficas de control utilizando StatTools 935

Capítulo 21 Análisis de decisiones 937

***Estadística en la práctica*. Ohio Edison Company 938**

21.1 Formulación del problema 939
Tablas de pagos 940
Árboles de decisión 940

21.2 Toma de decisiones con probabilidades 941
Método del valor esperado 941
Valor esperado de la información perfecta 943

21.3 Análisis de decisiones con información muestral 949
Árbol de decisión 950
Estrategia de decisión 951
Valor esperado de la información muestral 954

21.4 Cálculo de probabilidades mediante el teorema de Bayes 960

Resumen 964

Glosario 965

Fórmulas clave 966

Ejercicios complementarios 966

***Caso a resolver* Estrategia de defensa en una demanda 969**

Apéndice Introducción a PrecisionTree 970

Apéndice A Referencias y bibliografía 976

Apéndice B Tablas 978

Apéndice C Notación de la suma 1005

Apéndice D Soluciones a las autoevaluaciones y respuestas a los ejercicios de números pares 1007

Apéndice E Uso de las funciones de Excel 1062

Apéndice F Cálculo de los valores-*p* utilizando Minitab y Excel 1067

Índice analítico 1071

Prefacio

El propósito de *Estadística para negocios y economía* es proporcionar a los estudiantes, principalmente a quienes se preparan en las áreas de negocios y economía, una introducción conceptual al campo de la estadística. Su orientación se dirige a las aplicaciones y fue escrito tomando en cuenta las necesidades de los lectores que no cuentan con sólidos conocimientos de matemáticas; el requisito matemático para entenderlo es el conocimiento del álgebra.

Las aplicaciones del análisis de datos y la metodología estadística son parte integral de la organización y presentación del material del libro. El análisis y desarrollo de cada técnica se presentan en el escenario de una aplicación, cuyos resultados estadísticos permiten comprender las decisiones y las soluciones de los problemas.

Aunque el libro está orientado a las aplicaciones, se ha tenido cuidado de proporcionar un desarrollo metodológico sólido y de usar la notación convencional aceptada para el tema que se estudia. Por consiguiente, el lector encontrará que el texto proporciona una buena preparación para el estudio de material estadístico más avanzado. En el apéndice se incluye una bibliografía que servirá como guía para profundizar en el estudio de estos temas.

El libro introduce al estudiante a Minitab 15 y Microsoft® Office Excel 2007, y recalca el papel del software en la aplicación del análisis estadístico. Minitab fue incluido por constituir uno de los principales paquetes de software tanto en la enseñanza como en la práctica estadística. Excel no es un paquete para estadística, pero debido a su amplia disponibilidad y a su uso extendido, es importante que el lector comprenda las funciones estadísticas con que cuenta. Los procedimientos de Minitab y Excel se proporcionan en los apéndices del libro, de manera que los profesores tienen la flexibilidad de hacer tanto énfasis en la computadora como lo deseen para este curso.

Cambios en la 11a. ed.

Agradecemos la aceptación y respuesta positiva a las ediciones anteriores de este libro. Así, al hacer modificaciones para esta nueva edición, hemos mantenido el estilo de presentación y la legibilidad de las versiones previas. Los cambios significativos se resumen a continuación.

Revisiones del contenido

- **Actualización del capítulo 18 *Análisis de series de tiempo y elaboración de pronósticos*.** El capítulo fue reescrito por completo considerando el uso de patrones en una gráfica de serie de tiempo para seleccionar un método de elaboración de pronósticos apropiado. Comenzamos con una nueva sección 18.1 sobre los patrones de series de tiempo, seguida por la sección 18.2 sobre los métodos para medir la exactitud del pronóstico. La sección 18.3 estudia los promedios móviles y el suavizamiento exponencial, mientras que la 18.4 introduce métodos apropiados para una serie de tiempo que exhibe una tendencia. Aquí se ilustra cómo se usan el análisis de regresión y el suavizamiento exponencial lineal para la proyección de tendencias lineales, cómo se usa el análisis de regresión para modelar relaciones no lineales que involucran una tendencia cuadrática y un crecimiento exponencial. La sección 18.5 muestra, por tanto, cómo se usan las variables ficticias para modelar la estacionalidad en una ecuación de elaboración de pronósticos. La sección 18.6, por último, estudia la descomposición de las series de tiempo clásicas, incluyendo el concepto de desestacionalización de una serie de tiempo. Se incluye un apéndice nuevo sobre elaboración de pronósticos que usa el complemento StatTools de Excel y la mayoría de los ejercicios son nuevos o fueron actualizados.
- **Actualización del capítulo 19 *Métodos no paramétricos*.** El estudio de los métodos no paramétricos fue revisado y actualizado. Contrastamos cada método no paramétrico

con su contraparte paramétrica y explicamos que se requieren algunos supuestos para el procedimiento del primero. La prueba de signos subraya la importancia de la prueba para una mediana poblacional en las poblaciones con sesgo, donde la mediana a menudo es la medida preferida de ubicación central. La prueba de suma de rangos de Wilcoxon se usa para las pruebas de muestras relacionadas y para las pruebas sobre una mediana poblacional simétrica. Una nueva aplicación de muestras pequeñas de la prueba Mann-Whitney-Wilcoxon presenta la distribución de muestreo exacta del estadístico de prueba y se utiliza para explicar por qué la suma de rangos con signo es útil para probar la hipótesis de que dos poblaciones son idénticas. El capítulo concluye con la correlación de pruebas y rangos. Los nuevos apéndices incorporados describen cómo se usan Minitab, Excel y StatTools para implementar métodos no paramétricos. Ahora hay 27 bases de datos disponibles para facilitar la solución de los ejercicios por computadora.

- **Complemento StatTools para Excel.** Excel 2007 no contiene suficientes funciones estadísticas o herramientas de análisis de datos para realizar todos los procedimientos estudiados en el libro. StatTools es un complemento (*add-in*) comercial de Excel 2007, desarrollado por Palisades Corporation, que amplía la variedad de opciones estadísticas para los usuarios. En un apéndice del capítulo 1 se muestra cómo descargar e instalar StatTools, y la mayoría de los capítulos incluye un apéndice que muestra los pasos requeridos para realizar un procedimiento estadístico usando este complemento.

  Hemos sido muy cuidadosos en presentar el uso de StatTools de manera completamente opcional, de modo que los profesores que deseen impartir sus clases usando las herramientas estándar disponibles en Excel 2007 puedan seguir haciéndolo. Los usuarios que quieran otras capacidades estadísticas que no estén disponibles en Excel 2007 estándar ahora cuentan con acceso a un complemento de estadística estándar de la industria que los estudiantes podrán seguir usando en su lugar de trabajo.
- **Cambios en la terminología de los datos.** En la edición anterior los datos nominales y ordinales se clasificaron como cualitativos, y los datos de intervalos y de proporciones como cuantitativos. En esta edición, los datos nominales y ordinales se conocen como *datos categóricos*, los cuales utilizan etiquetas o nombres para identificar las categorías de elementos parecidos. Por tanto, creemos que el término *categórico* describe mejor este tipo de datos.
- **Introducción a la minería de datos.** Una sección nueva en el capítulo 1 introduce el campo relativamente nuevo de la minería de datos. Proporcionamos una breve descripción del tema y del concepto de almacén de datos. También se describe cómo se combinan los campos de la estadística y la ciencia de la computación para hacer la minería de datos operativa y valiosa.
- **Aspectos éticos en estadística.** Otra sección nueva en el capítulo 1 proporciona un análisis de los aspectos éticos cuando se presenta e interpreta información estadística.
- **Apéndice de Excel actualizado para la estadística descriptiva de tablas y gráficas.** El apéndice de Excel del capítulo 2 muestra cómo se usan las herramientas Chart Tools, PivotTable Report y PivotChart Report para mejorar las capacidades de exhibir estadística descriptiva en tablas y gráficas.
- **Análisis comparativo con diagramas de caja.** El tratamiento de diagramas de caja del capítulo 2 se ha ampliado para incluir comparaciones relativamente rápidas y fáciles de dos o más bases de datos. Los datos típicos de sueldos iniciales para las principales asignaturas de contabilidad, finanzas, administración y marketing se usan para ilustrar comparaciones de diagramas de caja multigrupo.
- **Material de muestreo actualizado.** La introducción del capítulo 7 fue actualizada y ahora incluye los conceptos de una población muestreada y un marco. Se ha esclarecido la diferencia entre el muestreo de una población finita y de una población infinita con el muestreo de un proceso usado para ilustrar la selección de una muestra aleatoria de una población infinita. Una sección de consejos prácticos recalca la importancia de obtener una correspondencia entre la población muestreada y la población objetivo.
- **Introducción actualizada de las pruebas de hipótesis.** La sección 9.1, "Formulación de las hipótesis nula y alternativa", fue actualizada y se desarrolló una base de lineamientos más adecuada para identificar ambas hipótesis. El contexto de la situación y el propósito de tomar la muestra son fundamentales. En situaciones donde la atención se

centra en encontrar evidencia para apoyar un hallazgo de investigación, la hipótesis de estudio es la hipótesis alternativa. En situaciones donde la atención se centra en cuestionar un supuesto, éste constituye la hipótesis nula.

- **Nuevo software PrecisionTree para el análisis de decisiones.** PrecisionTree es otro complemento de Excel desarrollado por Palisades Corporation, muy útil en el análisis de decisiones. El capítulo 21 contiene un apéndice nuevo que muestra cómo usarlo.
- **Nuevos casos resueltos.** Se incluyen cinco casos resueltos nuevos en esta edición, con lo cual su número total aumenta a 31. En el capítulo 3 se incluye uno sobre estadística descriptiva y en el capítulo 9 otro sobre la prueba de hipótesis. Tres casos resueltos nuevos se han agregado a la regresión en los capítulos 14, 15 y 16. Éstos proporcionan al lector la oportunidad de analizar bases de datos más grandes y preparar informes gerenciales con base en los resultados del análisis.
- **Actualización de la sección *Estadística en la práctica*.** Cada capítulo comienza con este recuadro que describe una aplicación de la metodología estadística estudiada en el mismo. Una novedad en esta edición son los artículos de Oceanwide Seafood, en el capítulo 4, y de la empresa de servicios de marketing dunnhumby, con sede en Londres, en el capítulo 15.
- **Ejemplos y ejercicios nuevos basados en datos reales.** Seguimos haciendo un esfuerzo significativo para actualizar nuestros ejemplos y ejercicios con datos reales y las fuentes de información estadística referidas más actuales. En esta edición hemos añadido aproximadamente 150 ejemplos y ejercicios nuevos basados en datos reales y en acreditadas fuentes. Con los datos de fuentes utilizadas también por *The Wall Street Journal, USA Today*, Barran's y otros, que hemos extraído de estudios reales, desarrollamos explicaciones y creamos ejercicios que muestran los diversos usos de la estadística en los negocios y la economía. Pensamos que al usar datos reales, los lectores se interesarán más en el material y podrán aprender tanto sobre la metodología estadística como sobre sus aplicaciones. Esta edición contiene más de 350 ejemplos y ejercicios basados en este tipo de información.

Características y pedagogía

Los autores han conservado muchas de las características que se presentaron en ediciones anteriores. Las más importantes se describen a continuación.

Ejercicios de métodos y de aplicaciones

Los ejercicios al final de cada sección se dividen en dos partes: métodos y aplicaciones. Los de métodos requieren que el estudiante use las fórmulas y realice los cálculos necesarios, y los de aplicaciones requieren que use el material del capítulo en situaciones reales. Por tanto, el estudiante primero se concentra en lo "esencial" de la computación y luego pasa a las sutilezas de la aplicación estadística y su interpretación.

Ejercicios de autoevaluación

Ciertos ejercicios se identifican como "Autoevaluación". Las soluciones completamente desarrolladas de estos ejercicios se incluyen en el apéndice D del libro. El estudiante puede intentar resolverlos y comprobar inmediatamente después la solución para evaluar su comprensión de los conceptos presentados en el capítulo.

Anotaciones al margen, notas y comentarios

Las anotaciones al margen que resaltan los puntos clave y proporcionan información adicional para el lector son una característica fundamental del libro. Estas anotaciones fueron diseñadas para resaltar y mejorar la comprensión de los términos y conceptos que se presentan en el texto.

Al final de cada sección se incluye un recuadro de Notas y comentarios, diseñado para proporcionar al lector información adicional que le permita comprender la metodología estadística y sus aplicaciones. Estas Notas y comentarios contienen advertencias o acotaciones de la metodología, recomendaciones para su aplicación, una breve descripción de consideraciones técnicas e información complementaria.

Archivos de datos que acompañan el libro

Más de 200 archivos de datos se incluyen en el sitio web del libro. Las bases de datos están disponibles tanto en formato de Minitab como en Excel. En el texto se usan iconos para identificarlas. Se incluyen bases de datos para todos los problemas resueltos, así como para los ejercicios grandes.

Agradecimientos

Un agradecimiento especial a Jeffrey D. Camm, de la University of Cincinnati, y a James J. Cochran, de Louisiana Tech University, por sus contribuciones a esta edición. Los profesores Camm y Cochran hicieron una gran aportación a los capítulos nuevos sobre elaboración de pronósticos y métodos no paramétricos. Además, contribuyeron con comentarios y sugerencias útiles para los casos a resolver, los ejercicios y artículos nuevos para *Estadística en la práctica*. También agradecemos a nuestros socios de empresas y de la industria que proporcionaron los artículos de *Estadística en la práctica*. Los reconocemos de manera individual en los créditos de cada uno de los artículos. Por último, también estamos en deuda con nuestro editor de Adquisiciones, Charles McCormick, Jr.; nuestra editora de Desarrollo, Maggie Kubale; nuestra gerente de Proyecto de contenido, Jacquelyn K. Featherly; nuestro gerente de Marketing, Bryant Chrzan, y otras personas que laboran en Cengage Learning por sus consejos y apoyo durante la preparación de este libro.

David R. Anderson
Dennis J. Sweeney
Thomas A. Williams

Acerca de los autores

David R. Anderson. Es profesor de análisis cuantitativo en el Colegio de Administración de Empresas de la University of Cincinnati. Nació en Grand Forks, Dakota del Norte, y obtuvo su licenciatura, maestría y doctorado en Purdue University. Ha colaborado como director del Departamento de Análisis Cuantitativo y Administración de Operaciones y como decano principal del Colegio de Administración de Empresas de la Universidad de Cincinnati. Además, fue coordinador del primer programa para ejecutivos del colegio.

En la Universidad de Cincinnati ha impartido la materia de introducción a la estadística a estudiantes de administración de empresas, así como cursos de posgrado sobre análisis de regresión, análisis multivariado y ciencias de la administración. También ha impartido cursos de estadística en el Departamento del Trabajo de Washington, D.C. Ha sido distinguido con nominaciones y premios de excelencia en la enseñanza al servicio de organizaciones estudiantiles.

Es coautor de 10 libros en las áreas de estadística, ciencias de la administración, programación lineal y administración de la producción y las operaciones. Es consultor activo en los campos del muestreo y los métodos estadísticos.

Dennis J. Sweeney. Es profesor de análisis cuantitativo y fundador del Centro de Mejoramiento de la Productividad en la University of Cincinnati. Nació en Des Moines, Iowa, y obtuvo su licenciatura en Drake University, y la maestría y doctorado en la Indiana University, donde le otorgaron una beca de investigación NDEA. Durante el periodo de 1978 a 1979 colaboró en el grupo de ciencias de la administración de Procter & Gamble, y de 1981 a 1982 fue profesor invitado de Duke University. También ocupó los puestos de director del Departamento de Análisis Cuantitativo y decano adjunto del Colegio de Administración de Empresas de la Universidad de Cincinnati.

Ha publicado más de 30 artículos y monografías en el área de ciencias de la administración y estadística. La National Science Foundation, IBM, Procter & Gamble, Federated Department Stores, Kroger y Cincinnati Gas & Electric han financiado sus trabajos de investigación, mismos que han sido publicados en *Management Science, Operations Research, Mathematical Programming, Decision Sciences* y otras revistas.

El profesor Sweeney es coautor de 10 libros en las áreas de estadística, ciencias de la administración, programación lineal y administración de la producción y las operaciones.

Thomas A. Williams. Es profesor de ciencias de la administración en el College of Business del Rochester Institute of Technology (RIT). Es originario de Elmira, Nueva York, y obtuvo su licenciatura en Clarkson University. Realizó sus estudios de posgrado en el Rensselaer Polytechnic Institute, donde obtuvo su maestría y doctorado.

Antes de integrarse al College of Business del RIT, el profesor Williams fue miembro del personal docente del Colegio de Administración de Empresas de la Universidad de Cincinnati durante siete años, donde desarrolló el programa de licenciatura en sistemas de información que más tarde coordinó. En el Rensselaer Polytechnic Institute fue el primer director del Departamento de Ciencias de las Decisiones. Imparte cursos sobre ciencias de la administración y estadística, así como cursos de posgrado sobre análisis de regresión y de decisiones.

Es coautor de 11 libros en las áreas de ciencias de la administración, estadística, administración de la producción y las operaciones, y matemáticas. Ha sido consultor de numerosas empresas *Fortune 500* y ha colaborado en proyectos que varían del uso de análisis de datos al desarrollo de modelos de regresión a gran escala.

Agradecimiento especial

Cengage Learning agradece de manera muy especial a los siguientes profesores e instituciones su invaluable apoyo y profesionalismo en el desempeño y éxito de esta obra en el mercado.

Centro Universitario de Ciencias Económico Administrativas
Universidad de Guadalajara

Angélica Beatriz Contreras Cuevas
Arturo Rafael Velázquez Patiño
Jorge Alberto Gutiérrez Limón
Héctor Arturo Caramon Loyo
Ricardo Solórzano Gutiérrez
José de Jesús Ponce García
Pedro Luis Celso Arellano
Héctor Luis del Toro Chávez
Jaime Bernardo Novoa Rojas
Salvador Sandoval Bravo
Juan Manuel Rodríguez Alfaro
Víctor Hugo Gualajara Estrada
Ramona Esmeralda Velázquez García
Juan Francisco Mejía García
Martín de la Cruz Casillas Romero
Jorge Martínez Olvera
Mario Alberto Naranjo González
Cornelio Cano Guzmán
José Antonio Domínguez González
María Bernardett Ochoa Hernández
Heriberto de Jesús Domínguez Rodríguez
Manuel Llontop Pisfil

Universidad del Valle de México-Campus Zapopan

Abel Vázquez Pérez
Laura Verónica Mendoza Sánchez
Irene Isabel Navarro González

Universidad del Valle de México-Campus Sur

Francisco Muñoz Zepeda
Hugo Alejandro Zavala García
Eduardo Jacobo Arroyo
Édgar Silva

Tecnológico de Monterrey. Campus Guadalajara

Margarita Orozco Gómez
Araceli Zavala Martínez
Cosme Zepeda Alatorre
Jorge Alberto Chávez Luna
Juan Ricardo Buenrostro Silva
María Luisa Olascoaga Cortina
María Guadalupe Lomelí Plascencia
Juan Francisco Corona Burgueño

Universidad Panamericana. Campus Guadalajara

Jesús Fernández Morán

Universidad Enrique Díaz de León

Miriam Camargo
Vladimir Ilich Campanelli
Rafael López Garibay
Giovanni Osvaldo Birueth

Universidad Tecnológica de Guadalajara

Manuel Cruz Serrano
Paulino Javier Domínguez Chávez
Ada Rocío Gallardo Enríquez
Emilio Delgado Ornelas
Francisco Carbajal Ramos
Eduardo Mejía González

Instituto Tecnológico de Lázaro Cárdenas

Ramón Mejía Rivera

Universidad Latina de América

Humberto Quintero Lizaola
Jaime Casiano Macías

Universidad Lasalle

Noé Sánchez Flores

Universidad del Valle de Atemajac

Miriam Sánchez Carmona
Silvia Martínez de León
Alejandro Ángeles Espino
Mónica del Carmen Juárez Valenzuela
Carmen Yolanda Álvarez Caballero
Giselle Andrade Hernández
María de los Ángeles Reyes Bañuelos
Leopoldo Cárdenas González
Juan Josué Morales Acosta
Ignacio Navarro Ruiz

Instituto Tecnológico de Estudios Superiores de Occidente

Sergio G. Mañón Espino
José Expectación Vázquez Arévalo

CAPÍTULO 1

Datos y estadística

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: *BUSINESSWEEK*

1.1 APLICACIONES EN NEGOCIOS Y ECONOMÍA
- Contabilidad
- Finanzas
- Marketing
- Producción
- Economía

1.2 DATOS
- Elementos, variables y observaciones
- Escalas de medición
- Datos categóricos y cuantitativos
- Datos de corte transversal y de series de tiempo

1.3 FUENTES DE DATOS
- Fuentes existentes
- Estudios estadísticos
- Errores en la adquisición de datos

1.4 ESTADÍSTICA DESCRIPTIVA

1.5 INFERENCIA ESTADÍSTICA

1.6 COMPUTADORAS Y ANÁLISIS ESTADÍSTICO

1.7 MINERÍA DE DATOS

1.8 LINEAMIENTOS ÉTICOS PARA LA PRÁCTICA ESTADÍSTICA

ESTADÍSTICA *en* LA PRÁCTICA

*BUSINESSWEEK**

NEW YORK, NEW YORK

Con una circulación global de más de 1 millón de ejemplares, *BusinessWeek* es la revista de negocios más leída en el mundo. Más de 200 reporteros y editores dedicados en 26 oficinas de todo el mundo producen una variedad de artículos de interés para la comunidad de los negocios y la economía. Además de reportajes especiales sobre temas de actualidad, la revista contiene secciones regulares sobre administración internacional, análisis económico, procesamiento de información y ciencia y tecnología. La información contenida en los reportajes especiales y las secciones regulares ayuda a los lectores a mantenerse al día en los desarrollos actuales y evalúa su impacto en los negocios y la economía bajo las condiciones actuales.

La mayoría de los números de *BusinessWeek* contiene un artículo de fondo sobre un tema de interés actual. Dichos artículos a menudo contienen hechos y resúmenes estadísticos que ayudan al lector a comprender la información de negocios y economía. Por ejemplo, el número del 23 de febrero de 2009 contenía un reportaje especial sobre la crisis hipotecaria; el número del 17 de marzo de 2009 incluía un análisis de cuándo comenzaría la recuperación del mercado de valores, y el número del 4 de mayo de 2009 tenía un reportaje especial sobre cómo hacer los recortes salariales menos dolorosos. Además, el suplemento semanal *BusinessWeek Investor* proporciona estadísticas sobre el estado de la economía, que incluyen índices de producción, precios de las acciones, fondos de inversión y tasas de interés.

BusinessWeek también utiliza información estadística en la administración de su propia empresa. Por ejemplo, una encuesta anual aplicada a los suscriptores permite a la empresa obtener sus datos demográficos, hábitos de lectura, compras probables, su estilo de vida, etc. Los directivos de *BusinessWeek* utilizan resúmenes estadísticos de la consulta para brindar un mejor servicio a sus suscriptores y anunciantes. Una encuesta reciente entre los estadounidenses reveló que 90% de los suscriptores de *BusinessWeek* utiliza una computadora personal en su hogar, y que 64% hace compras por computadora en el trabajo. Estas estadísticas alertaron a los directivos de la revista sobre el interés de los suscriptores en los nuevos avances en computación. Los resultados de la encuesta también se pusieron a disposición de los posibles anunciantes. El alto porcentaje de suscriptores que usan computadoras personales en el hogar y de los que realizan compras por Internet en su trabajo son un incentivo para que un fabricante de estos equipos considere anunciarse en *BusinessWeek*.



BusinessWeek utiliza hechos estadísticos y resúmenes en muchos de sus artículos. © Terri Miller/E-Visual Communications, Inc.

En este capítulo se estudian los tipos de datos de que se dispone para el análisis estadístico y se describe cómo se obtienen los datos. La estadística descriptiva y la inferencia estadística se presentan como medios para convertir los datos en información fácil de interpretar.

* Los autores agradecen a Charlene Trentham, gerente de investigación de *BusinessWeek*, por proporcionar este artículo para la sección Estadística en la práctica.

Es frecuente ver en los periódicos y las revistas las frases siguientes:

- La National Association of Realtors informó que el precio medio que pagan los compradores primerizos por una vivienda es de $165 000 (*The Wall Street Journal,* 11 de febrero de 2009).
- El presidente de la NCAA, Myles Brand, informó que los deportistas colegiales están obteniendo su título de licenciatura a tasas récord. Las cifras más recientes muestran que 79% de los estudiantes deportistas hombres y mujeres se gradúa (Associated Press, 15 de octubre de 2008).
- El tiempo medio que tarda una persona en llegar a su trabajo es de 25.3 minutos (U.S. Census Bureau, marzo de 2009).

- Un alto valor de 11% de las viviendas estadounidenses están vacías, un exceso creado por el auge habitacional y el colapso subsiguiente (*USA Today,* 13 de febrero de 2009).
- El precio medio nacional de la gasolina regular alcanzó los $4.00 por galón por primera vez en la historia (sitio web de Cable News Network, 8 de junio de 2008).
- Los Yankees de Nueva York perciben los sueldos más altos en las grandes ligas de beisbol. La nómina total es de $201 449 289, con un sueldo medio de $5 000 000 (*USA Today Salary Data Base,* abril de 2009).
- El promedio industrial Dow Jones cerró en 8 721 puntos (*The Wall Street Journal,* 2 de junio de 2009).

Los datos numéricos en las frases anteriores ($165 000, 79%, 25.3, 11%, $4.00, $201 449 289, $5 000 000 y 8 721) se llaman *estadísticas*. En este sentido, el término *estadística* se refiere a datos numéricos como promedios, medias, porcentajes e índices que nos ayudan a entender una variedad de situaciones de los negocios y la economía. Sin embargo, como verá más adelante, el campo, o materia, de la estadística abarca mucho más que los datos numéricos. En un sentido más amplio, la **estadística** se define como el arte y la ciencia de recolectar, analizar e interpretar datos. En particular en los negocios y la economía, la información que se obtiene a partir de la recolección, el análisis, la presentación y la interpretación de los datos permite a los administradores o gerentes y a quienes toman decisiones comprender mejor el entorno económico y de los negocios, y por tanto asumir mejores y más informadas decisiones. En este libro se enfatiza el uso de la estadística para la toma de decisiones en ambos ámbitos.

El capítulo 1 comienza con algunos ejemplos de aplicaciones de la estadística a los negocios y la economía. En la sección 1.2 se define el término *dato* y se introduce el concepto de banco de datos. Esta sección también presenta términos clave como *variables* y *observaciones*; estudia la diferencia entre datos cuantitativos y categóricos, e ilustra los usos de los datos de corte transversal y de series de tiempo. En la sección 1.3 se analiza cómo se obtienen los datos de fuentes existentes o por medio de estudios experimentales diseñados para obtener datos nuevos. El papel importante que Internet juega hoy día en la obtención de datos también se pone de relieve. Los usos de los datos en el desarrollo de la estadística descriptiva y la elaboración de inferencias estadísticas se describen en las secciones 1.4 y 1.5. Las últimas tres secciones tratan sobre la función de la computadora en el análisis estadístico, y presentan una introducción al campo relativamente nuevo de la minería de datos y un análisis de las pautas éticas para la práctica estadística. Al final del capítulo se incluye un apéndice con una introducción al complemento StatTools que se usa para ampliar las opciones estadísticas a los usuarios de Microsoft Excel.

1.1 Aplicaciones en negocios y economía

En el entorno global de los negocios y la economía de hoy, cualquiera tiene acceso a una vasta cantidad de información estadística. Los gerentes y líderes de decisiones más exitosos comprenden la información y saben cómo usarla de manera eficiente. En esta sección se proporcionan ejemplos que ilustran algunos usos de la estadística en los negocios y la economía.

Contabilidad

Las firmas contables públicas utilizan procedimientos de muestreo estadístico cuando realizan auditorías para sus clientes. Por ejemplo, suponga que una firma contable quiere determinar si el estado de cuenta de un cliente representa de manera precisa el monto real de las cuentas por cobrar. La gran cantidad de cuentas por cobrar individuales hace que la revisión y la validación de cada cuenta consuman demasiado tiempo y dinero. Como práctica común en este tipo de situaciones, el personal de auditoría selecciona un subconjunto de las cuentas llamado *muestra*. Después de revisar la precisión de la selección muestreada, los auditores llegan a una conclusión con respecto a si el monto de las cuentas por cobrar que aparece en el estado de cuenta del cliente es aceptable.

Finanzas

Los analistas financieros utilizan una variedad de información estadística como guía para sus recomendaciones de inversión. En el caso de las acciones, revisan diversos datos financieros que incluyen las razones precio/ganancias y el rendimiento de los dividendos. Al comparar la información para una acción con datos sobre los promedios del mercado de valores, un analista financiero puede formular una conclusión acerca de si una acción está sub o sobrevaluada. Por ejemplo, *Barron's* (18 de febrero de 2008) informó que la rentabilidad media por dividendo de las 30 acciones del promedio industrial Dow Jones fue de 2.45%. Altria Group mostró una rentabilidad por dividendo de 3.05%. En este caso la información estadística sobre la rentabilidad por dividendo indica que dicha empresa ofrece una rentabilidad mayor que el promedio para las acciones Dow Jones. Por tanto, un analista financiero podría concluir que Altria Group estaba subvaluada. Ésta y otra información sobre la compañía ayudan al analista a hacer una recomendación de comprar o vender las acciones, o esperar.

Marketing

Los escáneres electrónicos en las cajas de cobro de los establecimientos minoristas recolectan datos para una variedad de aplicaciones de investigación de mercados. Por ejemplo, proveedores de datos como ACNielsen e Information Resources, Inc. compran datos de los escáneres en puntos de venta como las tiendas de abarrotes, los procesan y luego venden resúmenes estadísticos a los fabricantes. Estos últimos gastan cientos de miles de dólares por categoría de producto para obtener este tipo de datos. Los fabricantes también compran datos y resúmenes estadísticos sobre actividades promocionales, como la fijación de precios especiales y el uso de exhibidores dentro de las tiendas. Los gerentes de marca pueden revisar las estadísticas de los escáneres y de la actividad promocional para comprender mejor la relación entre las actividades de promoción y las ventas. Estos análisis a menudo son útiles para establecer estrategias de marketing futuras para diversos productos.

Producción

El énfasis actual en la calidad hace que su control sea una aplicación importante de la estadística en la producción. Una variedad de gráficas estadísticas de control de calidad se usan para monitorear el resultado de un proceso de producción. En particular, una gráfica x barra sirve para monitorear el resultado medio. Suponga, por ejemplo, que una máquina llena envases con 12 onzas de una bebida refrescante. En forma periódica, un empleado de producción selecciona una muestra de envases y calcula el número medio de onzas en la muestra. Este promedio, o valor x barra, se traza en una gráfica x barra. Un valor trazado sobre el límite superior de control de la gráfica indica que hay un exceso en el llenado, y un valor trazado por debajo del límite inferior de control indica que el llenado es deficiente. El proceso se considera "bajo control" y permite continuar siempre que los valores x barra trazados se encuentren dentro de los límites de control superior e inferior de la gráfica. Si se interpreta de manera adecuada, una gráfica x barra ayuda a determinar cuándo es necesario hacer ajustes para corregir un proceso de producción.

Economía

Los economistas a menudo proporcionan pronósticos acerca del futuro de la economía o sobre algún otro aspecto relacionado. Utilizan una variedad de información estadística para elaborarlos. Por ejemplo, para pronosticar las tasas de inflación recurren a información estadística sobre indicadores como el índice de precios al consumidor, la tasa de desempleo y el uso de la capacidad de manufactura. Estos indicadores se introducen con frecuencia en modelos de pronóstico computarizados que predicen las tasas de inflación.

Las aplicaciones de la estadística, como las descritas en esta sección, son una parte integral de este libro. Estos ejemplos proporcionan una descripción general de gran diversidad de aplicaciones. Para complementar estos ejemplos, profesionales en el campo de los negocios y la economía aportaron artículos para la sección *Estadística en la práctica* al inicio del capítulo, donde se presenta el material que cubre su contenido. Dichas aplicaciones muestran la importancia de la estadística en una amplia variedad de situaciones de negocios y economía.

1.2 Datos

Los **datos** son los hechos y las cifras recabados, analizados y resumidos para su presentación e interpretación. Todos los datos recabados en un estudio en particular se conocen como **banco de datos** del estudio. La tabla 1.1 muestra un banco de datos que contiene información de 25 fondos de inversión que forman parte de *Morningstar Funds500* para 2008. Morningstar es una empresa que le sigue la pista a más de 7 000 fondos de inversión y prepara análisis detallados de 2 000 de ellos. Los analistas financieros e inversionistas individuales siguen sus recomendaciones al pie de la letra.

Elementos, variables y observaciones

Los **elementos** son las entidades a partir de las cuales se reúnen los datos. Para el banco de datos de la tabla 1.1, cada fondo de inversión es un elemento: sus nombres aparecen en la primera columna. Puesto que hay 25 fondos de inversión, el banco de datos contiene 25 elementos.

Una **variable** es una característica de interés para los elementos. El banco de datos de la tabla 1.1 incluye las cinco variables siguientes.

- *Fund Type* (*Tipo de fondo*). Tipo de fondo de inversión, etiquetado DE (capital nacional), IE (capital internacional) y FI (renta fija)
- *Net Asset Value* (*Valor de los activos netos en $*). Precio de cierre por acción al 31 de diciembre de 2007

TABLA 1.1 Banco de datos para 25 fondos de inversión

Fund Name	Fund Type	Net Asset Value ($)	5-Year Average Return (%)	Expense Ratio (%)	Morningstar Rank
American Century Intl. Disc	IE	14.37	30.53	1.41	3-star
American Century Tax-Free Bond	FI	10.73	3.34	0.49	4-star
American Century Ultra	DE	24.94	10.88	0.99	3-star
Artisan Small Cap	DE	16.92	15.67	1.18	3-star
Brown Cap Small	DE	35.73	15.85	1.20	4-star
DFA U.S. Micro Cap	DE	13.47	17.23	0.53	3-star
Fidelity Contrafund	DE	73.11	17.99	0.89	5-star
Fidelity Overseas	IE	48.39	23.46	0.90	4-star
Fidelity Sel Electronics	DE	45.60	13.50	0.89	3-star
Fidelity Sh-Term Bond	FI	8.60	2.76	0.45	3-star
Gabelli Asset AAA	DE	49.81	16.70	1.36	4-star
Kalmar Gr Val Sm Cp	DE	15.30	15.31	1.32	3-star
Marsico 21st Century	DE	17.44	15.16	1.31	5-star
Mathews Pacific Tiger	IE	27.86	32.70	1.16	3-star
Oakmark I	DE	40.37	9.51	1.05	2-star
PIMCO Emerg Mkts Bd D	FI	10.68	13.57	1.25	3-star
RS Value A	DE	26.27	23.68	1.36	4-star
T. Rowe Price Latin Am.	IE	53.89	51.10	1.24	4-star
T. Rowe Price Mid Val	DE	22.46	16.91	0.80	4-star
Thornburg Value A	DE	37.53	15.46	1.27	4-star
USAA Income	FI	12.10	4.31	0.62	3-star
Vanguard Equity-Inc	DE	24.42	13.41	0.29	4-star
Vanguard Sht-Tm TE	FI	15.68	2.37	0.16	3-star
Vanguard Sm Cp Idx	DE	32.58	17.01	0.23	3-star
Wasatch Sm Cp Growth	DE	35.41	13.98	1.19	4-star

Fuente. *Morningstar Funds500* (2008).

WEB archivo
Morningstar

Los bancos de datos como el de Morningstar están disponibles en inglés en el sitio web de este libro.

- 5-Year Average Return (*Rendimiento promedio de 5 años en %*). El rendimiento anual promedio del fondo durante los 5 años anteriores
- Expense Ratio (*Razón de gastos*). El porcentaje de activos deducidos en cada año fiscal de los gastos del fondo
- Morningstar Rank (*Calificación Morningstar*). La calificación general con estrellas ajustada al riesgo de cada fondo; las calificaciones de Morningstar varían de una baja de 1 estrella (1-Star) a una alta de 5 estrellas (5-Star)

En un estudio, las mediciones recabadas para cada elemento en cada variable proporcionan los datos. El conjunto de mediciones obtenido para un elemento en particular se llama **observación**. Al analizar de nuevo la tabla 1.1 vemos que el conjunto de mediciones para la primera observación (American Century Intl. Disc) es IE, 14.37, 30.53, 1.41 y 3-Star. El conjunto de mediciones para la segunda observación (American Century Tax-Free Bond) es FI, 10.73, 3.34, 0.49 y 4-Star, etc. Un banco de datos con 25 elementos contiene 25 observaciones.

Escalas de medición

La recolección de datos requiere una de las escalas de medición siguientes: nominal, ordinal, de intervalo o de razón. La escala de medición determina la cantidad de información contenida en los datos e indica la manera más apropiada de resumirlos y analizarlos estadísticamente.

Cuando los datos de una variable se componen de etiquetas o nombres utilizados para identificar un atributo del elemento, la escala de medición se considera una **escala nominal**. Por ejemplo, al observar los datos de la tabla 1.1 vemos que la escala de medición para la variable tipo de fondo es nominal, porque DE, IE y FI son etiquetas utilizadas para identificar la categoría o tipo de fondo. En tales casos se puede usar un código numérico o etiquetas no numéricas. Por ejemplo, para facilitar la recolección y preparación de los datos con el fin de introducirlos en una base de datos computarizada, podríamos usar un código numérico que establezca que 1 denota un capital nacional, 2 un capital internacional y 3 una renta fija. En este caso los valores numéricos 1, 2 y 3 identifican la categoría del fondo. La escala de medición es nominal a pesar de que los datos aparecen como valores numéricos.

La escala de medición de una variable se llama **escala ordinal** si los datos exhiben las propiedades de los datos nominales y su orden o clasificación es significativo. Por ejemplo, Eastside Automotive envía a los clientes un cuestionario diseñado para obtener datos sobre la calidad de su servicio de reparación de automóviles. Cada cliente califica el servicio de reparación como excelente, bueno o malo. Dado que los datos obtenidos son las etiquetas *excelente*, *bueno* o *malo*, poseen las cualidades de los datos nominales. Además, pueden clasificarse, u ordenarse, con respecto a la calidad en el servicio. Los datos registrados como excelente indican el mejor servicio, seguidos por bueno y luego por malo. Así, la escala de medición es ordinal. Como otro ejemplo, observe que la calificación de Morningstar para los datos de la tabla 1.1 es un dato ordinal. Proporciona una calificación de 1 a 5 estrellas basada en la evaluación del rendimiento ajustado al riesgo que proporciona el fondo. Los datos ordinales también pueden proporcionarse por medio de un código numérico, por ejemplo, su número de lista en clase.

En una **escala de intervalo** para una variable los datos presentan todas las propiedades de los datos ordinales, y el intervalo entre los valores se expresa en términos de una unidad de medida fija. Los datos de intervalo son siempre numéricos. Las calificaciones de la prueba de aptitudes Scholastic Aptitude Test (SAT) son un ejemplo de datos escala de intervalo. Por ejemplo, tres estudiantes que obtuvieron las calificaciones 620, 550 y 470 en una prueba de matemáticas llamada SAT pueden clasificarse u ordenarse en función del mejor al peor rendimiento. Además, las diferencias entre las puntuaciones son significativas. Por ejemplo, el estudiante 1 obtuvo $620 - 550 = 70$ puntos más que el alumno 2, mientras que éste obtuvo $550 - 470 = 80$ puntos más que el estudiante 3.

En una **escala de razón** para una variable los datos tienen todas las propiedades de los datos de intervalo, y la razón de los dos valores es significativa. Para la medición de variables como la distancia, la estatura, el peso y el tiempo se usa la escala de razón. Ésta requiere que se incluya un valor cero para indicar que en este punto no existe un valor para la variable. Por

ejemplo, considere el costo de un automóvil. Un valor cero para el costo indicaría que el vehículo no tiene costo, es gratis. Además, si se compara el costo de un automóvil de $30 000 con el costo de un segundo automóvil de $15 000, la propiedad de la razón muestra que el primero cuesta $30 000/$15 000 = 2 veces, o el doble, que el segundo.

Datos categóricos y cuantitativos

Los datos se clasifican como categóricos o cuantitativos. Los que se agrupan por categorías específicas se conocen como **datos categóricos**. Este tipo de datos utiliza una escala de medición que puede ser nominal u ordinal. Los que utilizan valores numéricos para indicar cuánto o cuántos se conocen como **datos cuantitativos**; éstos se obtienen usando la escala de medición ya sea de intervalo o de razón.

El método estadístico apropiado para resumir los datos depende de que los datos sean categóricos o cuantitativos.

Una **variable categórica** incluye datos categóricos y una **variable cuantitativa** comprende datos cuantitativos. El análisis estadístico apropiado para una variable en particular depende de que ésta sea categórica o cuantitativa. Si la variable es categórica, el análisis estadístico es muy limitado. Los datos categóricos se resumen mediante el conteo del número de observaciones en cada categoría o por medio del cálculo de la proporción de las observaciones en cada categoría. Sin embargo, aun cuando estos datos se identifican por medio de un código aritmético, operaciones como la suma, la resta, la multiplicación y la división no producen resultados que tengan sentido. En la sección 2.1 se estudian algunas maneras de resumir los datos categóricos.

Las operaciones aritméticas sí proporcionan resultados con sentido para las variables cuantitativas. Por ejemplo, los datos cuantitativos pueden sumarse y luego dividirse entre el número de observaciones para calcular el valor promedio, el cual tiene significado y se interpreta con facilidad. En general, se tienen más alternativas para el análisis estadístico cuando los datos son cuantitativos. La sección 2.2 y el capítulo 3 proporcionan maneras de resumir este tipo de datos.

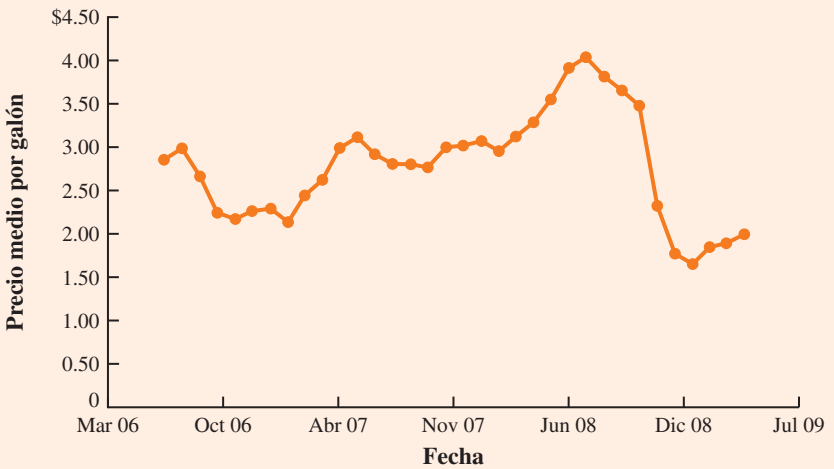
Datos de corte transversal y de series de tiempo

Para efectos del análisis estadístico es importante distinguir entre datos de corte transversal y datos de series de tiempo. Los **datos de corte transversal** son recabados en el mismo momento, o aproximadamente al mismo tiempo. Los de la tabla 1.1 son de corte transversal debido a que describen las cinco variables para los 25 fondos de inversión en el mismo momento. Los **datos de series de tiempo** o de series temporales son recabados a lo largo de varios periodos. Por ejemplo, la serie de tiempo de la figura 1.1 muestra el precio medio por galón de gasolina regular convencional en Estados Unidos entre 2006 y 2009. Observe que los precios más altos del hidrocarburo han tendido a presentarse en los meses del verano, con un promedio máximo histórico de $4.05 por galón en julio de 2008. Para enero de 2009, los precios de la gasolina habían alcanzado una marcada reducción de un mínimo en tres años de $1.65 por galón.

Las gráficas de los datos de series de tiempo a menudo se encuentran en publicaciones de negocios y economía; ayudan a los analistas a comprender lo que ocurrió en el pasado, identificar cualquier tendencia en el tiempo y proyectar niveles futuros para las series de tiempo. Las gráficas de este tipo pueden adoptar una variedad de formas, como muestra la figura 1.2. Con un poco de estudio, suelen ser fáciles de comprender e interpretar.

Por ejemplo, la gráfica (A) de la figura 1.2 muestra el índice promedio industrial Dow Jones de 1997 a 2009. En abril de 1997 el índice del mercado de valores generalizado era de cerca de 7 000. En los 10 años siguientes llegó a más de 14 000 en julio de 2007. Sin embargo, observe la marcada disminución en las series de tiempo después del máximo histórico de 2007. Para marzo de 2009, las malas condiciones económicas habían causado que el índice volviera al nivel de 7 000 de 1997. Este fue un periodo de temor y desaliento para los inversionistas. En junio de 2009 el indicador mostró una recuperación al alcanzar 8 700 puntos.

FIGURA 1.1 Precio promedio por galón para la gasolina regular convencional en Estados Unidos



Fuente. Energy Information Administration, U.S. Department of Energy, mayo de 2009.

La gráfica (B) muestra la utilidad neta de McDonald's Inc. desde 2003 hasta 2009. Las condiciones económicas en declive de 2008 y 2009 fueron realmente benéficas para la empresa, ya que su utilidad neta alcanzó un máximo histórico. Este crecimiento en la utilidad neta demostró que la firma estaba prosperando durante la crisis económica, cuando la gente empezó a restringir sus gastos y prefería las alternativas más económicas ofrecidas por McDonald's en lugar de los restaurantes tradicionales más costosos.

La gráfica (C) muestra la serie de tiempo para la tasa de ocupación de los hoteles en el sur de Florida durante un periodo de un año. Las tasas más altas, 95 y 98%, ocurren durante los meses de febrero y marzo, cuando el clima de la región es atractivo para los turistas. De hecho, la temporada de ocupación más alta para los hoteles del sur de Florida es de enero a abril de cada año. Por otra parte, observe las menores tasas de ocupación durante los meses de agosto a octubre, periodo en cual se encuentra el indicador más bajo de 50% durante septiembre. Las elevadas temperaturas y la temporada de huracanes son las razones principales de la caída en la ocupación de los hoteles durante este periodo.

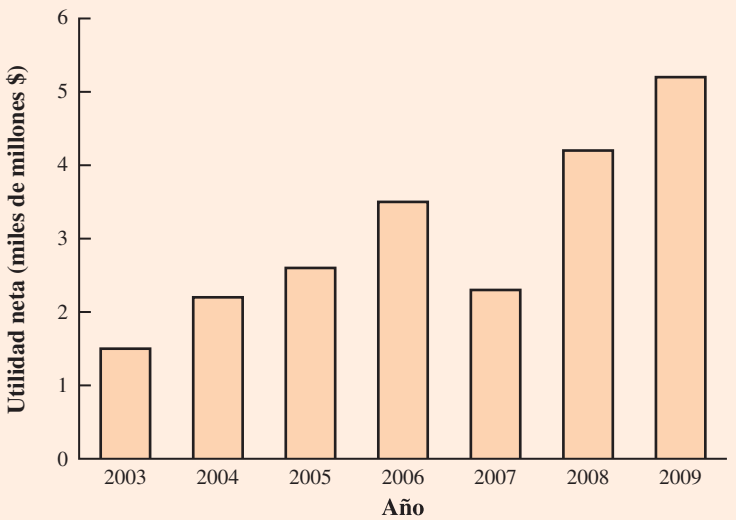
NOTAS Y COMENTARIOS

1. Una observación es el conjunto de mediciones obtenido para cada elemento de un banco de datos. Por consiguiente, el número de observaciones es siempre igual al número de elementos. El número de mediciones obtenidas para cada elemento es igual al número de variables. Por ende, el número total de elementos de datos se determina multiplicando el número de observaciones por el número de variables.
2. Los datos cuantitativos pueden ser discretos o continuos. Los datos cuantitativos que miden cuántos (por ejemplo, el número de llamadas recibidas en 5 minutos), son discretos. Los datos cuantitativos que miden cuánto (por ejemplo, el peso o el tiempo), son continuos debido a que no hay una separación entre los valores de datos posibles.

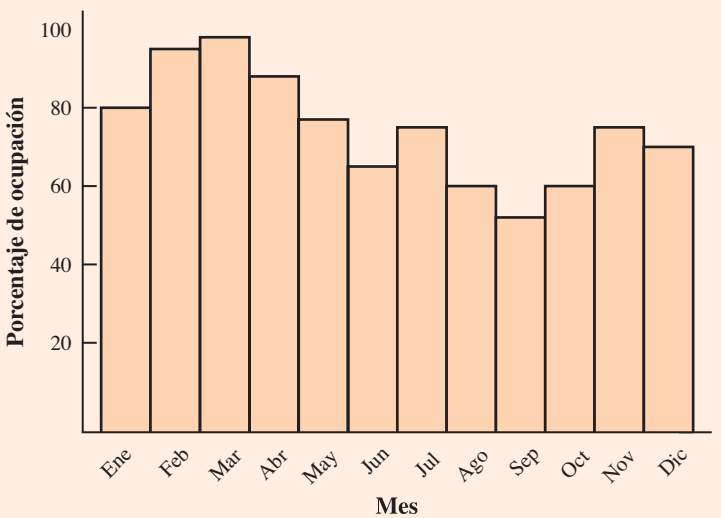
FIGURA 1.2 Varias gráficas de series de tiempo



(A) Promedio industrial Dow Jones



(B) Utilidad neta de McDonald's Inc.



(C) Tasa de ocupación de los hoteles del sur de Florida

1.3 Fuentes de datos

Los datos se obtienen de fuentes existentes o de encuestas y estudios experimentales diseñados para recabar datos nuevos.

Fuentes existentes

En algunos casos, los datos necesarios para una aplicación en particular ya existen. Las empresas mantienen una variedad de bases de datos sobre sus empleados, clientes y operaciones de negocios. Los datos sobre los sueldos, la edad y los años de experiencia de los empleados se obtienen por lo general de los registros internos del personal. Otros registros internos contienen datos sobre ventas, gastos de publicidad, costos de distribución, niveles de inventario y cantidades de producción. La mayoría de las empresas mantiene también datos detallados sobre sus clientes. La tabla 1.2 muestra algunos de los datos de que se dispone comúnmente a partir de los registros internos de una empresa.

Las organizaciones que se especializan en la recolección y el mantenimiento de datos proveen cantidades significativas de información económica y de negocios. Las empresas tienen acceso a estas fuentes de datos externos por medio de acuerdos o al comprarlos. Dun & Bradstreet, Bloomberg y Dow Jones & Company son tres firmas que ofrecen extensos servicios de bases de datos a sus clientes. ACNielsen e Information Resources, Inc. ha logrado el éxito en su negocio de recolección y procesamiento de datos que vende a anunciantes y fabricantes de productos.

También se obtienen datos de diversas asociaciones de la industria y de organizaciones de interés especial. Travel Industry Association of America mantiene información relacionada con viajes, como el número de turistas y los gastos de viaje por estado. Estos datos son de interés para las empresas y personas de la industria del ramo. El Graduate Management Admission Council cuenta con datos sobre calificaciones de exámenes, características de los estudiantes y programas sobre administración de educación universitaria. La mayoría de los datos provenientes de estos tipos de fuentes se proveen a usuarios calificados por un costo moderado.

La importancia de Internet como una fuente de datos e información estadística sigue creciendo. Casi todas las empresas tienen una página web que proporciona información general acerca de la organización, así como datos sobre ventas, número de empleados y de productos, el precio de los productos y sus especificaciones. Además, varias empresas se especializan en proveer información a través de Internet, gracias a lo cual se puede tener acceso a cotizaciones de acciones, precios de los platillos en los restaurantes, datos sobre sueldos y una variedad casi infinita de información.

Las agencias gubernamentales son otra fuente importante de datos existentes. Por ejemplo, el Departamento de Trabajo de Estados Unidos mantiene una gran cantidad de datos sobre las tasas de empleo, las tasas salariales, el porcentaje de la población activa y la afiliación a

TABLA 1.2 Ejemplos de datos disponibles de los registros internos de una empresa

Fuente	Algunos datos comúnmente disponibles
Registros de empleados	Nombre, domicilio, número de Seguro Social, sueldo, número de días de vacaciones, número de días de incapacidad y bonos
Registros de producción	Número de parte o de producto, cantidad producida, costo de mano de obra directa y costo de los materiales
Registros de inventarios	Número de parte o de producto, cantidad de unidades disponible, punto de reorden, lote económico y programa de descuentos
Registros de ventas	Número de producto, volumen de ventas, volumen de ventas por región y volumen de ventas por tipo de cliente
Registros de crédito	Nombre del cliente, domicilio, número telefónico, límite de crédito y saldo de las cuentas por cobrar
Perfiles de clientes	Edad, género, nivel de ingresos, número de miembros en la familia, domicilio y preferencias

TABLA 1.3 Ejemplos de datos disponibles de algunas agencias gubernamentales

Agencia gubernamental	Algunos datos disponibles
Oficina del Censo	Datos poblacionales, número de familias e ingresos por familia
Consejo de la Reserva Federal	Datos sobre la masa monetaria, crédito a plazo, tipos de cambio y tasas de descuento
Oficina de Administración y Presupuesto	Datos sobre ingresos, gastos y deudas del gobierno federal
Departamento de Comercio	Datos sobre la actividad comercial, valor de las remesas por industria, nivel de utilidades por industria e industrias en crecimiento y en declive
Oficina de Estadísticas Laborales	Gasto de los consumidores, ganancias por hora, tasa de desempleo, registros de seguridad y estadísticas internacionales

sindicatos. La tabla 1.3 lista algunas agencias gubernamentales seleccionadas y algunos de los datos que proporcionan. La mayoría de las dependencias que reúne y procesa datos también los pone a disposición de los usuarios por medio de un sitio web. La figura 1.3 muestra la página principal del sitio web de la Oficina del Censo de Estados Unidos.

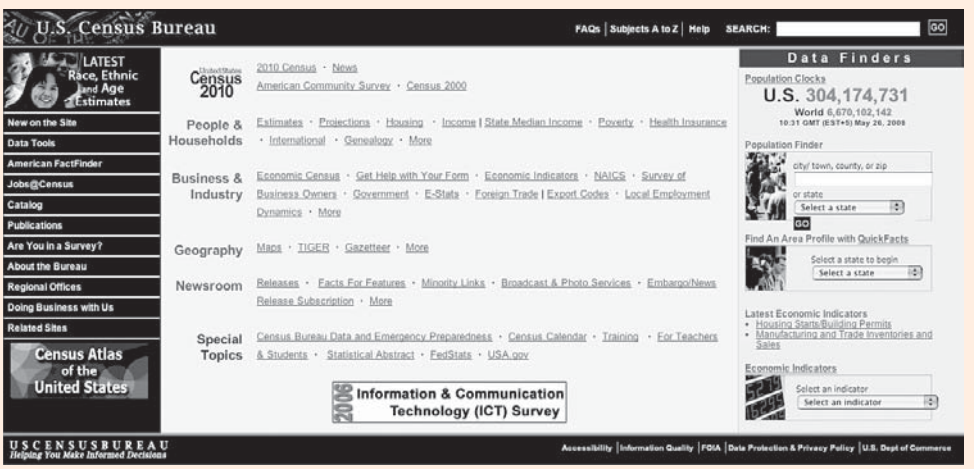
Estudios estadísticos

Se cree que el estudio estadístico experimental más grande jamás realizado es el experimento para la vacuna Salk *contra la polio del Servicio de Salud Pública efectuado en 1954. Se seleccionaron casi 2 millones de niños de 1o., 2o. y 3er. grados de enseñanza elemental de todo Estados Unidos.*

Algunas veces los datos necesarios para alguna aplicación en particular no están disponibles a través de las fuentes existentes. En estos casos suelen obtenerse mediante estudios estadísticos, los cuales se clasifican en *experimentales* u *observacionales*.

En un estudio experimental se identifica primero la variable de interés. Luego se toman una o más variables y se controlan para obtener datos de cómo influyen en la variable de interés. Por ejemplo, una compañía farmacéutica podría interesarse en realizar un experimento para enterarse de cómo afecta un nuevo medicamento la presión sanguínea. Ésta es la variable de interés en el estudio. El nivel de dosis del medicamento nuevo es otra variable que se espera que tenga un efecto causal en la presión sanguínea. Para obtener datos sobre el efecto del nuevo fármaco, los investigadores seleccionan una muestra de individuos. El nivel de dosis del medicamento está controlado, ya que a los distintos grupos de individuos se les suministran dosis diferentes.

FIGURA 1.3 Página principal de la Oficina del Censo de Estados Unidos



Antes y después se recaban datos sobre la presión sanguínea para cada grupo. El análisis estadístico de los datos experimentales ayuda a determinar el efecto del nuevo medicamento en la presión sanguínea.

Los estudios estadísticos no experimentales u observacionales de ninguna manera intentan controlar las variables de interés. Una encuesta es tal vez el tipo más común de estudio observacional. Por ejemplo, en una entrevista personal para una encuesta primero se identifican las preguntas de investigación, luego se diseña un cuestionario y después se administra a una muestra de individuos. Algunos restaurantes usan estudios observacionales para obtener datos sobre las opiniones de los clientes acerca de la calidad de la comida, la calidad en el servicio, la atmósfera del lugar, etc. En la figura 1.4 se aprecia un cuestionario de opinión del cliente usado por Chops City Grill en Naples, Florida. Observe que a los clientes que contestaron se les pide que califiquen 12 variables, como la experiencia general, la amabilidad de la *hostess*, el gerente (visita a la mesa), el servicio general, etc. Las categorías de respuesta de excelente, bueno, regular, aceptable y malo proporcionan datos categóricos que permiten a la gerencia de Chops City Grill mantener estándares altos para los alimentos y el servicio del restaurante.

Los estudios sobre los fumadores y los no fumadores son observacionales, debido a que los investigadores no determinan o controlan quién fuma y quién no.

Cualquiera que quiera usar datos y análisis estadísticos como apoyo en la toma de decisiones debe considerar el tiempo y el costo requeridos para obtenerlos. El uso de las fuentes existentes es recomendable cuando los datos deben obtenerse en un periodo relativamente breve. Si los datos importantes no se pueden recabar con facilidad de una fuente existente, debe tomarse el tiempo y el costo adicionales involucrados en su obtención. En todos los casos, quien

FIGURE 1.4 Cuestionario de opinión del cliente usado por el restaurante Chops City Grill en Naples, Florida

Chop's
CITY GRILL

Fecha: _____ Nombre del mesero: _____

Nuestros clientes son nuestra máxima prioridad. Por favor, tómese un momento para llenar esta encuesta que nos permitirá mejorar nuestro servicio con base en sus necesidades. Puede devolver esta tarjeta a la recepción o enviarla por correo. ¡Gracias!

ENCUESTA SOBRE EL SERVICIO	Excelente	Bueno	Promedio	Aceptable	Mediocre
Experiencia general	❑	❑	❑	❑	❑
Amabilidad de la *hostess*	❑	❑	❑	❑	❑
Gerente (visita a la mesa)	❑	❑	❑	❑	❑
Servicio general	❑	❑	❑	❑	❑
Profesionalismo	❑	❑	❑	❑	❑
Conocimiento del menú	❑	❑	❑	❑	❑
Amabilidad	❑	❑	❑	❑	❑
Selección de vinos	❑	❑	❑	❑	❑
Selección del menú	❑	❑	❑	❑	❑
Calidad de los alimentos	❑	❑	❑	❑	❑
Presentación de los alimentos	❑	❑	❑	❑	❑
Gastos estimados $	❑	❑	❑	❑	❑

¿Qué comentarios podría hacernos para mejorar nuestro restaurante?

_____

_____

Gracias, apreciamos sus comentarios. —El personal de Chops City Grill.

toma decisiones debe considerar la contribución del análisis estadístico al proceso de toma de decisiones. El costo de la adquisición de los datos y el análisis estadístico subsiguiente no deben exceder los ahorros generados por el uso de la información para tomar una mejor decisión.

Errores en la adquisición de los datos

Los gerentes deben estar siempre conscientes de la posibilidad de errores en los estudios estadísticos. El uso de datos erróneos es peor que no usar datos en lo absoluto. Una equivocación en la adquisición de los datos ocurre siempre que el valor de los datos obtenidos no es igual al valor verdadero o real que se obtendría con un procedimiento correcto. Estos errores pueden ocurrir de varias maneras. Por ejemplo, un entrevistador podría cometer una falla de registro, como una transposición de los números cuando escribe la edad de una persona, es decir, que anote 42 en vez de 24 años, o la persona que responde una pregunta durante una entrevista podría malinterpretarla y dar una respuesta incorrecta.

Los analistas de datos con experiencia tienen gran cuidado en recabar y registrar los datos para asegurarse de no cometer errores. Para verificar su consistencia interna se usan procedimientos especiales. Por ejemplo, estos procedimientos indicarían que el analista debe revisar la precisión de los datos para una persona entrevistada que indica que tiene 22 años de edad pero reporta 20 años de experiencia laboral. Los analistas también revisan los datos con valores inusualmente grandes o pequeños llamados *atípicos*, los cuales son candidatos a posibles errores. En el capítulo 3 se presentan algunos métodos que los expertos en estadística utilizan para identificar datos atípicos.

Los errores ocurren con frecuencia durante la adquisición de los datos. El uso a ciegas de cualquier dato con que se cuenta o el uso de los que se adquirieron con poco cuidado puede conducir a información y a decisiones erróneas. Por tanto, emprender acciones para adquirir datos precisos ayuda a garantizar que la información sea confiable y valiosa para la toma de decisiones.

1.4 Estadística descriptiva

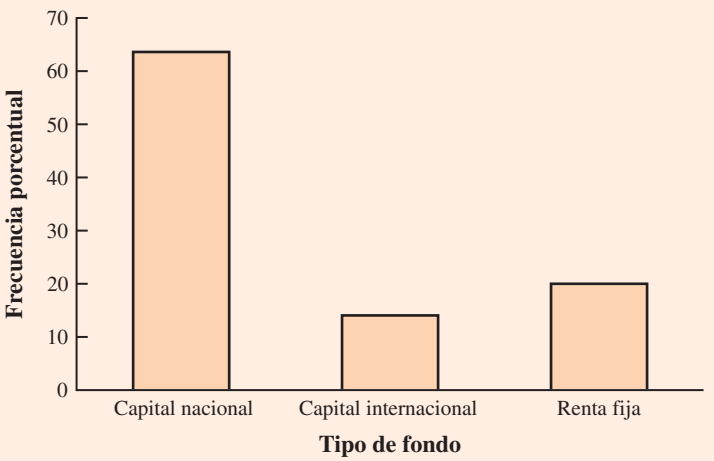
La mayor parte de la información estadística en periódicos, revistas, informes de empresas y otras publicaciones consiste en datos que se resumen y presentan en una forma fácil de comprender para el lector. Estos resúmenes de datos, que pueden ser tablas, gráficas o números, se conocen como **estadística descriptiva**.

Consulte de nuevo el banco de datos de la tabla 1.1 que muestra información referente a 25 fondos de inversión. Los métodos de la estadística descriptiva se pueden usar para elaborar resúmenes de la información de este banco de datos. Por ejemplo, en la tabla 1.4 se presenta un resumen de los datos para la variable categórica fund type (tipo de fondo). Un resumen gráfico de los mismos datos, llamado *gráfica de barras*, se muestra en la figura 1.5. Estos tipos de resúmenes gráficos y tabulares facilitan la interpretación de los datos. Al remitirnos a la tabla 1.4 y a la figura 1.5 podemos ver con facilidad que la mayoría de los fondos de inversión son del tipo domestic equity (capital nacional). Expresado en porcentajes, 64% son del tipo domestic equity, 16% son del tipo international equity (capital internacional) y 20% son del tipo fixed income (renta fija).

TABLA 1.4 Frecuencias y frecuencias de porcentaje para el tipo de fondo de inversión

Tipo de fondo de inversión	Frecuencia	Frecuencia porcentual
Capital nacional	16	64
Capital internacional	4	16
Renta fija	5	20
Totales	**25**	**100**

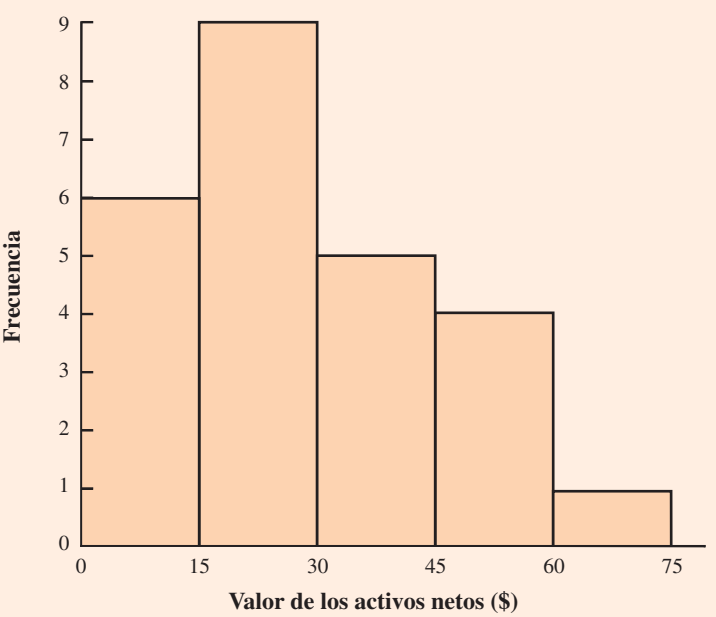
FIGURA 1.5 Gráfica de barras para el tipo de fondo de inversión



Un resumen gráfico de los datos para la variable cuantitativa net asset value (valor de los activos netos), llamado *histograma*, se muestra en la figura 1.6. El histograma facilita ver que los valores de activos netos varían de $0 a $75, con una mayor concentración entre $15 y $30. Sólo uno de los valores de los activos netos es mayor de $60.

Además de las tablas y las gráficas, para resumir los datos se usa la estadística descriptiva numérica. La medida estadística descriptiva numérica más común es el promedio, o media. Utilizando los datos sobre el rendimiento promedio de 5 años para los fondos de inversión de

FIGURA 1.6 Histograma del valor de los activos netos para 25 fondos de inversión



la tabla 1.1 podemos calcular el promedio al sumar los rendimientos para los 25 fondos de inversión y dividir la suma entre 25. Al hacerlo, obtenemos un rendimiento promedio de 5 años de 16.50%, el cual demuestra una medida de tendencia central, o posición central, de los datos para esa variable.

El interés en métodos eficaces para el desarrollo y la presentación de la estadística descriptiva es cada vez mayor. Los capítulos 2 y 3 se centran en los métodos tabular, gráfico y numérico de la estadística descriptiva.

1.5 Inferencia estadística

Numerosas situaciones requieren información sobre un grupo grande de elementos (personas, empresas, votantes, familias, productos, clientes, etc.), pero por razones de tiempo, costo y otras consideraciones sólo se pueden recabar datos de una pequeña porción del conjunto. El grupo de elementos más grande en un estudio en particular se llama **población**, y el grupo más pequeño se denomina **muestra**. Formalmente usamos las definiciones siguientes.

POBLACIÓN

Una población es el conjunto de todos los elementos de interés en un estudio en particular.

MUESTRA

Una muestra es un subconjunto de la población.

El gobierno estadounidense efectúa un censo cada 10 años. Las firmas de investigación de mercados realizan todos los días encuestas por muestreo.

El proceso de realizar una encuesta para recabar datos de toda una población se llama **censo**. El proceso de realizar una encuesta para recabar datos de una muestra se llama **encuesta por muestreo**. Como una de sus contribuciones importantes, la estadística usa datos de una muestra para hacer estimaciones y probar hipótesis sobre las características de una población mediante un proceso conocido como **inferencia estadística**.

Como ejemplo de inferencia estadística, considere el estudio realizado por Electronics Norris, la cual fabrica focos de alta luminosidad usados en una variedad de productos eléctricos. En un intento por aumentar la vida útil de los focos, el grupo de diseño desarrolló un filamento nuevo. En este caso, la población se define como todos los focos que pueden fabricarse con el filamento nuevo. Para evaluar sus ventajas, se fabricaron y probaron 200 focos con dicho aditamento. Los datos recabados de esta muestra señalan el número de horas de funcionamiento de cada foco antes de que el filamento se fundiera. Consulte la tabla 1.5.

Suponga que Norris quiere usar los datos muestrales para hacer una inferencia sobre las horas promedio de vida útil para la población de todos los focos que podrían producirse con el filamento nuevo. La adición de los 200 valores en la tabla 1.5 y la división del total entre 200 proporciona la vida útil promedio de la muestra para los focos: 76 horas. Podemos usar este resultado para estimar que la vida útil promedio para los focos en la población es de 76 horas. La figura 1.7 proporciona un resumen gráfico del proceso de inferencia estadística para Norris Electronics.

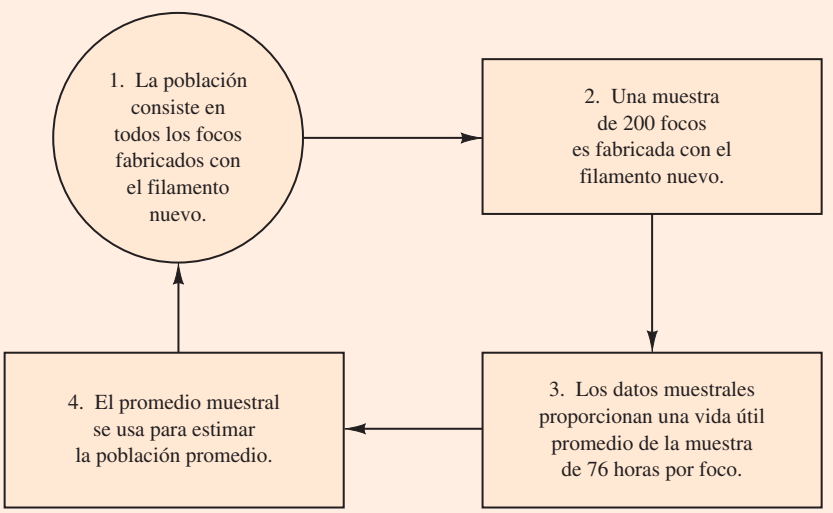
Siempre que los expertos en estadística usan una muestra para estimar una característica de interés de la población, proporcionan un enunciado de la calidad, o precisión, asociada con la estimación. Para el ejemplo de Norris, el experto en estadística podría afirmar que la estimación

TABLA 1.5 Horas hasta que el filamento se funde para una muestra de 200 focos en el ejemplo de Norris Electronics

WEB archivo
Norris

107	73	68	97	76	79	94	59	98	57
54	65	71	70	84	88	62	61	79	98
66	62	79	86	68	74	61	82	65	98
62	116	65	88	64	79	78	79	77	86
74	85	73	80	68	78	89	72	58	69
92	78	88	77	103	88	63	68	88	81
75	90	62	89	71	71	74	70	74	70
65	81	75	62	94	71	85	84	83	63
81	62	79	83	93	61	65	62	92	65
83	70	70	81	77	72	84	67	59	58
78	66	66	94	77	63	66	75	68	76
90	78	71	101	78	43	59	67	61	71
96	75	64	76	72	77	74	65	82	86
66	86	96	89	81	71	85	99	59	92
68	72	77	60	87	84	75	77	51	45
85	67	87	80	84	93	69	76	89	75
83	68	72	67	92	89	82	96	77	102
74	91	76	83	66	68	61	73	72	76
73	77	79	94	63	59	62	71	81	65
73	63	63	89	82	64	85	92	64	73

FIGURA 1.7 Proceso de inferencia estadística para el ejemplo de Norris Electronics



puntual de la vida útil promedio para la población de focos nuevos es de 76 horas con un margen de error de ± 4 horas. Por tanto, una estimación del intervalo de la vida útil promedio para todos los focos producidos con el filamento nuevo es de 72 a 80 horas. El experto también puede mencionar cuánta confianza tiene en que este intervalo contenga la población promedio.

1.6 Computadoras y análisis estadístico

Los bancos de datos de Minitab y Excel, y el complemento StatTools de Excel están disponibles en el sitio web de este libro.

Los expertos en estadística usan con frecuencia el *software* de computadora para realizar los cálculos estadísticos requeridos para grandes cantidades de datos. Por ejemplo, el cálculo de la vida útil promedio para los 200 focos en el ejemplo de Norris Electronics (tabla 1.5) sería muy tedioso sin una computadora. Para facilitar el uso de una computadora, muchos bancos de datos en este libro están disponibles en el sitio web que acompaña el volumen. Los archivos de datos pueden descargarse en formato de Minitab o de Excel. Además, el complemento StatTools de Excel puede descargarse del sitio web. Los apéndices de capítulo cubren paso a paso los procedimientos para usar Minitab, Excel y el complemento StatTools de Excel para implementar las técnicas estadísticas presentadas en el capítulo.

1.7 Minería de datos

Con la ayuda de lectores de tarjetas magnéticas, escáneres de código de barras y terminales punto de venta, la mayoría de las organizaciones obtiene cada día grandes cantidades de datos. Incluso para un restaurante local pequeño que usa monitores de pantalla táctil para introducir pedidos y manejar la facturación, la cantidad de datos recabados puede ser significativa. En el caso de los grandes minoristas, el gran volumen de datos recolectados es difícil de conceptuar, y averiguar cómo usarlos de manera eficiente para mejorar la rentabilidad constituye un reto. Por ejemplo, los minoristas masivos como Wal-Mart capturan datos de 20 a 30 millones de transacciones diarias; las compañías de telecomunicaciones como France Telecom y AT&T generan más de 300 millones de registros de llamadas por día, y Visa procesa 6 800 transacciones de pago por segundo o aproximadamente 600 millones de operaciones diarias. El almacenamiento y la administración de este tipo de datos es una tarea considerable.

El término *almacenamiento de datos* se utiliza para referirse al proceso de su captura, almacenamiento y mantenimiento. La capacidad de cómputo y las herramientas de recolección de datos han llegado al punto en que ahora es posible almacenar y recuperar en segundos cantidades sumamente grandes. El análisis de los datos del almacén puede resultar en decisiones que conducirán a estrategias nuevas y mayores beneficios para la organización.

El tema de la **minería de datos** trata de métodos relacionados con el desarrollo de información para la toma de decisiones útil de bases de datos grandes. Usando una combinación de procedimientos de estadística, matemáticas y ciencias de la computación, los analistas "extraen los datos" del almacén para convertirlos en información útil; de ahí el nombre *minería de datos.* El Dr. Kurt Thearling, un profesional líder en este campo, define la minería de datos como "la extracción automatizada de información predictiva de bases de datos (grandes)". Las dos palabras clave en la definición del Dr. Thearling son "automatizada" y "predictiva". Los sistemas de minería de datos representan el uso más eficiente de los procedimientos automatizados para extraer información de los datos usando sólo las consultas más generales o incluso vagas realizadas por el usuario. Y el *software* de minería de datos automatiza el proceso de descubrir información predictiva oculta que en el pasado requería análisis manual.

Las principales aplicaciones de la minería de datos las han realizado empresas con un fuerte enfoque en el consumidor, como los negocios minoristas, las organizaciones financieras y las compañías de comunicación. La minería de datos se ha usado de manera exitosa para ayudar a minoristas como Amazon y Barnes & Noble a determinar uno o más productos relacionados que también es probable que compren los clientes que ya han adquirido un artículo específico. Por tanto, cuando un cliente ingresa en el sitio web de una empresa y solicita un producto, el sitio web utiliza anuncios emergentes (*pop-ups*) para alertarlo acerca de los productos adicionales que es probable que compre. En otra aplicación, la minería de datos puede usarse para identificar a los consumidores que es probable que gasten más de $20 en un viaje de compras particular. Estos clientes pueden entonces identificarse como aquellos que reciben un mensaje de correo electrónico especial u ofertas con descuento por correo postal para animarlos a realizar su siguiente viaje de compras antes de la fecha de terminación del descuento.

La minería de datos es una tecnología que se basa en gran medida en metodología estadística como la regresión múltiple, la regresión logística y la correlación, pero se requiere una

Los métodos estadísticos juegan un papel importante en la minería de datos, tanto en términos del descubrimiento de relaciones en los datos como de la predicción de resultados futuros. Sin embargo, una cobertura rigurosa de la minería de datos y del uso de la estadística en la misma están fuera del alcance de este libro.

integración creativa de todos estos métodos y tecnologías de las ciencias de la computación que involucran la inteligencia artificial y el aprendizaje de máquinas para hacerla eficiente. Se requiere una inversión significativa en tiempo y dinero para implementar *software* de minería de datos comercial desarrollado por empresas como Oracle, Teradata y SAS. Los conceptos estadísticos introducidos en este libro serán útiles en la comprensión de la metodología estadística empleada por el *software* de minería de datos y le permitirán una mejor comprensión de la información estadística desarrollada.

Dado que los modelos estadísticos juegan un papel importante en el desarrollo de modelos predictivos en la minería de datos, muchas de las preocupaciones que inquietan a los expertos en el desarrollo de modelos estadísticos también son aplicables. Por ejemplo, una inquietud en cualquier estudio estadístico se relaciona con el problema de la confiabilidad del modelo. Encontrar uno que funcione bien para una muestra de datos en particular no necesariamente significa que pueda aplicarse de manera confiable a otros datos. Uno de los enfoques estadísticos comunes para la evaluación de la confiabilidad del modelo es dividir el banco de datos muestral en dos partes: un banco de datos de entrenamiento y un banco de datos de prueba. Si el modelo desarrollado usando los datos de entrenamiento puede predecir de manera precisa valores en los datos de prueba, decimos que es confiable. Una ventaja que la minería de datos tiene sobre la estadística clásica reside en que la enorme cantidad de datos disponible permite al *software* de minería de datos la partición del banco de datos, de modo que un modelo desarrollado para el banco de datos de entrenamiento pueda probarse para la confiabilidad de otros datos. En este sentido, la partición del banco de datos permite que la minería de datos desarrolle modelos y relaciones y luego observe de manera rápida si son repetibles y válidos con datos nuevos y diferentes. Por otra parte, una advertencia para las aplicaciones de minería de datos es que, con tantos datos disponibles, existe el peligro de sobreajuste (*overfitting*) del modelo al punto de que las asociaciones equivocadas y las conclusiones de causa y efecto parecen existir. La interpretación cuidadosa de la minería de datos se produce, y pruebas adicionales ayudarán a evitar esta dificultad.

1.8 Lineamientos éticos para la práctica estadística

El comportamiento ético es algo por lo que debemos luchar en todo lo que realizamos. Los problemas éticos surgen en la estadística debido a la importancia que ésta tiene en la recolección, el análisis, la presentación y la interpretación de los datos. En un estudio estadístico, el comportamiento poco ético puede adoptar una variedad de formas que incluyen el muestreo impropio, el análisis inapropiado de los datos, el desarrollo de gráficas erróneas, el uso de resúmenes estadísticos inadecuados o una interpretación tendenciosa de los resultados estadísticos.

A medida que empiece a adentrarse en su trabajo estadístico, le recomendamos que sea justo, meticuloso, objetivo y neutral cuando recabe los datos, realice los análisis, haga presentaciones orales y presente informes escritos que contengan información desarrollada. Como un consumidor de la estadística, usted también debe estar consciente de la posibilidad de que otros tengan un comportamiento poco ético hablando en tales términos. Cuando usted ve las estadísticas en los periódicos, la televisión, Internet, etc., es buena idea que muestre cierto escepticismo con respecto a la información y que esté siempre consciente de la fuente de donde provienen, así como de su propósito y objetividad.

La American Statistical Association,[1] la organización estadounidense profesional y líder en la rama de la estadística y para los expertos en la materia, elaboró el informe "Ethical Guidelines for Statistical Practice" para ayudar a los profesionales a tomar y comunicar decisiones éticas y ayudar a los estudiantes a aprender cómo realizar el trabajo estadístico con responsabilidad. El informe contiene 67 pautas organizadas en ocho áreas temáticas: Profesionalismo; Responsabilidades de los fundadores, clientes y empleadores; Responsabilidades en las publicaciones y el testimonio; Responsabilidades de los sujetos de investigación; Responsabilidades de los colegas del equipo de investigación; Responsabilidades con otros expertos o profesionales de la estadística; Responsabilidades respecto de acusaciones de falta de ética profesional, y Responsabilidades de los empleadores, incluidas las organizaciones, las personas, los abogados u otros clientes que emplean a profesionales de la estadística.

[1] American Statistical Association, "Ethical Guidelines for Statistical Practice", 1999.

Una de las pautas éticas en el área del profesionalismo aborda el problema de realizar múltiples pruebas hasta obtener el resultado deseado. Considere este ejemplo. En la sección 1.5 se comentó un estudio estadístico realizado por Norris Electronics que involucra una muestra de 200 focos de alta intensidad fabricados con un filamento nuevo. La vida útil promedio de la muestra, 76 horas, proporcionó una estimación de la vida útil promedio de todos los focos fabricados con el nuevo filamento. Sin embargo, considere lo siguiente: debido a que Norris seleccionó una muestra de focos, es razonable suponer que cualquier otra habría proporcionado una vida útil promedio diferente.

Suponga que la gerencia de Norris esperaba que los resultados le permitieran afirmar que la vida útil promedio de los focos nuevos era de 80 horas o más. Suponga además que la gerencia decidió continuar el estudio al fabricar y probar muestras repetidas de 200 focos con el filamento nuevo hasta obtener una muestra media de 80 horas o más. Si el estudio se repitiera suficientes veces, finalmente se podría conseguir una muestra, sólo por casualidad, que proporcionara el resultado deseado y permitiera a Norris hacer una afirmación como ésta. En este caso los consumidores cometerían un error al pensar que el producto nuevo es mejor de lo que en realidad es. Desde luego, este tipo de comportamiento es poco ético y representa un uso incorrecto grave de la estadística en la práctica.

Varias pautas éticas en las responsabilidades y publicaciones y en el área de testimonios se relacionan con problemas que involucran el manejo de datos. Por ejemplo, un experto en estadística debe representar todos los datos considerados en un estudio y explicar la(s) muestra(s) empleada(s) en realidad. En el estudio de Norris Electronics la vida útil promedio de los 200 focos de la muestra original es de 76 horas; este tiempo es considerablemente menor que las 80 horas o más que la gerencia esperaba obtener. Suponga ahora que después de revisar los resultados que arrojan una vida útil promedio de 76 horas, la empresa descarta todas las observaciones con 70 horas o menos hasta que el filamento se funde, supuestamente debido a que estos focos contienen imperfecciones causadas por problemas iniciales en el proceso de manufactura. Después de descartar estos focos, la vida útil promedio de los que restan en la muestra resulta ser de 82 horas. ¿Sospecharía de la afirmación de Norris de que la vida útil promedio de sus focos es de 82 horas?

Si los focos que funcionan 70 horas o menos hasta que se funden fueron descartados para sencillamente proporcionar una vida útil promedio de 82 horas, no hay duda de que descartar los focos con 70 horas o menos hasta que se funden es poco ético. Pero incluso si estos últimos contienen imperfecciones debido a los problemas iniciales en el proceso de manufactura y, como resultado, no deben haberse incluido en el análisis, el experto que realizó el estudio debe justificar todos los datos que se consideraron y explicar cómo se obtuvo la muestra empleada en realidad. Hacer lo contrario es potencialmente erróneo y constituiría un comportamiento poco ético tanto de la empresa como del experto.

Una pauta en los valores compartidos del informe de la American Statistical Association establece que los profesionales de la estadística deben evitar cualquier tendencia a sesgar su trabajo hacia resultados predeterminados. Este tipo de práctica poco ética se observa con frecuencia cuando se usan muestras poco representativas para hacer reclamos. Por ejemplo, en diversas áreas del país no está permitido fumar en restaurantes. Suponga, no obstante, que un miembro de un grupo de presión para la industria del tabaco entrevista a personas en restaurantes donde está permitido fumar con la finalidad de estimar el porcentaje de comensales que está a favor de consumir cigarrillos en este tipo de negocios. En este caso sostendríamos que el simple hecho de tomar una muestra sólo de las personas que comen en los restaurantes donde está permitido fumar ha sesgado los resultados. Si sólo se informan las conclusiones finales de un estudio como éste, los lectores que desconocen los detalles del sondeo (por ejemplo, que la muestra se recabó sólo en restaurantes donde se permite fumar) pueden ser engañados.

El alcance del informe de la American Statistical Association es amplio e incluye pautas éticas que son apropiadas no sólo para un experto en estadística, sino también para los consumidores de información de este tipo. Le sugerimos que lea el informe para obtener una mejor perspectiva de los problemas éticos a medida que continúe su estudio de la estadística, y adquirir experiencia para determinar cómo asegurar que las normas éticas se cumplen cuando usted empieza a usarla en la práctica.

Resumen

La estadística es el arte y la ciencia de recabar, analizar, presentar e interpretar los datos. Casi todos los estudiantes que cursan licenciaturas de negocios o economía requieren tomar por lo menos un curso de estadística. El capítulo comienza con la descripción de las aplicaciones estadísticas comunes para los negocios y la economía.

Los datos consisten en los hechos y las cifras que se recaban y analizan. Las cuatro escalas de medición empleadas para obtener datos sobre una variable en particular son las escalas nominal, ordinal, de intervalo y de razón. La escala de una variable es nominal cuando los datos consisten en etiquetas o nombres que se usan para identificar un atributo de un elemento. La escala es ordinal si los datos presentan las propiedades de los datos nominales y su orden o clasificación es significativo. La escala es de intervalo si los datos tienen las propiedades de los datos ordinales y el intervalo entre los valores se expresa en términos de una unidad de medida fija. Por último, se tiene una escala de razón cuando los datos presentan todas las propiedades de los datos de intervalo y la razón de dos valores es significativa.

Para fines prácticos del análisis estadístico, los datos pueden clasificarse como categóricos o cuantitativos. Los datos categóricos utilizan etiquetas o nombres para identificar un atributo de cada elemento. Asimismo, usan las escalas de medición, ya sea nominal u ordinal, y pueden ser numéricos o no numéricos. Los datos cuantitativos son valores numéricos que indican cuánto o cuántos, y utilizan las escalas de medición de intervalo o de razón. Las operaciones aritméticas ordinarias sólo tienen sentido si los datos son cuantitativos. De ahí que los cálculos estadísticos utilizados para los datos cuantitativos no siempre sean apropiados para los categóricos.

En las secciones 1.4 y 1.5 se introducen los temas de estadística descriptiva e inferencia estadística. La estadística descriptiva abarca los métodos tabular, gráfico y numérico usados para resumir los datos. El proceso de inferencia estadística usa los datos obtenidos de una muestra para realizar estimaciones o probar hipótesis acerca de las características de una población. Las últimas tres secciones del capítulo contienen información sobre el papel de las computadoras en el análisis estadístico, una introducción al campo relativamente nuevo de la minería de datos y un resumen de los lineamientos éticos para la práctica estadística.

Glosario

Banco de datos Todos los datos recabados en un estudio en particular.
Censo Una encuesta para recabar datos sobre toda la población.
Datos Hechos y cifras recabados, analizados y resumidos para su presentación e interpretación.
Datos categóricos Etiquetas o nombres usados para identificar un atributo de cada elemento. Los datos categóricos usan las escalas de medición nominal u ordinal y pueden ser numéricos o no numéricos.
Datos cuantitativos Valores numéricos que indican cuánto o cuántos de algo. Los datos cuantitativos se obtienen usando las escalas de medición de intervalo o de razón.
Datos de corte transversal Datos recabados en el mismo o aproximadamente el mismo punto en el tiempo.
Datos de series de tiempo Datos recabados durante varios periodos.
Elemento Entidades sobre las cuales se recaban los datos.
Encuesta de muestreo Una encuesta para recabar datos sobre una muestra.
Escala de intervalo Escala de medición para una variable si los datos demuestran las propiedades de los datos ordinales y el intervalo entre los valores se expresa en términos de una unidad de medida fija. Los datos de intervalo son siempre numéricos.
Escala de razón Escala de medición para una variable si los datos demuestran todas las propiedades de los datos de intervalo y la razón de dos valores es significativa. Los datos de razón son siempre numéricos.

Escala nominal Escala de medición para una variable cuando los datos son etiquetas o nombres usados para identificar un atributo de un elemento. Los datos nominales pueden ser numéricos o no numéricos.

Escala ordinal Escala de medición para una variable si los datos exhiben las propiedades de los datos nominales, y su orden o clasificación es significativo.

Estadística El arte y la ciencia de recabar, analizar, presentar e interpretar datos.

Estadística descriptiva Resúmenes de datos en forma de tabla, gráfica y números.

Inferencia estadística Proceso de usar datos obtenidos de una muestra para efectuar estimaciones o probar hipótesis acerca de las características de una población.

Minería de datos Proceso de utilizar procedimientos de la estadística y las ciencias de la computación para extraer información útil de bases de datos sumamente grandes.

Muestra Un subconjunto de la población.

Observación Conjunto de mediciones obtenido para un elemento en particular.

Población Conjunto de todos los elementos de interés en un estudio en particular.

Variable Una característica de interés para los elementos.

Variable categórica Variable con datos categóricos.

Variable cuantitativa Variable con datos cuantitativos.

Ejercicios complementarios

1. Comente las diferencias entre la estadística como hechos numéricos y la estadística como una disciplina o campo de estudio.

AUTO evaluación

2. El Departamento de Energía de Estados Unidos proporciona información sobre economía de combustible para diversos vehículos de motor. Una muestra de 10 automóviles se presenta en la tabla 1.6 (sitio web Fuel Economy, 22 de febrero de 2008). Los datos indican el tamaño del automóvil (compacto, mediano o grande), el número de cilindros del motor, las millas por galón en la ciudad, las millas por galón en autopista y el combustible recomendado (diesel, premium o regular).
   a) ¿Cuántos elementos hay en este banco de datos?
   b) ¿Cuántas variables hay en este banco de datos?
   c) ¿Cuáles variables son categóricas y cuáles son cuantitativas?
   d) ¿Qué tipo de escala de medición se usa para cada una de las variables?

AUTO evaluación

3. Consulte la tabla 1.6.
   a) ¿Cuántas millas por galón se consumen en la ciudad?
   b) En promedio, ¿cuántas millas más por galón se obtienen al conducir en autopista que en la ciudad?

TABLA 1.6 Información sobre economía de combustible para 10 automóviles

Vehículo	Tamaño	Cilindros	MPG* Ciudad	MPG Autopista	Combustible
Audi A8	Grande	12	13	19	Premium
BMW 328Xi	Compacto	6	17	25	Premium
Cadillac CTS	Mediano	6	16	25	Regular
Chrysler 300	Grande	8	13	18	Premium
Ford Focus	Compacto	4	24	33	Regular
Hyundai Elantra	Mediano	4	25	33	Regular
Jeep Grand Cherokee	Mediano	6	17	26	Diesel
Pontiac G6	Compacto	6	15	22	Regular
Toyota Camry	Mediano	4	21	31	Regular
Volkswagen Jetta	Compacto	5	21	29	Regular

* Millas por galón

TABLA 1.7 Datos para siete colegios y universidades

Escuela	Estado	Campus	Inversión ($ miles de millones)	% de solicitantes admitidos	División de la NCAA
Amherst College	Massachusetts	Pueblo: pequeño	1.7	18	III
Duke	Carolina del Norte	Ciudad: mediana	5.9	21	I-A
Universidad de Harvard	Massachusetts	Ciudad: mediana	34.6	9	I-AA
Swarthmore College	Pennsylvania	Suburbio: grande	1.4	18	III
Universidad de Pennsylvania	Pennsylvania	Ciudad: grande	6.6	18	I-AA
Williams College	Massachusetts	Pueblo: pequeño	1.9	18	III
Universidad de Yale	Connecticut	Ciudad: mediana	22.5	9	I-AA

c) ¿Qué porcentaje de los automóviles tiene motores de cuatro cilindros?
d) ¿Qué porcentaje de los vehículos usa combustible regular?

4. La tabla 1.7 muestra datos para siete colegios y universidades considerando la inversión (en miles de millones de dólares) y el porcentaje de solicitantes admitido (*USA Today,* 3 de febrero de 2008). El estado en que se localiza cada escuela, el campus y la División de NCAA para los equipos colegiales se obtuvieron del sitio web National Center of Education Statistics (22 de febrero de 2008).
   a) ¿Cuántos elementos hay en el banco de datos?
   b) ¿Cuántas variables hay en el banco de datos?
   c) ¿Cuáles variables son categóricas y cuáles son cuantitativas?

5. Considere el banco de datos de la tabla 1.7
   a) Calcule la inversión promedio para la muestra.
   b) Calcule el porcentaje promedio de solicitantes admitidos.
   c) ¿Qué porcentaje de las escuelas tiene equipos colegiales de división III de la NCAA?
   d) ¿Qué porcentaje de las escuelas tiene un campus en una ciudad mediana?

6. La revista *Foreign Affairs* realizó una encuesta para desarrollar un perfil de sus suscriptores (sitio web Foreign Affairs, 23 de febrero de 2008). Se formularon las preguntas siguientes.
   a) ¿Cuántas noches ha permanecido en un hotel en los 12 meses anteriores?
   b) ¿Dónde adquiere sus libros? Se listaron tres opciones: Bookstore, Internet y Book Club.
   c) ¿Posee o alquila un vehículo de lujo? (Sí o No).
   d) ¿Qué edad tiene?
   e) Para los viajes al extranjero realizados en los tres años pasados, ¿cuál fue su destino? Se listaron siete destinos internacionales.

   Comente si cada pregunta proporciona datos categóricos o cuantitativos.

7. El hotel Ritz-Carlton aplicó un cuestionario de opinión del cliente para obtener datos sobre el desempeño en sus servicios de comedor y entretenimiento (The Ritz-Carlton Hotel, Naples, Florida, febrero de 2006). Se pidió a los clientes que calificaran seis factores: bienvenida, servicio, alimentos, atractivo del menú, atmósfera y experiencia general. Se registraron datos para cada factor con las calificaciones de 1 para aceptable, 2 para normal, 3 para bueno y 4 para excelente.
   a) Las respuestas de los clientes proporcionaron datos para seis variables. ¿Estas variables son categóricas o cuantitativas?
   b) ¿Qué escala de medición se utiliza?

8. El programa *FinancialTimes*/Harris Poll es una encuesta mensual en línea para adultos de seis países de Europa y Estados Unidos. Una encuesta de enero incluyó a 1 015 adultos de Estados Unidos. Una de las preguntas fue: "¿Cómo calificaría usted al Federal Bank en el manejo de

los problemas de crédito en los mercados financieros?" Las respuestas posibles fueron excelente, bueno, aceptable, malo y muy malo (sitio web Harris Interactive, enero de 2008).

a) ¿De qué tamaño fue la muestra para esta encuesta?
b) ¿Los datos son categóricos o cuantitativos?
c) ¿Tendría más sentido usar promedios o porcentajes como resumen de los datos para esta pregunta?
d) De los encuestados en Estados Unidos, 10% dijo que el Federal Bank está efectuando un buen trabajo. ¿Cuántas personas proporcionaron esta respuesta?

9. El Departamento de Comercio informó que recibió las aplicaciones siguientes para el Premio Nacional de Calidad Malcolm Baldrige: 23 de las empresas de manufactura grandes, 18 de las empresas de servicios grandes y 30 de las pequeñas empresas.
   a) ¿El tipo de empresa es una variable categórica o cuantitativa?
   b) ¿Qué porcentaje de las aplicaciones proviene de las pequeñas empresas?

10. La encuesta a suscriptores de *The Wall Street Journal* (*WSJ*) (13 de octubre de 2003) formuló 46 preguntas sobre las características y los intereses de los lectores. Determine si cada una de las preguntas siguientes proporcionó datos categóricos o cuantitativos, e indique la escala de medición apropiada para cada uno.
   a) ¿Qué edad tiene?
   b) ¿Es usted hombre o mujer?
   c) ¿Cuándo empezó a leer el *WSJ*? ¿En secundaria, bachillerato, a principios de la carrera, a mitad de la carrera, a finales de la carrera o en el retiro?
   d) ¿Cuánto tiempo lleva en su empleo o puesto actual?
   e) ¿Qué tipo de vehículo está considerando para su compra siguiente? Nueve categorías de respuesta incluyen automóviles sedán, automóviles deportivos, vehículos todo terreno, minivans, etcétera.

11. Determine si cada una de las variables siguientes es categórica o cuantitativa, e indique su escala de medición.
   a) Ventas anuales.
   b) Tamaño de bebida refrescante (pequeño, mediano, grande).
   c) Clasificación de empleados (de GS1 a GS18).
   d) Utilidades por acción.
   e) Método de pago (efectivo, cheques, tarjeta de crédito).

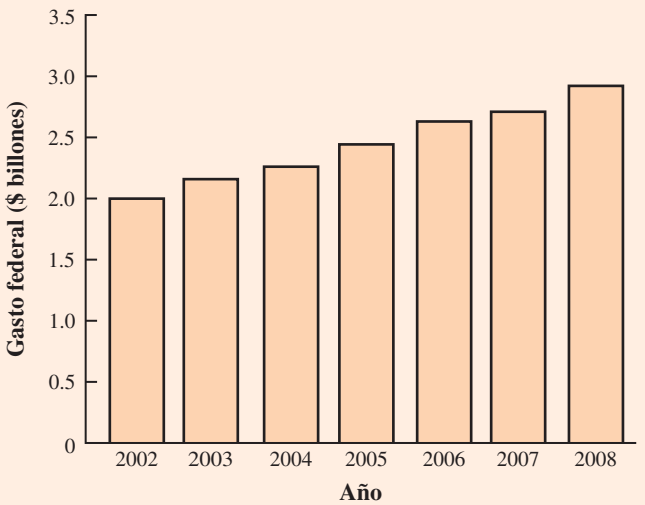
12. La agencia Hawaii Visitors Bureau recaba datos sobre los visitantes a Hawaii. Las preguntas siguientes se incluyeron entre las 16 formuladas en un cuestionario que se proporcionó a los pasajeros durante los vuelos entrantes de la aerolínea en junio de 2003.
   - Este viaje a Hawaii es mi: 1o., 2o., 3o., 4o., etcétera.
   - La razón principal para este viaje es: (10 categorías que incluyen vacaciones, convención, luna de miel).
   - Dónde planeo hospedarme: (11 categorías que incluyen hotel, departamento, parientes, acampar).
   - Días totales en Hawaii.

   a) ¿Qué población se estudia?
   b) ¿El uso de un cuestionario es una buena manera de llegar a la población de pasajeros en los vuelos de aerolíneas entrantes?
   c) Comente si cada una de las cuatro preguntas le proporcionará datos categóricos o cuantitativos.

AUTOevaluación

13. La figura 1.8 proporciona una gráfica de barras que muestra la cantidad de gasto federal para los años 2002 a 2008 (*USA Today*, 5 de febrero de 2008).
   a) ¿Cuál es la variable de interés?
   b) ¿Los datos son categóricos o cuantitativos?
   c) ¿Los datos son series de tiempo o de corte transversal?
   d) Comente sobre la tendencia en el gasto federal con respecto al tiempo.

FIGURA 1.8 Gastos federales



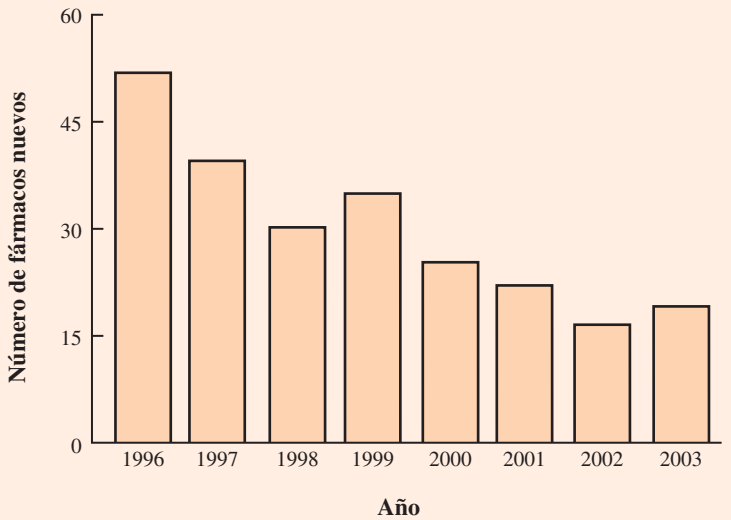
14. CSM Worldwide efectúa pronósticos de la producción global para todos los fabricantes de automóviles. Los siguientes datos de CSM muestran el pronóstico de la producción global de automóviles para General Motors, Ford, DaimlerChrysler y Toyota para los años 2004 a 2007 (*USA Today*, 21 de diciembre de 2005). Los datos están en millones de vehículos.

Fabricante	2004	2005	2006	2007
General Motors	8.9	9.0	8.9	8.8
Ford	7.8	7.7	7.8	7.9
DaimlerChrysler	4.1	4.2	4.3	4.6
Toyota	7.8	8.3	9.1	9.6

a) Elabore una gráfica de series de tiempo para los años 2004 a 2007 que muestre el número de vehículos fabricados por cada compañía automotriz. Muestre las series de tiempo para los cuatro fabricantes en la misma gráfica.
b) General Motors ha sido el líder indiscutible en la producción de automóviles desde 1931. ¿Qué muestra la gráfica de serie de tiempo sobre cuál es la compañía automotriz más importante del mundo? Comente.
c) Elabore una gráfica de barras que muestre los vehículos producidos por los fabricantes de automóviles usando los datos de 2007. ¿Esta gráfica se basa en datos de corte transversal o de series de tiempo?

15. La Food and Drug Administration (Administración de Alimentos y Fármacos, FDA) reportó el número de fármacos nuevos aprobados durante un periodo de ocho años (*The Wall Street Journal*, 12 de enero de 2004). La figura 1.9 muestra una gráfica de barras que resume el número de medicamentos nuevos aprobado cada año.
a) ¿Los datos son categóricos o cuantitativos?
b) ¿Los datos son de series de tiempo o de corte transversal?
c) ¿Cuántos medicamentos nuevos se aprobaron en 2003?
d) ¿En qué año se aprobó el menor número de fármacos nuevos? ¿Cuántos fueron?
e) Comente la tendencia en el número de medicamentos nuevos aprobados por la FDA durante el periodo de ocho años.

FIGURA 1.9 Número de fármacos nuevos aprobados por la FDA



16. La Oficina de Información del Departamento de Energía de Estados Unidos proporcionó datos de series de tiempo para el precio promedio de gasolina regular convencional en dólares por galón entre julio de 2006 y junio de 2009 (sitio web Energy Information Administration, junio de 2009). Use Internet para obtener el precio medio por galón de gasolina regular convencional desde junio de 2009.
    a) Amplíe la gráfica de la serie de tiempo mostrada en la figura 1.1 (pág. 8).
    b) ¿Qué interpretaciones puede hacer acerca del precio por galón de gasolina regular convencional desde junio de 2009?
    c) ¿La serie de tiempo sigue mostrando un incremento en verano en el precio promedio por galón? Explique por qué.
17. Un gerente de una corporación grande recomienda que se otorgue un aumento de sueldo de $10 000 para evitar que un subordinado valioso se vaya a otra empresa. ¿Qué fuentes de datos internas y externas podrían usarse para decidir si es apropiado este incremento?
18. Una encuesta a 430 viajeros de negocios reveló que 155 de ellos usaron una agencia para hacer sus arreglos de viaje (*USA Today*, 20 de noviembre de 2003).
    a) Desarrolle una estadística descriptiva que se pueda usar para estimar el porcentaje de todos los viajeros de negocios que contratan una agencia de viajes para hacer sus arreglos respectivos.
    b) La encuesta reveló que la manera más frecuente en que los viajeros de negocios hacen sus arreglos de viaje es por medio de un sitio de viajes en línea. Si 44% de los encuestados preparó sus arreglos de esta manera, ¿cuántos de los 430 viajeros de negocios usaron un sitio de viajes en línea?
    c) ¿Los datos sobre cómo se hacen los arreglos de viaje son categóricos o cuantitativos?
19. Un estudio sobre los suscriptores de *BusinessWeek* en Norteamérica recabó datos de una muestra de 2 861 clientes. El 59% de los encuestados indicó un ingreso anual de $75 000 o más, y 50% informó tener una tarjeta de crédito de American Express.
    a) ¿Cuál es la población de interés en este estudio?
    b) ¿El ingreso anual es una variable categórica o cuantitativa?
    c) ¿La propiedad de una tarjeta American Express es una variable categórica o cuantitativa?
    d) ¿Este estudio involucra datos de corte transversal o de series de tiempo?
    e) Describa cualquier inferencia estadística que *BusinessWeek* podría hacer sobre la base de la encuesta.

20. Una consulta a 131 administradores de inversiones en la encuesta Big Money de *Barron* reveló lo siguiente:
    - 43% de los gerentes se clasificó a sí mismo como a la alza o muy a la alza en el mercado de valores.
    - El rendimiento promedio esperado durante los 12 meses siguientes para los valores de renta variable fue 11.2%.
    - El 21% seleccionó la asistencia médica como el sector con más probabilidades de dirigir el mercado en los 12 meses siguientes.
    - Cuando se les pidió que estimaran cuánto tiempo tardarían las acciones de tecnología y telecomunicaciones en reanudar un crecimiento sostenido, la respuesta promedio de los gerentes fue 2.5 años.

    a) Cite dos estadísticas descriptivas.
    b) Desarrolle una inferencia sobre la población de todos los administradores de inversiones con respecto al rendimiento promedio esperado sobre los valores de renta variable durante los 12 meses siguientes.
    c) Haga una inferencia sobre el tiempo que tardarán las acciones de tecnología y telecomunicaciones en reanudar un crecimiento sostenido.

21. Un estudio de investigación médica de siete años reveló que las mujeres cuyas madres tomaron el fármaco DES durante el embarazo, respecto de las mujeres cuyas madres no tomaron el fármaco, tenían el doble de probabilidad de desarrollar anormalidades en el tejido que podrían provocar cáncer.
    a) Este estudio involucró la comparación de dos poblaciones. ¿Cuáles fueron éstas?
    b) ¿Dónde supone usted que se obtuvieron los datos: en una encuesta o en un experimento?
    c) Para la población de mujeres cuyas madres tomaron el fármaco DES durante el embarazo, una muestra de 3 980 mujeres mostró que 63 desarrollaron anormalidades en el tejido que podrían provocar cáncer. Proporcione una estadística descriptiva que podría usarse para estimar el número de mujeres por cada 1000 en esta población que presentan anormalidades en el tejido.
    d) Para la población de mujeres cuyas madres no tomaron el fármaco DES durante el embarazo, ¿cuál es la estimación del número de mujeres por cada 1000 que esperarían presentar anormalidades en el tejido?
    e) Los estudios médicos usan con frecuencia una muestra relativamente grande (en este caso, 3 980). ¿Por qué?

22. La firma Nielsen encuestó a consumidores de 47 mercados de Europa, Asia-Pacífico, el continente americano y el Oriente Medio con el propósito de establecer cuáles factores son los más importantes para determinar dónde realizan sus compras. Utilizando una escala de 1 (baja) a 5 (alta), el factor con mayor calificación fue *gran valor por su dinero*, con una calificación media de 4.32 puntos. El factor que calificó en segundo lugar fue *mejor selección de marcas y productos de alta calidad*, con una calificación promedio de 3.78 puntos, y el factor con menor calificación fue *utiliza bolsas y empaques reciclables*, con una calificación promedio de 2.71 (sitio web de Nielsen, 24 de febrero de 2008). Suponga que una cadena de tiendas de abarrotes lo contrató para que realice un estudio parecido para determinar qué factores consideran los clientes de la cadena en Charlotte, Carolina del Norte, que son los más importantes para determinar dónde efectúan sus compras.
    a) ¿Cuál es la población para la encuesta que usted realizará?
    b) ¿Cómo recolectará los datos para este estudio?

23. Nielsen Media Research efectúa encuestas semanales de los programas de televisión que se ven en todo Estados Unidos, y publica datos tanto de la audiencia como de la participación de mercado. El índice de audiencia de Nielsen es el porcentaje de familias con televisión que ve un programa, mientras que la participación de mercado es el porcentaje de familias que ve un programa entre aquellas con televisión en uso. Por ejemplo, los resultados de Nielsen Media Research para la Serie Mundial de Beisbol de 2003 entre los Yankees de Nueva York y los Marlins de Florida reportó una audiencia de 12.8% y una participación de 22% (Associated Press, 27 de octubre de 2003). Por tanto, 12.8% de las familias con televisión y 22% de las familias con televisión en uso vieron la Serie Mundial. Con base en la audiencia y la información de participación de los programas más importantes, Nielsen publica una calificación semanal tanto de programas de televisión como de las cuatro cadenas principales: ABC, CBS, NBC y Fox.
    a) ¿Qué intenta medir Nielsen Media Research?
    b) ¿Cuál es la población?
    c) ¿Por qué se usaría una muestra en esta situación?
    d) ¿Qué tipos de decisiones o acciones se basan en las calificaciones de Nielsen?

TABLA 1.8 Banco de datos para 25 acciones shadow

Company	Exchange	Ticker Symbol	Market Cap ($ millions)	Price/ Earnings Ratio	Gross Profit Margin (%)
DeWolfe Companies	AMEX	DWL	36.4	8.4	36.7
North Coast Energy	OTC	NCEB	52.5	6.2	59.3
Hansen Natural Corp.	OTC	HANS	41.1	14.6	44.8
MarineMax, Inc.	NYSE	HZO	111.5	7.2	23.8
Nanometrics Incorporated	OTC	NANO	228.6	38.0	53.3
TeamStaff, Inc.	OTC	TSTF	92.1	33.5	4.1
Environmental Tectonics	AMEX	ETC	51.1	35.8	35.9
Measurement Specialties	AMEX	MSS	101.8	26.8	37.6
SEMCO Energy, Inc.	NYSE	SEN	193.4	18.7	23.6
Party City Corporation	OTC	PCTY	97.2	15.9	36.4
Embrex, Inc.	OTC	EMBX	136.5	18.9	59.5
Tech/Ops Sevcon, Inc.	AMEX	TO	23.2	20.7	35.7
ARCADIS NV	OTC	ARCAF	173.4	8.8	9.6
Qiao Xing Universal Tele	OTC	XING	64.3	22.1	30.8
Energy West Incorporated	OTC	EWST	29.1	9.7	16.3
Barnwell Industries, Inc.	AMEX	BRN	27.3	7.4	73.4
Innodata Corporation	OTC	INOD	66.1	11.0	29.6
Medical Action Industries	OTC	MDCI	137.1	26.9	30.6
Instrumentarium Corp.	OTC	INMRY	240.9	3.6	52.1
Petroleum Development	OTC	PETD	95.9	6.1	19.4
Drexler Technology Corp.	OTC	DRXR	233.6	45.6	53.6
Gerber Childrenswear Inc.	NYSE	GCW	126.9	7.9	25.8
Gaiam, Inc.	OTC	GAIA	295.5	68.2	60.7
Artesian Resources Corp.	OTC	ARTNA	62.8	20.5	45.5
York Water Company	OTC	YORW	92.2	22.9	74.2

WEB archivo
Shadow02

24. Una muestra de las calificaciones obtenidas en los exámenes parciales de cinco estudiantes mostró los resultados siguientes: 72, 65, 82, 90 y 76. ¿Cuáles de los enunciados listados enseguida son correctos y cuáles deben considerarse demasiado generalizados?
    a) La calificación promedio de los exámenes parciales para la muestra de cinco estudiantes es 77.
    b) La calificación promedio de los exámenes parciales para los cinco estudiantes que presentaron el examen es 77.
    c) Una estimación de la calificación promedio de los exámenes parciales para todos los estudiantes que presentaron el examen es 77.
    d) Más de la mitad de los estudiantes que presentaron este examen obtendrá una calificación de entre 70 y 85.
    e) Si otros cinco estudiantes se incluyen en la muestra, obtendrán calificaciones de entre 65 y 90.

25. La tabla 1.8 muestra un banco de datos que contiene información para 25 de las acciones shadow rastreadas por la American Association of Individual Investors. Las shadow son acciones comunes de empresas pequeñas que no siguen de cerca los analistas de Wall Street. El banco de datos también está en el sitio web del libro, en el archivo llamado Shadow02.
    a) ¿Cuántas variables hay en el banco de datos?
    b) ¿Cuáles de las variables son categóricas y cuáles son cuantitativas?
    c) Para la variable Exchange (intercambio) muestre la frecuencia y la frecuencia de porcentaje de AMEX, NYSE y OTC. Elabore una gráfica de barras parecida a la de la figura 1.5 para la variable intercambio.
    d) Muestre la distribución de frecuencia para el Gross Profit Margin (margen de utilidad bruta) usando los cinco intervalos siguientes: 0–14.9, 15–29.9, 30–44.9, 45–59.9 y 60–74. Elabore un histograma parecido al de la figura 1.6.
    e) ¿Cuál es la razón promedio Price/Earnings Ratio (precio/utilidades)?

Apéndice Una introducción a StatTools

StatTools es un complemento profesional que amplía las capacidades estadísticas disponibles con Microsoft Excel. StatTools puede descargarse del sitio web de este libro.

Excel no contiene funciones o herramientas de análisis de datos para realizar todos los procedimientos estadísticos estudiados en el libro. StatTools es un complemento de estadística para Microsoft Excel que amplía la variedad de opciones estadísticas y gráficas para los usuarios de Excel. La mayoría de los capítulos incluye un apéndice al final que muestra los pasos requeridos para realizar un procedimiento estadístico usando StatTools. Para aquellos que desean hacer un uso más exhaustivo del software, StatTools ofrece una excelente función de ayuda llamada Help. Este sistema incluye explicaciones detalladas de las opciones del análisis de datos, así como descripciones y definiciones de los tipos de resultados proporcionados.

Cómo empezar a usar StatTools

Usted puede descargar StatTools desde el sitio web de este libro e instalarlo en su computadora. Después de descargar el software, realice los pasos siguientes para usarlo como complemento de Excel.

Paso 1. Haga clic en el botón **Start** en la barra de tareas y luego apunte a **All programs**.
Paso 2. Apunte a la carpeta llamada **Palisade Decision Tools**.
Paso 3. Haga clic en **StatTools for Excel**.

Estos pasos abrirán Excel y añadirán la ficha StatTools al lado de la ficha Complementos en la cinta Excel. De manera opcional, si usted ya está trabajando en Excel, estos pasos harán que StatTools aparezca como una opción disponible en el programa.

Uso de StatTools

Antes de realizar cualquier análisis estadístico, debemos crear un banco de datos de StatTools usando el Data Set Manager de este programa. Usemos la hoja de trabajo de Excel para el banco de datos de los fondos de inversión de la tabla 1.1 para mostrar cómo se hace esto. Los pasos siguientes muestran cómo crear un banco de datos de StatTools para la información de los fondos de inversión.

Paso 1. Abra el archivo de Excel llamado Morningstar.
Paso 2. Seleccione cualquier celda del banco de datos (por ejemplo, la celda A1).
Paso 3. Haga clic en la ficha **StatTools** en la cinta de opciones.
Paso 4. En el grupo **Data** haga clic en **Data Set Manager**.
Paso 5. Cuando StatTools le pregunte si quiere sumar el rango $A$1:$F$26 como un banco de datos StatTools nuevo, haga clic en **Yes**.
Paso 6. Cuando el cuadro de diálogo Data Set Manager de StatTools aparezca, haga clic en **OK**.

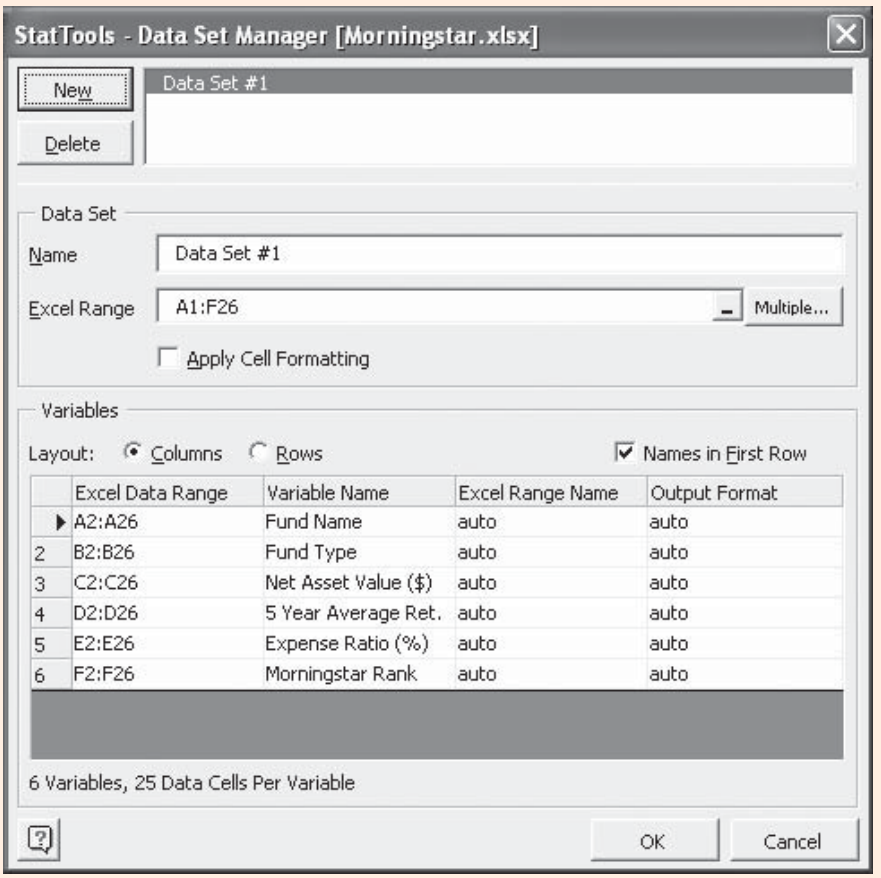
La figura 1.10 muestra el cuadro de diálogo Data Set Manager de StatTools que aparece en el paso 6. En forma predeterminada, el nombre del nuevo banco de datos de StatTools es Data Set #1. Usted puede reemplazar Data Set #1 en el paso 6 con un nombre más descriptivo. Y, si usted selecciona la opción Apply Cell Format, las etiquetas de la columna se resaltarán en azul y todo el banco de datos tendrá bordes interiores y exteriores. Usted siempre puede seleccionar el Data Set Manager en cualquier momento en su análisis para hacer este tipo de cambios.

Configuración recomendada de la aplicación

StatTools permite al usuario especificar algunas de las características de la aplicación que controlan funciones como dónde se muestra el resultado estadístico y cómo se realizan los cálculos. Los pasos siguientes muestran cómo tener acceso al cuadro de diálogo Application Settings de StatTools.

Paso 1. Haga clic en la ficha **StatTools** en la cinta de opciones.
Paso 2. En el **Tools Group**, haga clic en **Utilities**.
Paso 3. Elija **Application Settings** de la lista de opciones.

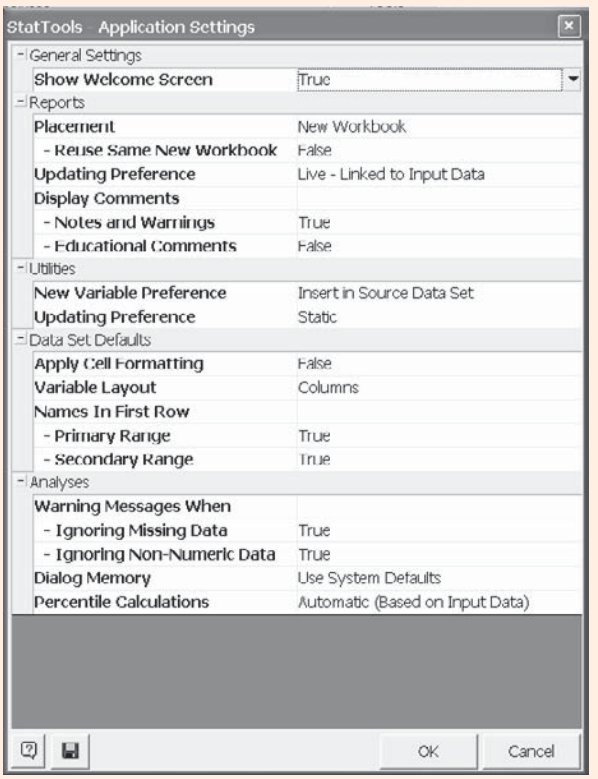
FIGURA 1.10 Cuadro de diálogo Data Set Manager de Stattools



La figura 1.11 muestra que el cuadro de diálogo StatTools-Application Settings tiene cinco secciones: General Settings, Reports, Utilities, Data Set Defaults y Analyses. Veamos cómo hacer cambios en la sección Reports del cuadro de diálogo.

La figura 1.11 muestra que la opción Placement seleccionada actualmente es **New Workbook**. Usando esta opción, el resultado de StatTools se desplegará en un libro de trabajo nuevo. Pero suponga que usted quiere desplegar el resultado de StatTools en el libro de trabajo actual (activo). Si usted hace clic en las palabras **New Workbook**, una flecha que apunta hacia abajo aparecerá a la derecha. Al hacer clic en esta flecha se despliega una lista de todas las opciones de colocación, incluido el libro de trabajo activo (**Active Workbook**); le recomendamos usar esta opción. La figura 1.11 también muestra que la opción Updating Preferences en la sección Reports actualmente es **Live–Linked to Input Data**. Con la actualización de Live, en cualquier momento uno o más valores de datos se modifican, StatTools cambia automáticamente el resultado producido previamente; también recomendamos usar esta opción. Observe que hay dos alternativas disponibles bajo Display Comments: **Notes and Warnings** (notas y advertencias) y **Educational Comments** (comentarios educativos). Dado que ambas proporcionan notas e información útiles respecto del resultado, le recomendamos usar las dos. Por tanto, para

FIGURA 1.11 Cuadro de diálogo Application Settings de StatTools



incluir comentarios educativos como parte del resultado de StatTools, usted tendrá que cambiar el valor de False para Educational Comments por verdadero (True).

El cuadro de diálogo StatTools–Application Settings contiene otras funciones que le permiten personalizar la manera en que usted desea operar StatTools. Puede aprender más sobre estas funciones al seleccionar la opción Help del grupo Tools, o al hacer clic en el icono ubicado en la esquina inferior izquierda del cuadro de diálogo. Cuando haya terminado de efectuar cambios en la configuración de la aplicación, haga clic en OK en la parte inferior del cuadro de diálogo y luego en Yes cuando StatTools le pregunte si quiere guardar la nueva configuración de la aplicación.

CAPÍTULO 2

Estadística descriptiva: presentaciones tabulares y gráficas

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: COLGATE-PALMOLIVE COMPANY

2.1 RESUMEN DE DATOS CUALITATIVOS
- Distribución de frecuencia
- Distribuciones de frecuencia relativa y frecuencia porcentual
- Gráficas de barras y circulares

2.2 RESUMEN DE DATOS CUANTITATIVOS
- Distribución de frecuencia
- Distribuciones de frecuencia relativa y frecuencia porcentual
- Diagramas de puntos
- Histograma
- Distribuciones acumuladas
- Ojiva

2.3 ANÁLISIS DE DATOS EXPLORATORIOS: EL DIAGRAMA DE TALLO Y HOJA

2.4 TABULACIONES CRUZADAS Y DIAGRAMAS DE DISPERSIÓN
- Tabulación cruzada
- La paradoja de Simpson
- Diagrama de dispersión y línea de tendencia

ESTADÍSTICA *en* LA PRÁCTICA

COLGATE-PALMOLIVE COMPANY*

NEW YORK, NEW YORK

Colgate-Palmolive inició como una tienda pequeña de jabones y velas en la ciudad de Nueva York en 1806. Hoy la empresa emplea a más de 40 000 personas que trabajan en más de 200 países y territorios de todo el mundo. Aunque es más conocida por sus marcas Colgate, Palmolive, Ajax y Fab, la empresa también comercializa los productos de Mennen, Hill's Science Diet y Hill's Prescription Diet.

Colgate-Palmolive aplica la estadística en su programa de aseguramiento de la calidad para los productos de detergente para ropa. Una preocupación radica en que el cliente quede satisfecho con la cantidad de contenido que se incluye en cada paquete. Los empaques de cada categoría se llenan con la misma cantidad de detergente en peso, pero el volumen varía dependiendo de la densidad del polvo. Por ejemplo, si ésta es alta, se necesita un volumen menor del producto para alcanzar el peso especificado en el empaque. Como resultado, cuando el consumidor lo abre, le parece que le falta detergente.

Para controlar el problema de la alta densidad del detergente en polvo, se han establecido límites aceptables para este nivel. De manera periódica se toman muestras estadísticas del producto y se mide la densidad de cada muestra. Luego se proporcionan resúmenes de los datos a los operarios para que emprendan acciones correctivas en caso necesario con el fin de mantener la densidad dentro de las especificaciones de calidad deseadas.

En la tabla y figura adjuntas se presenta una distribución de frecuencia y un histograma de las densidades de 150 muestras tomadas durante un periodo de una semana. Los niveles de densidad superiores a 0.40 son inaceptablemente altos. La distribución de frecuencia y el histograma indican que la operación cumple con las especificaciones de calidad, ya que todas las densidades son menores o iguales a 0.40. Cuando los gerentes leen estos resúmenes estadísticos quedan complacidos con la calidad del proceso de producción del detergente.

En este capítulo usted aprenderá acerca de los métodos tabulares y gráficos de la estadística descriptiva, por ejemplo las distribuciones de frecuencia, las gráficas de barras, los histogramas, los diagramas de tallo y hoja y las tabulaciones cruzadas, entre otros. El objetivo de estos métodos es resumir los datos de modo que sea fácil entenderlos e interpretarlos.

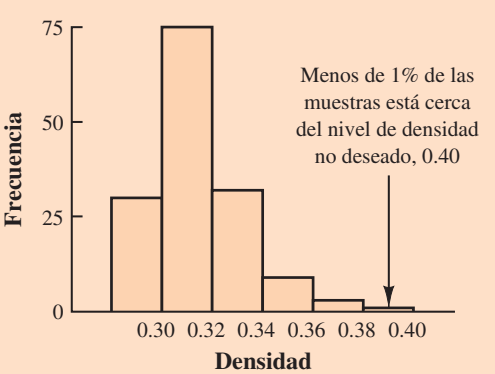


Los resúmenes gráficos ayudan a seguir la pista de la demanda de los productos de Colgate-Palmolive.
© Victor Fisher/ Bloomberg News/Landov.

Distribución de frecuencia de los datos de densidad

Densidad	Frecuencia
0.29–0.30	30
0.31–0.32	75
0.33–0.34	32
0.35–0.36	9
0.37–0.38	3
0.39–0.40	1
Total	150

Histograma de los datos de densidad



* Los autores agradecen a William R. Fowle, gerente de Aseguramiento de la Calidad de Colgate-Palmolive, por proporcionar este artículo.

Como se indicó en el capítulo 1, los datos se clasifican en cualitativos o cuantitativos. Los **datos cualitativos** utilizan etiquetas o nombres para identificar las categorías de elementos similares. Los **datos cuantitativos** son valores numéricos que indican cuánto o cuántos.

Este capítulo presenta los métodos tabulares y gráficos de uso común para resumir datos cualitativos y cuantitativos. Los resúmenes tabulares y gráficos de los datos pueden encontrarse en informes anuales, artículos periodísticos y estudios de investigación. Todos estamos expuestos a este tipo de presentaciones. Por tanto, es importante entender cómo se elaboran y cómo deben interpretarse. Comencemos con los métodos tabulares y gráficos para resumir los datos que se refieren a una sola variable. La última sección presenta los métodos para resumir datos cuando lo que interesa es la relación entre dos variables.

El software moderno para estadística cuenta con numerosas funciones para resumir datos y elaborar presentaciones gráficas. Minitab y Excel son dos paquetes que se utilizan mucho. En los apéndices del capítulo se mencionan algunas de sus funciones.

2.1 Resumen de datos cualitativos

Distribución de frecuencia

Con el fin de explicar cómo se usan los métodos tabulares y gráficos para resumir datos cualitativos, comenzaremos con la definición de **distribución de frecuencia**.

DISTRIBUCIÓN DE FRECUENCIA

Una distribución de frecuencia es un resumen tabular de datos que muestra el número (frecuencia) de elementos en cada una de varias clases que no se superponen.

Con el ejemplo siguiente se explica la elaboración e interpretación de una distribución de frecuencia para datos cualitativos. Coke Classic, Diet Coke, Dr. Pepper, Pepsi y Sprite son cinco bebidas refrescantes conocidas. Suponga que los datos de la tabla 2.1 presentan la bebida refrescante seleccionada en una muestra de 50 bebidas adquiridas.

WEB archivo
SoftDrink

TABLA 2.1 Datos de una muestra de 50 bebidas refrescantes adquiridas

Coke Classic	Sprite	Pepsi
Diet Coke	Coke Classic	Coke Classic
Pepsi	Diet Coke	Coke Classic
Diet Coke	Coke Classic	Coke Classic
Coke Classic	Diet Coke	Pepsi
Coke Classic	Coke Classic	Dr. Pepper
Dr. Pepper	Sprite	Coke Classic
Diet Coke	Pepsi	Diet Coke
Pepsi	Coke Classic	Pepsi
Pepsi	Coke Classic	Pepsi
Coke Classic	Coke Classic	Pepsi
Dr. Pepper	Pepsi	Pepsi
Sprite	Coke Classic	Coke Classic
Coke Classic	Sprite	Dr. Pepper
Diet Coke	Dr. Pepper	Pepsi
Coke Classic	Pepsi	Sprite
Coke Classic	Diet Coke	

TABLA 2.2
Distribución de frecuencia de la compra de bebidas refrescantes

Bebida refrescante	Frecuencia
Coke Classic	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5
Total	50

Para elaborar una distribución de frecuencia con estos datos, se cuenta el número de veces que cada bebida refrescante aparece en la tabla 2.1. La Coke Classic se registra 19 veces, la Diet Coke 8, Dr. Pepper 5, Pepsi 13 y Sprite 5. Estos conteos se resumen en la tabla 2.2.

Esta distribución de frecuencia proporciona un resumen de cómo se distribuyen las 50 bebidas refrescantes adquiridas entre las cinco marcas. El resumen ofrece una visión más clara que los datos originales mostrados en la tabla 2.1. Al observar la distribución de frecuencia, Coke Classic destaca como la bebida refrescante más vendida, Pepsi como la segunda, Diet Coke la tercera, y Sprite y Dr. Pepper empatan en el cuarto lugar. La distribución de frecuencia resume la información acerca de la popularidad de las cinco bebidas.

Distribuciones de frecuencia relativa y frecuencia porcentual

Una distribución de frecuencia muestra el número (la frecuencia) de elementos en cada una de varias clases que no se superponen. Sin embargo, lo que interesa a menudo es la proporción, o el porcentaje, de elementos en cada clase. La *frecuencia relativa* de una clase es igual a la fracción o proporción de elementos que pertenecen a cada clase. Para un conjunto de datos con n observaciones, la frecuencia relativa de cada clase se determina como sigue.

FRECUENCIA RELATIVA

$$\text{Frecuencia relativa de una clase} = \frac{\text{frecuencia de la clase}}{n} \qquad (2.1)$$

La *frecuencia porcentual* de una clase es la frecuencia relativa multiplicada por 100.

Una **distribución de frecuencia relativa** proporciona un resumen tabular de los datos que indica la frecuencia relativa de cada clase. Una **distribución de frecuencia porcentual** resume la frecuencia porcentual de los datos para cada clase. La tabla 2.3 muestra las distribuciones de frecuencia relativa y frecuencia porcentual para los datos de las bebidas refrescantes. En ella se aprecia que la frecuencia relativa de la Coke Classic es 19/50 = 0.38, la frecuencia relativa de la Diet Coke es 8/50 = 0.16, etc. En la distribución de frecuencia porcentual se aprecia que 38% de las bebidas refrescantes adquiridas fue de Coke Classic, 16% de Diet Coke, etc. También se observa que 38% + 26% + 16% = 80% de las bebidas refrescantes compradas fue de las tres marcas principales de la muestra.

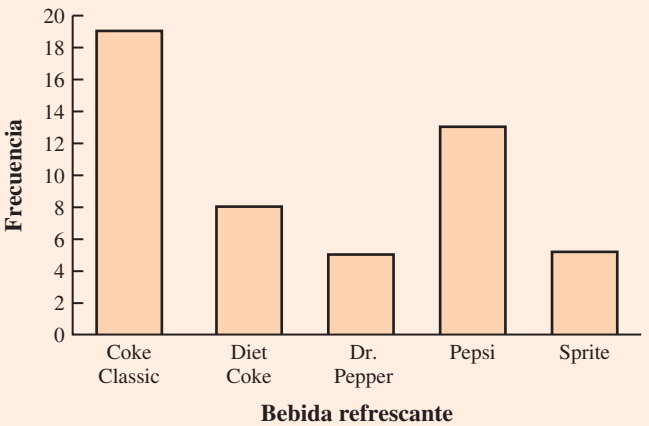
Gráficas de barras y circulares

Una **gráfica de barras** es un dispositivo gráfico que se usa para representar los datos cualitativos resumidos en una distribución de frecuencia, de frecuencia relativa o de frecuencia porcentual. En un eje de la gráfica (por lo general el horizontal) se especifican las etiquetas utilizadas para las clases (categorías). En el otro eje de la gráfica (por lo general el vertical)

TABLA 2.3 Distribuciones de frecuencia relativa y frecuencia porcentual de la compra de bebidas refrescantes

Bebida refrescante	Frecuencia relativa	Frecuencia porcentual
Coke Classic	0.38	38
Diet Coke	0.16	16
Dr. Pepper	0.10	10
Pepsi	0.26	26
Sprite	0.10	10
Total	1.00	100

FIGURA 2.1 Gráfica de barras de la compra de bebidas refrescantes

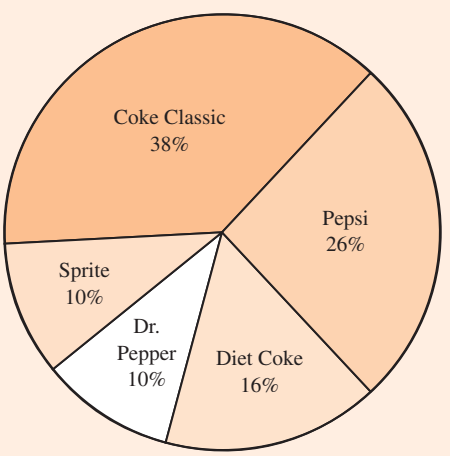


Las gráficas de barras se usan en las aplicaciones de control de calidad para identificar las principales causas de los problemas. Cuando las barras se acomodan en orden descendente de altura y de izquierda a derecha, colocando primero la causa que ocurre con más frecuencia, la gráfica de barras se llama diagrama de Pareto. *Recibe este nombre en honor de su fundador, Wilfredo Pareto, un economista italiano.*

se coloca una escala de frecuencia, de frecuencia relativa o de frecuencia porcentual. Luego se traza una barra de ancho fijo arriba de cada etiqueta de clase y se extiende su longitud hasta llegar a la frecuencia, la frecuencia relativa o la frecuencia porcentual de la clase. Para los datos cualitativos, las barras deben estar a cierta distancia para recalcar el hecho de que cada clase está separada. La figura 2.1 muestra una gráfica de barras de la distribución de frecuencia de las 50 bebidas refrescantes adquiridas. Note cómo la presentación muestra que la Coke Classic, la Pepsi y la Diet Coke son las marcas preferidas.

La **gráfica circular o de pastel** es otro dispositivo gráfico que presenta las distribuciones de frecuencia relativa y de frecuencia porcentual para datos cualitativos. Para elaborarla, primero se traza un círculo que represente todos los datos. Luego se usan las frecuencias relativas para subdividir el círculo en sectores, o partes, que corresponden a la frecuencia relativa de cada clase. Por ejemplo, como un círculo contiene 360 grados y la Coke Classic muestra una frecuencia relativa de 0.38, el sector de la gráfica circular etiquetado como Coke Classic mide 0.38(360) = 136.8 grados. El sector etiquetado como Diet Coke mide 0.16(360) = 57.6 grados. Cálculos parecidos para las demás clases producen la gráfica circular de la figura 2.2. Los

FIGURA 2.2 Gráfica circular de la compra de bebidas refrescantes



valores numéricos mostrados para cada sector pueden ser frecuencias, frecuencias relativas o frecuencias porcentuales.

NOTAS Y COMENTARIOS

1. A menudo, el número de clases en una distribución de frecuencia es igual al número de categorías encontradas en los datos, como ocurre con los de la compra de bebidas refrescantes en esta sección. Los datos se refieren sólo a cinco marcas, y para cada una se definió una clase de distribución de frecuencia separada. Si los datos hicieran referencia a todas las bebidas, se requerirían muchas categorías, la mayoría de las cuales tendría un número pequeño de bebidas refrescantes adquiridas. La mayoría de los expertos en estadística recomienda que las clases con frecuencias menores se agrupen en una clase agregada llamada "otro". Las que presentan frecuencias de 5% o menos se tratan de esta manera.
2. La suma de las frecuencias en cualquier distribución de frecuencia es siempre igual al número de observaciones. La suma de las frecuencias relativas en cualquier distribución de frecuencia relativa es siempre igual a 1.00, y la de los porcentajes en una distribución de frecuencia porcentual es siempre igual a 100.

Ejercicios

Métodos

1. La respuesta a una pregunta tiene tres alternativas: A, B y C. Una muestra de 120 respuestas proporciona 60 A, 24 B y 36 C. Muestre las distribuciones de frecuencia y de frecuencia relativa.
2. Se tiene una distribución de frecuencia relativa parcial.

Clase	Frecuencia relativa
A	0.22
B	0.18
C	0.40
D	

   a) ¿Cuál es la frecuencia relativa de la clase D?
   b) El tamaño de la muestra total es 200. ¿Cuál es la frecuencia de la clase D?
   c) Muestre la distribución de frecuencia.
   d) Muestre la distribución de frecuencia porcentual.

AUTOevaluación

3. Un cuestionario proporciona 58 respuestas Sí, 42 No y 20 sin opinión.
   a) En la elaboración de una gráfica circular, ¿cuántos grados mediría la sección del círculo que corresponde a las respuestas Sí?
   b) ¿Cuántos grados mediría la sección del círculo que corresponde a las respuestas No?
   c) Dibuje una gráfica circular.
   d) Elabore una gráfica de barras.

Aplicaciones

WEB archivo
BestTV

4. Los cuatro programas principales de televisión con mayor audiencia fueron *La ley y el orden* (*LyO*), *CSI*, *Sin rastro* y *Esposas desesperadas* (ED) (Nielsen Media Research, 1 de enero de 2007). A continuación se proporcionan los datos que incluyen los programas preferidos en una muestra de 50 espectadores.

ED	*CSI*	*ED*	*CSI*	*LyO*
Sin rastro	*CSI*	*LyO*	*Sin rastro*	*CSI*
CSI	*ED*	*Sin rastro*	*CSI*	*ED*
LyO	*LyO*	*LyO*	*CSI*	*ED*
CSI	*ED*	*ED*	*LyO*	*CSI*
ED	*Sin rastro*	*CSI*	*Sin rastro*	*ED*
ED	*CSI*	*CSI*	*LyO*	*CSI*
LyO	*CSI*	*Sin rastro*	*Sin rastro*	*ED*
LyO	*CSI*	*CSI*	*CSI*	*ED*
CSI	*ED*	*Sin rastro*	*Sin rastro*	*LyO*

a) ¿Estos datos son cualitativos o cuantitativos?
b) Proporcione las distribuciones de frecuencia y frecuencia porcentual.
c) Elabore una gráfica de barras y una gráfica circular.
d) Según la muestra, ¿qué programa de televisión tiene la mayor audiencia? ¿Cuál es el segundo?

WEB archivo
Names

5. En orden alfabético, los seis apellidos más comunes en Estados Unidos son Brown, Davis, Johnson, Jones, Smith y Williams (*The World Almanac*, 2006). Suponga que una muestra de 50 personas con uno de estos apellidos proporcionó los datos siguientes.

Brown	Williams	Williams	Williams	Brown
Smith	Jones	Smith	Johnson	Smith
Davis	Smith	Brown	Williams	Johnson
Johnson	Smith	Smith	Johnson	Brown
Williams	Davis	Johnson	Williams	Johnson
Williams	Johnson	Jones	Smith	Brown
Johnson	Smith	Smith	Brown	Jones
Jones	Jones	Smith	Smith	Davis
Davis	Jones	Williams	Davis	Smith
Jones	Johnson	Brown	Johnson	Davis

Resuma los datos mediante la elaboración de lo siguiente.
a) Distribuciones de frecuencia relativa y frecuencia porcentual.
b) Una gráfica de barras.
c) Una gráfica circular.
d) Con base en estos datos, ¿cuáles son los tres apellidos más comunes?

WEB archivo
Networks

6. El rating de audiencia televisiva de Nielsen Media Research mide el porcentaje de propietarios de un televisor que ven un programa en particular. El programa con mayor rating en la historia de la televisión estadounidense fue el último episodio especial de *M*A*S*H*, transmitido el 28 de febrero de 1983. Un rating de 60.2 indicó que 60.2% de los televidentes lo vio. Nielsen Media Research proporcionó la lista de los 50 programas con mayor rating en la historia de la televisión (*The New York Times Almanac*, 2006). Los datos siguientes muestran la cadena que produjo cada uno de los 50 programas con mayor rating.

ABC	ABC	ABC	NBC	CBS
ABC	CBS	ABC	ABC	NBC
NBC	NBC	CBS	ABC	NBC
CBS	ABC	CBS	NBC	ABC
CBS	NBC	NBC	CBS	NBC
CBS	CBS	CBS	NBC	NBC
FOX	CBS	CBS	ABC	NBC
ABC	ABC	CBS	NBC	NBC
NBC	CBS	NBC	CBS	CBS
ABC	CBS	ABC	NBC	ABC

a) Elabore una distribución de frecuencia, una distribución de frecuencia porcentual y una gráfica de barras para los datos.

b) ¿Cuál o cuáles cadenas han transmitido los programas de televisión con los mayores ratings? Compare el desempeño de ABC, CBS y NBC.

AUTO evaluación

7. Waterfront Steakhouse de Leverock, en Madeira Beach, Florida, utiliza un cuestionario para preguntar a sus clientes cómo califican el servicio, la calidad de los alimentos, los cocteles, los precios y la atmósfera del restaurante. Cada característica se califica en una escala de excepcional (O), muy bueno (V), bueno (G), promedio (A) y malo (P). Use la estadística descriptiva para resumir los datos siguientes recabados sobre la calidad de la comida. ¿Qué piensa sobre las calificaciones de la calidad en el restaurante?

G	O	V	G	A	O	V	O	V	G	O	V	A
V	O	P	V	O	G	A	O	O	O	G	O	V
V	A	G	O	V	P	V	O	O	G	O	O	V
O	G	A	O	V	O	O	G	V	A	G		

8. Los datos de una muestra de 55 miembros del Salón de la Fama de Béisbol en Cooperstown, Nueva York, se presentan enseguida. Cada observación indica la posición principal jugada por los famosos del Salón de la Fama: pitcher (P), catcher (H), primera base (1), segunda base (2), tercera base (3), parador en corto (S), jardinero izquierdo (L), jardinero central (C) y jardinero derecho (R).

L	P	C	H	2	P	R	1	S	S	1	L	P	R	P
P	P	P	R	C	S	L	R	P	C	C	P	P	R	P
2	3	P	H	L	P	1	C	P	P	P	S	1	L	R
R	1	2	H	S	3	H	2	L	P					

a) Use las distribuciones de frecuencia y de frecuencia relativa para resumir los datos.
b) ¿Qué posición proporciona el mayor número de jugadores en el Salón de la Fama?
c) ¿Qué posición proporciona el menor número de jugadores?
d) ¿Qué posición de jardinero (L, C o R) proporciona la mayoría de jugadores en el Salón de la Fama?
e) Compare los jugadores de cuadro (1, 2, 3 y S) con los jugadores de campo (L, C y R).

9. El proyecto Tendencias demográficas y sociales del Pew Research Center encontró que 46% de los adultos estadounidenses preferiría vivir en un tipo distinto de comunidad que donde reside ahora (Pew Research Center, 29 de enero de 2009). La encuesta nacional preguntó a 2 260 adultos: "¿En dónde vive ahora?" y "¿Cuál considera que es la comunidad ideal?" Las opciones de respuesta fueron ciudad (C), suburbio (S), ciudad pequeña (T) o comunidad rural (R). Una muestra de 100 personas se proporciona enseguida.

WEB archivo
LivingArea

¿En dónde vive ahora?

S	T	R	C	R	R	T	C	S	T	C	S	C	S	T
S	S	C	S	S	T	T	C	C	S	T	C	S	T	C
T	R	S	S	T	C	S	C	T	C	T	C	T	C	R
C	C	R	T	C	S	S	T	S	C	C	C	R	S	C
S	S	C	C	S	C	R	T	T	T	C	R	T	C	R
C	T	R	R	C	T	C	C	R	T	T	R	S	R	T
T	S	S	S	S	S	C	C	R	T					

¿Cuál considera que es la comunidad ideal?

S	C	R	R	R	S	T	S	S	T	T	S	C	S	T
C	C	R	T	R	S	T	T	S	S	C	C	T	T	S
S	R	C	S	C	C	S	C	R	C	T	S	R	R	R
C	T	S	T	T	T	R	R	S	C	C	R	R	S	S
S	T	C	T	T	C	R	T	T	T	C	T	T	R	R
C	S	R	T	C	T	C	C	T	T	T	R	C	R	T
T	C	S	S	C	S	T	S	S	R					

a) Proporcione una distribución de frecuencia porcentual para cada pregunta.
b) Trace una gráfica de barras para cada pregunta.
c) ¿En dónde vive ahora la mayoría de los adultos?
d) ¿Cuál considera la mayoría de los adultos que es la comunidad ideal?

e) ¿Qué cambios en las áreas habitables esperaría usted ver si las personas se mudan de donde viven actualmente a su comunidad ideal?

WEB archivo
FedBank

10. La *Financial Times/Harris* es una encuesta mensual en línea de adultos de seis países de Europa y Estados Unidos. La consulta realizada en enero de 2008 incluyó las respuestas de 1 015 adultos. Una de las preguntas formuladas fue: "¿Cómo calificaría al Federal Bank en el manejo de los problemas crediticios de los mercados financieros?" Las respuestas posibles fueron excelente, bueno, justo, malo y terrible (sitio web Harris Interactive, enero de 2008). Las 1 015 respuestas para esta pregunta pueden encontrarse en el archivo de datos llamado FedBank.
   a) Construya una distribución de frecuencia.
   b) Prepare una distribución de frecuencia porcentual.
   c) Elabore una gráfica de barras para la distribución de frecuencia porcentual.
   d) Comente cómo piensan los adultos estadounidenses que el Federal Bank está manejando los problemas crediticios en los mercados financieros.
   e) En España se preguntó a 1 114 adultos "¿Cómo calificaría usted al Banco Central Europeo en el manejo de los problemas crediticios en los mercados financieros?" La distribución de frecuencia porcentual obtenida es la siguiente.

Calificación	Frecuencia porcentual
Excelente	0
Bueno	4
Justo	46
Malo	40
Terrible	10

Compare los resultados obtenidos en España con los resultados de Estados Unidos.

2.2 Resumen de datos cuantitativos

Distribución de frecuencia

TABLA 2.4
Duración de la auditoría de fin de año (en días)

12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

Como se definió en la sección 2.1, una distribución de frecuencia es un resumen tabular de los datos que muestra el número (frecuencia) de elementos en cada una de varias clases que no se superponen. Esta definición es válida para los datos tanto cuantitativos como cualitativos. Sin embargo, con los datos cuantitativos debemos ser más cuidadosos al definir clases que no se superponen y que se utilizarán en la distribución de frecuencia.

Por ejemplo, considere los datos cuantitativos de la tabla 2.4. Éstos presentan el tiempo en días necesario para completar las auditorías de final de año para una muestra de 20 clientes de Sanderson and Clifford, una pequeña firma de contadores públicos. Los tres pasos necesarios para definir las clases de una distribución de frecuencia con los datos cuantitativos son los siguientes:

1. Determine el número de clases que no se superponen.
2. Defina el ancho de cada clase.
3. Determine los límites de clase.

WEB archivo
Audit

Ahora se demostrarán estos pasos mediante el desarrollo de una distribución de frecuencia para los datos de duración de la auditoría de la tabla 2.4.

Número de clases Las clases se forman mediante la especificación de los rangos que se usarán para agrupar los datos. Como regla general, se recomienda utilizar entre 5 y 20 clases. En el caso de un número pequeño de elementos de datos se pueden utilizar cinco o seis clases para resumir los datos. Si se tienen muchos elementos, se requiere un número grande de clases. La idea es utilizar suficientes clases para mostrar la variación en los datos, pero no demasiadas si sólo se tienen algunos elementos. Dado que el número de elementos de datos en la tabla 2.4 es relativamente pequeño ($n = 20$), se eligió elaborar una distribución de frecuencia con cinco clases.

Asignar el mismo ancho a las clases reduce la posibilidad de interpretaciones inadecuadas de los usuarios.

Ancho de clase El segundo paso en la elaboración de una distribución de frecuencia para datos cuantitativos es elegir el ancho de las clases. Como regla general, es recomendable que sea el mismo para todas. Por tanto, el número y el ancho de clase no son decisiones independientes. Un número grande de clases significa un ancho de clase menor, y viceversa. Para determinar un ancho de clase aproximado, primero se identifican los valores de datos mayores y menores. Luego, una vez especificado el número de clases deseado, se utiliza la expresión siguiente para determinar el ancho de clase aproximado.

$$\text{Ancho de clase aproximado} = \frac{\text{valor de datos mayor} - \text{valor de datos menor}}{\text{número de clases}} \qquad \textbf{(2.2)}$$

El ancho de clase aproximado que se obtiene por la ecuación (2.2) se redondea a un valor más conveniente con base en la preferencia de la persona que elabora la distribución de frecuencia. Por ejemplo, un ancho de 9.28 podría redondearse a 10, sencillamente porque 10 es un ancho de clase más adecuado para la presentación de una distribución de frecuencia.

Para los datos que involucran la duración de la auditoría al final del año, el valor de datos mayor es 33 y el valor de datos menor es 12. Dado que se decidió resumir los datos en cinco clases, usando la ecuación (2.2) se obtiene un ancho de clase aproximado de $(33 - 12)/5 = 4.2$. Por tanto, se toma la decisión de redondear y usar un ancho de clase de cinco días en la distribución de frecuencia.

Ninguna frecuencia de datos es mejor para un conjunto de datos. Distintas personas pueden elaborar distribuciones de frecuencia diferentes, pero igualmente aceptables. La meta es mostrar el agrupamiento natural y la variación en los datos.

En la práctica, el número y ancho de clases apropiados se determinan por prueba y error. Una vez que se elige un número de clases determinado, la ecuación (2.2) se usa para encontrar el ancho de clase aproximado. El proceso se repite para los diferentes números de clases. En última instancia, el analista recurre a su juicio para determinar la combinación del número y ancho de clases que proporcionan la mejor distribución de frecuencia para resumir los datos.

En el caso de los datos de duración de la auditoría de la tabla 2.4, después de decidir utilizar cinco clases, cada una con un ancho de cinco días, el paso siguiente es especificar los límites de clase para cada una.

Límites de clase Deben elegirse de modo que cada elemento de datos pertenezca a una y sólo una de las clases. El *límite de clase inferior* identifica el valor de datos menor asignado a la clase. El *límite de clase superior* identifica el valor de datos mayor asignado a la clase. En la elaboración de distribuciones de frecuencia para datos cualitativos no se necesita especificar los límites de clase, debido a que cada elemento de datos corresponde de manera natural a una clase separada. Pero con los datos cuantitativos, como en el caso de la duración de las auditorías de la tabla 2.4, se necesitan los límites de clase para determinar a dónde pertenece cada valor de datos.

Utilizando los datos de duración de la auditoría de la tabla 2.4, se selecciona 10 días como el límite de clase inferior y 14 días como el límite de clase superior de la primera clase. Ésta se denota como 10–14 en la tabla 2.5. El valor de datos menor, 12, se incluye en la clase 10–14. Luego se selecciona 15 días como el límite de clase inferior y 19 días como límite superior de la clase siguiente. Enseguida se prosigue con la definición de los límites superior e inferior para obtener un total de cinco clases: 10–14, 15–19, 20–24, 25–29 y 30–34. El valor de datos mayor, 33, se incluye en la clase 30–34. La diferencia entre los límites inferiores de las clases adyacentes es el ancho de clase. Utilizando los primeros dos límites inferiores, 10 y 15, se observa que el ancho de clase es $15 - 10 = 5$.

TABLA 2.5

Distribución de frecuencia para los datos de duración de la auditoría

Duración de la auditoría (días)	Frecuencia
10–14	4
15–19	8
20–24	5
25–29	2
30–34	1
Total	20

Una vez determinados el número, ancho y límites de clase se obtiene una distribución de frecuencia mediante el conteo del número de valores de datos que pertenecen a cada clase. Por ejemplo, los datos de la tabla 2.4 muestran que cuatro valores, 12, 14, 14 y 13, pertenecen a la clase 10–14. Por tanto, la frecuencia de la clase 10–14 es 4. Al continuar con este proceso de conteo para las clases 15–19, 20–24, 25–29 y 30–34 se obtiene la distribución de frecuencia de la tabla 2.5. Esta distribución permite observar lo siguiente:

1. Las duraciones de las auditorías que ocurren con más frecuencia están en la clase 15–19 días. Ocho de las 20 duraciones de las auditorías pertenecen a esta clase.
2. Sólo una auditoría requirió 30 o más días.

Es posible formular otras conclusiones, dependiendo de los intereses de la persona que observa la distribución de frecuencia. Su utilidad estriba en que permite comprender los datos, lo que no se logra fácilmente con la simple observación de éstos en su forma desorganizada original.

TABLA 2.6 Distribuciones de frecuencia relativa y frecuencia porcentual para los datos de duración de la auditoría

Duración de la auditoría (días)	Frecuencia relativa	Frecuencia porcentual
10–14	0.20	20
15–19	0.40	40
20–24	0.25	25
25–29	0.10	10
30–34	0.05	5
Total	1.00	100

Punto medio de clase En algunas aplicaciones se desea conocer los puntos medios de las clases en una distribución de frecuencia de datos cuantitativos. El **punto medio de clase** es el valor medio entre los límites de clase inferior y superior. En el caso de los datos de duración de la auditoría, los puntos medios de las cinco clases son 12, 17, 22, 27 y 32.

Distribuciones de frecuencia relativa y frecuencia porcentual

Las distribuciones de frecuencia relativa y de frecuencia porcentual para los datos cuantitativos se definen de la misma manera que para los datos cualitativos. Primero, recuerde que la frecuencia relativa es la proporción de las observaciones que pertenecen a una clase. Si se tienen n observaciones:

$$\text{Frecuencia relativa de la clase} = \frac{\text{frecuencia de la clase}}{n}$$

La frecuencia porcentual de una clase es la frecuencia relativa multiplicada por 100.

Con base en la frecuencia de las clases de la tabla 2.5 y el ancho $n = 20$, en la tabla 2.6 se muestra la distribución de frecuencia relativa y la distribución de frecuencia porcentual de los datos de duración de la auditoría. Observe que 0.40 de las auditorías, o 40%, requirió de 15 a 19 días, y sólo 0.05, o 5%, requirió 30 o más días. De nuevo, las interpretaciones y elementos de comprensión adicionales se obtienen usando la tabla 2.6.

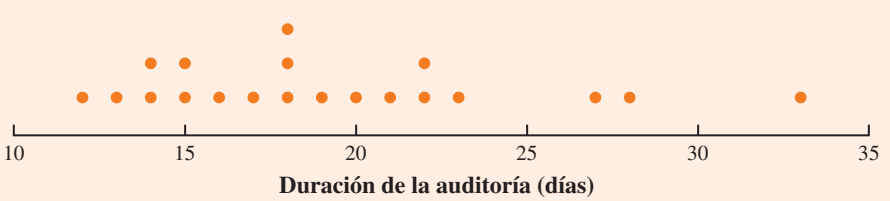
Diagrama de puntos

Uno de los resúmenes gráficos de datos más sencillos es el **diagrama de puntos**. El eje horizontal muestra el rango de los datos. Cada valor se representa por medio de un punto colocado sobre este eje. La figura 2.3 es el diagrama de puntos de los datos de duración de la auditoría de la tabla 2.4. Los tres puntos ubicados por encima de 18 sobre el eje horizontal indican que una duración de la auditoría de 18 días ocurrió tres veces. Los diagramas de puntos muestran los detalles de los datos y son útiles para comparar la distribución de los datos de dos o más variables.

Histograma

El **histograma** es una presentación gráfica común de los datos cuantitativos. Este resumen gráfico se elabora para datos previamente resumidos, ya sea mediante una distribución de frecuencia,

FIGURA 2.3 Diagrama de puntos para los datos de duración de la auditoría



de frecuencia relativa o de frecuencia porcentual. Para elaborar un histograma, la variable de interés se coloca sobre el eje horizontal y la frecuencia de que se trate, sobre el eje vertical. La frecuencia, la frecuencia relativa o la frecuencia porcentual de cada clase se indica con el trazo de un rectángulo, cuya base está determinada por los límites de clase sobre el eje horizontal, y cuya altura es la frecuencia, la frecuencia relativa o la frecuencia porcentual correspondiente.

La figura 2.4 es un histograma de los datos de duración de la auditoría. Observe que la clase con mayor frecuencia se muestra por medio del rectángulo que aparece encima de la clase 15–19 días. La altura del rectángulo indica que la frecuencia de esta clase es 8. Un histograma de la distribución de frecuencia relativa o porcentual de estos datos se ve igual que el de la figura 2.4, con la excepción de que el eje vertical se etiqueta con los valores de la frecuencia relativa o porcentual respectiva.

Como muestra esta figura, los rectángulos adyacentes de un histograma están en contacto uno con otro. A diferencia de una gráfica de barras, no hay una separación natural entre los rectángulos de las clases adyacentes. Este formato es la convención usual para los histogramas. Debido a que las clases de los datos de duración de la auditoría se establecen como 10–14, 15–19, 20–24, 25–29 y 30–34, parecería que se requieren espacios de una unidad entre las clases, de 14 a 15, de 19 a 20, de 24 a 25 y de 29 a 30. Sin embargo, estos espacios se eliminan cuando se elabora este tipo de gráfico. La eliminación de los espacios entre las clases de un histograma de los datos de duración de la auditoría sirve para mostrar que todos los valores entre el límite inferior de la primera clase y el límite superior de la última clase son posibles.

Uno de los usos más importantes del histograma es proporcionar información acerca de la forma de una distribución. La figura 2.5 presenta cuatro histogramas elaborados a partir de distribuciones de frecuencia relativa. El panel A lo muestra para un conjunto de datos moderadamente sesgado a la izquierda. Se dice que un histograma está sesgado a la izquierda si su cola se extiende más hacia esta dirección. Éste es típico para las calificaciones de exámenes: no hay calificaciones superiores a 100%, la mayoría es superior a 70%, y sólo algunas son realmente bajas. El panel B muestra el histograma para un conjunto de datos moderadamente sesgado a la derecha. Se dice que está sesgado a la derecha si su cola se extiende más hacia esta dirección. Un ejemplo de este tipo son los histogramas de datos, como los precios de las viviendas: algunas casas costosas crean el sesgo de la cola hacia la derecha.

El histograma C es simétrico. En este tipo, la cola izquierda imita la forma de la cola derecha. Los histogramas de datos que se encuentran en las aplicaciones nunca son perfectamente simétricos, pero en muchas aplicaciones lo son en forma aproximada. Con los datos de las calificaciones de la prueba SAT, las estaturas y los pesos de las personas, y así por el estilo, se generan histogramas aproximadamente simétricos. El histograma D está muy sesgado a la derecha. Éste se elaboró a partir de los datos sobre la cantidad de compras que los clientes realizaron a lo largo de un día en una tienda de ropa para dama. Con los datos de las aplicaciones a los negocios y la economía, a menudo se elaboran histogramas sesgados a la derecha. Por ejemplo,

FIGURA 2.4 Histograma de los datos de duración de la auditoría

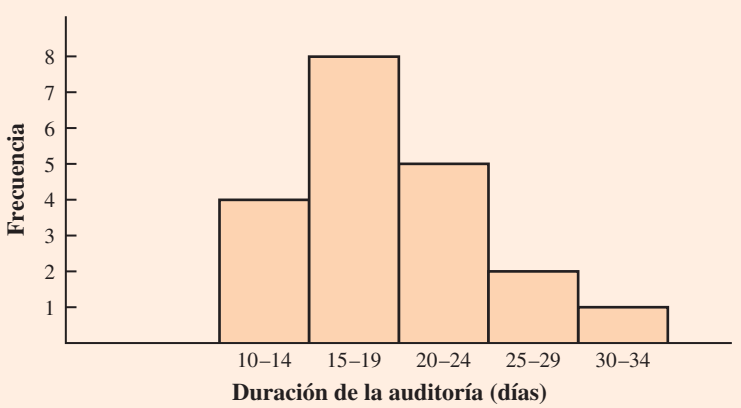
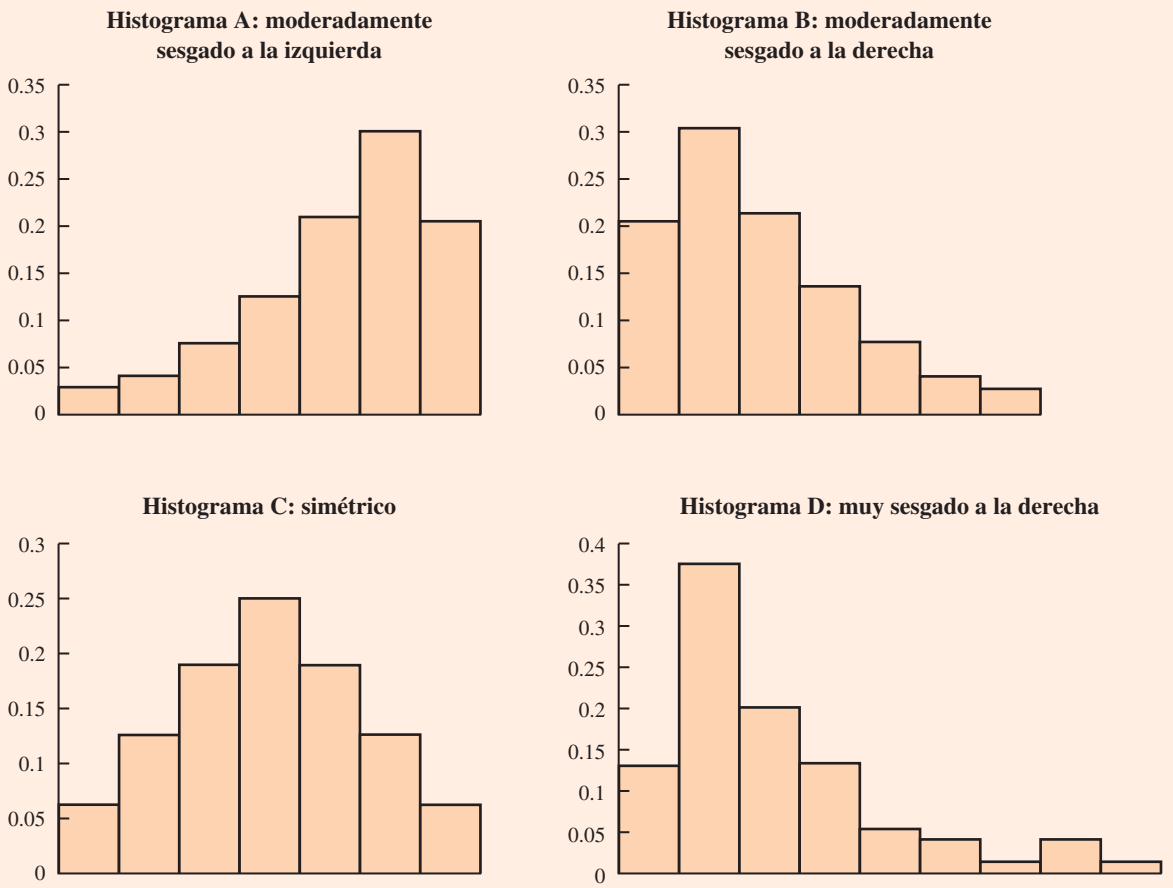


FIGURA 2.5 Histogramas que muestran diferentes niveles de sesgo



los datos sobre los precios de la vivienda, los sueldos, los montos de las compras, etc., con frecuencia generan gráficas de este tipo.

Distribuciones acumuladas

Una variación de la distribución de frecuencia que proporciona otro resumen tabular de los datos cuantitativos es la **distribución de frecuencia acumulada**. Ésta utiliza el número, los anchos y los límites de clases desarrollados para la distribución de frecuencia. Sin embargo, en vez de indicar la frecuencia de cada clase, la distribución de frecuencia acumulada muestra el número de elementos de datos con valores *menores o iguales que el límite de clase superior de cada clase*. Las primeras dos columnas de la tabla 2.7 proporcionan la distribución de frecuencia acumulada de los datos de duración de la auditoría.

Para comprender cómo se determinan las frecuencias acumuladas, considere la clase con la descripción "Menos o igual que 24". La frecuencia acumulada de esta clase es sencillamente la suma de las frecuencias de todas las clases con los valores de datos menores o iguales que 24. Para la distribución de frecuencia de la tabla 2.5, la suma de las frecuencias de las clases 10–14, 15–19 y 20–24 indica que $4 + 8 + 5 = 17$ valores de datos son menores o iguales que 24. Por consiguiente, la frecuencia acumulada para esta clase es 17. Además, la distribución de frecuencia acumulada de la tabla 2.7 indica que se completaron cuatro auditorías en 14 días o menos y 19 auditorías en 29 días o menos.

TABLA 2.7 Distribuciones de frecuencia acumulada, frecuencia relativa acumulada y frecuencia porcentual acumulada para los datos de duración de la auditoría

Duración de la auditoría (días)	Frecuencia acumulada	Frecuencia relativa acumulada	Frecuencia porcentual acumulada
Menos o igual que 14	4	0.20	20
Menos o igual que 19	12	0.60	60
Menos o igual que 24	17	0.85	85
Menos o igual que 29	19	0.95	95
Menos o igual que 34	20	1.00	100

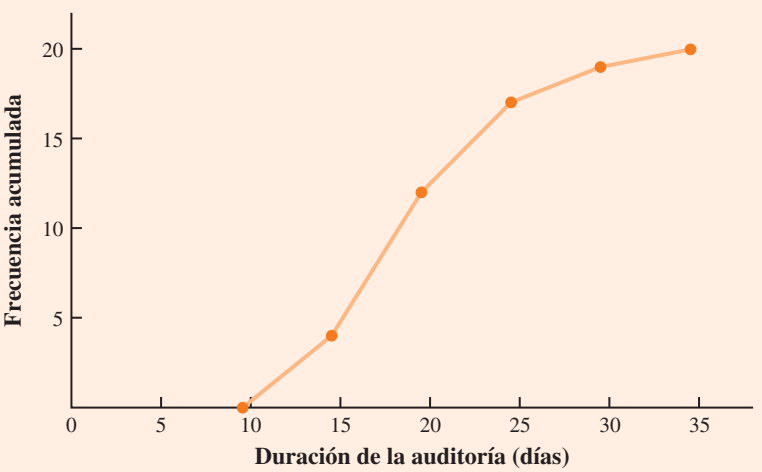
Para finalizar, observamos que una **distribución de frecuencia relativa acumulada** muestra la proporción de elementos de datos, y una **distribución de frecuencia porcentual acumulada** muestra el porcentaje de elementos de datos con valores inferiores o iguales al límite superior de cada clase. La distribución de frecuencia relativa acumulada se calcula con la suma de las frecuencias relativas en la distribución de frecuencia relativa o mediante la división de las frecuencias acumuladas entre el número total de elementos. Utilizando el último enfoque, las frecuencias relativas acumuladas de la columna 3 de la tabla 2.7 se obtuvieron al dividir las frecuencias acumuladas de la columna 2 entre el número total de elementos ($n = 20$). Las frecuencias porcentuales acumuladas se calcularon de nuevo multiplicando las frecuencias relativas por 100. Las distribuciones de frecuencias relativa y porcentual acumuladas muestran que 0.85 de las auditorías, u 85%, se completó en 24 días o menos; 0.95 de las auditorías, o 95%, se completó en 29 días o menos, y así sucesivamente.

Ojiva

La gráfica de una distribución acumulada, llamada **ojiva**, muestra los valores de datos sobre el eje horizontal, y ya sea las frecuencias acumuladas, las frecuencias relativas acumuladas o las frecuencias porcentuales acumuladas, sobre el eje vertical. La figura 2.6 ilustra la ojiva de las frecuencias acumuladas de los datos de duración de la auditoría de la tabla 2.7.

La ojiva se trazó marcando el punto correspondiente a la frecuencia acumulada de cada clase. Dado que las clases para los datos de duración de la auditoría son 10–14, 15–19, 20–24, etc., aparecen vacíos de una unidad entre 14 y 15, 19 y 20, etc. Éstos se eliminan al trazar pun-

FIGURA 2.6 Ojiva de los datos de duración de la auditoría



tos a medio camino entre los límites de clase. Por tanto, 14.5 se usa para la clase 10–14; 19.5 para la clase 15–19, y así sucesivamente. La clase "Menos o igual que 14", con una frecuencia acumulada de 4, se representa en la ojiva de la figura 2.6 por el punto localizado en 14.5 sobre el eje horizontal y 4 sobre el eje vertical. La clase "Menos o igual que 19", con una frecuencia acumulada de 12, se representa por medio del punto localizado en 19.5 sobre el eje horizontal y 12 sobre el eje vertical. Note que se trazó otro punto en el extremo izquierdo de la ojiva. Este punto la inicia, lo cual indica que debajo de la clase 10–14 no hay ningún valor de datos. Este punto se localiza en 9.5 sobre el eje horizontal y 0 sobre el eje vertical. Los puntos trazados se conectan por medio de líneas rectas para completar la ojiva.

NOTAS Y COMENTARIOS

1. Una gráfica de barras y un histograma son en esencia lo mismo; ambos son representaciones gráficas de los datos de una distribución de frecuencia. Un histograma es sólo una gráfica de barras sin separación entre éstas. Para algunos datos cuantitativos discretos es apropiada una separación entre las barras. Considere, por ejemplo, el número de materias a las cuales se inscribe un estudiante universitario. Los datos sólo pueden asumir valores enteros. Los valores intermedios como 1.5, 2.73, etc., no son posibles. No obstante, con datos cuantitativos continuos, como la duración de la auditoría de la tabla 2.4, una separación entre las barras no es adecuada.
2. Los valores apropiados para los límites de clase con datos cuantitativos dependen del nivel de precisión de éstos. Por ejemplo, con los datos de duración de la auditoría de la tabla 2.4 los límites empleados fueron valores enteros. Si se redondean a la décima más cercana de un día (p. ej., 12.3, 14.4, etc.), entonces los límites se establecerían en décimas de días. Por ejemplo, la primera clase sería 10.0–14.9. Si los datos se registraran a la centésima más cercana de un día (p. ej., 12.34, 14.45, etc.), los límites se fijarían en centésimas de días. Por ejemplo, la primera clase sería 10.00–14.99.
3. Una clase de *extremo abierto* requiere sólo un límite de clase inferior o un límite de clase superior. Por ejemplo, en los datos de la tabla 2.4, suponga que dos de las auditorías han tardado 58 y 65 días. En vez de continuar con las clases de ancho 5 con las clases 35–39, 40–44, 45–49, etc., podríamos simplificar la frecuencia de distribución para mostrar una clase de extremo abierto de "35 o más", la cual tendría una frecuencia de 2. Con más frecuencia la clase de extremo abierto aparece en el extremo superior de la distribución; a veces en el extremo inferior, y ocasionalmente en ambos extremos.
4. La última entrada de una distribución de frecuencia acumulada es siempre igual al número total de observaciones, mientras que en una distribución de frecuencia relativa acumulada siempre es igual a 1.00, a la vez que la última entrada en una distribución de frecuencia porcentual acumulada siempre es igual a 100.

Ejercicios

Métodos

11. Considere los datos siguientes.

14	21	23	21	16
19	22	25	16	16
24	24	25	19	16
19	18	19	21	12
16	17	18	23	25
20	23	16	20	19
24	26	15	22	24
20	22	24	22	20

WEB archivo
Frequency

a) Elabore una distribución de frecuencia usando las clases 12–14, 15–17, 18–20, 21–23 y 24–26.

b) Desarrolle una distribución de frecuencia relativa y una distribución de frecuencia porcentual usando las clases del inciso *a*).

AUTOevaluación

12. Considere la distribución de frecuencia siguiente.

Clase	Frecuencia
10–19	10
20–29	14
30–39	17
40–49	7
50–59	2

Elabore una distribución de frecuencia acumulada y una distribución de frecuencia relativa acumulada.

13. Elabore un histograma y una ojiva para los datos del ejercicio 12.

14. Considere los datos siguientes.

8.9	10.2	11.5	7.8	10.0	12.2	13.5	14.1	10.0	12.2
6.8	9.5	11.5	11.2	14.9	7.5	10.0	6.0	15.8	11.5

a) Elabore un diagrama de puntos.
b) Desarrolle una distribución de frecuencia.
c) Elabore una distribución de frecuencia porcentual.

Aplicaciones

AUTOevaluación

15. El personal del consultorio de un médico estudió los tiempos de espera de los pacientes que llegaron al consultorio con una solicitud de servicio de urgencia. Se recabaron los datos siguientes con los tiempos de espera en minutos durante un periodo de un mes.

2 5 10 12 4 4 5 17 11 8 9 8 12 21 6 8 7 13 18 3

Use las clases 0–4, 5–9, etc. para realizar lo siguiente:
a) La distribución de frecuencia.
b) La distribución de frecuencia relativa.
c) La distribución de frecuencia acumulada.
d) La distribución de frecuencia relativa acumulada.
e) ¿Qué proporción de pacientes que necesita un servicio de emergencia espera 9 minutos o menos?

16. Una escasez de candidatos ha requerido que se paguen sueldos más altos y se ofrezcan beneficios adicionales para atraer y retener a los superintendentes de los distritos escolares. Los datos siguientes muestran el sueldo base anual (miles) para los superintendentes en 20 distritos del área más grande de Rochester, Nueva York (*The Rochester Democrat and Chronicle*, 10 de febrero de 2008).

187	184	174	185
175	172	202	197
165	208	215	164
162	172	182	156
172	175	170	183

Use las clases 150–159, 160–169, etc. para elaborar lo siguiente.
a) La distribución de frecuencia.
b) La distribución de frecuencia porcentual.
c) La distribución de frecuencia porcentual acumulada.
d) Un histograma para el sueldo base anual.
e) ¿Los datos parecen estar sesgados? Explique por qué.
f) ¿Qué porcentaje de los superintendentes ganan más de $200 000?

17. El promedio industrial Dow Jones (DJIA) sufrió una de sus poco frecuentes reorganizaciones de empresas cuando Cisco Systems y Travelers reemplazaron a General Motors y Citigroup (*The Wall Street Journal*, 8 de junio de 2009). En ese momento, los precios por acción para las 30 empresas en el DJIA eran los que se listan en el cuadro de la siguiente página.

WEB archivo
DJIAprices

Company	$/Share	Company	$/Share
3M	61	IBM	107
Alcoa	11	Intel	16
American Express	25	J.P. Morgan Chase	35
AT&T	24	Johnson & Johnson	56
Bank of America	12	Kraft Foods	27
Boeing	52	McDonald's	59
Caterpillar	38	Merck	26
Chevron	69	Microsoft	22
Cisco Systems	20	Pfizer	14
Coca-Cola	49	Procter & Gamble	53
DuPont	27	Travelers	43
ExxonMobil	72	United Technologies	56
General Electric	14	Verizon	29
Hewlett-Packard	37	Wal-Mart Stores	51
Home Depot	24	Walt Disney	25

a) ¿Cuál es el precio por acción más alto? ¿Cuál es el precio por acción más bajo?
b) Utilizando un ancho de clase de 10, elabore una distribución de frecuencia de los datos.
c) Prepare un histograma e interprételo, incluyendo una discusión de su forma general, el rango del precio medio y el rango de precios más frecuente.
d) Consulte *The Wall Street Journal* u otra publicación para encontrar el precio actual por acción de estas empresas. Elabore un histograma de los datos y comente cualquier cambio desde junio de 2009. ¿Qué empresa ha tenido el mayor incremento en el precio por acción? ¿Cuál ha tenido la disminución más grande?

WEB archivo
Holiday

18. La investigación de NRF/BIG proporcionó los resultados de una encuesta de gastos vacacionales de los consumidores (*USA Today*, 20 de diciembre de 2005). Los datos siguientes indican la cantidad en dólares de gastos vacacionales para una muestra de 25 consumidores.

1 200	850	740	590	340
450	890	260	610	350
1 780	180	850	2 050	770
800	1 090	510	520	220
1 450	280	1 120	200	350

a) ¿Cuál es el gasto vacacional más bajo? ¿Y el más alto?
b) Utilice un ancho de clase de $250 para preparar una distribución de frecuencia y una distribución de frecuencia porcentual para los datos.
c) Elabore un histograma y comente la forma de la distribución.
d) ¿Qué observaciones puede plantear sobre los gastos vacacionales?

19. El correo electrónico no solicitado y el *spam* afectan la productividad de los empleados de oficina. Una encuesta de InsightExpress monitoreó a dichos empleados para determinar el tiempo improductivo por día dedicado a correo electrónico no solicitado y *spam* (*USA Today*, 13 de noviembre de 2003). Los datos siguientes presentan una muestra del tiempo en minutos dedicado a esta tarea.

2	4	8	4
8	1	2	32
12	1	5	7
5	5	3	4
24	19	4	14

Resuma los datos mediante la elaboración de lo siguiente:
a) Una distribución de frecuencia (clases 1–5, 6–10, 11–15, 16–20, etcétera).
b) Una distribución de frecuencia relativa.
c) Una distribución de frecuencia acumulada.
d) Una distribución de frecuencia relativa acumulada.
e) Una ojiva.
f) ¿Qué porcentaje de empleados de oficina pasó 5 minutos o menos en correo electrónico no solicitado o *spam*? ¿Qué porcentaje pasó más de 10 minutos al día en esta tarea?

20. *Golf Digest 50* publica una lista de los 50 golfistas profesionales con el ingreso anual total más alto. Éste representa la suma tanto de las ganancias en campo como de las obtenidas fuera de éste. Tiger Woods se clasificó en primer lugar con un ingreso total anual de $122 millones. Sin embargo, casi $100 millones de este total provinieron de actividades *fuera de campo*, como las promociones de productos y la apariencia personal. Los 10 golfistas profesionales con el ingreso fuera de campo más alto se muestran en la tabla siguiente (sitio web Golf Digest, febrero de 2008).

Nombre	Ingreso fuera de campo (miles de dólares)
Tiger Woods	99 800
Phil Mickelson	40 200
Arnold Palmer	29 500
Vijay Singh	25 250
Ernie Els	24 500
Greg Norman	24 000
Jack Nicklaus	20 750
Sergio Garcia	14 500
Michelle Wie	12 500
Jim Furyk	11 000

WEB archivo
OffCourse

El ingreso fuera de campo de los 50 golfistas profesionales en el *Golf Digest 50* puede encontrarse en el sitio web del libro. Los datos se proporcionan en miles de dólares. Use las clases de 0–999, 5 000–9 999, 10 000–14 999, etc. para responder las preguntas siguientes. Incluya una clase de extremo abierto de 50 000 o más como la clase de ingresos más grande.

a) Elabore una distribución de frecuencia y una distribución de frecuencia porcentual del ingreso anual fuera de campo de los 50 golfistas profesionales.
b) Elabore un histograma para estos datos.
c) Comente la forma de la distribución del ingreso fuera de campo.
d) ¿Cuál es la clase de ingresos fuera de campo más frecuente para los 50 golfistas profesionales? Usando sus resúmenes tabulares y gráficos, ¿qué observaciones adicionales puede hacer sobre el ingreso fuera de campo de estos deportistas?

21. El informe *Nielsen Home Technology Report* proporciona información sobre tecnología en el hogar y su uso. Los datos siguientes registran las horas de uso de computadoras personales durante una semana para una muestra de 50 personas.

WEB archivo
Computer

4.1	1.5	10.4	5.9	3.4	5.7	1.6	6.1	3.0	3.7
3.1	4.8	2.0	14.8	5.4	4.2	3.9	4.1	11.1	3.5
4.1	4.1	8.8	5.6	4.3	3.3	7.1	10.3	6.2	7.6
10.8	2.8	9.5	12.9	12.1	0.7	4.0	9.2	4.4	5.7
7.2	6.1	5.7	5.9	4.7	3.9	3.7	3.1	6.1	3.1

Resuma los datos al elaborar lo siguiente:

a) Una distribución de frecuencia (utilice un ancho de clase de 3 horas).
b) Una distribución de frecuencia relativa.
c) Un histograma.
d) Una ojiva.
e) Comente qué indican los datos sobre el uso de computadoras personales en casa.

2.3 Análisis de datos exploratorios: el diagrama de tallo y hoja

Las técnicas del **análisis de datos exploratorio** consisten en una aritmética simple y gráficas fáciles de elaborar que pueden usarse para resumir los datos rápidamente. Una de ellas, conocida como **diagrama de tallo y hoja**, puede utilizarse para mostrar simultáneamente tanto la clasificación como la forma de un conjunto de datos.

TABLA 2.8 Número de preguntas respondidas correctamente en una prueba de aptitudes

112	72	69	97	107
73	92	76	86	73
126	128	118	127	124
82	104	132	134	83
92	108	96	100	92
115	76	91	102	81
95	141	81	80	106
84	119	113	98	75
68	98	115	106	95
100	85	94	106	119

WEB archivo
ApTest

Para ilustrar el uso de un diagrama de tallo y hoja, considere los datos de la tabla 2.8. Éstos fueron obtenidos de una prueba de aptitudes de 150 preguntas a 50 personas entrevistadas recientemente para un puesto en Haskens Manufacturing. Los datos indican el número de preguntas respondidas correctamente.

Para desarrollar un diagrama de tallo y hoja, primero se colocan los dígitos principales de cada valor de datos a la izquierda de una línea vertical. A la derecha de dicha línea se registra el último dígito para cada valor de datos. Con base en la fila superior de los datos de la tabla 2.8 (112, 72, 69, 97 y 107), las primeras cinco entradas en la construcción de un diagrama de tallo y hoja serían como sigue:

6	9
7	2
8	
9	7
10	7
11	2
12	
13	
14	

Por ejemplo, el valor del dato 112 muestra los dígitos principales 11 a la izquierda de la línea y el último dígito 2 a la derecha. De modo parecido, el valor de datos 72 muestra el dígito principal 7 a la izquierda y el último dígito 2 a la derecha de la línea. Si se continúa colocando el último dígito de cada valor de datos en la línea que corresponde a sus dígitos principales, el resultado es el siguiente.

6	9 8
7	2 3 6 3 6 5
8	6 2 3 1 1 0 4 5
9	7 2 2 6 2 1 5 8 8 5 4
10	7 4 8 0 2 6 6 0 6
11	2 8 5 9 3 5 9
12	6 8 7 4
13	2 4
14	1

Con esta organización de los datos, la colocación de los dígitos en cada línea en el orden de clasificación es simple. Hacerlo así proporciona el diagrama de tallo y hoja mostrado aquí.

6	8 9
7	2 3 3 5 6 6
8	0 1 1 2 3 4 5 6
9	1 2 2 2 4 5 5 6 7 8 8
10	0 0 2 4 6 6 6 7 8
11	2 3 5 5 8 9 9
12	4 6 7 8
13	2 4
14	1

Los números a la izquierda de la línea vertical (6, 7, 8, 9, 10, 11, 12, 13 y 14) forman el *tallo*, y cada dígito a la derecha de la línea es una *hoja*. Por ejemplo, considere la primera fila con un valor de tallo de 6 y las hojas de 8 y 9.

6	8 9

Esta fila indica que los dos valores de datos tienen un primer dígito 6. Las hojas muestran que los valores de datos son 68 y 69. De modo parecido, la segunda fila

7	2 3 3 5 6 6

indica que los seis valores de datos tienen un primer dígito 7. Las hojas muestran que los valores de datos son 72, 73, 73, 75, 76 y 76.

Para enfocarse en la forma indicada por el diagrama de tallo y hoja, usemos un rectángulo que contenga las hojas de cada tallo. Al hacerlo así se obtiene lo siguiente.

6	8 9
7	2 3 3 5 6 6
8	0 1 1 2 3 4 5 6
9	1 2 2 2 4 5 5 6 7 8 8
10	0 0 2 4 6 6 6 7 8
11	2 3 5 5 8 9 9
12	4 6 7 8
13	2 4
14	1

La rotación de esta página en sentido contrario a las manecillas del reloj sobre su lado proporciona una imagen de los datos similar a un histograma con las clases 60–69, 70–79, 80–89, etcétera.

Aunque puede parecer que el diagrama de tallo y hoja ofrece la misma información que un histograma, tiene dos ventajas principales.

1. El diagrama de tallo y hoja es más fácil de elaborar a mano.
2. Dentro de un intervalo de clase, proporciona más información que el histograma, debido a que el tallo y la hoja muestran los datos actuales.

Justamente como una distribución de frecuencia o un histograma no tienen un número absoluto de clases, ningún diagrama de tallo y hoja tiene un número absoluto de fila o tallos. Si creemos que nuestro diagrama original condensó demasiado los datos, podemos extenderlo fácilmente usando dos o más tallos para cada dígito principal. Por ejemplo, para usar dos tallos por cada

En un diagrama de tallo y hoja extendido, siempre que un valor de tallo se establece dos veces, el primer valor corresponde a los valores de hoja de 0–4, y el segundo valor corresponde a los valores de hoja de 5–9.

dígito principal, colocaríamos todos los valores de datos que terminan en 0, 1, 2, 3 y 4 en una fila, y todos los valores que terminan en 5, 6, 7, 8 y 9 en una segunda fila. El diagrama de tallo y hoja siguiente ilustra este enfoque.

6	8 9
7	2 3 3
7	5 6 6
8	0 1 1 2 3 4
8	5 6
9	1 2 2 2 4
9	5 5 6 7 8 8
10	0 0 2 4
10	6 6 6 7 8
11	2 3
11	5 5 8 9 9
12	4
12	6 7 8
13	2 4
13	
14	1

Observe que los valores 72, 73 y 73 tienen hojas en el rango de 0–4 y se muestran con el primer valor de tallo de 7. Los valores 75, 76 y 76 tienen hojas en el rango de 5–9 y se registran con el segundo valor de tallo de 7. Este diagrama de tallo y hoja extendido es parecido a una distribución de frecuencia con intervalos de 65–69, 70–74, 75–79, etcétera.

El ejemplo anterior mostró un diagrama de tallo y hoja para los datos con hasta tres dígitos. Este tipo de diagramas para datos con más de tres dígitos es posible. Por ejemplo, considere los datos siguientes sobre el número de hamburguesas vendidas por un restaurante de comida rápida durante cada una de 15 semanas.

1 565	1 852	1 644	1 766	1 888	1 912	2 044	1 812
1 790	1 679	2 008	1 852	1 967	1 954	1 733	

Un diagrama de tallo y hoja de estos datos se presenta a continuación.

Unidad de hoja = 10

15	6
16	4 7
17	3 6 9
18	1 5 5 8
19	1 5 6
20	0 4

Un solo dígito se utiliza para definir cada hoja en un diagrama de tallo y hoja. La unidad de hoja indica cómo multiplicar los números de tallo y hoja con la finalidad de aproximar los datos originales. Las unidades de hoja pueden ser 100, 10, 1, 0.1, etcétera.

Observe que se usa un solo dígito para definir cada hoja y que sólo los primeros tres dígitos de cada valor de datos se han utilizado para construir el diagrama. En la parte superior del diagrama hemos especificado que la unidad de hoja = 10. Para ilustrar cómo interpretar los valores, considere el primer tallo, 15, y su hoja asociada, 6. Al combinar estos números, obtenemos 156. Para reconstruir una aproximación de los valores de datos originales, debemos multiplicar este número por 10, el valor de la *unidad de hoja*. Por tanto, $156 \times 10 = 1\,560$ es una aproximación del valor de datos original utilizado para construir el diagrama de tallo y hoja. Aunque no es posible reconstruir el valor de datos exacto a partir de este diagrama, la convención de manejar un sólo dígito para cada hoja permite que el diagrama se construya para datos que tienen muchos dígitos. Para diagramas donde la unidad de hoja no se muestra, se asume que esta unidad es igual a 1.

Ejercicios

Métodos

22. Elabore un diagrama de tallo y hoja para los datos siguientes.

70	72	75	64	58	83	80	82
76	75	68	65	57	78	85	72

AUTOevaluación

23. Diseñe un diagrama de tallo y hoja para los datos siguientes.

11.3	9.6	10.4	7.5	8.3	10.5	10.0
9.3	8.1	7.7	7.5	8.4	6.3	8.8

24. Elabore un diagrama de tallo y hoja para los datos siguientes. Use una unidad de hoja de 10.

1 161	1 206	1 478	1 300	1 604	1 725	1 361	1 422
1 221	1 378	1 623	1 426	1 557	1 730	1 706	1 689

Aplicaciones

AUTOevaluación

25. Una psicóloga desarrolló una nueva prueba de inteligencia para adultos, la cual se aplicó a 20 individuos; se obtuvieron los datos siguientes.

114	99	131	124	117	102	106	127	119	115
98	104	144	151	132	106	125	122	118	118

Elabore un diagrama de tallo y hoja para los datos.

26. La Asociación Estadounidense de Inversionistas Individuales realiza una encuesta anual de corredores de descuento. Los precios siguientes corresponden a una muestra de 24 corredores o (brokers) (*AAII Journal*, enero de 2003). Los dos tipos de negociaciones son una negociación asistida por un corredor de 100 acciones a $50 por acción y una negociación en línea de 500 acciones a $50 por acción.

WEBarchivo

Broker

Broker	Broker-Assisted 100 Shares at $50/Share	Online 500 Shares at $50/Share	Broker	Broker-Assisted 100 Shares at $50/Share	Online 500 Shares at $50/Share
Accutrade	30.00	29.95	Merrill Lynch Direct	50.00	29.95
Ameritrade	24.99	10.99	Muriel Siebert	45.00	14.95
Bank of America	54.00	24.95	NetVest	24.00	14.00
Brown & Co.	17.00	5.00	Recom Securities	35.00	12.95
Charles Schwab	55.00	29.95	Scottrade	17.00	7.00
CyberTrader	12.95	9.95	Sloan Securities	39.95	19.95
E*TRADE Securities	49.95	14.95	Strong Investments	55.00	24.95
First Discount	35.00	19.75	TD Waterhouse	45.00	17.95
Freedom Investments	25.00	15.00	T. Rowe Price	50.00	19.95
Harrisdirect	40.00	20.00	Vanguard	48.00	20.00
Investors National	39.00	62.50	Wall Street Discount	29.95	19.95
MB Trading	9.95	10.55	York Securities	40.00	36.00

a) Redondee los precios comerciales al dólar más cercano y elabore un diagrama de tallo y hoja para 100 acciones a $50 por acción. Comente qué aprendió acerca de los precios negociados asistidos por un corredor.

b) Redondee los precios negociados al dólar más cercano y elabore un diagrama de tallo y hoja extendido para 500 acciones en línea a $50 por acción. Comente qué aprendió sobre los precios negociados en línea.

27. La mayoría de las estaciones de esquí ofrece programas familiares que proporcionan instrucción de esquí y *snowboard* para niños. Las clases típicas constan de cuatro a seis horas en la nieve con un instructor certificado. A continuación se presenta la tarifa diaria para una lección de grupo de 15 estaciones (*The Wall Street Journal*, 20 de enero de 2006).

Estación	Ubicación	Tarifa diaria	Estación	Ubicación	Tarifa diaria
Beaver Creek	Colorado	$137	Okemo	Vermont	$ 86
Deer Valley	Utah	115	Park City	Utah	145
Diamond Peak	California	95	Butternut	Massachusetts	75
Heavenly	California	145	Steamboat	Colorado	98
Hunter	Nueva York	79	Stowe	Vermont	104
Mammoth	California	111	Sugar Bowl	California	100
Mount Sunapee	New Hampshire	96	Whistler-Blackcomb	British Columbia	104
Mount Bachelor	Oregon	83			

a) Elabore un diagrama de tallo y hoja para los datos.
b) Interprete el diagrama en términos de qué indica la tarifa diaria para los programas de instrucción de esquí y *snowboard*.

28. En el minimaratón de Naples, Florida (13.1 millas), de 2004 se registraron 1 228 corredores (*Naples Daily News*, 17 de enero de 2004). La competencia se celebró en seis grupos de edades. Los datos siguientes muestran las edades de 40 individuos que participaron en la carrera.

WEB archivo
Marathon

49	33	40	37	56
44	46	57	55	32
50	52	43	64	40
46	24	30	37	43
31	43	50	36	61
27	44	35	31	43
52	43	66	31	50
72	26	59	21	47

a) Muestre un diagrama de tallo y hoja.
b) ¿Qué grupo de edad tuvo el mayor número de corredores?
c) ¿Qué edad se registró con mayor frecuencia?
d) Un artículo de portada del *Naples Daily News* destacó el número de corredores de "veintitantos". ¿Qué porcentaje de deportistas estaba en este grupo de edad? ¿Cuál se supone que fue el enfoque del artículo?

2.4

Tabulaciones cruzadas y diagramas de dispersión

Las tabulaciones cruzadas y los diagramas de dispersión se usan para resumir datos de una manera en que revelan la relación entre dos variables.

Las tabulaciones cruzadas y los diagramas de dispersión se utilizan para resumir datos de una manera en que revelan la relación entre dos variables. Hasta ahora este capítulo se ha centrado en los métodos tabular y gráfico utilizados con el fin de resumir los datos para una *variable a la vez*. A menudo un gerente o quien toma decisiones requiere métodos de esa índole que le ayuden a comprender la *relación entre dos variables*. La tabulación cruzada y los diagramas de dispersión son dos métodos de este tipo.

Tabulación cruzada

Una **tabulación cruzada** es un resumen de los datos para dos variables. Ilustremos su uso al considerar la aplicación siguiente con base en los datos de Zagat's Restaurant Review. Se reunieron los datos sobre la calificación de calidad y precio de las comidas para una muestra de 300 restaurantes ubicados en el área de Los Ángeles. La tabla 2.9 muestra los datos para los primeros 10 negocios. Se informan los datos sobre la calificación de calidad de un restaurante y el precio de comidas típico. La calificación de calidad es una variable cualitativa con categorías de calificación de bueno, muy bueno y excelente. El precio de la comida es una variable cuantitativa que varía de $10 a $49.

Una tabulación cruzada de los datos para esta aplicación se muestra en la tabla 2.10. Las etiquetas del margen superior izquierdo definen las clases para las dos variables. En el margen izquierdo, las etiquetas de las filas buena, muy buena y excelente corresponden a las tres clases de la variable de calificación de la calidad. En el margen superior, las etiquetas de columna

TABLA 2.9 Calificación de calidad y precio de la comida para 300 restaurantes de Los Ángeles

WEB archivo
Restaurante

Restaurant	Quality Rating	Meal Price ($)
1	Buena	18
2	Muy buena	22
3	Buena	28
4	Excelente	38
5	Muy buena	33
6	Buena	28
7	Muy buena	19
8	Muy buena	11
9	Muy buena	23
10	Buena	13
.	.	.
.	.	.
.	.	.

($10–19, $20–29, $30–39 y $40–49) corresponden a las cuatro clases de la variable precio de la comida. Cada restaurante de la muestra exhibe una calificación de calidad y un precio de alimentos. Por tanto, cada uno se asocia con una celda que aparece en una de las filas y en una de las columnas de la tabulación cruzada. Por ejemplo, el restaurante 5 se identifica con una calificación de muy buena calidad y un precio de comida de $33. Este negocio pertenece a la celda que intercepta la fila 2 y la columna 3 de la tabla 2.10. Al elaborar una tabulación cruzada, sencillamente contamos el número de restaurantes que pertenece a cada una de las celdas en la tabla.

Al revisar la tabla 2.10 observamos que el mayor número de restaurantes en la muestra (64) tiene calificación de muy buena calidad y un precio de comida en el rango de $20–29. Sólo dos tienen una calificación de excelente y un precio de alimentos en el rango de $10-19. Se pueden hacer interpretaciones parecidas de otras frecuencias. Además, observe que los márgenes derecho e inferior de la tabulación cruzada proporcionan por separado la distribución de frecuencia para la calificación de la calidad y el precio de la comida. De la distribución de frecuencia en el margen derecho, observamos que los datos sobre las calificaciones de calidad muestran 84 restaurantes buenos, 150 muy buenos y 66 excelentes. De modo parecido, el margen inferior ilustra la distribución de frecuencia para la variable del precio de la comida.

Al dividir los totales en el margen derecho de la tabulación cruzada entre el total para esa columna, se obtienen las distribuciones de frecuencia relativa y porcentual para la variable de calificación de la calidad.

Calificación de la calidad	Frecuencia relativa	Frecuencia porcentual
Buena	0.28	28
Muy buena	0.50	50
Excelente	0.22	22
Total	1.00	100

TABLA 2.10 Tabulación cruzada de la calificación de calidad y el precio de la comida para 300 restaurantes de Los Ángeles

Calificación de calidad	Precio de la comida				
	$10–19	$20–29	$30–39	$40–49	Total
Buena	42	40	2	0	84
Muy buena	34	64	46	6	150
Excelente	2	14	28	22	66
Total	78	118	76	28	300

De la distribución de frecuencia porcentual, vemos que 28% de los restaurantes fue calificado como bueno, 50% muy bueno y 22% excelente.

Al dividir los totales en la fila inferior de la tabulación cruzada entre el total para esa fila se obtiene una distribución de frecuencia relativa y otra porcentual para la variable del precio de la comida.

Precio de la comida	Frecuencia relativa	Frecuencia porcentual
$10–19	0.26	26
$20–29	0.39	39
$30–39	0.25	25
$40–49	0.09	9
Total	1.00	100

Note que la suma de los valores de cada columna no se adiciona exactamente al total de la columna debido a que los valores se redondean. De la distribución de frecuencia porcentual vemos que 26% de los precios de la comida está en la clasificación más baja ($10–19), 39% en la siguiente clase más alta, etcétera.

Las distribuciones de frecuencia relativa y porcentual elaboradas a partir de los márgenes proporcionan información acerca de cada una de las variables en forma individual, pero no arrojan luz sobre la relación entre las variables. El valor principal de una tabulación cruzada radica en los elementos que ofrece para comprender esta relación. Una revisión de la tabulación cruzada en la tabla 2.10 revela que los precios más altos de la comida se asocian con los restaurantes de mayor calidad, y los precios más bajos se ligan con los de menor calidad.

La conversión de las entradas en una tabulación cruzada en porcentajes de fila y de columna puede proporcionar más elementos para comprender la relación entre las dos variables. Para los porcentajes de fila, los resultados de dividir cada frecuencia entre su total de filas correspondiente en la tabla 2.10, se presentan en la tabla 2.11. Cada fila de esta tabla es una distribución de frecuencia porcentual del precio de la comida para una de las categorías de la calificación de calidad. De los restaurantes con la calificación de calidad menor (buena), los porcentajes mayores son para los menos caros (50% tiene precios de comida de $10–19, y 47.6% de $20–29). De los restaurantes con la calificación de calidad más alta (excelente), los porcentajes mayores son para los más caros (42.4% tiene precios de comida de $30–39, y 33.4% de $40–49). Por tanto, seguimos observando que los alimentos más caros se asocian con los restaurantes de mayor calidad.

La tabulación cruzada es de uso común al examinar la relación entre dos variables. En la práctica, los informes finales para muchos estudios estadísticos incluyen un número grande de tablas de este tipo. En la encuesta de restaurantes de Los Ángeles, la tabulación cruzada se basa en una variable cualitativa (calificación de la calidad) y una variable cuantitativa (precio de la comida). Las tabulaciones cruzadas también pueden elaborarse cuando ambas variables son cuantitativas y cuando ambas variables son cualitativas. Sin embargo, cuando se usan las cuantitativas, primero se deben crear clases para los valores de la variable. Por ejemplo, en el caso de los restaurantes agrupamos los precios de la comida en cuatro clases ($10–19, $20–29, $30–39 y $40–49).

TABLA 2.11 Porcentajes de fila para cada categoría de calificación de la calidad

Calificación de calidad	Precio de la comida				
	$10–19	$20–29	$30–39	$40–49	Total
Buena	50.0	47.6	2.4	0.0	100
Muy buena	22.7	42.7	30.6	4.0	100
Excelente	3.0	21.2	42.4	33.4	100

La paradoja de Simpson

Los datos en dos o más tabulaciones cruzadas se combinan o se agrupan con frecuencia para producir una tabulación cruzada de resumen que muestra cómo se relacionan las variables. En estos casos, debemos tener cuidado al formular una conclusión con base en los datos agregados, ya que ésta puede invertirse si estudiamos los datos no agregados. La revocación de las conclusiones basada en los datos agregados y no agregados se llama *paradoja de Simpson*. Para ilustrar la **paradoja de Simpson** considere un ejemplo que involucra el análisis de los veredictos de dos jueces en dos cortes distintas.

Los jueces Ron Luckett y Dennis Kendall presidieron los casos llevados en el tribunal de primera instancia y en el tribunal municipal en los tres años pasados. Algunos de los veredictos que emitieron se revocaron. En la mayoría de estos casos el tribunal de apelaciones ratificó las sentencias originales, pero en otros las revocó. Para cada juez se elaboró una tabulación cruzada con base en dos variables: el veredicto (ratificado o revocado) y el tipo de tribunal (primera instancia y municipal). Suponga que las dos tabulaciones cruzadas se combinaron al agregar los datos del tipo de tribunal. La tabulación cruzada agregada resultante contiene dos variables: el veredicto (ratificado o revocado) y el juez (Luckett o Kendall). Esta tabulación muestra el número de apelaciones en las cuales se ratificó la sentencia, el número y en las cuales se revocó para ambos jueces. La tabla siguiente ilustra estos resultados junto con los porcentajes de columna en el paréntesis al lado de cada valor.

	Juez		
Veredicto	Luckett	Kendall	Total
Ratificado	129 (86%)	110 (88%)	239
Revocado	21 (14%)	15 (12%)	36
Total (%)	150 (100%)	125 (100%)	275

Una revisión de los porcentajes de la columna muestra que 86% de los veredictos se ratificó para el juez Luckett, mientras que 88% se ratificó para el juez Kendall. A partir de esta tabulación cruzada agregada concluimos que Kendall está efectuando un mejor trabajo debido a que un mayor porcentaje de sus sentencias ha sido ratificado.

Las tabulaciones cruzadas no agregadas muestran los casos juzgados por Luckett y Kendall en cada tribunal; los porcentajes de columna se registran en el paréntesis al lado de cada valor.

	Juez Luckett		
Veredicto	Tribunal de primera instancia	Tribunal municipal	Total
Ratificado	29 (91%)	100 (85%)	129
Revocado	3 (9%)	18 (15%)	21
Total (%)	32 (100%)	118 (100%)	150

	Juez Kendall		
Veredicto	Tribunal de primera instancia	Tribunal municipal	Total
Ratificado	90 (90%)	20 (80%)	110
Revocado	10 (10%)	5 (20%)	15
Total (%)	100 (100%)	25 (100%)	125

A partir de la tabulación cruzada y los porcentajes de columna para el juez Luckett, observamos que los veredictos se ratificaron en 91% de los casos del tribunal de primera instancia y en 85% de los casos del tribunal municipal. De la tabulación cruzada y los porcentajes de columna para Kendall, los veredictos se mantienen en 90% de los casos del tribunal de primera instancia y en 80% de los correspondientes al tribunal municipal. Por tanto, cuando desagregamos los datos, observamos que Luckett tiene un mejor registro debido a que el mayor porcentaje de sus veredictos se mantiene en ambos tribunales. Este resultado contradice la conclusión a la que llegamos con la tabulación cruzada de los datos agregados que mostraron que Kendall tenía un mejor registro. Esta revocación de las conclusiones con base en los datos agregados y desagregados ilustra la paradoja de Simpson.

La tabulación cruzada original se obtuvo al agregar los datos en las tabulaciones cruzadas separadas para los dos tribunales. Note que para ambos jueces el porcentaje de apelaciones que dio como resultado revocaciones fue mucho mayor en el tribunal municipal que en el tribunal de primera instancia. Debido a que Luckett procesó un porcentaje mucho más alto de sus casos en el tribunal municipal, los datos agregados favorecieron al juez Kendall. No obstante, cuando miramos las tabulaciones cruzadas para los dos tribunales por separado, Luckett muestra el mejor registro. Por consiguiente, para la tabulación cruzada original, el *tipo de tribunal* es una variable oculta que no puede ignorarse cuando se evalúan los registros de los dos jueces.

Dada la posibilidad de la paradoja de Simpson, dése cuenta de que la conclusión o interpretación puede revocarse dependiendo de si usted está viendo datos de tabulación cruzada desagregados o agregados. Antes de formular una conclusión, tal vez quiera investigar si la forma agregada o desagregada de la tabulación proporciona la mejor comprensión y conclusión. En particular, cuando la tabulación cruzada involucra datos agregados, usted debe investigar si una variable oculta podría afectar los resultados, ya que las tabulaciones separadas o desagregadas proporcionan una comprensión y una conclusión diferentes y posiblemente mejores.

Diagrama de dispersión y línea de tendencia

Un **diagrama de dispersión** es una presentación gráfica de la relación entre dos variables cuantitativas, y una **línea de tendencia** es aquella que proporciona una aproximación de la relación. Como ejemplo, considere la relación publicidad/ventas para una tienda de estéreos y equipos de sonido en San Francisco. Durante los tres meses pasados, en 10 ocasiones la tienda usó los comerciales de televisión de fin de semana para promover las ventas en sus establecimientos. Los gerentes quieren investigar si existe una relación entre el número de comerciales transmitidos y las ventas en la tienda durante la semana siguiente. Los datos muestrales de las 10 semanas con las ventas en cientos de dólares se registran en la tabla 2.12.

La figura 2.7 ilustra el diagrama de dispersión y la línea de tendencia[1] para los datos de la tabla 2.12. El número de comerciales (x) se observa en el eje horizontal y las ventas (y) en el eje vertical. Para la semana 1, $x = 2$ y $y = 50$. Un punto con esas coordenadas se traza en el diagrama de dispersión. Puntos similares se trazan en las otras nueve semanas. Observe que en dos semanas se transmitió un comercial, durante dos de las semanas se mostraron dos comerciales, etcétera.

El diagrama de dispersión completo de la figura 2.7 indica una relación positiva entre el número de comerciales y las ventas. Las ventas más altas se asocian con un número mayor de anuncios publicitarios. La relación no es perfecta, ya que no todos los puntos están en una línea recta; sin embargo, el patrón general de los puntos y la línea de tendencia sugieren que en general la relación es positiva.

TABLA 2.12 Datos muestrales para la tienda de estéreos y equipos de sonido

WEB archivo
Stereo

Week	Number of Commercials x	Sales ($100s) y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

[1] La ecuación de la línea de tendencia es $y = 36.15 + 4.95x$. La pendiente de la línea de tendencia es 4.95 y el intercepto en y (el punto donde la línea intercepta el eje y) es 36.15. Comentaremos con detalle la interpretación de la pendiente y el intercepto en y para una línea de tendencia lineal en el capítulo 14, cuando estudiemos la regresión lineal simple.

FIGURA 2.7 Diagrama de dispersión y línea de tendencia para la tienda de estéreos y equipos de sonido

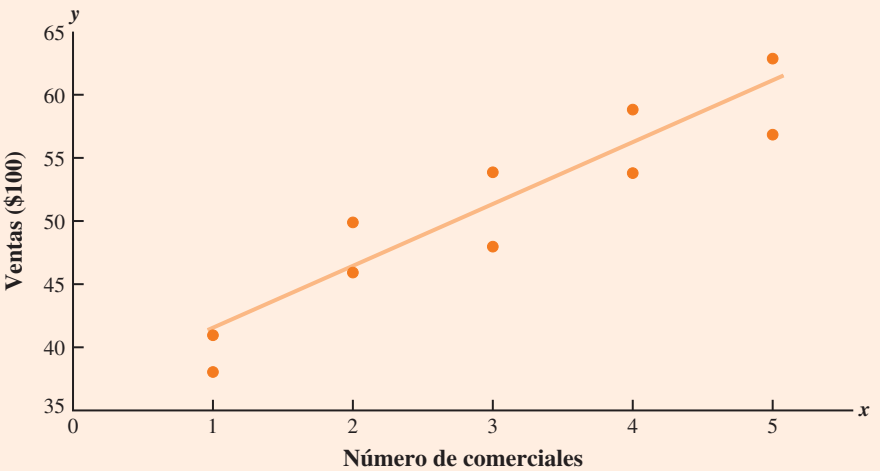
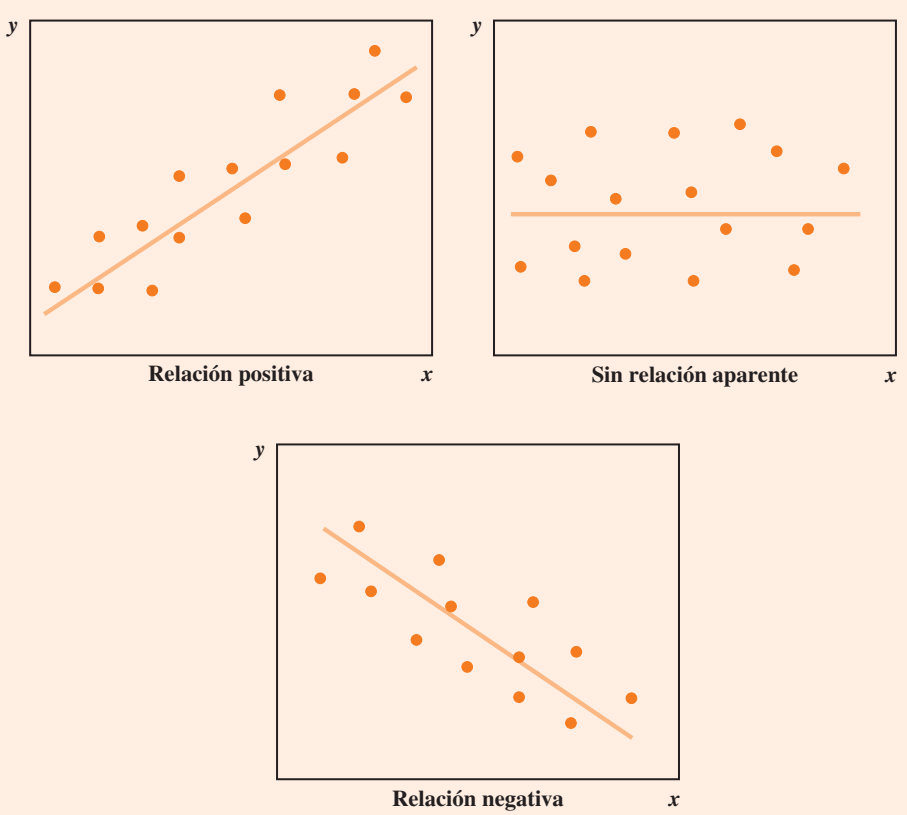


FIGURA 2.8 Tipos de relaciones representados por los diagramas de dispersión



Algunos patrones generales de diagramas de dispersión y los tipos de relación que sugieren se muestran en la figura 2.8. El panel superior izquierdo representa una relación positiva parecida a la del ejemplo del número de comerciales y ventas. En el panel superior derecho, el diagrama de dispersión no muestra una relación aparente entre las variables. El panel inferior representa una relación negativa donde y tiende a disminuir a medida que x aumenta.

Ejercicios

Métodos

AUTOevaluación

29. Los datos siguientes corresponden a 30 observaciones que involucran dos variables cualitativas, x y y. Las categorías para x son A, B y C; las categorías para y son 1 y 2.

WEB archivo
Crosstab

Observation	x	y	Observation	x	y
1	A	1	16	B	2
2	B	1	17	C	1
3	B	1	18	B	1
4	C	2	19	C	1
5	B	1	20	B	1
6	C	2	21	C	2
7	B	1	22	B	1
8	C	2	23	C	2
9	A	1	24	A	1
10	B	1	25	B	1
11	A	1	26	C	2
12	B	1	27	C	2
13	C	2	28	A	1
14	C	2	29	B	1
15	C	2	30	B	2

a) Elabore una tabulación cruzada para los datos, con x como la variable de la fila y y como la variable de la columna.
b) Calcule los porcentajes de la fila.
c) Calcule los porcentajes de la columna.
d) ¿Cuál es la relación, si existe, entre x y y.

AUTOevaluación

30. Las siguientes 20 observaciones son para dos variables cuantitativas, x y y.

WEB archivo
Scatter

Observation	x	y	Observation	x	y
1	−22	22	11	−37	48
2	−33	49	12	34	−29
3	2	8	13	9	−18
4	29	−16	14	−33	31
5	−13	10	15	20	−16
6	21	−28	16	−3	14
7	−13	27	17	−15	18
8	−23	35	18	12	17
9	14	−5	19	−20	−11
10	3	−3	20	−7	−22

a) Elabore un diagrama de dispersión para la relación entre x y y.
b) ¿Cuál es la relación, si existe, entre x y y?

Aplicaciones

31. La tabulación cruzada siguiente muestra el ingreso familiar por nivel educativo del jefe de familia (*Statistical Abstract of the United States, 2008*).

	Ingresos familiares ($1 000)					
Nivel educativo	Menos de 25	25.0–49.9	50.0–74.9	75.0–99.9	100 o más	Total
Sin educación media	4 207	3 459	1 389	539	367	9 961
Eduación media	4 917	6 850	5 027	2 637	2 668	22 099
Educación superior inconclusa	2 807	5 258	4 678	3 250	4 074	20 067
Educación superior	885	2 094	2 848	2 581	5 379	13 787
Maestría o doctorado	290	829	1 274	1 241	4 188	7 822
Total	13 106	18 490	15 216	10 248	16 676	73 736

a) Calcule los porcentajes de la fila e identifique la distribución de frecuencia porcentual para las familias donde el jefe tiene educación media y donde posee educación superior.

b) ¿Qué porcentaje de familias dirigidas por una persona con educación media gana $75 000 o más? ¿Qué porcentaje de familias encabezadas por una persona con educación superior gana $75 000 o más?

c) Elabore histogramas de frecuencia porcentual de los ingresos de familias dirigidas por personas con educación media y con educación superior. ¿La relación entre los ingresos familiares y el nivel educativo es evidente?

32. Consulte de nuevo la tabulación cruzada de los ingresos familiares por nivel educativo que se muestra en el ejercicio 31.

a) Calcule los porcentajes de columna e identifique la distribución de frecuencia porcentual mostrada. ¿Qué proporción de los jefes de familia no cuenta con educación media?

b) ¿Qué porcentaje de las familias que ganan $100 000 o más están encabezadas por una persona que cuenta con una maestría o un doctorado? ¿Qué porcentaje de las familias dirigidas por una persona con alguno de esos grados gana más de $100 000? ¿Por qué difieren estos dos porcentajes?

c) Compare la distribución de frecuencia porcentual de las familias que ganan "menos de 25", "100 o más" y del "total". Comente la relación entre el ingreso familiar y el nivel educativo del jefe de familia.

33. Recientemente, la gerencia de Oak Tree Golf Course ha recibido algunas quejas sobre el estado de los *greens*. Varios jugadores se quejaron de que son muy rápidos. En vez de reaccionar a los comentarios de unos cuantos, la asociación de golf aplicó una encuesta a 100 hombres y 100 mujeres golfistas. Los resultados se resumen a continuación.

Golfistas hombres

	Estado de los *greens*	
Handicap	Demasiado rápidos	Bien
Menos de 15	10	40
15 o más	25	25

Golfistas mujeres

	Estado de los *greens*	
Handicap	Demasiado rápidos	Bien
Menos de 15	1	9
15 o más	39	51

a) Combine en una estas dos tablas de contingencia con las etiquetas de fila *hombres* y *mujeres*, y las etiquetas de columna *demasiado rápidos* y *bien*. ¿Qué grupo muestra el porcentaje más alto que afirma que los *greens* son demasiado rápidos?

b) Consulte las tabulaciones cruzadas iniciales. Para los jugadores de handicap bajo (los mejores), ¿cuál grupo (hombres o mujeres) muestra el porcentaje más alto que afirma que los *greens* son demasiado rápidos?

c) Consulte las tabulaciones cruzadas iniciales. Para los jugadores de handicap alto, ¿cuál grupo (hombres o mujeres) muestra el porcentaje más alto que afirma que los *greens* son demasiado rápidos?

d) ¿Qué conclusiones puede formular acerca de las preferencias de los hombres y las mujeres respecto de la rapidez de los *greens*? ¿Las conclusiones del inciso *a*) en comparación con las de los incisos *b*) y *c*) son consistentes? Explique cualquier incongruencia aparente.

34. La tabla 2.13 de la siguiente página muestra un conjunto de datos con información para 45 fondos de inversión que son parte del *Morningstar Funds500* de 2008. El conjunto de datos incluye las cinco variables siguientes:

Tipo de fondo. El tipo de fondo, etiquetado como DE (capital nacional), IE (capital internacional) y FI (renta fija)

Valor neto de los activos. El precio de cierre por acción

Rendimiento promedio a 5 años (%). El rendimiento promedio anual para el fondo durante los cinco años anteriores

Razón de gastos (%). El porcentaje de activos deducido cada año fiscal para los gastos de fondos

Calificación Morningstar. La calificación, con estrellas, del riesgo ajustada para cada fondo; las calificaciones de Morningstar varían de una baja de 1 estrella (1-star) a una alta de 5 estrellas (5-stars)

a) Elabore una tabulación cruzada de los datos sobre el tipo de fondo (filas) y el rendimiento promedio anual durante los cinco años anteriores (columnas). Utilice clases de 0–9.99; 10–19.99; 20–29.99; 30–39.99; 40–49.99, y 50–59.99 para el rendimiento promedio de 5 años (%).

b) Prepare una distribución de frecuencia para los datos sobre el tipo de fondo.

c) Elabore una distribución de frecuencia para los datos sobre el rendimiento promedio de 5 años (%).

d) ¿Cómo ayudó la tabulación cruzada para la preparación de la distribución de frecuencias en los incisos *b*) y *c*)?

e) ¿Qué conclusiones puede formular sobre el tipo de fondo y el rendimiento promedio sobre los 5 años anteriores?

35. Consulte los datos de la tabla 2.13.

a) Elabore una tabulación cruzada de los datos sobre el tipo de fondo (filas) y la razón de gastos (columnas). Use las clases de 0.25–0.49; 0.50–0.74; 0.75–0.99; 1.00–1.24, y 1.25–1.49 para la razón de gastos (%).

b) Prepare una distribución de frecuencia porcentual para la razón de gastos (%).

c) ¿Qué conclusiones puede formular acerca del tipo de fondo y la razón de gastos?

36. Consulte los datos de la tabla 2.13.

a) Elabore un diagrama de dispersión con un rendiminto promedio de cinco años (%) sobre el eje horizontal y el valor neto de los activos ($) sobre el eje vertical.

b) Comente la relación, si existe, entre las variables.

37. La *Guía de economía de combustible* del Departamento de Energía de Estados Unidos proporciona datos sobre la eficiencia de combustible para automóviles y camiones (sitio web Fuel Economy, 22 de febrero de 2008). Una porción de los datos para 311 automóviles compactos, medianos y grandes se muestra en la tabla 2.14. El conjunto de datos contiene las variables siguientes:

Tamaño: compacto, mediano y grande

Desplazamiento: tamaño del motor en litros

Cilindros: número de cilindros en el motor

Tracción: delantera (F), trasera (R) y en las cuatro llantas (4)

Tipo de combustible: premium (P) o regular (R)

Mi/gal en ciudad: calificación de la eficiencia del combustible para uso del automóvil en la ciudad en términos de millas por galón

Mi/gal en autopista: calificación de la eficiencia del combustible para uso del automóvil en autopista en términos de millas por galón

El conjunto de datos completo está contenido en el archivo llamado FuelData08.

a) Prepare una tabulación cruzada de los datos de tamaño (filas) y las millas por galón en autopista (columnas). Use las clases 15–19, 20–24, 25–29, 30–34 y 35–39 para mi/gal en autopista.

b) Comente la relación entre el tamaño y las millas por galón en autopista.

TABLA 2.13 Datos financieros para una muestra de 45 fondos de inversión

WEB archivo
MutualFunds

Fund Name	Fund Type	Net Asset Value ($)	5-Year Average Return (%)	Expense Ratio (%)	Morningstar Rank
Amer Cent Inc & Growth Inv	DE	28.88	12.39	0.67	2-Star
American Century Intl. Disc	IE	14.37	30.53	1.41	3-Star
American Century Tax-Free Bond	FI	10.73	3.34	0.49	4-Star
American Century Ultra	DE	24.94	10.88	0.99	3-Star
Ariel	DE	46.39	11.32	1.03	2-Star
Artisan Intl Val	IE	25.52	24.95	1.23	3-Star
Artisan Small Cap	DE	16.92	15.67	1.18	3-Star
Baron Asset	DE	50.67	16.77	1.31	5-Star
Brandywine	DE	36.58	18.14	1.08	4-Star
Brown Cap Small	DE	35.73	15.85	1.20	4-Star
Buffalo Mid Cap	DE	15.29	17.25	1.02	3-Star
Delafield	DE	24.32	17.77	1.32	4-Star
DFA U.S. Micro Cap	DE	13.47	17.23	0.53	3-Star
Dodge & Cox Income	FI	12.51	4.31	0.44	4-Star
Fairholme	DE	31.86	18.23	1.00	5-Star
Fidelity Contrafund	DE	73.11	17.99	0.89	5-Star
Fidelity Municipal Income	FI	12.58	4.41	0.45	5-Star
Fidelity Overseas	IE	48.39	23.46	0.90	4-Star
Fidelity Sel Electronics	DE	45.60	13.50	0.89	3-Star
Fidelity Sh-Term Bond	FI	8.60	2.76	0.45	3-Star
Fidelity	DE	39.85	14.40	0.56	4-Star
FPA New Income	FI	10.95	4.63	0.62	3-Star
Gabelli Asset AAA	DE	49.81	16.70	1.36	4-Star
Greenspring	DE	23.59	12.46	1.07	3-Star
Janus	DE	32.26	12.81	0.90	3-Star
Janus Worldwide	IE	54.83	12.31	0.86	2-Star
Kalmar Gr Val Sm Cp	DE	15.30	15.31	1.32	3-Star
Managers Freemont Bond	FI	10.56	5.14	0.60	5-Star
Marsico 21st Century	DE	17.44	15.16	1.31	5-Star
Mathews Pacific Tiger	IE	27.86	32.70	1.16	3-Star
Meridan Value	DE	31.92	15.33	1.08	4-Star
Oakmark I	DE	40.37	9.51	1.05	2-Star
PIMCO Emerg Mkts Bd D	FI	10.68	13.57	1.25	3-Star
RS Value A	DE	26.27	23.68	1.36	4-Star
T. Rowe Price Latin Am.	IE	53.89	51.10	1.24	4-Star
T. Rowe Price Mid Val	DE	22.46	16.91	0.80	4-Star
Templeton Growth A	IE	24.07	15.91	1.01	3-Star
Thornburg Value A	DE	37.53	15.46	1.27	4-Star
USAA Income	FI	12.10	4.31	0.62	3-Star
Vanguard Equity-Inc	DE	24.42	13.41	0.29	4-Star
Vanguard Global Equity	IE	23.71	21.77	0.64	5-Star
Vanguard GNMA	FI	10.37	4.25	0.21	5-Star
Vanguard Sht-Tm TE	FI	15.68	2.37	0.16	3-Star
Vanguard Sm Cp Idx	DE	32.58	17.01	0.23	3-Star
Wasatch Sm Cp Growth	DE	35.41	13.98	1.19	4-Star

TABLA 2.14 Datos de la eficiencia de combustible para 311 automóviles

WEB archivo
FuelData08

Car	Size	Displacement	Cylinders	Drive	Fuel Type	City MPG	Hwy MPG
1	Compacto	3.1	6	4	P	15	25
2	Compacto	3.1	6	4	P	17	25
3	Compacto	3.0	6	4	P	17	25
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
161	Mediano	2.4	4	F	R	22	30
162	Mediano	2.0	4	F	P	19	29
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
310	Grande	3.0	6	F	R	17	25
311	Grande	3.0	6	F	R	18	25

c) Elabore una tabulación cruzada de los datos de tracción (filas) y las millas por galón en la ciudad (columnas). Use las clases 5–9, 10–14, 15–19, 20–24, 25–29, 30–34 y 35–39 para las millas por galón en la ciudad.

d) Comente la relación entre la tracción y las millas por galón en la ciudad.

e) Prepare una tabulación cruzada de los datos sobre el tipo de combustible (filas) y las millas por galón en la ciudad (columnas). Use las clases 5–9, 10–14, 15–19, 20–24, 25–29, 30–34 y 35–39 para las millas por galón en la ciudad.

f) Comente la relación entre el tipo de combustible y las millas por galón en la ciudad.

38. Remítase al ejercicio 37 y a los datos en el archivo FuelData08.

a) Elabore una tabulación cruzada de los datos sobre el desplazamiento (filas) y las millas por galón en autopista (columnas). Use las clases 1.0–2.9, 3.0–4.9 y 5.0–6.9 para el desplazamiento. Use las clases 15–19, 20–24, 25–29, 30–34 y 35–39 para las millas por galón en autopista.

b) Comente la relación, si existe, entre el desplazamiento y las millas por galón en autopista.

c) Elabore un diagrama de dispersión de los datos sobre el desplazamiento y las millas por galón en autopista. Use el eje vertical para las millas.

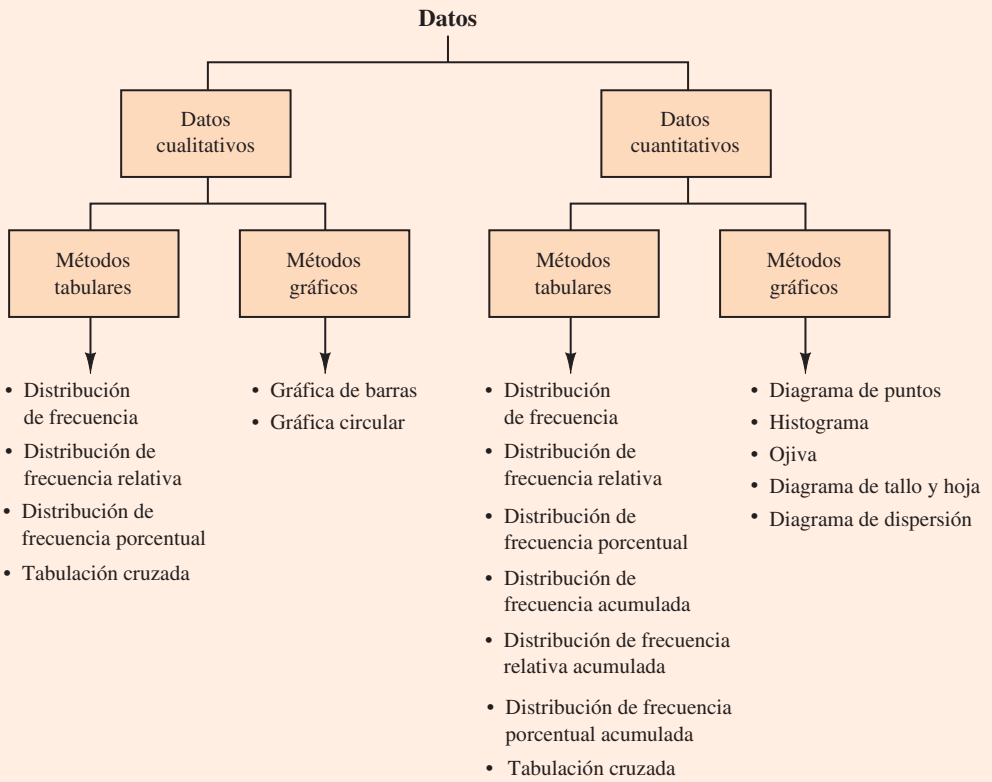
d) ¿Qué indica el diagrama de dispersión elaborado en el inciso *c*) sobre la relación, si existe, entre el desplazamiento y las millas por galón en autopista?

e) En la investigación de la relación entre el desplazamiento y las millas por galón en autopista usted elaboró un resumen tabular de los datos (tabulación cruzada) y un resumen gráfico (diagrama de dispersión). En este caso, ¿cuál método prefiere? Explique sus razones.

Resumen

Con frecuencia es difícil interpretar directamente un conjunto de datos, incluso si es pequeño, en la forma en que se recolecta. Los métodos tabulares y gráficos proporcionan procedimientos para la organización y el resumen de los datos de modo que los patrones se inviertan y los datos se interpreten con más facilidad. Las distribuciones de frecuencia, las distribuciones de frecuencia relativa, las distribuciones de frecuencia porcentual, las gráficas de barras y las gráficas circulares se presentan como procedimientos tabulares y gráficos para el resumen de datos cualitativos. Las distribuciones de frecuencia, de frecuencia relativa y de frecuencia porcentual, así como los histogramas, las distribuciones de frecuencia acumulada, las distribuciones de frecuencia relativa acumulada, las distribuciones de frecuencia porcentual acumulada y las ojivas constituyen maneras de resumir datos cuantitativos. El diagrama de tallo y hoja proporciona una técnica de análisis explicativo de los datos que también se utiliza para resumir datos cuantitativos. La tabulación cruzada constituye un método tabular para resumir datos para dos variables. El diagrama de dispersión se plantea como un método gráfico para mostrar la relación entre dos variables cuantitativas. La figura 2.9 muestra los métodos tabulares y gráficos expuestos en este capítulo.

FIGURA 2.9 Métodos tabulares y gráficos para el resumen de datos



Con conjuntos de datos grandes, el software de computadora es fundamental para la elaboración de resúmenes tabulares y gráficos de los datos. En los apéndices del capítulo se muestra cómo se usan Minitab, Excel y StatTools para este propósito.

Glosario

Análisis de datos exploratorios Métodos que utilizan aritmética sencilla y gráficas fáciles de trazar para resumir los datos de manera rápida.

Datos cualitativos Etiquetas o nombres que sirven para identificar categorías de elementos parecidos.

Datos cuantitativos Valores numéricos que indican cuánto o cuántos.

Diagrama de dispersión Presentación gráfica de la relación entre dos variables cuantitativas. Una variable se muestra en el eje horizontal y la otra en el eje vertical.

Diagrama de puntos Dispositivo gráfico que resume los datos según el número de puntos arriba de cada valor de datos en el eje horizontal.

Diagrama de tallo y hoja Técnica de análisis de datos exploratorios que clasifica de manera simultánea los órdenes de los datos cuantitativos y permite comprender la forma de la distribución.

Distribución de frecuencia Resumen tabular de los datos que muestra el número (frecuencia) de los valores de datos en cada una de varias clases que no se superponen.

Distribución de frecuencia acumulada Resumen tabular de datos cuantitativos que muestra el número de valores de datos que son menores o iguales que el límite de clase superior de cada clase.

Distribución de frecuencia porcentual Resumen tabular de los datos que muestra el porcentaje de valores de datos en cada una de varias clases que no se superponen.

Distribución de frecuencia porcentual acumulada Resumen tabular de los datos cuantitativos que muestra el porcentaje de los valores de datos que son menores o iguales que el límite de clase superior de cada clase.
Distribución de frecuencia relativa Resumen tabular de los datos que registra la fracción o proporción de los valores de datos en cada una de varias clases que no se superponen.
Distribución de frecuencia relativa acumulada Resumen tabular de los datos cuantitativos que muestra la fracción o proporción de los valores de datos que son menores o iguales al límite de clase superior de cada clase.
Gráfica circular Dispositivo gráfico para representar resúmenes de datos con base en la subdivisión de un círculo en sectores que corresponden a la frecuencia relativa de cada clase.
Gráfica de barras Dispositivo para representar datos cualitativos previamente resumidos en una distribución de frecuencia, distribución de frecuencia relativa o distribución de frecuencia porcentual.
Histograma Presentación gráfica de una distribución de frecuencia, distribución de frecuencia relativa o distribución de frecuencia porcentual de datos cuantitativos, elaborada mediante la colocación de los intervalos de clase en el eje horizontal y las frecuencias, frecuencias relativas o frecuencias porcentuales en el eje vertical.
Línea de tendencia Línea que proporciona una aproximación de la relación entre dos variables.
Ojiva Gráfica de una distribución acumulada.
Paradoja de Simpson Conclusiones obtenidas de dos o más tabulaciones cruzadas separadas que pueden revocarse cuando los datos se agregan en una sola tabulación cruzada.
Punto medio de clase Valor intermedio entre los límites de clase inferior y superior.
Tabulación cruzada Resumen tabular de los datos para dos variables. Las clases para una de las variables se representan por medio de filas; las clases para la otra variable se representan por medio de columnas.

Fórmulas clave

Frecuencia relativa

$$\frac{\text{Frecuencia de la clase}}{n} \tag{2.1}$$

Ancho de clase aproximado

$$\frac{\text{Valor de datos mayor/valor de datos menor}}{\text{número de clases}} \tag{2.2}$$

Ejercicios complementarios

39. El Instituto de Investigación de Educación Superior de la Universidad de California en Los Ángeles (UCLA) proporciona estadísticas sobre las asignaturas más populares entre los estudiantes de primer año que asisten a la universidad. Las cinco materias principales son arte y humanidades (A), administración de empresas (B), ingeniería (E), profesional (P) y ciencias sociales (S) (*The New York Times Almanac*, 2006). Una amplia variedad de otras asignaturas principales (O), que incluyen biología, física, ciencias de la computación y educación, se agrupan juntas. Las asignaturas principales seleccionadas por una muestra de 64 universitarios de primer año se presentan a continuación.

S	P	P	O	B	E	O	E	P	O	O	B	O	O	O	A
O	E	E	B	S	O	B	O	A	O	E	O	E	O	B	P
B	A	S	O	E	A	B	O	S	S	O	O	E	B	O	B
A	E	B	E	A	A	P	O	O	E	O	B	B	O	P	B

WEB archivo
Major

a) Elabore una distribución de frecuencia y una distribución de frecuencia porcentual.
b) Trace una gráfica de barras.

c) ¿Qué porcentaje de estudiantes de primer año seleccionó una de las cinco asignaturas principales más populares?
d) Cuál es la materia principal más popular para los universitarios de primer año? ¿Qué porcentaje de estudiantes la seleccionó?

WEB archivo
GMSales

40. General Motors tuvo un participación de 23% en la industria automotriz, con ventas que provienen de ocho divisiones: Buick, Cadillac, Chevrolet, GMC, Hummer, Pontiac, Saab y Saturn (*Forbes*, 22 de diciembre de 2008). El conjunto de datos de GMSales registra las ventas para una muestra de 200 vehículos de General Motors. La división por vehículo se proporciona para cada venta.
a) Prepare la distribución de frecuencia y la distribución de frecuencia porcentual de las ventas por división para General Motors.
b) Muestre una gráfica de barras de la distribución de frecuencia porcentual.
c) ¿Cuál división de General Motors fue la empresa líder en ventas? ¿Cuál fue el porcentaje de ventas para esta división? ¿Fue la división más importante de General Motors? Explique por qué.
d) Debido a la recesión actual, los altos precios de la gasolina y la disminución en las ventas de automóviles, General Motors enfrentó una bancarrota en 2009. Se anticiparon por ende el rescate financiero del gobierno y una restructuración de la empresa. Las expectativas eran que ésta no podía seguir operando las ocho divisiones. Con base en el porcentaje de ventas, ¿cuál de las ocho divisiones parecería ser la mejor candidata a ser suspendida por General Motors? ¿Y cuáles divisiones parecerían ser las candidatas menos probables para suspensión?

41. El rendimiento de dividendos es el dividendo anual pagado por una empresa expresado como un porcentaje del precio de la acción (dividendo/precio de la acción × 100). El rendimiento de dividendos para las empresas del promedio industrial Dow Jones se muestra en la tabla 2.15 (*The Wall Street Journal*, 8 de junio de 2009).
a) Elabore una distribución de frecuencia y una distribución de frecuencia porcentual.
b) Prepare un histograma.
c) Comente la forma de la distribución.
d) ¿Qué indican los resúmenes tabulares y gráficos sobre los rendimientos de dividendos entre las empresas del promedio industrial Dow Jones?
e) ¿Cuál empresa tiene el dividendo más alto producido? Si las acciones de ésta se venden actualmente a $20 por acción y usted compra 500, ¿cuánto ingreso por dividendos generará esta inversión en un año?

42. Aproximadamente 1.5 estudiantes de secundaria y bachillerato presentan cada año el examen de aptitudes escolares (scholastic aptitude test, SAT), y casi 80% de los colegios y universidades sin políticas de admisión abiertas utilizan estas calificaciones en la toma de decisiones

TABLA 2.15 Rendimiento de dividendos para las empresas del promedio industrial Dow Jones

WEB archivo
DYield

Company	Dividend Yield %	Company	Dividend Yield %
3M	3.6	IBM	2.1
Alcoa	1.3	Intel	3.4
American Express	2.9	J.P. Morgan Chase	0.5
AT&T	6.6	Johnson & Johnson	3.6
Bank of America	0.4	Kraft Foods	4.4
Boeing	3.8	McDonald's	3.4
Caterpillar	4.7	Merck	5.5
Chevron	3.9	Microsoft	2.5
Cisco Systems	0.0	Pfizer	4.2
Coca-Cola	3.3	Procter & Gamble	3.4
DuPont	5.8	Travelers	3.0
ExxonMobil	2.4	United Technologies	2.9
General Electric	9.2	Verizon	6.3
Hewlett-Packard	0.9	Wal-Mart Stores	2.2
Home Depot	3.9	Walt Disney	1.5

de ingreso (College Board, marzo de 2009). La versión actual del SAT incluye tres partes: comprensión de lectura, matemáticas y redacción. Una calificación combinada perfecta para las tres partes es 2 400. Una muestra de calificaciones para el SAT combinado de tres partes es la siguiente.

WEB archivo
NewSAT

1 665	1 525	1 355	1 645	1 780
1 275	2 135	1 280	1 060	1 585
1 650	1 560	1 150	1 485	1 990
1 590	1 880	1 420	1 755	1 375
1 475	1 680	1 440	1 260	1 730
1 490	1 560	940	1 390	1 175

a) Elabore una distribución de frecuencia y un histograma. Comience con la primera clase en 800 y utilice un ancho de clase de 200.
b) Comente la forma de la distribución.
c) ¿Qué otras observaciones puede hacer acerca de las calificaciones del SAT con base en los resúmenes gráfico y tabular?

43. Los Acereros de Pittsburgh derrotaron a los Cardenales de Arizona 27 a 23 en el Super Bowl 43 del futbol americano. Con esta victoria, su sexto campeonato, los Acereros de Pittsburgh se convirtieron en el equipo con más triunfos en los 43 años de historia del evento (*Tampa Tribune*, 2 de febrero de 2009). El Super Bowl se ha celebrado en ocho estados diferentes de Estados Unidos: Arizona (AZ), California (CA), Florida (FL), Georgia (GA), Louisiana (LA), Michigan (MI), Minnesota (MN) y Texas (TX). Los datos de la tabla siguiente muestran el estado donde se celebraron los Super Bowls y el margen de puntos de la victoria para el equipo ganador.

WEB archivo
SuperBowl

Super Bowl	State	Won By Points	Super Bowl	State	Won By Points	Super Bowl	State	Won By Points
1	CA	25	16	MI	5	31	LA	14
2	FL	19	17	CA	10	32	CA	7
3	FL	9	18	FL	19	33	FL	15
4	LA	16	19	CA	22	34	GA	7
5	FL	3	20	LA	36	35	FL	27
6	FL	21	21	CA	19	36	LA	3
7	CA	7	22	CA	32	37	CA	27
8	TX	17	23	FL	4	38	TX	3
9	LA	10	24	LA	45	39	FL	3
10	FL	4	25	FL	1	40	MI	11
11	CA	18	26	MN	13	41	FL	12
12	LA	17	27	CA	35	42	AZ	3
13	FL	4	28	GA	17	43	FL	4
14	CA	12	29	FL	23			
15	LA	17	30	AZ	10			

a) Elabore una distribución de frecuencia y una gráfica de barra para el estado donde se celebró el Super Bowl.
b) ¿A qué conclusiones llega a partir del resumen del inciso *a*)? ¿Cuál es el porcentaje de Super Bowls celebrados en los estados de Florida o California? ¿Qué porcentaje se celebró en los estados del norte o de clima frío?
c) Muestre un diagrama de tallo y hoja para el margen de puntos de victoria para el equipo ganador. Elabore un histograma.
d) ¿Qué conclusiones obtiene a partir de su resumen del inciso c)? ¿Qué porcentaje de Super Bowls han sido partidos cerrados con un margen de victoria menor que 5 puntos? ¿Qué porcentaje se ha ganado por 20 puntos o más?
e) El partido de Super Bowl más cerrado tuvo lugar cuando los Gigantes de Nueva York derrotaron a los Bills de Búfalo. ¿Dónde se realizó este partido y cuál fue el margen de victoria del equipo ganador? El margen de puntos más grande en la historia del Super Bowl ocurrió cuando los 49's de San Francisco derrotaron a los Broncos de Denver. ¿En dónde se celebró este partido y cuál fue el margen de victoria del equipo ganador?

44. Los datos de la Oficina del Censo de Estados Unidos proporcionan la población por estado en millones de personas (*The World Almanac*, 2006).

WEB archivo
Population

State	Population	State	Population	State	Population
Alabama	4.5	Louisiana	4.5	Ohio	11.5
Alaska	0.7	Maine	1.3	Oklahoma	3.5
Arizona	5.7	Maryland	5.6	Oregon	3.6
Arkansas	2.8	Massachusetts	6.4	Pennsylvania	12.4
California	35.9	Michigan	10.1	Rhode Island	1.1
Colorado	4.6	Minnesota	5.1	South Carolina	4.2
Connecticut	3.5	Mississippi	2.9	South Dakota	0.8
Delaware	0.8	Missouri	5.8	Tennessee	5.9
Florida	17.4	Montana	0.9	Texas	22.5
Georgia	8.8	Nebraska	1.7	Utah	2.4
Hawaii	1.3	Nevada	2.3	Vermont	0.6
Idaho	1.4	New Hampshire	1.3	Virginia	7.5
Illinois	12.7	New Jersey	8.7	Washington	6.2
Indiana	6.2	New Mexico	1.9	West Virginia	1.8
Iowa	3.0	New York	19.2	Wisconsin	5.5
Kansas	2.7	North Carolina	8.5	Wyoming	0.5
Kentucky	4.1	North Dakota	0.6		

a) Elabore una distribución de frecuencia, una distribución de frecuencia porcentual y un histograma. Utilice un ancho de clase de 2.5 millones.
b) Comente la tendencia en la distribución.
c) ¿Qué observaciones puede hacer acerca de la población de los 50 estados?

45. *Drug Store News* (septiembre de 2002) proporcionó datos sobre las ventas farmacéuticas anuales para los minoristas de farmacias líderes en Estados Unidos. Los datos siguientes muestran las ventas anuales en millones de dólares.

Minorista	Ventas	Minorista	Ventas
Ahold USA	$ 1 700	Medicine Shoppe	$ 1 757
CVS	12 700	Rite-Aid	8 637
Eckerd	7 739	Safeway	2 150
Kmart	1 863	Walgreens	11 660
Kroger	3 400	Wal-Mart	7 250

a) Elabore un diagrama de tallo y hoja.
b) Identifique los niveles de ventas anuales para los minoristas de farmacia pequeños, medianos y grandes.
c) ¿Cuáles son los dos minoristas más grandes?

46. Las temperaturas alta y baja diarias para 20 ciudades se listan a continuación (*USA Today*, 3 de marzo de 2006).

WEB archivo
CityTemp

City	High	Low	City	High	Low
Albuquerque	66	39	Los Ángeles	60	46
Atlanta	61	35	Miami	84	65
Baltimore	42	26	Minneapolis	30	11
Charlotte	60	29	Nueva Orleáns	68	50
Cincinnati	41	21	Oklahoma City	62	40
Dallas	62	47	Phoenix	77	50
Denver	60	31	Portland	54	38
Houston	70	54	St. Louis	45	27
Indianapolis	42	22	San Francisco	55	43
Las Vegas	65	43	Seattle	52	36

a) Elabore un diagrama de tallo y hoja de las temperaturas altas.
b) Elabore un diagrama de tallo y hoja para las temperaturas bajas.
c) Compare los dos diagramas y comente sobre la diferencia entre las tem-peraturas altas y bajas.
d) Proporcione una distribución de frecuencia tanto para las temperaturas altas como para las bajas.

47. Consulte el conjunto de datos para las temperaturas altas y bajas de las 20 ciudades del ejercicio 46.
a) Elabore un diagrama de dispersión para mostrar la relación entre las dos variables: temperatura alta y temperatura baja.
b) Comente sobre la relación entre ambas temperaturas.

48. Una de las preguntas en una encuesta de *Financial Times*/Harris Poll fue: "¿Qué tanto está a favor o en contra de un impuesto mayor sobre las emisiones de carbono de los automóviles?" Las respuestas posibles fueron totalmente a favor, más a favor que en contra, más en contra que a favor y totalmente en contra. La tabulación cruzada siguiente muestra las respuestas obtenidas para 5 372 adultos encuestados en cuatro países de Europa y en Estados Unidos (sitio web de Harris Interactive, 27 de febrero de 2008).

	País					
Nivel de apoyo	**Gran Bretaña**	**Italia**	**España**	**Alemania**	**Estados Unidos**	**Total**
Totalmente a favor	337	334	510	222	214	1 617
Más a favor que en contra	370	408	355	411	327	1 871
Más en contra que a favor	250	188	155	267	275	1 135
Totalmente en contra	130	115	89	211	204	749
Total	1 087	1 045	1 109	1 111	1 020	5 372

a) Elabore una distribución de frecuencia porcentual para la variable del nivel de apoyo. ¿Piensa usted que los resultados muestran apoyo a un impuesto más alto sobre la emisión de carbono de los automóviles?
b) Elabore una distribución de frecuencia porcentual para la variable país.
c) ¿El nivel de apoyo entre los adultos en los países europeos difiere del que se manifiesta en Estados Unidos? Explique por qué.

49. Western University sólo tiene una beca para jugadoras de softbol femenil para el próximo año. Las dos finalistas que la universidad está considerando son Allison Fealey y Emily Janson. El personal técnico ha concluido que la velocidad y las habilidades defensivas de las dos jugadoras son prácticamente idénticas, y que la decisión final se basará en la que tenga el mejor promedio de bateo. Las tabulaciones cruzadas del desempeño de bateo de cada jugadora en la sencundaria y el bachillerato son las siguientes.

	Allison Fealey	
Resultado	**Junior**	**Senior**
Hit	15	75
Sin hit	25	175
Total de bateos	40	250

	Emily Janson	
Resultado	**Junior**	**Senior**
Hit	70	35
Sin hit	130	85
Total de bateos	200	120

El promedio de bateo se calcula al dividir el número de hits de un jugador por el número total de turnos al bate. Los promedios de bateo se representan como un número decimal con tres lugares después del punto decimal.
a) Calcule el promedio de bateo de cada jugadora en la secundaria. Luego calcule el promedio de bateo en el bachillerato. Utilice este análisis para responder cuál jugadora debe obtener la beca. Explique sus razones.

b) Combine o agregue los datos para la secundaria y el bachillerato en una tabulación cruzada de la manera siguiente.

	Jugadora	
Resultado	Fealey	Janson
Hit		
Sin hit		
Total de bateos		

Calcule el promedio de bateo de cada jugadora para los dos años combinados. Utilice este análisis para responder cuál de ellas debe recibir la beca. Explique sus razones.

c) ¿Sus recomendaciones de los incisos *a*) y *b*) son congruentes? Justifique cualquier falta aparente de congruencia.

50. En una encuesta de edificios comerciales realizada por la Cincinnati Gas & Electric Company se preguntaba cuál era el combustible utilizado en la calefacción principal y en qué año se construyó el edificio. Una tabulación cruzada parcial de los hallazgos se presenta a continuación.

Año de construcción	Tipo de combustible				
	Electricidad	Gas natural	Petróleo	Propano	Otros
1973 o antes	40	183	12	5	7
1974–1979	24	26	2	2	0
1980–1986	37	38	1	0	6
1987–1991	48	70	2	0	1

a) Complete la tabulación cruzada mostrando el total de las filas y el total de las columnas.

b) Elabore las distribuciones de frecuencia del año de construcción y del tipo de combustible.

c) Desarrolle una tabulación cruzada que muestre los porcentajes de columna.

d) Elabore una tabulación cruzada que muestre los porcentajes de fila.

e) Comente la relación entre el año de construcción y el tipo de combustible.

51. La tabla 2.16 incluye una porción de los datos contenidos en el archivo Fortune, y lista los datos sobre el capital de los accionistas, el valor de mercado y las utilidades para una muestra de 50 empresas *Fortune 500*.

TABLA 2.16 Datos para una muestra de 50 empresas *Fortune 500*

WEB archivo
Fortune

Company	Stockholders' Equity ($1 000s)	Market Value ($1 000s)	Profit ($1 000s)
AGCO	982.1	372.1	60.6
AMP	2 698.0	12 017.6	2.0
Apple Computer	1 642.0	4 605.0	309.0
Baxter International	2 839.0	21 743.0	315.0
Bergen Brunswick	629.1	2 787.5	3.1
Best Buy	557.7	10 376.5	94.5
Charles Schwab	1 429.0	35 340.6	348.5
.	.	.	.
.	.	.	.
.	.	.	.
Walgreen	2 849.0	30 324.7	511.0
Westvaco	2 246.4	2 225.6	132.0
Whirlpool	2 001.0	3 729.4	325.0
Xerox	5 544.0	35 603.7	395.0

a) Elabore una tabulación cruzada para las variables del capital de los accionistas (Stockholders' Equity) y de las utilidades (Profit). Use las clases 0–200, 200–400, . . . , 1 000–1 200 para las utilidades, y las clases 0–1 200, 1 200–2 400, . . . , 4 800–6 000 para el capital de los accionistas.
b) Calcule los porcentajes de fila para la tabulación cruzada que elaboró para el inciso *a*).
c) ¿Qué relación observa, si hay alguna, entre las utilidades y el capital de los accionistas?

52. Remítase al conjunto de datos de la tabla 2.16.
a) Elabore una tabulación cruzada para las variables valor de mercado (Market Value) y utilidades (Profit).
b) Calcule los porcentajes de fila para su tabulación cruzada del inciso *a*).
c) Comente sobre cualquier relación entre las variables.

53. Consulte el conjunto de datos de la tabla 2.16.
a) Trace un diagrama de dispersión que muestre la relación entre las variables utilidades y capital de los accionistas.
b) Comente acerca de cualquier relación entre las variables.

54. Consulte el conjunto de datos de la tabla 2.16.
a) Elabore un diagrama de dispersión que muestre la relación entre las variables valor de mercado y capital de los accionistas.
b) Comente sobre cualquier relación entre las variables.

Caso a resolver 1 Pelican Stores

Pelican Stores, una división de National Clothing, es una cadena de tiendas de ropa para mujer que opera en todo Estados Unidos. La cadena lanzó recientemente una promoción en la que se enviaron cupones de descuento a los clientes de otras tiendas de National Clothing. Los datos recabados de una muestra de 100 transacciones de tarjetas de crédito en Pelican Stores durante un día, mientras la promoción estuvo vigente, se encuentran en el archivo llamado *PelicanStores*. La tabla 2.17 muestra una parte del conjunto de datos.

El método de pago Proprietary Card se refiere a los cargos realizados con una tarjeta de National Clothing. A los clientes que efectuaron una compra utilizando un cupón de descuento se les llama *clientes de promoción* y a los que hicieron una compra pero no usaron un cupón de descuento se les llama *clientes habituales*. Dado que los cupones promocionales no se enviaron a los clientes regulares de Pelican Stores, la gerencia considera las ventas realizadas a personas que presentaron los cupones como ventas que de lo contrario no se hubieran efectuado. Por supuesto, Pelican también espera que los clientes de promoción sigan comprando en sus tiendas.

WEB archivo
PelicanStores

TABLA 2.17 Datos para una muestra de 100 compras con tarjeta de crédito en Pelican Stores

Customer	Type of Customer	Items	Net Sales	Method of Payment	Gender	Marital Status	Age
1	Regular	1	39.50	Discover	Male	Married	32
2	Promotional	1	102.40	Proprietary Card	Female	Married	36
3	Regular	1	22.50	Proprietary Card	Female	Married	32
4	Promotional	5	100.40	Proprietary Card	Female	Married	28
5	Regular	2	54.00	MasterCard	Female	Married	34
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
96	Regular	1	39.50	MasterCard	Female	Married	44
97	Promotional	9	253.00	Proprietary Card	Female	Married	30
98	Promotional	10	287.59	Proprietary Card	Female	Married	52
99	Promotional	2	47.60	Proprietary Card	Female	Married	30
100	Promotional	1	28.44	Proprietary Card	Female	Married	44

La mayoría de las variables mostradas en la tabla 2.17 se explican por sí mismas, pero dos de ellas requieren una aclaración.

Artículos (Items)	Número total de productos adquiridos
Ventas netas (Net Sales)	Monto total ($) cargado a la tarjeta de crédito

A la gerencia de Pelican le gustaría usar estos datos muestrales para enterarse de su base de clientes y evaluar la promoción que consiste en los cupones de descuento.

Informe gerencial

Use los métodos tabular y gráfico de la estadística descriptiva para ayudar a la gerencia a elaborar un perfil de los clientes y evaluar la compaña promocional. Como mínimo, su informe debe incluir lo siguiente:

1. Distribución de frecuencia porcentual para las variables clave.
2. Una gráfica de barras o circular que muestre el número de compras del cliente atribuible al método de pago.
3. Una tabulación cruzada del tipo de cliente (regular o promocional) contra las ventas netas. Comente cualquier similitud o diferencia que se presente.
4. Un diagrama de dispersión para explorar la relación entre las ventas netas (Net sales) y la edad (Age) de los clientes.

Caso a resolver 2 Industria del cine

La industria fílmica estadounidense es un negocio muy competitivo. Más de 50 estudios producen un total de 300 a 400 películas nuevas cada año, y el éxito financiero de cada una varía considerablemente. Las ventas brutas del fin de semana de estreno (en millones de dólares), las ventas brutas totales (Total Gross Sales) (en millones de dólares), el número de salas (Number of Theaters) donde se exhibe la película y el número de semanas en que ésta permaneció entre las primeras 60 (Weeks in the Top 60) en ventas brutas son variables comunes utilizadas para medir el éxito de una cinta. Los datos recabados de una muestra de 100 películas producidas en 2005 se incluyen en el archivo llamado *Movies*. La tabla 2.18 muestra los datos de las primeras 10 películas de este archivo.

Informe gerencial

Use los métodos tabulares y gráficos de la estadística descriptiva para conocer cómo estas variables contribuyen al éxito de una película. Incluya en su informe los puntos que se indican en la siguiente página.

TABLA 2.18 Datos del desempeño de 10 películas

WEB archivo
Movies

Motion Picture	Opening Gross Sales ($millions)	Total Gross Sales ($millions)	Number of Theaters	Weeks in Top 60
Coach Carter	29.17	67.25	2574	16
Ladies in Lavender	0.15	6.65	119	22
Batman Begins	48.75	205.28	3858	18
Unleashed	10.90	24.47	1962	8
Pretty Persuasion	0.06	0.23	24	4
Fever Pitch	12.40	42.01	3275	14
Harry Potter and the Goblet of Fire	102.69	287.18	3858	13
Monster-in-Law	23.11	82.89	3424	16
White Noise	24.11	55.85	2279	7
Mr. and Mrs. Smith	50.34	186.22	3451	21

1. Los resúmenes tabulares y gráficos para cada una de las cuatro variables junto con un análisis de cada resumen que proporcionen información sobre la industria del cine.
2. Un diagrama de dispersión para explorar la relación entre las ventas brutas totales y las ventas brutas de estreno. Explíquelo.
3. Un diagrama de dispersión para explorar la relación entre las ventas brutas totales y el número de cines. Coméntelo.
4. Un diagrama de dispersión para explorar la relación entre las ventas brutas totales y el número de semanas en que la película permaneció entre las primeras 60. Coméntelo.

Apéndice 2.1 Uso de Minitab para presentaciones tabulares y gráficas

Minitab ofrece amplias capacidades para elaborar resúmenes tabulares y gráficos de los datos. En este apéndice se muestra cómo se usa para elaborar varios resúmenes gráficos y tabulares de una tabulación cruzada. Los métodos gráficos presentados incluyen el diagrama de puntos, el histograma, el diagrama de tallo y hoja, el diagrama de dispersión y la tabulación cruzada.

Diagrama de puntos

WEB archivo
Audit

Para esta demostración se utilizan los datos de duración de la auditoría de la tabla 2.4. Los datos están en la columna C1 de una hoja de trabajo de Minitab. Los pasos siguientes generarán un diagrama de puntos.

Paso 1. Seleccione el menú **Graph** y elija **Dotplot**.
Paso 2. Seleccione **One Y, Simple** y haga clic en **OK**.
Paso 3. Cuando el cuadro de diálogo Dotplot-One Y, Simple aparezca:
Introduzca C1 en el cuadro **Graph Variables**.
Haga clic en **OK**.

Histograma

WEB archivo
Audit

En esta sección se muestra cómo elaborar un histograma con las frecuencias sobre el eje vertical usando los datos de duración de la auditoría de la tabla 2.4. Los datos se encuentran en la columna C1 de la hoja de trabajo de Minitab. Los pasos siguientes generarán un histograma para la duración de las auditorías.

Paso 1. Seleccione el menú **Graph**.
Paso 2. Elija **Histogram**.
Paso 3. Seleccione **Simple** y haga clic en **OK**.
Paso 4. Cuando el cuadro de diálogo Histogram-Simple se abra:
Introduzca C1 en el cuadro **Graph Variables**.
Haga clic en **OK**.
Paso 5. Cuando aparezca el histograma:
Coloque el puntero del mouse sobre cualquiera de las barras.
Haga doble clic.
Paso 6. Cuando el cuadro de diálogo Edit Bars (editar barras) aparezca:
Haga clic en la ficha **Binning**.
Seleccione **Cutpoint** para el tipo de intervalo (Interval Type).
Seleccione **Midpoint/Cutpoint positions** para la definición del intervalo (Interval Definition).
Introduzca 10:35/5 en el cuadro **Midpoint/Cutpoint positions**.*
Haga clic en **OK**.

* La entrada 10:35/5 indica que 10 es el valor inicial para el histograma, 35 es el valor final para el histograma y 5 es el ancho de clase.

Observe que Minitab también proporciona la opción de aumentar o disminuir el eje *x* de modo que los valores numéricos aparezcan en los puntos medios de los rectángulos del histograma. Si desea activar esta opción, modifique el paso 6 para incluir **Select Midpoint** para el tipo de intervalo e introduzca 12:32/5 en el cuadro **Midpoint/Cutpoint positions**. Estos pasos producen el mismo histograma con los puntos medios de los rectángulos rotulados como 12, 17, 22, 27 y 32.

Diagrama de tallo y hoja

WEB archivo
ApTest

Los datos de la prueba de aptitudes de la tabla 2.8 se usan para mostrar la elaboración de un diagrama de tallo y hoja. Los datos se encuentran en la columna C1 de la hoja de trabajo de Minitab. Los pasos siguientes generarán el diagrama de tallo y hoja extendido que se muestra en la sección 2.3.

Paso 1. Seleccione el menú **Graph**.
Paso 2. Elija **Stem-and-Leaf**.
Paso 3. Cuando el cuadro de diálogo Stem-and-Leaf aparezca:
Introduzca C1 en el cuadro **Graph Variables**.
Haga clic en **OK**.

Diagrama de dispersión

WEB archivo
Stereo

Los datos de la tienda de estéreos y equipos de sonido de la tabla 2.12 se usan para mostrar la elaboración de un diagrama de dispersión. Las semanas están numeradas del 1 al 10 en la columna C1; los datos del número de comerciales están en la columna C2, y los datos de las ventas en la columna C3 de la hoja de trabajo de Minitab. Los pasos siguientes generan el diagrama de dispersión mostrado en la figura 2.7.

Paso 1. Seleccione el menú **Graph**.
Paso 2. Elija **Scatterplot**.
Paso 3. Seleccione **Simple** y haga clic en **OK**.
Paso 4. Cuando el cuadro de diálogo Scatterplot-Simple se abra:
Introduzca C3 bajo **Y variables** y C2 bajo **X variables**.
Haga clic en **OK**.

Tabulación cruzada

WEB archivo
Restaurant

Para esta demostración se utilizan los datos de la revisión del restaurante Zagat's, parte de los cuales se presentan en la tabla 2.9. Los restaurantes se numeran del 1 al 300 en la columna C1 de la hoja de trabajo de Minitab. Las calificaciones de calidad están en la columna C2 y los precios de los alimentos en la columna C3.

Minitab sólo puede crear una tabulación cruzada para variables cualitativas, y el precio de la comida es una variable cuantitativa. Así que primero necesitamos codificar los datos de la segunda variable al especificar la clase a la cual pertenece el precio de cada comida. Los pasos siguientes codificarán los datos del precio de los alimentos para crear cuatro clases en la columna C4: \$10–19, \$20–29, \$30–39 y \$40–49.

Paso 1. Seleccione el menú **Data**.
Paso 2. Elija **Code**.
Paso 3. Elija **Numeric to Text**.
Paso 4. Cuando el cuadro de diálogo Code-Numeric to Text aparezca:
Introduzca C3 en el cuadro **Code data from columns**.
Dé enter a C4 en el cuadro **Store coded data in columns**.
Introduzca 10:19 en el primer cuadro **Original values** y \$10-19 en el cuadro **New** adyacente.
Introduzca 20:29 en el segundo cuadro **Original values** y \$20-29 en el cuadro **New** adyacente.

Introduzca 30:39 en el tercer cuadro **Original values** y $30-39 en el cuadro **New** adyacente.
Introduzca 40:49 en el cuarto cuadro **Original values** y $40-49 en el cuadro **New** adyacente.
Haga clic en **OK**.

En la columna C4 aparecerá la categoría de precio asociada con cada precio de comida de la columna C3. Ahora podemos elaborar una tabulación cruzada para la calificación de la calidad y las categorías de precio de los alimentos usando los datos de las columnas C2 y C4. Los pasos siguientes crearán una tabulación cruzada que contiene la misma información que la mostrada en la tabla 2.10.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Tables**.
Paso 3. Elija **Cross Tabulation and Chi-Square**.
Paso 4. Cuando el cuadro de diálogo Cross Tabulation and Chi-Square se abra:
Introduzca C2 en el cuadro **For rows** y C4 en el cuadro **For columns**.
Seleccione **Counts** bajo **Display**.
Haga clic en **OK**.

Apéndice 2.2 Uso de Excel para presentaciones tabulares y gráficas

Excel ofrece muchas capacidades para elaborar resúmenes de datos tabulares y gráficos. En este apéndice se muestra cómo se usa este programa para elaborar una distribución de frecuencia, una gráfica de barras, una gráfica circular, un histograma, un diagrama de dispersión y una tabulación cruzada. Veremos el uso de tres de las herramientas más poderosas de Excel para el análisis de datos: herramientas de graficación y los informes de tablas dinámicas y gráficas dinámicas.

Distribución de frecuencia y gráfica de barras de datos cualitativos

En esta sección se muestra cómo se usa Excel para elaborar una distribución de frecuencia y una gráfica de barras de datos cualitativos. Ilustramos cómo se utilizan los datos sobre las compras de bebidas refrescantes de la tabla 2.1.

Distribución de frecuencia Comenzamos por mostrar el uso de la función COUNTIF para elaborar una distribución de frecuencia de los datos de la tabla 2.1. Vuelva a observar la figura 2.10 a medida que se describan los pasos requeridos. La hoja de trabajo de la fórmula (donde aparecen las funciones y las fórmulas usadas) se coloca en segundo plano, y la hoja de trabajo de valores (donde se muestran los resultados obtenidos mediante las funciones y las fórmulas) aparece en primer plano.

WEB archivo
SoftDrink

La etiqueta "Brand Purchased" y los datos para la compra de 50 bebidas refrescantes se encuentran en las celdas A1:A51. También se introdujeron las etiquetas "Soft Drink" y "Frequency" en las celdas C1:D1. Los cinco nombres de bebidas refrescantes se introducen en las celdas C2:C6. La función COUNTIF de Excel se usa entonces para contar el número de veces que aparece cada bebida en las celdas A2:A51. Siga estos pasos:

Paso 1. Seleccione la celda D2.
Paso 2. Introduzca =COUNTIF($A$2:$A$51,C2).
Paso 3. Copie la celda D2 a las celdas D3:D6.

La hoja de trabajo de la fórmula de la figura 2.10 muestra las fórmulas de las celdas que se insertan al aplicar estos pasos. La hoja de trabajo de valores registra los valores calculados por las fórmulas de las celdas y presenta la misma distribución de frecuencia que se elaboró en la tabla 2.2.

FIGURA 2.10 Distribución de frecuencia de la compra de bebidas refrescantes elaborada mediante la función COUNTIF de Excel

	A	B	C	D	E
1	**Brand Purchased**		**Soft Drink**	**Frequency**	
2	Coke Classic		Coke Classic	=COUNTIF($A$2:$A$51,C2)	
3	Diet Coke		Diet Coke	=COUNTIF($A$2:$A$51,C3)	
4	Pepsi		Dr. Pepper	=COUNTIF($A$2:$A$51,C4)	
5	Diet Coke		Pepsi	=COUNTIF($A$2:$A$51,C5)	
6	Coke Classic		Sprite	=COUNTIF($A$2:$A$51,C6)	
7	Coke Classic				
8	Dr. Pepper				
9	Diet Coke				
10	Pepsi				
45	Pepsi				
46	Pepsi				
47	Pepsi				
48	Coke Classic				
49	Dr. Pepper				
50	Pepsi				
51	Sprite				
52					

Nota. *Las filas 11-44 están ocultas.*

	A	B	C	D	E
1	**Brand Purchased**		**Soft Drink**	**Frequency**	
2	Coke Classic		Coke Classic	19	
3	Diet Coke		Diet Coke	8	
4	Pepsi		Dr. Pepper	5	
5	Diet Coke		Pepsi	13	
6	Coke Classic		Sprite	5	
7	Coke Classic				
8	Dr. Pepper				
9	Diet Coke				
10	Pepsi				
45	Pepsi				
46	Pepsi				
47	Pepsi				
48	Coke Classic				
49	Dr. Pepper				
50	Pepsi				
51	Sprite				
52					

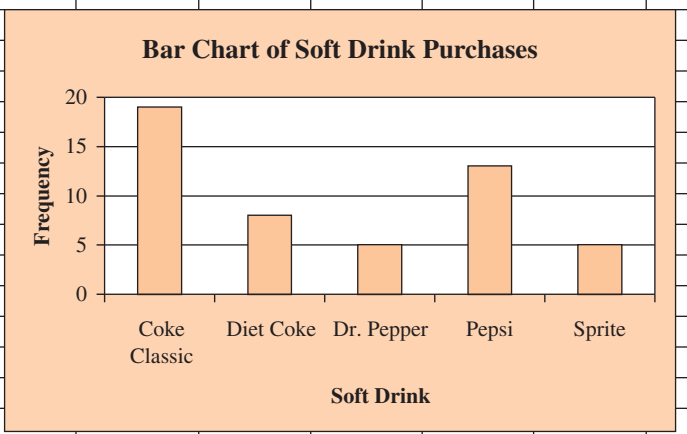
WEB archivo
SoftDrink

Gráfica de barras Aquí se explica cómo se utilizan las herramientas de graficación de Excel para elaborar una gráfica de barras de los datos de las bebidas refrescantes. Consulte la distribución de frecuencia mostrada en la hoja de trabajo de valores de la figura 2.10. La gráfica de barras que se diseñará es una extensión de esta hoja de trabajo. La gráfica de barras y la hoja de trabajo elaboradas se presentan en la figura 2.11. Los pasos para realizarlas son los siguientes.

Paso 1. Seleccione las celdas C2:D6.
Paso 2. Haga clic en la ficha **Insert** de la cinta de opciones.
Paso 3. En el grupo **Charts**, haga clic en **Column**.
Paso 4. Cuando aparezca la lista de subtipos de gráficas de columna:
Vaya a la sección **2-D Column**.
Haga clic en la gráfica del extremo izquierdo, **Clustered Column**.
Paso 5. En el grupo **Chart Layouts**, haga clic en el botón **More** (la flecha que apunta hacia abajo con una línea sobre ella) para ver todas las opciones.
Paso 6. Elija el **Layout 9**.
Paso 7. Seleccione **Chart Title** y remplace el nombre de la gráfica con **BarChart of Soft Drink Purchases**.
Paso 8. Elija **Horizontal (Category) Axis Title** y remplácelo con **Soft Drink**.
Paso 9. Seleccione **Vertical (Value) Axis Title** y sustitúyalo con **Frequency**.
Paso 10. Haga clic con el botón secundario en **Series 1 Legend Entry**.
Haga clic en **Delete**.
Paso 11. Haga clic con el botón secundario en el eje vertical.
Haga clic en **Format Axis**.

FIGURA 2.11 Gráfica de barras de la compra de bebidas refrescantes elaborada con las herramientas de graficación de Excel

	A	B	C	D	E	F	G	H	I
1	**Brand Purchased**		**Soft Drink**	**Frequency**					
2	Coke Classic		Coke Classic	19					
3	Diet Coke		Diet Coke	8					
4	Pepsi		Dr. Pepper	5					
5	Diet Coke		Pepsi	13					
6	Coke Classic		Sprite	5					
7	Coke Classic								
8	Dr. Pepper								
9	Diet Coke								
10	Pepsi								
11	Pepsi								
12	Coke Classic								
13	Dr. Pepper								
14	Sprite								
15	Coke Classic								
16	Diet Coke								
17	Coke Classic								
18	Coke Classic								
19	Sprite								
20	Coke Classic								
50	Pepsi								
51	Sprite								
52									



Paso 12. Cuando el cuadro de diálogo Format Axis se abra:
Vaya a la sección **Axis Options**.
Seleccione **Fixed** para **Major Unit** e introduzca 5.0 en el cuadro correspondiente.
Haga clic en **Close**.

La gráfica de barras resultante se muestra en la figura 2.11.* Excel puede producir una gráfica circular para los datos de la compra de bebidas refrescantes de una manera parecida. La principal diferencia radica en que en el paso 3 haría clic en **Pie** en el grupo **Charts**. Existen varios estilos de gráficas circulares.

Distribución de frecuencia e histograma para datos cuantitativos

En una sección posterior de este apéndice se describe cómo usar PivotTable Report de Excel para elaborar una tabulación cruzada.

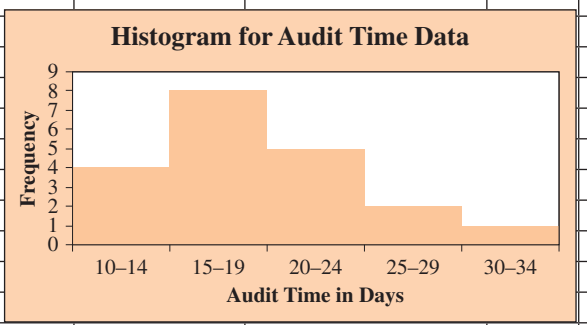
El informe de tabla dinámica de Excel (PivotTable Report) es una herramienta interactiva que permite resumir los datos de manera rápida en una variedad de maneras, que incluyen la elaboración de una distribución de frecuencia para datos cuantitativos. Una vez que se crea una distribución de frecuencia utilizando el informe de tabla dinámica, entonces se usan las herramientas de graficación de Excel para preparar el histograma correspondiente. No obstante, el informe de gráfico dinámico de Excel permite elaborar de forma simultánea una distribución de frecuencia y un histograma. Ilustraremos este procedimiento usando los datos de duración de la auditoría de la tabla 2.4. La etiqueta "Audit Time" y los 20 valores de la duración de la auditoría se introdujeron en las celdas A1:A21 de la hoja de cálculo de Excel. Los pasos listados en la siguiente página describen cómo usar el informe de gráfico dinámico de Excel para preparar una distribución de frecuencia y un histograma de dichos datos. Vuelva a observar la figura 2.12 cuando se describan los pasos correspondientes.

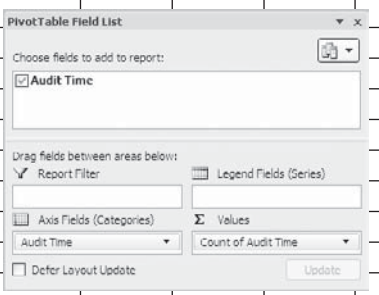
WEB archivo
Audit

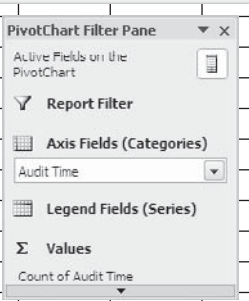
* El tamaño de la gráfica de barras de la figura 2.11 puede modificarse. Cambiar el tamaño de una gráfica en Excel no es difícil. Primero seleccione la gráfica. Los controladores de tamaño aparecerán en el borde de ésta. Haga clic en los controladores y arrástrelos para modificar el tamaño de la figura como lo desee.

FIGURA 2.12 Uso del informe de gráfico dinámico de Excel para elaborar una distribución de frecuencia y un histograma de los datos de duración de la auditoría

	A	B	C	D	E	F	G	H	I	J
1	**Audit Time**		**Row Labels**	**Count of Audit Time**						
2	12		10–14	4						
3	15		15–19	8						
4	20		20–24	5						
5	22		25–29	2						
6	14		30–34	1						
7	14		**Grand Total**	**20**						
8	15									
9	27									
10	21									
11	18									
12	19									
13	18									
14	22									
15	33									
16	16									
17	18									
18	17									
19	23									
20	28									
21	13									
22										







Paso 1. Haga clic en la ficha **Insert** de la cinta de opciones.
Paso 2. En el grupo **Tables**, haga clic en la palabra **PivotTable**.
Paso 3. Elija **PivotChart** de las opciones que aparecen.
Paso 4. Cuando el cuadro de diálogo **Create PivotTable with PivotChart** aparezca:
Seleccione **Select a table or range**.
Introduzca A1:A21 en el cuadro **Table/Range**.
Elija **Existing Worksheet** como la ubicación para la PivotTable y PivotChart.
Introduzca C1 en el cuadro **Location**.
Haga clic en **OK**.
Paso 5. En **PivotTable Field List**, vaya a **Choose Fields to add to report**.
Arrastre el campo **Audit Time** al área **Axis Fields** (Categories).
Arrastre el campo **Audit Time** al área **Values**.
Paso 6. Haga clic en **Sum of Audit Time** en la sección **Values**.
Paso 7. Dé clic en **Value Field Settings** de la lista de opciones que aparece.
Paso 8. Cuando el cuadro de diálogo Value Field Settings se abra:
Bajo **Summarize value field by**, elija **Count**.
Haga clic en **OK**.
Paso 9. Cierre la **PivotTable Field List**.
Paso 10. Haga clic con el botón secundario del mouse en la celda C2 en el informe de PivotTable o en cualquier otra celda que contenga una duración de la auditoría.
Paso 11. Elija **Group** de la lista de opciones que aparezca.
Paso 12. Cuando se abra el cuadro de diálogo **Grouping**.
Introduzca 10 en el cuadro **Starting at**.

Introduzca 34 en el cuadro **Ending at**.
Introduzca 5 en el cuadro **By**.
Haga clic en **OK** (aparecerá un gráfico dinámico).

Paso 13. Haga clic dentro del **PivotChart** resultante.
Paso 14. Haga clic en la ficha **Design** en la cinta de opciones.
Paso 15. En el grupo **Chart Layouts**, haga clic en el botón **More** (la flecha que apunta hacia abajo con una línea sobre ella) para mostrar todas las opciones.
Paso 16. Elija **Layout 8**.
Paso 17. Seleccione el **Chart Title** y remplácelo con **Histogram forAudit Time Data**.
Paso 18. Seleccione **Horizontal (Category) Axis Title** y sustitúyalo con **Audit Time in Days**.
Paso 19. Elija el título **Vertical (Value) Axis Title** y remplácelo con **Frequency**.

La figura 2.12 muestra los informes de tabla dinámica y gráfico dinámico resultantes. Observamos que el informe de tabla dinámica proporciona la distribución de frecuencia de los datos de duración de la auditoría y el informe de gráfico dinámico proporciona el histograma correspondiente. Si lo desea, puede cambiar las etiquetas de cualquier celda en la distribución de frecuencia al seleccionar la celda y teclear la etiqueta nueva.

Tabulación cruzada

El informe de tabla dinámica de Excel es una manera excelente de resumir los datos para dos o más variables de forma simultánea. Se explicará el uso de este informe al mostrar cómo elaborar una tabulación cruzada de las calificaciones de calidad y los precios de la comida de la muestra de 300 restaurantes de Los Ángeles. Se usarán los datos del archivo llamado *Restaurant*; las etiquetas "Restaurant", "Quality Rating" (calificación de la calidad) y "Meal Price ($)" (precio de la comida) se introdujeron en las celdas A1:C1 de la hoja de cálculo como se aprecia en la figura 2.13. Los datos de cada uno de los restaurantes de la muestra se introdujeron en las celdas B2:C301.

FIGURA 2.13 Hoja de cálculo de Excel que contiene datos de los restaurantes

WEB archivo
Restaurant

	A	B	C	D
1	Restaurant	Quality Rating	Meal Price ($)	
2	1	Good	18	
3	2	Very Good	22	
4	3	Good	28	
5	4	Excellent	38	
6	5	Very Good	33	
7	6	Good	28	
8	7	Very Good	19	
9	8	Very Good	11	
10	9	Very Good	23	
11	10	Good	13	
292	291	Very Good	23	
293	292	Very Good	24	
294	293	Excellent	45	
295	294	Good	14	
296	295	Good	18	
297	296	Good	17	
298	297	Good	16	
299	298	Good	15	
300	299	Very Good	38	
301	300	Very Good	31	
302				

Nota. *Las filas 12-291 están ocultas.*

Si se desea utilizar el informe de tabla dinámica para elaborar una tabulación cruzada, se requiere realizar tres tareas: mostrar la lista de campos de la tabla dinámica inicial y el informe de tabla dinámica; establecer la lista de campos de la tabla dinámica, y finalizar el informe de tabla dinámica. Estas tareas se describen enseguida.

Mostrar la lista de campos de la tabla dinámica inicial y el informe de tabla dinámica. Se requieren tres pasos para mostrar la lista de campos inicial y el informe de tabla dinámica.

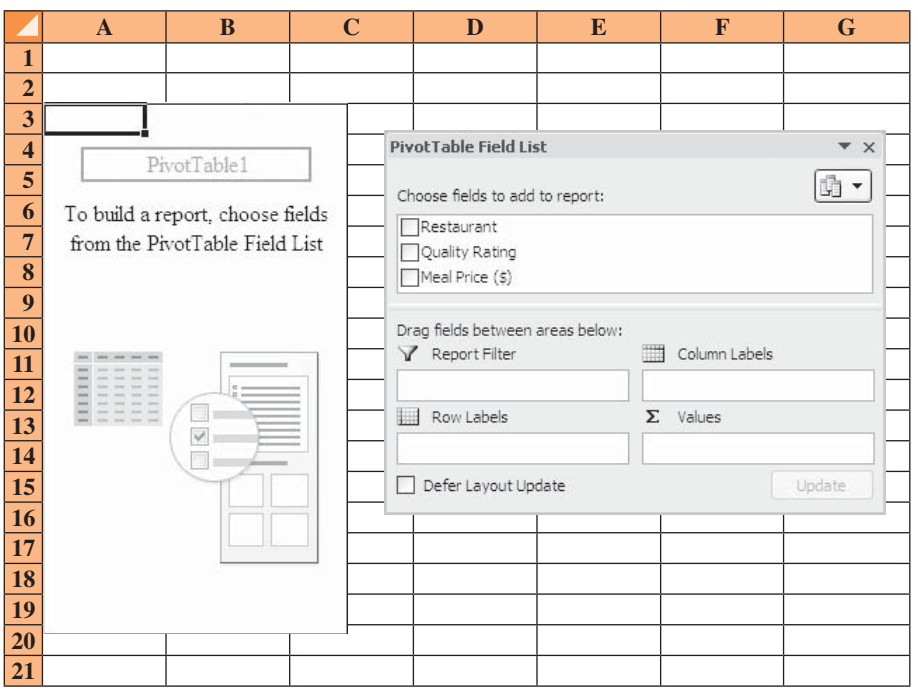
Paso 1. Haga clic en la ficha **Insert** en la cinta de opciones.
Paso 2. En el grupo **Tables**, haga clic en el icono sobre la palabra PivotTable.
Paso 3. Cuando el cuadro de diálogo **Create PivotTable** aparezca:
Elija **Select a Table or Range**.
Introduzca A1:C301 en el cuadro **Table/Range**.
Elija **New Worksheet** como la ubicación para PivotTable Report.
Haga clic en **OK**.

La lista de campo inicial de tabla dinámica y el informe de tabla dinámica se muestran en la figura 2.14.

Configuración de la lista de campos inicial de tabla dinámica. Excel considera cada una de las tres columnas de la figura 2.13 [etiquetadas como Restaurant, Quality Rating y Meal Price ($)] como un campo. Los campos se eligen para representar filas, columnas o valores en el cuerpo del informe de tabla dinámica. Los pasos siguientes muestran cómo utilizar la lista de campos de tabla dinámica de Excel para asignar el campo Quality Rating a las filas, el campo Meal Price ($) a las columnas y el campo Restaurant al cuerpo del informe de la tabla dinámica.

Paso 1. En **PivotTable Field List**, vaya a **Choose Fields to add to report**.
Arrastre el campo **Quality Rating** a la sección **Row Labels**.
Arrastre el campo **Meal Price ($)** a la sección **Column Labels**.
Arrastre el campo **Restaurant** a la sección **Values**.

FIGURA 2.14 Lista de campo inicial e informe de campo de PivotTable para los datos del restaurante



Paso 2. Haga clic con el botón secundario en **Sum of Restaurant** en la sección **Values**.
Paso 3. Haga clic en la opción **Value Field Settings** de la lista de opciones que se despliega.
Paso 4. Cuando el cuadro de diálogo **Value Field Settings** se abra:
Bajo **Summarize value field by**, elija **Count**.
Haga clic en **OK**.

La figura 2.15 muestra la lista de campos de tabla dinámica completada y una porción de la hoja de trabajo de tabla dinámica tal como aparece.

Finalizar el informe de tabla dinámica. Para completar el informe de tabla dinámica se necesita agrupar las columnas que representan los precios de la comida y colocar las etiquetas de fila para la calificación de la calidad en el orden apropiado. Considere los pasos siguientes para hacerlo.

Paso 1. Haga clic con el botón secundario en la celda B4 o en cualquier otra que contenga precios de comida.
Paso 2. Elija **Group** en la lista de opciones que aparece.
Paso 3. Cuando el cuadro de diálogo **Grouping** se abra:
Introduzca 10 en el cuadro **Starting at**.
Introduzca 49 en el cuadro **Ending at**.
Introduzca 10 en el cuadro **By**.
Haga clic en **OK**.
Paso 4. Haga clic con el botón secundario en **Excellent** en la celda A5.
Paso 5. Elija **Move** y haga clic en **Move "Excellent" to End**.

El informe de tabla dinámica final se muestra en la figura 2.16. Observe que contiene la misma información que la tabulación cruzada de la tabla 2.10.

Diagrama de dispersión

Las herramientas de graficación de Excel se utilizaron para elaborar un diagrama de dispersión y una línea de tendencia de los datos de la tienda de estéreos y equipos de sonido presentados

FIGURA 2.15 Lista de campos de la tabla dinámica completada y una porción del informe de tabla dinámica de los datos del restaurante (las columnas H:AK están ocultas)

	A	B	C	D	E	F	G	AL	AM	AN	AO
1											
2											
3	**Count of Restaurant**	**Column Labels**									
4	**Row Labels**	**10**	**11**	**12**	**13**	**14**	**15**	**47**	**48**	**Grand Total**	
5	Excellent				1			2	2	66	
6	Good	6	4	3	3	2	4			84	
7	Very Good	1	4	3	5	6	1		1	150	
8	**Grand Total**	**7**	**8**	**6**	**9**	**8**	**5**	**2**	**3**	**300**	
9											
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											

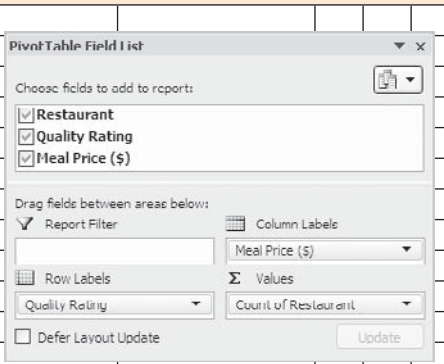


FIGURA 2.16 Informe de tabla dinámica final de los datos del restaurante

	A	B	C	D	E	F	G
1							
2							
3	**Count of Restaurant**	**Column Labels**					
4	**Row Labels**	**10–19**	**20–29**	**30–39**	**40–49**	**Grand Total**	
5	Good	42	40	2	1	84	
6	Very Good	34	64	46	6	150	
7	Excellent	2	14	28	25	66	
8	**Grand Total**	**78**	**118**	**76**	**28**	**300**	
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							

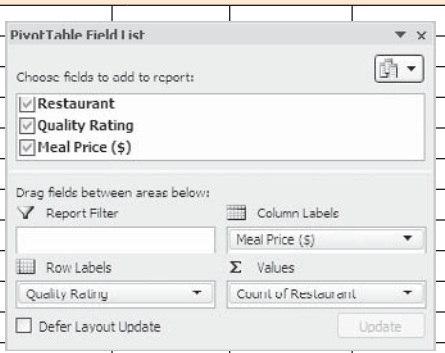


FIGURA 2.17 Diagrama de dispersión de la tienda de estéreos y equipos de sonido usando las herramientas de graficación de excel

	A	B	C	D	E	F	G	H
1	**Week**	**No. of Commercials**	**Sales Volume**					
2	1	2	50					
3	2	5	57					
4	3	1	41					
5	4	3	54					
6	5	4	54					
7	6	1	38					
8	7	5	63					
9	8	3	48					
10	9	4	59					
11	10	2	46					
12								
13								
14								
15								
16								
17								
18								
19								
20								

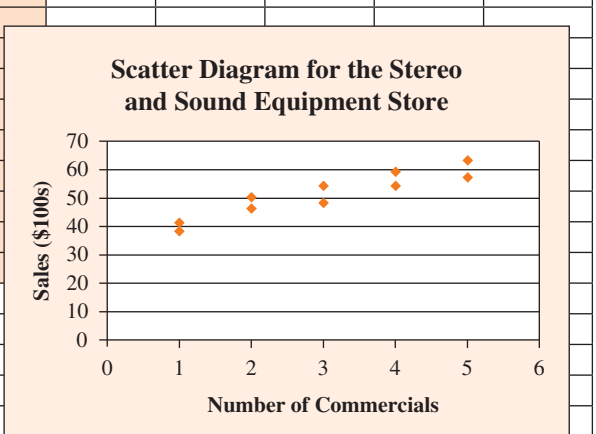
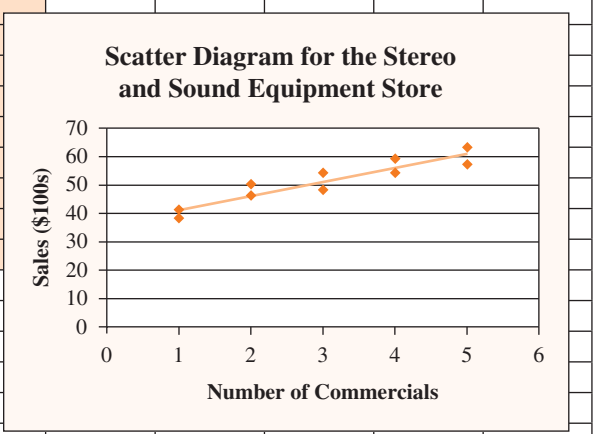


FIGURA 2.18 Diagrama de dispersión y línea de tendencia de la tienda de estéreos y equipos de sonido usando las herramientas de graficación de Excel

	A	B	C	D	E	F	G	H
1	Week	No. of Commercials	Sales Volume					
2	1	2	50					
3	2	5	57					
4	3	1	41					
5	4	3	54					
6	5	4	54					
7	6	1	38					
8	7	5	63					
9	8	3	48					
10	9	4	59					
11	10	2	46					
12								
13								
14								
15								
16								
17								
18								
19								
20								



en la tabla 2.12. Vuelva a observar las figuras 2.17 y 2.18 a medida que se describan los pasos correspondientes. Usaremos los datos del archivo llamado *Stereo*; las etiquetas Week, No. of Commercials y Sales Volume se introdujeron en las celdas A1:C1 de la hoja de cálculo. Los datos de cada una de las 10 semanas se introdujeron en las celdas B2:C11. Los pasos siguientes describen cómo usar las herramientas de graficación de Excel para producir un diagrama de dispersión de los datos.

Paso 1. Seleccione las celdas B2:C11.
Paso 2. Haga clic en la ficha **Insert** en la cinta de opciones.
Paso 3. En el grupo **Charts**, haga clic en **Scatter**.
Paso 4. Cuando la lista de subtipos de diagramas de dispersión se abra, haga clic en **Scatter with only Markers** (la tabla de la esquina superior izquierda).
Paso 5. En el grupo **Chart Layouts**, haga clic en **Layout 1**.
Paso 6. Elija el **Chart Title** y remplácelo con **Scatter Diagram for the Stereo and Sound Equipment Store**.
Paso 7. Seleccione **Horizontal (Value) Axis Title** y remplácelo con **Number of Commercials**.
Paso 8. Seleccione **Vertical (Value) Axis Title** y remplácelo con **Sales ($100s)**.
Paso 9. Haga clic con el botón secundario en **Series 1 Legend Entry** y haga clic en **Delete**.

La hoja de cálculo de la figura 2.17 muestra el diagrama de dispersión producido por Excel. Los pasos siguientes describen cómo añadir una línea de tendencia.

Paso 1. Coloque el puntero del *mouse* sobre cualquier punto de datos en el diagrama de dispersión y haga clic con el botón secundario para mostrar una lista de opciones.
Paso 2. Elija **Add Trendline**.
Paso 3. Cuando el cuadro de diálogo **Format Trendline** se abra:
Seleccione **Trendline Options**.
Elija **Linear** de la lista **Trend/Regression Type**.
Haga clic en **Close**.

La hoja de cálculo de la figura 2.18 muestra el diagrama de dispersión con la línea de tendencia agregada.

Apéndice 2.3 Uso de StatTools para presentaciones tabulares y gráficas

En este apéndice se explica cómo se utiliza StatTools para elaborar un histograma y un diagrama de dispersión.

Histograma

Recurrimos a los datos de duración de la auditoría en la tabla 2.4 para la explicación. Comience usando Data Set Manager para crear un conjunto de datos StatTools para esos datos por medio del procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes generarán un histograma.

WEB archivo
Audit

Paso 1. Haga clic en la ficha **StatTools** en la cinta de opciones.
Paso 2. En **Analyses Group**, haga clic en **Summary Graphs**.
Paso 3. Elija la opción **Histogram**.
Paso 4. Cuando el cuadro de diálogo StatTools–Histogram se abra:
En la sección **Variables**, seleccione **Audit Time**.
En la sección **Options**:
Introduzca 5 en el cuadro **Number of Bins**.
Introduzca 9.5 en el cuadro **Histogram Minimum**.
Introduzca 34.5 en el cuadro **Histogram Maximum**.
Elija **Categorical** en el cuadro **X-Axis**.
Elija **Frequency** en el cuadro **Y-Axis**.
Haga clic en **OK**.

Aparecerá un histograma para los datos de duración de la auditoría parecido al de la figura 2.12. La única diferencia es que el histograma elaborado usando StatTools muestra los puntos medios de clase en el eje horizontal.

Diagrama de dispersión

WEB archivo
Stereo

Para mostrar la elaboración de un diagrama de dispersión con StatTools se utilizan los datos de los estéreos y los equipos de sonido de la tabla 2.12. Comience con Data Set Manager para crear un conjunto de datos StatTools de estos datos mediante el procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes generarán un diagrama de dispersión.

Paso 1. Haga clic en la ficha **StatTools** en la cinta de opciones.
Paso 2. En **Analyses Group**, haga clic en **Summary Graphs**.
Paso 3. Elija la opción **Scatterplot**.
Paso 4. Cuando aparezca el cuadro de diálogo StatTools–Scatterplot:
En la sección **Variables**,
En la columna etiquetada **X**, seleccione **No. of Commercials**.
En la columna etiquetada **Y**, elija **Sales Volume**.
Haga clic en **OK**.

Aparecerá un diagrama de dispersión parecido al de la figura 2.17.

CAPÍTULO 3

Estadística descriptiva: medidas numéricas

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: SMALL FRY DESIGN

3.1 MEDIDAS DE POSICIÓN O LOCALIZACIÓN
- Media
- Mediana
- Moda
- Percentiles
- Cuartiles

3.2 MEDIDAS DE VARIABILIDAD
- Rango
- Rango intercuartílico
- Varianza
- Desviación estándar
- Coeficiente de variación

3.3 MEDIDAS DE LA FORMA DE LA DISTRIBUCIÓN, POSICIÓN RELATIVA Y DETECCIÓN DE OBSERVACIONES ATÍPICAS
- Forma de la distribución
- Valor *z*
- Teorema de Chebyshev
- Regla empírica
- Detección de observaciones atípicas

3.4 ANÁLISIS EXPLORATORIO DE DATOS
- Resumen de cinco números
- Diagrama de caja

3.5 MEDIDAS DE ASOCIACIÓN ENTRE DOS VARIABLES
- Covarianza
- Interpretación de la covarianza
- Coeficiente de correlación
- Interpretación del coeficiente de correlación

3.6 MEDIA PONDERADA Y TRABAJO CON DATOS AGRUPADOS
- Media ponderada
- Datos agrupados

ESTADÍSTICA *en* LA PRÁCTICA

SMALL FRY DESIGN*

SANTA ANA, CALIFORNIA

Small Fry Design, fundada en 1997, es una compañía de juguetes y accesorios que diseña e importa productos para niños. La línea de artículos de la empresa incluye osos de peluche, móviles, juguetes musicales, sonajas y cobertores de seguridad, y presenta diseños de juguetes de alta calidad para bebé con un énfasis en los colores, las texturas y los sonidos. Los productos se diseñan en Estados Unidos y se fabrican en China.

Small Fry Design emplea a representantes independientes para la venta de sus productos a minoristas de muebles infantiles, tiendas de accesorios y ropa para niños, negocios de regalos, tiendas departamentales exclusivas e importantes compañías de ventas por catálogo. En la actualidad, los productos de Small Fry Design se distribuyen en más de 1 000 puntos de venta minoristas en todo Estados Unidos.

La administración del flujo de efectivo es una de las actividades más importantes para la operación diaria de esta empresa. Garantizar que dicho flujo entrante sea suficiente para cumplir con las obligaciones de deudas tanto corrientes como a corto plazo puede significar la diferencia entre el éxito y el fracaso. Un factor crítico en la administración del flujo de efectivo es el análisis y control de las cuentas por cobrar. Al medir el tiempo promedio de cobro y el valor monetario de las facturas pendientes, la gerencia puede predecir la disponibilidad de efectivo y monitorear los cambios en el estado de las cuentas por cobrar. La empresa estableció las metas siguientes: la antigüedad promedio de las facturas pendientes no debe exceder los 45 días y el valor de las facturas con una antigüedad mayor a 60 días no debe exceder 5% del valor de todas las cuentas por cobrar.

En un resumen reciente del estado de las cuentas por cobrar se proporcionó la siguiente estadística descriptiva para la antigüedad de las facturas pendientes.

Media	40 días
Mediana	35 días
Moda	31 días



Móvil "Rey de la selva" de Small Fry Design.
© Joe-Higgins/South-Western.

La interpretación de estos datos estadísticos muestra que el tiempo promedio de cobro de una factura es de 40 días. La mediana señala que la mitad de estos documentos permanece pendiente 35 días o más. La moda de 31 días, el tiempo de cobro de una factura más frecuente, indica que el lapso más común en que ésta permanece pendiente es de 31 días. El resumen estadístico indica también que sólo 3% del valor de todas las cuentas por cobrar tiene un tiempo de cobro de más de 60 días. Con base en la información estadística, la gerencia quedó satisfecha, dado que las cuentas por cobrar y el flujo de efectivo entrante estaban bajo control.

En este capítulo aprenderá a calcular e interpretar algunas de las medidas estadísticas que utiliza Small Fry Design. Además de la media, la mediana y la moda, aprenderá otros datos de estadística descriptiva, como el rango, la varianza, la desviación estándar, los percentiles y la correlación. Estas medidas numéricas ayudan a la comprensión e interpretación de los datos.

* Los autores agradecen a John A. McCarthy, presidente de Small Fry Design, por proporcionar este artículo para *Estadística en la práctica.*

En el capítulo 2 se estudiaron las presentaciones tabulares y gráficas utilizadas para resumir los datos. En este capítulo se presentan varias medidas numéricas que proporcionan otras opciones para la misma tarea.

Primero se verá el desarrollo de medidas numéricas para conjuntos de datos que constan de una sola variable. Cuando un conjunto de datos contiene más de una variable, las mismas medidas numéricas se calculan por separado para cada variable. Sin embargo, en el caso de dos variables, se desarrollarán también medidas de la relación entre éstas.

Se presentan las medidas numéricas de posición, dispersión, forma y asociación. Si las medidas se calculan para los datos de una muestra, se les llama **estadístico muestral**. Si se calculan para los datos de una población, se les llama **parámetros poblacionales**. En la inferencia estadística, un estadístico muestral se conoce como **estimador puntual** del parámetro poblacional correspondiente. En el capítulo 7 se verá con más detalle el proceso de la estimación puntual.

En los tres apéndices del capítulo se explica cómo se usan Minitab, Excel y StatTools para calcular las medidas numéricas descritas en el capítulo.

3.1 Medidas de posición o localización

Media

La **media**, o valor medio, es quizá la medida de ubicación más importante para una variable, pues proporciona una medida de la ubicación central de los datos. Si los datos son para una muestra, la media se denota por $\bar{x}$; si son para una población, se denota por la letra griega μ.

En las fórmulas estadísticas se acostumbra denotar el valor de la primera observación de la variable x mediante x_1, el valor de la segunda observación de la variable x por medio de x_2, y así sucesivamente. En general, el valor de la i-ésima observación de la variable x se representa por medio de x_i. Si se tiene una muestra con n observaciones, la fórmula para la media muestral es la siguiente.

La media muestral $\bar{x}$ es un estadístico muestral.

MEDIA MUESTRAL

$$\bar{x} = \frac{\Sigma x_i}{n} \qquad (3.1)$$

En la fórmula anterior, el numerador es la suma de los valores de las n observaciones. Es decir,

$$\Sigma x_i = x_1 + x_2 + \cdots + x_n$$

La letra griega Σ es el signo de sumatoria.

Para ilustrar el cálculo de una media muestral, considere los datos siguientes sobre el tamaño del grupo para una muestra de cinco grupos de estudiantes universitarios.

46 54 42 46 32

La notación x_1, x_2, x_3, x_4, x_5 se utiliza para representar el número de estudiantes en cada uno de los cinco grupos.

$$x_1 = 46 \qquad x_2 = 54 \qquad x_3 = 42 \qquad x_4 = 46 \qquad x_5 = 32$$

Por consiguiente, para calcular la media muestral se escribe

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{46 + 54 + 42 + 46 + 32}{5} = 44$$

El tamaño de grupo de la media muestral es 44 estudiantes.

Otro ejemplo del cálculo de una media muestral se da en la situación siguiente. Suponga que una oficina de colocación de empleos a nivel universitario envió un cuestionario a una muestra de licenciados en administración de empresas recién egresados solicitando información sobre

TABLA 3.1 Sueldos mensuales iniciales para una muestra de 12 licenciados en administración de empresas recién egresados

Graduate	Monthly Starting Salary ($)	Graduate	Monthly Starting Salary ($)
1	3450	7	3490
2	3550	8	3730
3	3650	9	3540
4	3480	10	3925
5	3355	11	3520
6	3310	12	3480

WEB archivo
StartSalary

los sueldos mensuales iniciales. La tabla 3.1 exhibe los datos reunidos. El sueldo mensual inicial medio para la muestra de 12 licenciados en administración de empresas se calcula como sigue:

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + \cdots + x_{12}}{12}$$

$$= \frac{3450 + 3550 + \cdots + 3480}{12}$$

$$= \frac{42480}{12} = 3540$$

La ecuación (3.1) ilustra cómo se calcula la media para una muestra con n observaciones. La fórmula para determinar la media de una población es la misma, pero se usa una notación diferente para indicar que se está trabajando con toda la población. El número de observaciones en una población se denota por N y el símbolo para la media poblacional es μ.

La media muestral $\bar{x}$ es un estimador puntual de la media poblacional μ.

MEDIA POBLACIONAL

$$\mu = \frac{\Sigma x_i}{N} \qquad (3.2)$$

Mediana

La **mediana** es otra medida de ubicación central; es el valor de en medio cuando los datos están acomodados en orden ascendente (del valor menor al valor mayor). Con un número impar de observaciones, la mediana es el valor de en medio. Con un número par, no hay valor de en medio. En este caso se sigue la convención y la mediana se define como el promedio de los valores de las dos observaciones de en medio. Por conveniencia, la definición de la mediana se replantea como sigue.

MEDIANA

Ordene los datos de forma ascendente (del valor menor al valor mayor).

a) Para un número impar de observaciones, la mediana es el valor de en medio.
b) Para un número par de observaciones, la mediana es el promedio de los dos valores de en medio.

Esta definición se aplica para calcular la mediana de los tamaños de grupo para la muestra de cinco grupos de estudiantes universitarios. Al ordenar los datos de forma ascendente se obtiene la lista siguiente.

$$32 \quad 42 \quad 46 \quad 46 \quad 54$$

Dado que $n = 5$ es impar, la mediana es el valor de en medio. Por tanto, la mediana del tamaño de grupo es 46 estudiantes. Aun cuando este conjunto de datos contiene dos observaciones con valores de 46, cada una se trata de forma separada cuando los datos se acomodan en orden ascendente.

Suponga además que se calcula la mediana de los sueldos iniciales para los 12 licenciados en administración de empresas de la tabla 3.1. Primero se acomodan los datos en orden ascendente.

$$3\,310 \quad 3\,355 \quad 3\,450 \quad 3\,480 \quad 3\,480 \quad \underbrace{3\,490 \quad 3\,520}_{\text{Los dos valores de en medio}} \quad 3\,540 \quad 3\,550 \quad 3\,650 \quad 3\,730 \quad 3\,925$$

Como $n = 12$ es par, se identifican los dos valores de en medio: la mediana es el promedio de estos dos valores.

$$\text{Mediana} = \frac{3\,490 + 3\,520}{2} = 3\,505$$

La mediana es la medida de posición más empleada para los datos de los ingresos anuales y el valor de propiedad, debido a que algunos ingresos o valores de propiedad muy grandes pueden inflar la media. En tales casos, la mediana es la medida preferida de posición central.

Aunque la media es la medida de posición central de uso más común, en algunas situaciones se prefiere la mediana, ya que los valores de datos muy pequeños y muy grandes influyen en la media. Por ejemplo, suponga que uno de los licenciados recién graduados (tabla 3.1) tenía un sueldo inicial de $10 000 al mes (tal vez la empresa es propiedad de su familia). Si se cambia el sueldo mensual inicial más alto de la tabla 3.1 de $3 925 a $10 000 y se vuelve a calcular la media, la media muestral pasa de $3 540 a $4 046. Sin embargo, la mediana de $3 505 permanece igual, ya que $3 490 y $3 520 siguen siendo los dos valores de en medio. Si el sueldo inicial es sumamente alto, la mediana proporciona una mejor medida de posición central que la media. Al hacer una generalización, se afirma que siempre que un conjunto de datos contiene valores extremos, la mediana suele ser la medida preferida de posición central.

Moda

Una tercera medida de posición es la **moda**. Se define de la manera siguiente.

MODA

La moda es el valor que ocurre con mayor frecuencia.

Para ilustrar cómo identificar la moda, considere el tamaño de grupo de la muestra de cinco grupos de estudiantes universitarios. El único valor que ocurre más de una vez es el 46. Debido a que se presenta con una frecuencia de 2, que es la frecuencia más grande, se le considera la moda. Como otro ejemplo, considere la muestra de sueldos iniciales de los licenciados en administración de empresas. El único sueldo mensual inicial que ocurre más de una vez es $3 480. Dado que este valor tiene la frecuencia mayor, es la moda.

Hay situaciones en que la frecuencia mayor ocurre en dos o más valores diferentes; cuando esto sucede, existe más de una moda. Si los datos contienen exactamente dos modas, se dice que son *bimodales.* Si contienen más de dos, se dice que son *multimodales.* En estos casos, la moda casi nunca se presenta debido a que listar tres o más no resulta particularmente útil para describir la posición de los datos.

Percentiles

Un **percentil** proporciona información sobre cómo se distribuyen los datos en el intervalo del valor menor al valor mayor. Para datos que no contienen muchos valores repetidos, el percentil p-ésimo los divide en dos partes. Alrededor de p por ciento de las observaciones tiene valores menores que el percentil p-ésimo y cerca de $(100 - p)$ por ciento de las observaciones tiene valores mayores que el percentil p-ésimo. Éste se define formalmente del modo siguiente.

PERCENTIL

El percentil p-ésimo es un valor tal que *por lo menos* p por ciento de las observaciones es menor o igual que este valor, y *por lo menos* $(100 - p)$ por ciento de las observaciones es mayor o igual que este valor.

Los colegios y universidades suelen reportar los resultados de los exámenes de admisión en términos de percentiles. Por ejemplo, suponga que un solicitante obtiene una puntuación bruta de 54 en la parte verbal de un examen de admisión. Esta información no dice mucho acerca del desempeño que este estudiante tuvo en relación con otros que presentaron el mismo examen. Sin embargo, si la puntuación bruta de 54 corresponde al percentil 70, se sabe que aproximadamente 70% de los estudiantes obtuvo una puntuación menor a la de esta persona y alrededor de 30% alcanzó una puntuación mayor a la de esta persona.

El procedimiento siguiente se usa para calcular el p-ésimo percentil.

La ejecución de estos pasos facilita el cálculo de percentiles.

CÁLCULO DEL p-ÉSIMO PERCENTIL

Paso 1. Ordene los datos de modo ascendente (del valor menor al valor mayor).

Paso 2. Calcule un índice i

$$i = \left(\frac{p}{100}\right)n$$

donde p es el percentil de interés y n es el número de observaciones.

Paso 3. *a*) Si i *no es un entero, redondéelo.* El entero siguiente *mayor* que i denota la posición del p-ésimo percentil.

b) Si i *es un entero,* el p-ésimo percentil es el promedio de los valores en las posiciones i e $i + 1$.

Como ejemplo de este procedimiento, se determinará el percentil 85 para los datos de los sueldos iniciales mensuales de la tabla 3.1.

Paso 1. Ordene los datos de modo ascendente.

3 310 3 355 3 450 3 480 3 480 3 490 3 520 3 540 3 550 3 650 3 730 3 925

Paso 2.

$$i = \left(\frac{p}{100}\right)n = \left(\frac{85}{100}\right)12 = 10.2$$

Paso 3. Como i no es un entero, se *redondea.* La posición del percentil 85 es el siguiente entero mayor que 10.2, es decir, la posición 11.

Observe de nuevo los datos: el percentil 85 es el valor de datos en la posición 11, o 3 730.

Como otro ejemplo de este procedimiento, considere el cálculo del percentil 50 para los datos de los sueldos iniciales. Al aplicar el paso 2 se obtiene

$$i = \left(\frac{50}{100}\right)12 = 6$$

Dado que i es un entero, el paso 3*b*) establece que el percentil 50 es el promedio de los valores sexto y séptimo; por tanto, el percentil 50 es (3 490 + 3 520)/2 = 3 505. Observe que el *percentil 50 coincide con la mediana.*

Cuartiles

Los cuartiles son sencillamente percentiles específicos; por tanto, los pasos para calcular los percentiles se aplican directamente en el cálculo de cuartiles.

A menudo es recomendable dividir los datos en cuatro partes, cada una de las cuales contiene aproximadamente un cuarto, o 25% de las observaciones. La figura 3.1 muestra una distribución de datos dividida en cuatro partes. Los puntos de división se conocen como **cuartiles** y son definidos como:

Q_1 = primer cuartil, o percentil 25
Q_2 = segundo cuartil, o percentil 50 (también la mediana)
Q_3 = tercer cuartil, o percentil 75

Los datos sobre los sueldos iniciales mensuales se acomodan de nuevo en orden ascendente. Ya se identificó Q_2, el segundo cuartil (mediana), como 3 505.

3 310 3 355 3 450 3 480 3 480 3 490 3 520 3 540 3 550 3 650 3 730 3 925

El cálculo de los cuartiles Q_1 y Q_3 requiere el uso de la regla para obtener los percentiles 25 y 75. Estos cálculos son los siguientes.

Para obtener Q_1,

$$i = \left(\frac{p}{100}\right)n = \left(\frac{25}{100}\right)12 = 3$$

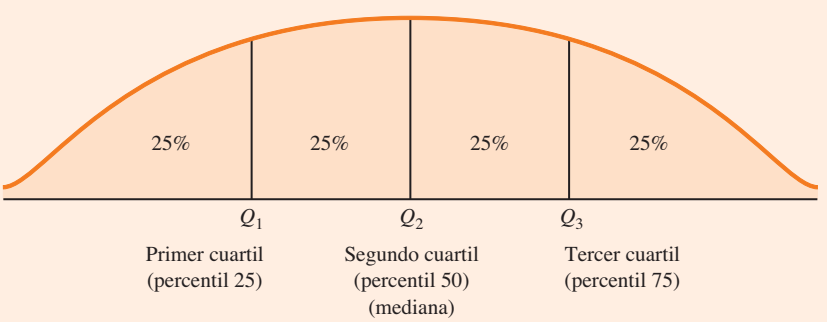
Como i es un entero, el paso 3 *b*) indica que el primer cuartil, o percentil 25, es el promedio del tercer y cuarto valores de los datos; por tanto, $Q_1 = (3\,450 + 3\,480)/2 = 3\,465$.

Para obtener Q_3,

$$i = \left(\frac{p}{100}\right)n = \left(\frac{75}{100}\right)12 = 9$$

Una vez más, dado que i es un entero, el paso 3*b*) indica que el tercer cuartil, o percentil 75, es el promedio del noveno y décimo valores de los datos; es decir, $Q_3 = (3\,550 + 3\,650)/2 = 3\,600$.

FIGURA 3.1 Posición de los cuartiles



Los cuartiles dividen los datos de los sueldos iniciales en cuatro partes, de las cuales cada una contiene 25% de las observaciones.

3 310 3 355 3 450 | 3 480 3 480 3 490 | 3 520 3 540 3 550 | 3 650 3 730 3 925

$Q_1 = 3\,465$ $Q_2 = 3\,505$ (mediana) $Q_3 = 3\,600$

Los cuartiles se definieron como los percentiles 25, 50 y 75; de ahí que se calculen de la misma manera que los percentiles. Sin embargo, a veces se usan otras convenciones para calcularlos, por lo que los valores reales reportados para los cuartiles pueden variar ligeramente, dependiendo de la convención utilizada. No obstante, el objetivo de todos los procedimientos es dividir los datos en cuatro partes iguales.

NOTAS Y COMENTARIOS

Cuando un conjunto de datos contiene valores extremos es preferible utilizar la mediana más que la media como medida de la ubicación central. Otra medida que se emplea a veces cuando hay valores extremos es la *media recortada.* Ésta se obtiene al eliminar un porcentaje de los valores menores y mayores de un conjunto de datos y luego calcular la media de los valores restantes. Por ejemplo, la media recortada al 5% se obtiene al eliminar 5% de los valores menores y 5% de los valores mayores de los datos y luego calcular la media de los valores restantes. Si se usa la muestra con $n = 12$ sueldos iniciales, $0.05(12) = 0.6$. El redondeo de este valor a 1 indica que la media recortada al 5% elimina el valor 1 menor y el valor 1 mayor. La media recortada al 5% utilizando las 10 observaciones restantes es 3 524.50.

Ejercicios

Métodos

1. Considere una muestra con los datos 10, 20, 12, 17 y 16. Calcule la media y la mediana.
2. Asuma una muestra con los datos 10, 20, 21, 17, 16 y 12. Calcule la media y la mediana.
3. Considere una muestra con los datos 27, 25, 20, 15, 30, 34, 28 y 25. Calcule los percentiles 20, 25, 65 y 75.

AUTO evaluación

4. Considere una muestra con los datos 53, 55, 70, 58, 64, 57, 53, 69, 57, 68 y 53. Calcule la media, la mediana y la moda.

Aplicaciones

5. El índice Dow Jones de viajes informó cuánto pagan los viajeros de negocios por una noche en una habitación de hotel en las principales ciudades estadounidenses (*The Wall Street Journal,* 16 de enero de 2004). Las tarifas promedio de una habitación por noche para 20 ciudades son las siguientes:

WEB archivo

Hotels

Atlanta	$163	Minneapolis	$125
Boston	177	New Orleans	167
Chicago	166	New York	245
Cleveland	126	Orlando	146
Dallas	123	Phoenix	139
Denver	120	Pittsburgh	134
Detroit	144	San Francisco	167
Houston	173	Seattle	162
Los Ángeles	160	St. Louis	145
Miami	192	Washington, D.C.	207

a) ¿Cuál es la tarifa media de una habitación por noche?
b) ¿Cuál es la mediana de las tarifas de una habitación por noche?
c) ¿Cuál es la moda?
d) ¿Cuál es el primer cuartil?
e) ¿Cuál es el tercer cuartil?

6. Durante la temporada de basquetbol colegial de la NCAA 2007-2008 en Estados Unidos, los equipos de basquetbol varonil intentaron un número récord de tiros de 3 puntos, que promedió 19.07 tiros por partido (Associated Press Sports, 24 de enero de 2009). Al tratar de desalentar tantos tiros de 3 puntos y estimular a los estudiantes a hacer más jugadas, el comité de reglas de la NCAA movió la línea de tiro de 3 puntos de 19 pies, 9 pulgadas a 20 pies, 9 pulgadas al inicio de la temporada 2008-2009. En la tabla siguiente se aprecian los tiros de 3 puntos realizados y los encestes para una muestra de 19 partidos de basquetbol durante la temporada de referencia.

WEB archivo
3Points

3-Point Shots	Shots Made	3-Point Shots	Shots Made
23	4	17	7
20	6	19	10
17	5	22	7
18	8	25	11
13	4	15	6
16	4	10	5
8	5	11	3
19	8	25	8
28	5	23	7
21	7		

a) ¿Cuál es la media del número de tiros de 3 puntos realizados por partido?
b) ¿Cuál es la media del número de tiros de 3 puntos encestados por partido?
c) Al usar la línea de 3 puntos más cercana, los jugadores encestaban 35.2% de sus tiros. ¿Qué porcentaje de tiros encestan desde la nueva línea de 3 puntos?
d) ¿Cuál fue el impacto del cambio de reglas de la NCAA que retrocedió la línea de tiro a 20 pies, 9 pulgadas para la temporada 2008-2009? ¿Estaría usted de acuerdo con el artículo de Associated Press Sports que establece que "El retroceso de la línea de tiro de 3 puntos no ha cambiado drásticamente el juego"? Explique por qué.

7. El ingreso por donativos es una parte vital de los presupuestos anuales en los colegios y universidades. Un estudio realizado por los directivos administrativos de la Asociación Nacional de Colegios y Universidades informó que 435 instituciones encuestadas recibieron un total de $413 mil millones en donaciones. Las 10 universidades más ricas se listan a continuación (*The Wall Street Journal,* 27 de enero de 2009). Los montos se proporcionan en miles de millones de dólares.

Universidad	Donativo (miles de millones de dólares)	Universidad	Donativo (miles de millones de dólares)
Columbia	7.2	Princeton	16.4
Harvard	36.6	Stanford	17.2
MIT	10.1	Texas	16.1
Michigan	7.6	Texas A&M	6.7
Northwestern	7.2	Yale	22.9

a) ¿Cuál es la media de los donativos para estas universidades?
b) ¿Cuál es la mediana de los donativos?
c) ¿Cuál es la moda de estos apoyos?
d) Calcule el primer y el tercer cuartiles.

e) ¿Cuál es el donativo total para estas 10 universidades? Éstas representan 2.3% de los 435 colegios y universidades encuestados, ¿qué porcentaje del total de $413 mil millones en donativos recibieron?

f) *The Wall Street Journal* reportó que durante un periodo reciente de cinco meses, un declive económico ocasionó que los donativos disminuyeran 23%. ¿Cuál es la estimación en dólares de la reducción en los donativos totales que recibieron estas 10 universidades? Dada esta situación, ¿cuáles son algunos pasos que usted esperaría que los administradores universitarios tomaran en consideración?

AUTO evaluación

8. El costo de las compras que realizaron los consumidores, como vivienda unifamiliar, gasolina, servicios de Internet, declaración de impuestos y hospitalización fue difundido en un artículo de *The Wall Street Journal* (2 de enero de 2007). Los datos muestrales típicos sobre el costo de la declaración de impuestos por servicios tales como H&R Block se muestran en seguida.

WEB archivo
TaxCost

120	230	110	115	160
130	150	105	195	155
105	360	120	120	140
100	115	180	235	255

a) Calcule la media, la mediana y la moda.

b) Determine el primer y el tercer cuartiles.

c) Calcule e interprete el percentil 90.

9. Datos de la Asociación Nacional de Agentes Inmobiliarios de Estados Unidos muestran que las ventas de vivienda fueron las más bajas en 10 años (Associated Press, 24 de diciembre de 2008). A continuación se presentan los datos muestrales con el precio de venta representativo para las casas usadas y las nuevas. Los datos se expresan en miles de dólares.

Casas usadas	315.5	202.5	140.2	181.3	470.2	169.9	112.8	230.0	177.5
Casas nuevas	275.9	350.2	195.8	525.0	225.3	215.5	175.0	149.5	

a) ¿Cuál es la mediana de los precios de venta de las casas usadas?

b) ¿Cuál es la mediana de los precios de venta de las viviendas nuevas?

c) ¿Cuáles casas tienen la mediana de los precios de venta más alta: las usadas o las nuevas? ¿Cuál es la diferencia entre la mediana de los precios de venta?

d) Hace un año la mediana de los precios de venta de las casas usadas era de $208.4 mil y la de los precios de venta de las casas nuevas era de $249 mil. Calcule el cambio porcentual en la mediana de los precios de venta de unos y otros inmuebles durante un periodo de un año. ¿Cuáles viviendas tienen el cambio porcentual mayor en la mediana de los precios de venta: las usadas o las nuevas?

10. Un panel de economistas proporcionó pronósticos de la economía estadounidense para los primeros seis meses de 2007 (*The Wall Street Journal,* 2 de enero de 2007). Los cambios porcentuales en el producto interno bruto (PIB) pronosticados por 30 economistas son los siguientes.

WEB archivo
Economy

2.6	3.1	2.3	2.7	3.4	0.9	2.6	2.8	2.0	2.4
2.7	2.7	2.7	2.9	3.1	2.8	1.7	2.3	2.8	3.5
0.4	2.5	2.2	1.9	1.8	1.1	2.0	2.1	2.5	0.5

a) ¿Cuál es el pronóstico mínimo para el cambio porcentual en el PIB? ¿Cuál es el pronóstico máximo?

b) Calcule la media, la mediana y la moda.

c) Calcule el primer y el tercer cuartiles.

d) ¿Los economistas proporcionaron una perspectiva optimista o pesimita de la economía estadounidense? Comente.

11. En un experimento automotriz sobre millaje y consumo de gasolina se aplicó una prueba de circulación a 13 automóviles a lo largo de 300 millas tanto en ciudad como en autopista. Los datos siguientes se obtuvieron para el rendimiento en millas por galón.

Ciudad	16.2	16.7	15.9	14.4	13.2	15.3	16.8	16.0	16.1	15.3	15.2	15.3	16.2
Autopista	19.4	20.6	18.3	18.6	19.2	17.4	17.2	18.6	19.0	21.1	19.4	18.5	18.7

Use la media, la mediana y la moda para señalar cuál es la diferencia en el rendimiento para la circulación en ciudad y en autopista.

12. Walt Disney Company compró Pixar Animation Studios, Inc. por 7 400 millones de dólares (sitio web de CNN Money, 24 de enero de 2006). Las películas animadas producidas por Disney y Pixar durante los 10 años previos a la compra se listan en la tabla siguiente. Los ingresos de taquilla (Revenue) se proporcionan en millones de dólares. Calcule el ingreso total, la media, la mediana y los cuartiles para comparar el éxito de taquilla de las películas producidas por ambas empresas. ¿Los estadísticos sugieren por lo menos una de las razones por las que Disney se interesó en comprar Pixar? Comente.

WEB archivo

Disney

Disney Movies	Revenue ($millions)	Pixar Movies	Revenue ($millions)
Pocahontas	346	*Toy Story*	362
Hunchback of Notre Dame	325	*A Bug's Life*	363
Hercules	253	*Toy Story 2*	485
Mulan	304	*Monsters, Inc.*	525
Tarzan	448	*Finding Nemo*	865
Dinosaur	354	*The Incredibles*	631
The Emperor's New Groove	169		
Lilo & Stitch	273		
Treasure Planet	110		
The Jungle Book 2	136		
Brother Bear	250		
Home on the Range	104		
Chicken Little	249		

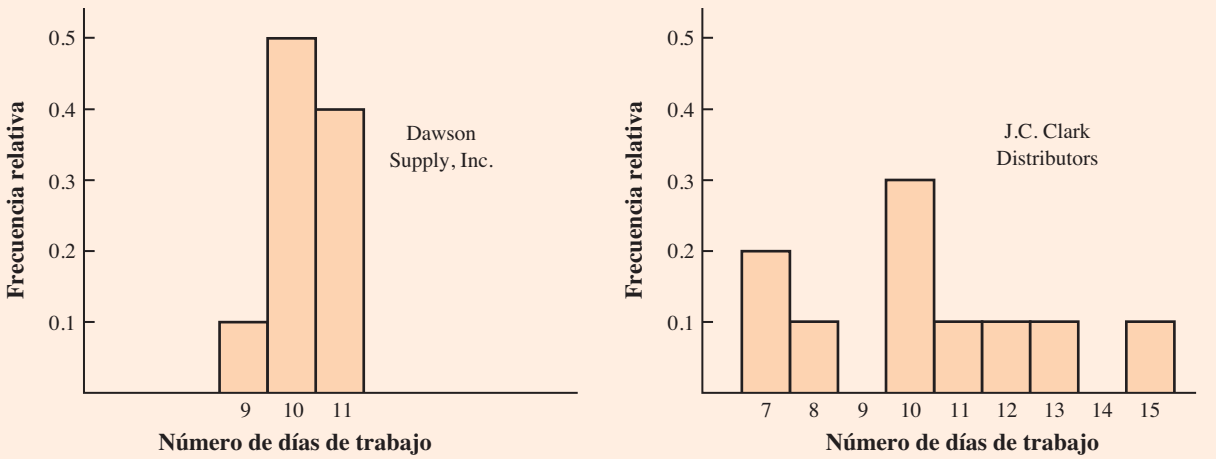
3.2 Medidas de variabilidad

La variabilidad en los plazos de entrega genera incertidumbre en la planeación de la producción. Los métodos presentados en esta sección ayudan a medir y entender la variabilidad.

Además de las medidas de posición, con frecuencia es conveniente considerar las medidas de variabilidad o dispersión. Por ejemplo, suponga que usted es un agente de compras de una empresa manufacturera grande y que coloca con regularidad pedidos con dos proveedores diferentes. Después de varios meses de operación, se da cuenta de que el número medio de días necesario para que ambos surtan los pedidos es de 10 días. Los histogramas que resumen el número de días de trabajo requeridos para que los proveedores suministren los pedidos se muestran en la figura 3.2. Aunque el número medio de días es 10 para los dos proveedores, ¿ambos muestran el mismo grado de confiabilidad en cuanto a efectuar las entregas a tiempo? Note la dispersión, o variabilidad, en los plazos de entrega indicados por los histogramas. ¿Qué proveedor prefiere usted?

Para la mayoría de las empresas es importante recibir a tiempo los materiales y suministros para sus procesos. Los plazos de entrega de 7 u 8 días mostrados para J.C. Clark Distributors podrían considerarse favorables, sin embargo, algunos plazos largos de 13 a 15 días podrían resultar desastrosos en términos de mantener ocupada a la fuerza de trabajo y la producción

FIGURE 3.2 Datos históricos que muestran el número de días requerido para surtir los pedidos



dentro de lo programado. Este ejemplo ilustra una situación en la que la variabilidad en los tiempos de entrega puede ser una consideración primordial al seleccionar a un proveedor. Para la mayoría de los agentes de compra, la menor variabilidad mostrada por Dawson Supply, Inc. lo haría el preferido.

Ahora se verá a la revisión de algunas medidas de variabilidad de uso común.

Rango

La medida de variabilidad más sencilla es el **rango**.

RANGO

$$\text{Rango} = \text{valor mayor} - \text{valor menor}$$

Revise los datos sobre los sueldos iniciales para los licenciados en administración de empresas recién egresados que hemos venido trabajando de la tabla 3.1. El sueldo inicial mayor es de 3 925 y el menor es de 3 310. El rango es $3\,925 - 3\,310 = 615$.

Aun cuando el rango es la medida de variabilidad más fácil de calcular, pocas veces se usa como la única medida debido a que se basa sólo en dos de las observaciones y, por tanto, los valores extremos influyen mucho en él. Suponga que uno de los licenciados recién egresados recibe un sueldo inicial de $10 000 al mes. En este caso, el rango sería $10\,000 - 3\,310 = 6\,690$ en vez de 615. Este valor mayor para el rango no describe con claridad la variabilidad de los datos debido a que 11 de los 12 sueldos iniciales se agrupan estrechamente entre 3 310 y 3 730.

Rango intercuartílico

Una medida de la variabilidad que supera la dependencia sobre los valores extremos es el **rango intercuartílico (RIC)**. Esta medida de la variabilidad es la diferencia entre el tercer cuartil, Q_3, y el primer cuartil, Q_1. En otras palabras, el rango intercuartílico es el rango de la media de 50% de los datos.

RANGO INTERCUARTÍLICO

$$\text{RIC} = Q_3 - Q_1 \tag{3.3}$$

Para los datos sobre los sueldos mensuales iniciales, los cuartiles son $Q_3 = 3\,600$ y $Q_1 = 3\,465$. Por tanto, el rango intercuartílico es $3\,600 - 3\,465 = 135$.

Varianza

La **varianza** es una medida de la variabilidad que utiliza todos los datos. Se basa en la diferencia entre el valor de cada observación (x_i) y la media. La diferencia entre cada x_i y la media ($\bar{x}$ para una muestra; μ para una población) se llama *desviación respecto de la media.* Para una muestra, una desviación respecto de la media se escribe $(x_i - \bar{x})$; para una población, se escribe $(x_i - \mu)$. Si se desea calcular la varianza, las desviaciones respecto de la media *se elevan al cuadrado.*

Si los datos pertenecen a una población, el promedio de las desviaciones elevadas al cuadrado se llama *varianza poblacional*, la cual se denota por medio del símbolo griego σ^2. Para una población de N observaciones con una media poblacional μ, la definición de la varianza poblacional es la siguiente.

VARIANZA POBLACIONAL

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N} \tag{3.4}$$

En la mayoría de las aplicaciones estadísticas, los datos que se analizan provienen de una muestra. Cuando se calcula una varianza muestral, a menudo lo que interesa es usarla para estimar la varianza poblacional σ^2. Aunque una explicación detallada está más allá del alcance de este libro, puede mostrarse que si la suma de las desviaciones respecto de la media al cuadrado se divide entre $n - 1$, y no entre n, la varianza muestral resultante proporciona un estimador insesgado de la varianza poblacional. Por esta razón, la *varianza muestral,* denotada por s^2, se define como sigue.

La varianza muestral s^2 es el estimador de la varianza poblacional σ^2.

VARIANZA MUESTRAL

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} \tag{3.5}$$

Para ilustrar el cálculo de la varianza muestral se usarán los datos sobre los tamaños de grupo de la muestra de cinco grupos de estudiantes universitarios presentada en la sección 3.1. Un resumen de los datos, que incluye el cálculo de las desviaciones respecto de la media y los cuadrados de las desviaciones respecto de la media, se aprecia en la tabla 3.2. La suma de los cuadrados de estas desviaciones es $\Sigma(x_i - \bar{x})^2 = 256$. Por ende, si $n - 1 = 4$, la varianza muestral es

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{256}{4} = 64$$

Antes de proseguir, observe que las unidades asociadas con la varianza muestral suelen causar confusión. Debido a que los valores que se suman para calcular la varianza, $(x_i - \bar{x})^2$, están elevados al cuadrado, las unidades asociadas con la varianza muestral también están *elevadas*

TABLA 3.2 Cálculo de desviaciones respecto de la media y desviaciones cuadradas respecto de la media de los datos de tamaños de grupo

Número de estudiantes en el grupo (x_i)	Tamaño de grupo medio ($\bar{x}$)	Desviación respecto de la media ($x_i - \bar{x}$)	Desviación cuadrada respecto de la media $(x_i - \bar{x})^2$
46	44	2	4
54	44	10	100
42	44	−2	4
46	44	2	4
32	44	−12	144
		0	256
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

La varianza es útil para comparar la variabilidad de dos o más variables.

al cuadrado. Por ejemplo, la varianza muestral para los datos del tamaño de grupo es $s^2 = 64$ (estudiantes)^2. Las unidades cuadradas asociadas con la varianza dificultan obtener una comprensión e interpretación intuitiva del valor numérico de ésta. Se recomienda considerarla como una medida útil en la comparación de la cantidad de variabilidad para dos o más variables. En una comparación de las variables, aquella con la varianza más grande muestra la mayor variabilidad. Una interpretación del valor de la varianza tal vez no sea necesaria.

Como otra ilustración del cálculo de una varianza muestral, considere los sueldos iniciales listados en la tabla 3.1 para los 12 licenciados en administración de empresas. En la sección 3.1 se observa que la media muestral de los sueldos es de 3 540. El cálculo de la varianza muestral ($s^2 = 27\,440.91$) se muestra en la tabla 3.3.

TABLA 3.3 Cálculo de la varianza muestral para los datos de los sueldos iniciales

Sueldo mensual (x_i)	Media muestral ($\bar{x}$)	Desviación respecto de la media ($x_i - \bar{x}$)	Desviación cuadrada respecto de la media $(x_i - \bar{x})^2$
3 450	3 540	−90	8 100
3 550	3 540	10	100
3 650	3 540	110	12 100
3 480	3 540	−60	3 600
3 355	3 540	−185	34 225
3 310	3 540	−230	52 900
3 490	3 540	−50	2 500
3 730	3 540	190	36 100
3 540	3 540	0	0
3 925	3 540	385	148 225
3 520	3 540	−20	400
3 480	3 540	−60	3 600
		0	301 850
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

Usando la ecuación (3.5),

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{301\,850}{11} = 27\,440.91$$

En las tablas 3.2 y 3.3 se aprecian la suma de las desviaciones sobre la media y la suma de las desviaciones cuadradas sobre la media. Para cualquier conjunto de datos, la suma de las desviaciones sobre la media *siempre será igual a cero.* Note que en esas tablas, $\Sigma(x_i - \bar{x}) = 0$. Las desviaciones positivas y negativas se cancelan entre sí, ocasionando que la suma de las desviaciones sobre la media sea igual a cero.

Desviación estándar

La **desviación estándar** se define como la raíz cuadrada positiva de la varianza. Siguiendo la notación que se adoptó para las varianzas muestral y poblacional, se usa s para denotar la desviación estándar muestral y σ para denotar la desviación estándar poblacional. La desviación estándar se deriva de la varianza de la manera siguiente.

La desviación estándar muestral s *es el estimador de la desviación estándar poblacional* σ.

DESVIACIÓN ESTÁNDAR

$$\text{Desviación estándar muestral} = s = \sqrt{s^2} \tag{3.6}$$

$$\text{Desviación estándar poblacional} = \sigma = \sqrt{\sigma^2} \tag{3.7}$$

Recuerde que la varianza muestral para los tamaños de grupo de la muestra de cinco grupos de estudiantes es $s^2 = 64$. Por tanto, la desviación estándar muestral es $s = \sqrt{64} = 8$. Para los datos sobre los sueldos iniciales, la desviación estándar muestral es $s = \sqrt{27\,440.91} = 165.65$.

La desviación estándar es más fácil de interpretar que la varianza debido a que se mide en las mismas unidades que los datos.

¿Qué se gana al convertir la varianza en la desviación estándar correspondiente? Recuerde que las unidades asociadas con la varianza están elevadas al cuadrado. Por ejemplo, la varianza muestral para los datos sobre los sueldos iniciales de los licenciados en administración de empresas recién egresados es $s^2 = 27\,440.91$ (dólares)^2. Debido a que la desviación estándar es la raíz cuadrada de la varianza, las unidades de esta última, los dólares al cuadrado, se convierten en dólares en la desviación estándar. Por consiguiente, la desviación estándar de los datos de los sueldos iniciales es $165.65. En otras palabras, ésta se mide en las mismas unidades que los datos originales; por esta razón la desviación estándar se compara más fácilmente con la media y con otros estadísticos que se miden en las mismas unidades que los datos originales.

Coeficiente de variación

En algunas situaciones nos interesa la estadística descriptiva que indique qué tan grande es la desviación estándar con respecto a la media. Esta medida se llama **coeficiente de variación**, y se expresa por lo general como un porcentaje.

El coeficiente de variación es una medida relativa de la variabilidad; mide la desviación estándar con respecto a la media.

COEFICIENTE DE VARIACIÓN

$$\left(\frac{\text{desviación estándar}}{\text{media}} \times 100\right)\% \tag{3.8}$$

Para los datos de los tamaños de grupo, se encontró una media muestral de 44 y una desviación estándar muestral de 8. El coeficiente de variación es $[(8/44) \times 100]\% = 18.2\%$. Expresado con palabras, el coeficiente de variación indica que la desviación estándar muestral es 18.2% del valor de la media muestral. Para los datos de los sueldos iniciales con una media muestral de 3 540 y una desviación estándar muestral de 165.65, el coeficiente de variación, $[(165.65/3\,540) \times 100]\% = 4.7\%$, señala que la desviación estándar muestral es sólo 4.7% del valor de la media muestral. En general, el coeficiente de variación es un estadístico útil para comparar la variabilidad de las variables que tienen tanto desviaciones estándar como medias distintas.

NOTAS Y COMENTARIOS

1. El *software* y las hojas de cálculo para estadística se usan para obtener los estadísticos descriptivos presentados en este capítulo. Una vez que los datos se introducen en una hoja de cálculo, bastan unos comandos sencillos para generar el resultado deseado. En los tres apéndices del capítulo se explica cómo usar Minitab, Excel y StatTools para obtener estadísticos descriptivos.
2. La desviación estándar es una medida de uso común para el riesgo asociado con la inversión en acciones y fondos de acciones (*BusinessWeek,* 17 de enero de 2000). Proporciona una medida de cómo fluctúan los rendimientos mensuales en torno al rendimiento medio a largo plazo.
3. Cuando los valores de la media muestral $\bar{x}$ y los valores de los cuadrados de las desviaciones $(x_i - \bar{x})^2$ se redondean, se pueden introducir errores en la calculadora al obtener la varianza y la desviación estándar. Para reducir los errores de redondeo, se recomienda trabajar por lo menos con seis dígitos significativos durante los cálculos intermedios. La varianza o la desviación estándar resultantes pueden redondearse después a menos dígitos.
4. Una fórmula opcional para el cálculo de la varianza muestral es

$$s^2 = \frac{\Sigma x_i^2 - n\bar{x}^2}{n - 1}$$

donde $\Sigma x_i^2 = x_1^2 + x_2^2 + \cdots x_n^2$.

Ejercicios

Métodos

13. Considere una muestra con los datos 10, 20, 12, 17 y 16. Calcule el rango y el rango intercuartílico.
14. Asuma una muestra con los datos 10, 20, 12, 17 y 16. Determine la varianza y la desviación estándar.
15. Considere una muestra con los datos 27, 25, 20, 15, 30, 34, 28 y 25. Calcule el rango, el rango intercuartílico, la varianza y la desviación estándar.

AUTO evaluación

Aplicaciones

AUTO evaluación

16. Las puntuaciones que obtuvo un jugador de boliche en seis partidos fueron 182, 168, 184, 190, 170 y 174. Usando estos datos como una muestra, calcule los estadísticos descriptivos siguientes:
    a) Rango
    b) Varianza
    c) Desviación estándar
    d) Coeficiente de variación
17. Un sistema de teatro en casa (*home theater*) es la manera más fácil y económica de proporcionar sonido ambiental para un centro de entretenimiento en el hogar. Enseguida se presenta una muestra de precios (*Consumer Reports Buying Guide*, 2004) para modelos con y sin reproductor de DVD.

Modelos con reproductor de DVD	Precio	Modelos sin reproductor de DVD	Precio
Sony HT-1800DP	$450	Pioneer HTP-230	$300
Pioneer HTD-330DV	300	Sony HT-DDW750	300
Sony HT-C800DP	400	Kenwood HTB-306	360
Panasonic SC-HT900	500	RCA RT-2600	290
Panasonic SC-MTI	400	Kenwood HTB-206	300

    a) Calcule el precio medio de los modelos con reproductor de DVD y el precio medio de los modelos sin reproductor de DVD. ¿Cuál es el precio adicional que se paga por tener un reproductor de DVD en el sistema de teatro en casa?
    b) Calcule el rango, la varianza y la desviación estándar de las dos muestras. ¿Qué le dice esta información sobre los precios de los modelos con y sin reproductor de DVD?

18. Las tarifas de renta de automóviles por día para una muestra de siete ciudades del este de Estados Unidos son las siguientes (*The Wall Street Journal,* 16 de enero de 2004).

Ciudad	Tarifa diaria
Boston	$43
Atlanta	35
Miami	34
Nueva York	58
Orlando	30
Pittsburgh	30
Washington, D.C.	36

a) Calcule la media, la varianza y la desviación estándar de estas tarifas.
b) En una muestra similar de siete ciudades del oeste de Estados Unidos se obtuvo una media muestral de las tarifas de renta de automóviles de $38 por día. La varianza y la desviación estándar fueron 12.3 y 3.5, respectivamente. Comente la diferencia entre las tarifas de renta de las ciudades del este y del oeste de Estados Unidos.

19. *Los Ángeles Times* informa el índice de calidad del aire de varias zonas del sur de California. Una muestra de valores de este índice en Pomona proporcionó los datos siguientes: 28, 42, 58, 48, 45, 55, 60, 49 y 50.
a) Calcule el rango y el rango intercuartílico.
b) Calcule la varianza muestral y la desviación estándar muestral.
c) Una muestra de lecturas del índice de calidad del aire de Anaheim proporcionó una media muestral de 48.5, una varianza muestral de 136 y una desviación estándar muestral de 11.66. ¿Qué comparaciones puede hacer entre la calidad del aire en Pomona y en Anaheim sobre la base de estos estadísticos descriptivos?

20. Los datos siguientes se utilizaron para elaborar los histogramas del número de días requerido para que Dawson Supply, Inc. y J.C. Clark Distributors surtan pedidos (figura 3.2).

Días de entrega de Dawson Supply	11	10	9	10	11	11	10	11	10	10
Días de entrega de Clark Distributors	8	10	13	7	10	11	10	7	15	12

Use el rango y la desviación estándar para apoyar la observación anterior de que Dawson Supply proporciona los tiempos de entrega más consistentes y confiables.

21. ¿Cómo se comparan los costos de abarrotes en Estados Unidos? Usando una canasta básica que contiene 10 artículos que incluyen carne, leche, pan, huevos, café, papas, cereal y jugo de naranja, la revista *Where to Retire* calculó el costo de la canasta básica en seis ciudades y seis comunidades de jubilados en todo Estados Unidos (*Where to Retire,* noviembre/diciembre de 2003). Los datos con el costo de la canasta básica al dólar más cercano son los siguientes.

Ciudad	Costo	Comunidad de jubilados	Costo
Buffalo, NY	$33	Biloxi-Gulfport, MS	$29
Des Moines, IA	27	Asheville, NC	32
Hartford, CT	32	Flagstaff, AZ	32
Los Ángeles, CA	38	Hilton Head, SC	34
Miami, FL	36	Fort Myers, FL	34
Pittsburgh, PA	32	Santa Fe, NM	31

a) Calcule la media, la varianza y la desviación estándar para la muestra de ciudades y la muestra de las comunidades de jubilados.
b) ¿Qué observaciones puede hacer con base en las dos muestras?

WEB archivo
BackToSchool

22. La Federación Nacional de Minoristas informó que los estudiantes universitarios de primer año gastan más en artículos de regreso a clases que cualquier otro grupo universitario (*USA Today,* 4 de agosto de 2006). El archivo BackToSchool contiene una base de datos muestrales que compara los gastos de regreso a clases de 25 estudiantes de primer año y 20 del último año.
   a) ¿Cuál es el gasto medio de regreso a clases de cada grupo? ¿Los datos son consistentes con el informe de la Federación Nacional de Minoristas?
   b) ¿Cuál es el rango de los gastos de cada grupo?
   c) ¿Cuál es el rango intercuartílico para cada grupo?
   d) ¿Cuál es la desviación estándar de los gastos de cada grupo?
   e) ¿Qué gastos de regreso a clases muestran más variación: los de los estudiantes de primer año o los de los universitarios de último año?

23. Las puntuaciones anotadas por un golfista amateur en el campo de golf de Bonita Fairways, en Bonita Springs, Florida, durante 2005 y 2006 son los siguientes.

Temporada 2005	74	78	79	77	75	73	75	77
Temporada 2006	71	70	75	77	85	80	71	79

   a) Use la media y la desviación estándar para evaluar el desempeño del golfista durante el periodo de dos años.
   b) ¿Cuál es la principal diferencia en su desempeño entre 2005 y 2006? ¿Qué mejora, si la hay, puede verse en las puntuaciones de 2006?

24. Los corredores de un equipo de atletismo universitario registraron los siguientes tiempos para los carreras de cuarto de milla y de milla (los tiempos están en minutos).

Tiempos de cuarto de milla	0.92	0.98	1.04	0.90	0.99
Tiempos de milla	4.52	4.35	4.60	4.70	4.50

   Después de ver esta muestra de tiempos, uno de los entrenadores comentó que los corredores de cuarto de milla registraron tiempos más consistentes. Utilice la desviación estándar y el coeficiente de variación para resumir la variabilidad de los datos. ¿El uso del coeficiente de variación indica que el comentario del entrenador es correcto?

3.3 Medidas de la forma de la distribución, posición relativa y detección de observaciones atípicas

Se han descrito varias medidas de ubicación y variabilidad para los datos. Además de éstas, es importante tener una medida de la forma de la distribución. En el capítulo 2 se vio que un histograma proporciona una representación gráfica de la forma de una distribución. Una medida numérica importante de la forma de una distribución es el **sesgo**.

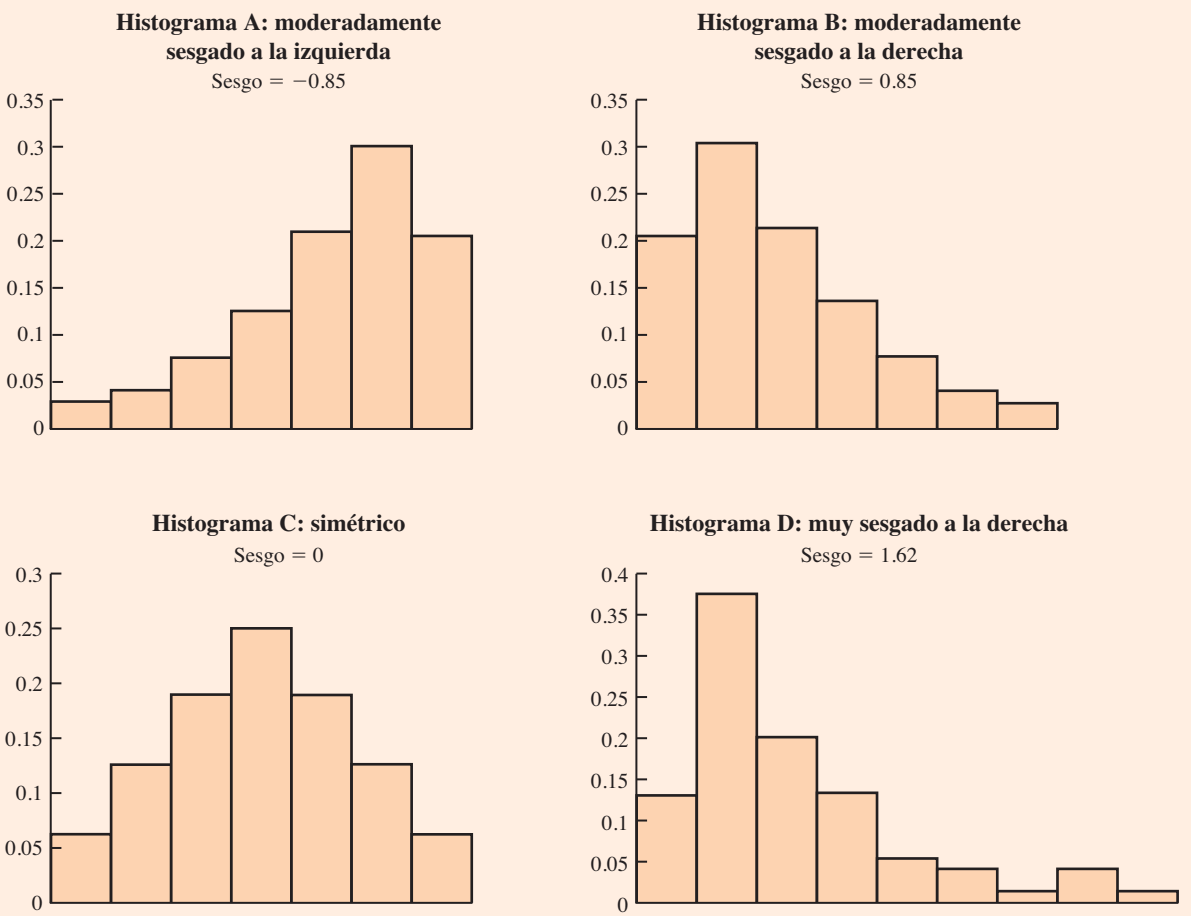
Forma de la distribución

En la figura 3.3 aparecen cuatro histogramas elaborados a partir de distribuciones de frecuencia relativa. Los histogramas A y B están moderadamente sesgados. El A está sesgado a la izquierda; su sesgo es −0.85. El B está sesgado a la derecha; su sesgo es +0.85. El histograma C es simétrico; su sesgo es cero. El D está muy sesgado a la derecha; su sesgo es 1.62. La fórmula empleada para calcular el sesgo es un tanto compleja.[1] Sin embargo, éste se obtiene fácilmente

[1] La fórmula para el sesgo de datos muestrales es:

$$\text{Sesgo} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

FIGURA 3.3 Histogramas que muestran el sesgo de cuatro distribuciones



utilizando software para estadística. Para datos sesgados a la izquierda, el sesgo es negativo; para datos sesgados a la derecha, el sesgo es positivo. Si los datos son simétricos, el sesgo es cero.

En una distribución simétrica, la media y la mediana son iguales. Cuando los datos están sesgados positivamente, la media por lo general será mayor que la mediana; cuando están sesgados negativamente, la media será menor que la mediana. Los datos usados para construir el histograma D corresponden a las compras de los clientes de una tienda de ropa femenina. El monto medio de compra es $77.60 y la mediana del monto de compra es $59.70. Los pocos montos de compra grandes tienden a incrementar la media, mientras que a la mediana no le afectan. Cuando los datos están muy sesgados, se prefiere la mediana como medida de ubicación.

Valor *z*

Además de las medidas de posición, variabilidad y forma, también interesa la posición relativa de los valores dentro de un conjunto de datos. Las medidas de posición relativa ayudan a determinar a qué distancia de la media está un valor determinado.

A partir de la media y la desviación estándar se puede determinar la posición relativa de cualquier observación. Suponga que se tiene una muestra de n observaciones, con los valores

denotados por $x_1, x_2, \ldots, x_n$. Asimismo, suponga que la media muestral, $\bar{x}$, y la desviación estándar muestral, s, ya se calcularon. Asociado con cada valor, x_i, hay otro valor llamado **valor z**. La ecuación (3.9) muestra cómo se calcula la puntuación z para cada x_i.

VALOR z

$$z_i = \frac{x_i - \bar{x}}{s} \tag{3.9}$$

Donde

z_i = valor z para x_i
$\bar{x}$ = media muestral
s = desviación estándar muestral

El valor z se llama *valor estandarizado*. El valor z, z_i, puede interpretarse como el *número de desviaciones estándar que x_i se encuentra de la media $\bar{x}$*. Por ejemplo, $z_1 = 1.2$ indicaría que x_1 es 1.2 desviaciones estándar mayor que la media muestral. De modo parecido, $z_2 = -0.5$ indicaría que x_2 es 0.5, o 1/2 desviaciones estándar menor que la media muestral. Un valor z mayor que cero ocurre para observaciones con un valor mayor que la media, y un valor z menor que cero ocurre para observaciones con un valor menor que la media. Un valor z de cero indica que el valor de la observación es igual a la media.

El valor z para cualquier observación puede interpretarse como una medida de la posición relativa de la observación en un conjunto de datos. Por tanto, se dice que las observaciones de dos conjuntos de datos diferentes con el mismo valor z tienen la misma posición relativa en términos de que presentan igual número de desviaciones estándar de la media.

Los valores z para los datos de los tamaños de grupo se calculan en la tabla 3.4. Recuerde la media muestral previamente calculada, $\bar{x} = 44$, y la desviación estándar muestral, $s = 8$. El valor z de -1.50 de la quinta observación indica que ésta es la más alejada de la media: está 1.50 desviaciones estándar por debajo de la media.

Teorema de Chebyshev

El **teorema de Chebyshev** permite hacer afirmaciones acerca de la proporción de los valores de datos que deben estar dentro de un número específico de desviaciones estándar de la media.

TABLA 3.4 Valores z de los datos de tamaños de grupo

Número de estudiantes en la clase (x_i)	Desviación respecto de la media ($x_i - \bar{x}$)	Valor z $\left(\frac{x_i - \bar{x}}{s}\right)$
46	2	2/8 = 0.25
54	10	10/8 = 1.25
42	−2	−2/8 = −0.25
46	2	2/8 = 0.25
32	−12	−12/8 = −1.50

TEOREMA DE CHEBYSHEV

Por lo menos $(1 - 1/z^2)$ de los valores de datos debe estar dentro de z desviaciones estándar de la media, donde z es cualquier valor mayor que 1.

A continuación se mencionan algunas implicaciones de este teorema cuando $z = 2$, 3 y 4 desviaciones estándar.

- Por lo menos 0.75, o 75%, de los datos debe estar dentro de $z = 2$ desviaciones estándar de la media.
- Al menos 0.89, u 89%, de los datos debe estar dentro de $z = 3$ desviaciones estándar de la media.
- Por lo menos 0.94, o 94%, de los datos debe estar dentro de $z = 4$ desviaciones estándar de la media.

Como ejemplo del uso del teorema de Chebyshev, suponga que las calificaciones obtenidas en los exámenes parciales por 100 estudiantes universitarios en un curso de estadística para negocios tenían una media de 70 y una desviación estándar de 5. ¿Cuántos alumnos obtuvieron una calificación de entre 60 y 80 en los exámenes? ¿Cuántos obtuvieron calificaciones de entre 58 y 82?

Para calificaciones entre 60 y 80, observe que 60 está dos desviaciones estándar por abajo de la media, y 80 está dos desviaciones estándar por encima de la media. Usando el teorema de Chebyshev se ve que como mínimo 0.75, o por lo menos 75% de las observaciones debe tener valores dentro de dos desviaciones estándar de la media. Por tanto, 75% de los estudiantes como mínimo debió obtener una calificación de entre 60 y 80.

En el teorema de Chebyshev se requiere $z > 1$; pero no es necesario que z sea un número entero.

Si las calificaciones de los exámenes están entre 58 y 82, observe que $(58 - 70)/5 = -2.4$ indica que 58 está a 2.4 desviaciones estándar por debajo de la media y que $(82 - 70)/5 = +2.4$ indica que 82 está a 2.4 desviaciones estándar por encima de la media. Al aplicar el teorema de Chebyshev con $z = 2.4$, tenemos

$$\left(1 - \frac{1}{z^2}\right) = \left(1 - \frac{1}{(2.4)^2}\right) = 0.826$$

Al menos 82.6% de los estudiantes debe obtener calificaciones de entre 58 y 82 en los exámenes.

Regla empírica

La regla empírica se basa en la distribución de probabilidad normal, la cual se estudia en el capítulo 6. La distribución normal se utiliza ampliamente en todo el libro.

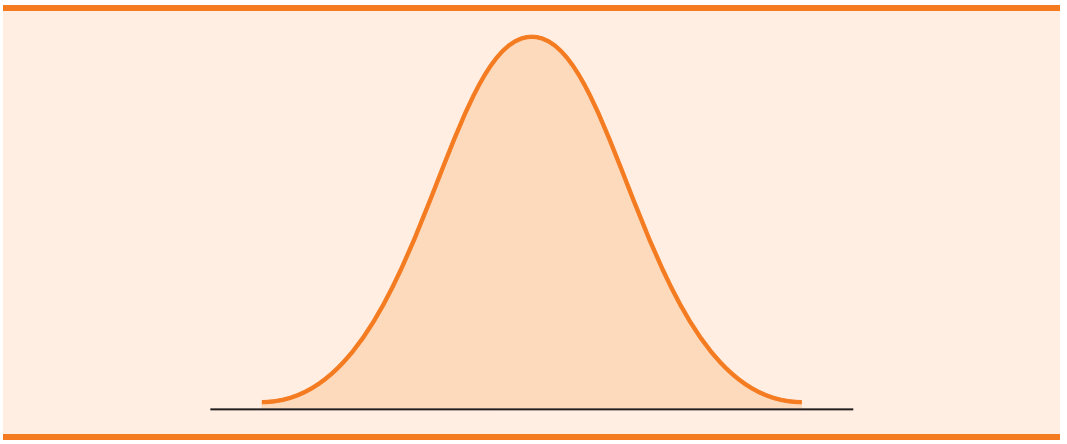
Una de las ventajas del teorema de Chebyshev estriba en que se aplica a cualquier conjunto de datos sin importar su forma de distribución. De hecho, podría usarse con cualquiera de las distribuciones de la figura 3.3. Sin embargo, en muchas aplicaciones prácticas los conjuntos de datos exhiben una distribución simétrica con forma de pila o de campana, como se aprecia en la figura 3.4. Cuando se piensa que los datos se aproximan a esta distribución, la **regla empírica** se usa para determinar el porcentaje de valores de datos que deben estar dentro de un número específico de desviaciones estándar de la media.

REGLA EMPÍRICA

Cuando los datos tienen una distribución en forma de campana:

- Aproximadamente 68% de los valores de datos estará dentro de una desviación estándar de la media.
- Aproximadamente 95% de los valores de datos estará dentro de dos desviaciones estándar de la media.
- Casi todos los valores de datos deben estar dentro de tres desviaciones estándar de la media.

FIGURA 3.4 Distribución simétrica con forma de pila o de campana



Por ejemplo, los envases de detergente líquido se llenan automáticamente en una línea de producción. Los pesos de llenado suelen tener una distribución en forma de campana. Si el peso medio de llenado es de 16 onzas y la desviación estándar de 0.25 onzas, se utiliza la regla empírica para formular las conclusiones siguientes.

- Aproximadamente 68% de los envases llenos pesará entre 15.75 y 16.25 onzas (dentro de una desviación estándar de la media).
- Aproximadamente 95% de los envases llenos pesará entre 15.50 y 16.50 onzas (dentro de dos desviaciones estándar de la media).
- Casi todos los envases llenos pesarán entre 15.25 y 16.75 onzas (dentro de tres desviaciones estándar de la media).

Detección de observaciones atípicas

Un conjunto de datos a veces tiene una o más observaciones con valores inusualmente grandes o sumamente pequeños. Estos valores extremos se llaman **observaciones atípicas**. Los expertos en estadística experimentados emprenden acciones para identificar observaciones atípicas y luego revisan cada una con detalle. Una observación atípica suele ser un valor de datos que se registró incorrectamente; si esto ocurre, el error se corrige antes de un análisis posterior. También puede ser una observación que se introdujo de forma incorrecta en el conjunto de datos; si este es el caso, se elimina. Por último, puede consistir en un valor de datos inusual que se registró correctamente y pertenece al conjunto de datos. En tal caso, debe conservarse.

Es una buena idea buscar observaciones atípicas antes de tomar decisiones basadas en el análisis de datos. Suelen cometerse errores en el registro y la introducción de los datos en la computadora. Las observaciones atípicas no necesariamente tienen que eliminarse, pero debe verificarse qué tan exactas y apropiadas son.

Los valores estandarizados (puntuaciones z), se utilizan para identificar observaciones atípicas. Recuerde que la regla empírica permite concluir que cuando los datos tienen una distribución en forma de campana, casi todos los valores de datos están dentro de tres desviaciones estándar de la media. Por tanto, al usar puntuaciones z para identificar observaciones extremas, se recomienda tomar en cuenta como una observación atípica cualquier valor de datos con una puntuación z menor que -3 o mayor que $+3$. La exactitud de estos valores debe verificarse y determinar si pertenecen al conjunto de datos.

Vuelva a observar las puntuaciones z de los datos sobre los tamaños de grupo de la tabla 3.4. La puntuación z de -1.50 muestra que el tamaño del quinto grupo está más alejado de la media. Sin embargo, este valor estandarizado está dentro de los límites de -3 y $+3$ para las observaciones atípicas. Por esta razón, la puntuación z no indica que las observaciones atípicas estén presentes en los datos de los tamaños de clase.

NOTAS Y COMENTARIOS

1. El teorema de Chebyshev es aplicable a cualquier conjunto de datos y se utiliza para establecer el número mínimo de valores de datos que estará dentro de cierto número de desviaciones estándar de la media. Si se sabe que los datos tienen una forma aproximada de campana, se puede decir más.

Por ejemplo, la regla empírica permite afirmar que *aproximadamente* 95% de los valores de datos estará dentro de dos desviaciones estándar de la media; el teorema de Chebyshev sólo permite concluir que por lo menos 75% de estos valores estará dentro de ese intervalo.

2. Antes de analizar un conjunto de datos, los expertos en estadística efectúan varias revisiones para confirmar su validez. En un estudio grande no es raro que se cometan errores en el registro de los valores de datos o al introducirlos en una computadora. La identificación de las observaciones es una herramienta empleada para verificar la validez de los datos.

Ejercicios

Métodos

25. Considere una muestra con los datos 10, 20, 12, 17 y 16. Calcule el valor z de cada una de estas cinco observaciones.

26. Suponga una muestra con una media de 500 y una desviación estándar de 100. ¿Cuáles son los valores z de los datos siguientes: 520, 650, 500, 450 y 280?

AUTO evaluación

27. Considere una muestra con una media de 30 y una desviación estándar de 5. Utilice el teorema de Chebyshev para determinar el porcentaje de datos que se encuentra dentro de cada uno de los rangos siguientes.
    a) 20 a 40
    b) 15 a 45
    c) 22 a 38
    d) 18 a 42
    e) 12 a 48

28. Suponga que los datos tienen una distribución con forma de campana, una media de 30 y una desviación estándar de 5. Use la regla empírica para determinar el porcentaje de los datos que está dentro de cada uno de los rangos siguientes.
    a) 20 a 40
    b) 15 a 45
    c) 25 a 35

Aplicaciones

AUTO evaluación

29. Los resultados de una encuesta nacional revelaron que, en promedio, los adultos duermen 6.9 horas por noche. Imagine que la desviación estándar es de 1.2 horas.
    a) Use el teorema de Chebyshev para calcular el porcentaje de personas que duermen entre 4.5 y 9.3 horas.
    b) Con el teorema de Chebyshev calcule ahora el porcentaje que duerme entre 3.9 y 9.9 horas.
    c) Suponga que el número de horas de sueño sigue una distribución con forma de campana. Utilice la regla empírica para calcular el porcentaje de personas que duerme entre 4.5 y 9.3 horas por día. ¿Cómo se compara este resultado con el valor obtenido con el teorema de Chebyshev en el inciso *a*)?

30. La Oficina de Información Energética reportó que el precio medio por galón de gasolina de grado regular es de $2.05 (Energy Information Administration, mayo de 2009). Suponga que la desviación estándar es $0.10 y que el precio al detalle (o al menudeo) por galón tiene una distribución con forma de campana.
    a) ¿Qué porcentaje de gasolina de grado regular se vendió entre $1.95 y $2.15 por galón?
    b) ¿Qué porcentaje se vendió entre $1.95 y $2.25 por galón?
    c) ¿Qué porcentaje de gasolina de grado regular se vendió por más de $2.25 por galón?

31. El promedio nacional para la sección de matemáticas del examen de aptitudes escolares (College Board's Scholastic Aptitude Test, SAT) es 515 (*The World Almanac,* 2009). El Consejo Universitario vuelve a escalar en forma periódica las calificaciones del examen de tal manera que la desviación estándar sea aproximadamente 100. Responda las preguntas siguientes usando una distribución con forma de campana y la regla empírica para las calificaciones del examen verbal.

a) ¿Qué porcentaje de estudiantes obtuvo una calificación en el SAT verbal mayor que 615?
b) ¿Qué porcentaje obtuvo una calificación en el SAT verbal mayor que 715?
c) ¿Qué porcentaje de alumnos logró una calificación entre 415 y 515?
d) ¿Qué porcentaje obtuvo una calificación entre 315 y 615?

32. Los altos costos del mercado de bienes raíces en California han ocasionado que las familias que no pueden darse el lujo de comprar casas más grandes consideren los cobertizos de los patios traseros como una opción de ampliación. Muchos están usando las estructuras de sus patios para construir sus estudios, salas de arte y áreas de pasatiempos, así como para almacenamiento adicional. El precio medio de una estructura de tablillas de madera para patio trasero hecha a la medida es de $3 100 (*Newsweek,* 29 de septiembre de 2003). Suponga que la desviación estándar es $1 200.
a) ¿Cuál es el valor *z* para una estructura de patio trasero que cuesta $2 300?
b) ¿Cuál es el valor *z* para una estructura que cuesta $4 900?
c) Interprete los valores *z* en los incisos *a*) y *b*). Comente si alguna debe considerarse una observación atípica.
d) El artículo de *Newsweek* describió una combinación de oficina en el cobertizo del patio trasero construida con $13 000 en Albany, California. ¿Esta estructura debe considerarse una observación atípica? Explique por qué.

33. Florida Power & Light (FP&L) Company ha gozado de la reputación de reparar rápidamente un sistema eléctrico después de las tormentas. Sin embargo, durante las temporadas de huracanes de 2004 y 2005 la realidad fue otra: el método comprobado de la empresa para las reparaciones de emergencia ya no fue lo suficientemente bueno (*The Wall Street Journal,* 16 de enero de 2006). Los datos siguientes muestran los días requeridos para restablecer el servicio eléctrico después de siete huracanes durante los años de referencia.

Huracán	Días para restablecer el servicio
Charley	13
Frances	12
Jeanne	8
Dennis	3
Katrina	8
Rita	2
Vilma	18

Con base en esta muestra de siete huracanes, calcule los estadísticos descriptivos siguientes.
a) Media, mediana y moda.
b) Rango y desviación estándar.
c) ¿*Vilma* debe considerarse una observación atípica en términos de los días requeridos para restablecer el servicio eléctrico?
d) Los siete huracanes ocasionaron 10 millones de interrupciones en el servicio a los clientes. ¿Los estadísticos indican que FP&L debe considerar la necesidad de mejorar su método de reparaciones del sistema eléctrico? Comente.

34. Una muestra de puntuaciones de 10 partidos de basquetbol colegial de la NCAA proporcionó los datos siguientes (*USA Today,* 26 de enero de 2004).

WEB archivo
NCAA

Winning Team	Points	Losing Team	Points	Winning Margin
Arizona	90	Oregon	66	24
Duke	85	Georgetown	66	19
Florida State	75	Wake Forest	70	5
Kansas	78	Colorado	57	21
Kentucky	71	Notre Dame	63	8
Louisville	65	Tennessee	62	3
Oklahoma State	72	Texas	66	6

Winning Team	Points	Losing Team	Points	Winning Margin
Purdue	76	Michigan State	70	6
Stanford	77	Southern Cal	67	10
Wisconsin	76	Illinois	56	20

a) Calcule la media y la desviación estándar de los puntos anotados por el equipo ganador.
b) Suponga que los puntos anotados por los equipos triunfadores en todos los partidos de la NCAA siguen una distribución con forma de campana. Utilizando la media y la desviación estándar obtenidas en el inciso *a*), estime el porcentaje de los partidos de la NCAA en los cuales el equipo ganador anota 84 puntos o más. Calcule el porcentaje de los partidos de la NCAA en los cuales el equipo triunfador anota más de 90 puntos.
c) Calcule la media y la desviación estándar del margen de victoria. ¿Los datos contienen observaciones atípicas? Explique por qué.

35. *Consumer Reports* publica reseñas y calificaciones de una variedad de productos en su sitio web. A continuación se presenta una muestra de 20 sistemas de bocinas y sus calificaciones, las cuales varían en una escala de 1 a 5, en la que 5 es la mejor.

WEB archivo
Speakers

Speaker	Rating	Speaker	Rating
Infinity Kappa 6.1	4.00	ACI Sapphire III	4.67
Allison One	4.12	Bose 501 Series	2.14
Cambridge Ensemble II	3.82	DCM KX-212	4.09
Dynaudio Contour 1.3	4.00	Eosone RSF1000	4.17
Hsu Rsch. HRSW12V	4.56	Joseph Audio RM7si	4.88
Legacy Audio Focus	4.32	Martin Logan Aerius	4.26
Mission 73li	4.33	Omni Audio SA 12.3	2.32
PSB 400i	4.50	Polk Audio RT12	4.50
Snell Acoustics D IV	4.64	Sunfire True Subwoofer	4.17
Thiel CS1.5	4.20	Yamaha NS-A636	2.17

a) Calcule la media y la mediana.
b) Estime el primer y el tercer cuartiles.
c) Calcule la desviación estándar.
d) El sesgo de estos datos es −1.67. Comente la forma de la distribución.
e) ¿Cuáles son las puntuaciones z asociadas con Allison One y Omni Audio?
f) ¿Los datos contienen observaciones atípicas? Explique.

3.4 Análisis exploratorio de datos

En el capítulo 2 se introdujo el diagrama de tallo y hoja como una técnica de análisis exploratorio de datos. Recuerde que dicho análisis permite usar operaciones aritméticas simples y representaciones gráficas fáciles de dibujar para resumir los datos. En esta sección continúa el análisis exploratorio de datos considerando resúmenes de cinco números y diagramas de caja.

Resumen de cinco números

En un **resumen de cinco números**, los cinco siguientes se usan para resumir los datos.

1. Valor menor
2. Primer cuartil (Q_1)
3. Mediana (Q_2)
4. Tercer cuartil (Q_3)
5. Valor mayor

La manera más fácil de elaborar un resumen de cinco números es colocar primero los datos en orden ascendente. Una vez hecho esto es fácil identificar el valor menor, los tres cuartiles y el valor mayor. Los sueldos mensuales de inicio mostrados en la tabla 3.1 para la muestra de 12 licenciados en administración de empresas recién egresados se repiten aquí en orden ascendente.

3 310 3 355 3 450 | 3 480 3 480 3 490 | 3 520 3 540 3 550 | 3 650 3 730 3 925

$Q_1 = 3\,465$ $Q_2 = 3\,505$ (mediana) $Q_3 = 3\,600$

La mediana de 3 505 y los cuartiles $Q_1 = 3\,465$ y $Q_3 = 3\,600$ se calcularon en la sección 3.1. Al revisar los datos se observa un valor menor de 3 310 y un valor mayor de 3 925. Por tanto, el resumen de cinco números para los datos de los sueldos iniciales es 3 310, 3 465, 3 505, 3 600 y 3 925. Entre los números adyacentes de un resumen de cinco números se encuentra aproximadamente un cuarto, o 25%, de las observaciones.

Diagrama de caja

Un **diagrama de caja** es un resumen gráfico de los datos basado en un resumen de cinco números. La clave para elaborar de un diagrama de caja es el cálculo de la mediana y los cuartiles Q_1 y Q_3. El rango intercuartílico, RIC $= Q_3 - Q_1$, también se utiliza. En la figura 3.5 se aprecia el diagrama de cuadro de los datos de los sueldos mensuales iniciales. Los pasos que se siguen para elaborarlo se presentan a continuación.

1. Se traza una caja con sus extremos ubicados en el primer y tercer cuartiles. Para los datos de los sueldos iniciales, $Q_1 = 3\,465$ y $Q_3 = 3\,600$. Este cuadro contiene la mitad, 50%, de los datos.
2. Se traza una línea vertical en el cuadro donde se ubica la mediana (3 505 para los datos de los sueldos iniciales).
3. Al usar el rango intercuartílico, RIC $= Q_3 - Q_1$, se localizan los *límites*. Para el diagrama de caja los límites son 1.5(RIC) por debajo de Q_1 y 1.5(RIC) por encima de Q_3. Para los datos de los sueldos, RIC $= Q_3 - Q_1 = 3\,600 - 3\,465 = 135$. Por tanto, los límites son $3\,465 - 1.5(135) = 3\,262.5$ y $3\,600 + 1.5(135) = 3\,802.5$. Los datos fuera de estos límites se consideran *observaciones atípicas.*
4. Las líneas punteadas de la figura 3.5 se llaman *bigotes.* Éstos se trazan desde los extremos de la caja hasta los valores menor y mayor *dentro de los límites* calculados en el paso 3. Por tanto, los bigotes terminan en los valores de los sueldos de 3 310 y 3 730.
5. Por último, la ubicación de cada observación atípica se señala con un asterisco (símbolo *). En la figura 3.5 se aprecia una observación, 3 925.

Los diagramas de caja proporcionan otra manera de identificar observaciones atípicas. Sin embargo, no necesariamente identifican los mismos valores que aquellos con una puntuación z menor que −3 o mayor que +3. Cualquiera de los dos procedimientos o ambos pueden usarse.

En la figura 3.5 se trazaron líneas que ilustran la posición de los límites superior e inferior, cómo se calculan los límites y dónde se ubican. Aunque los límites siempre se calculan, no se trazan

FIGURA 3.5 Diagrama de caja de los datos de los sueldos iniciales con líneas que muestran los límites superior e inferior

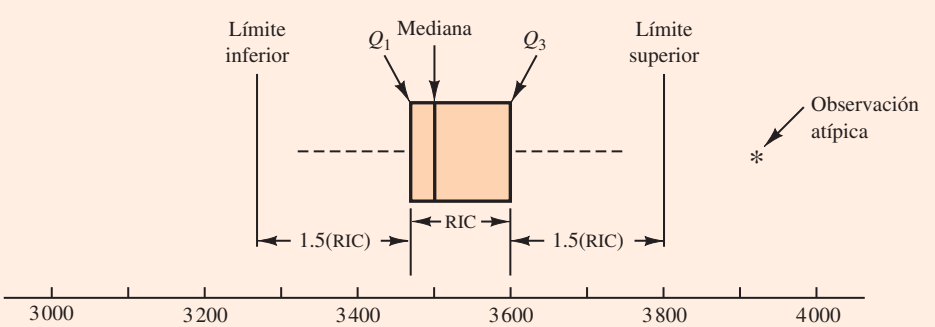
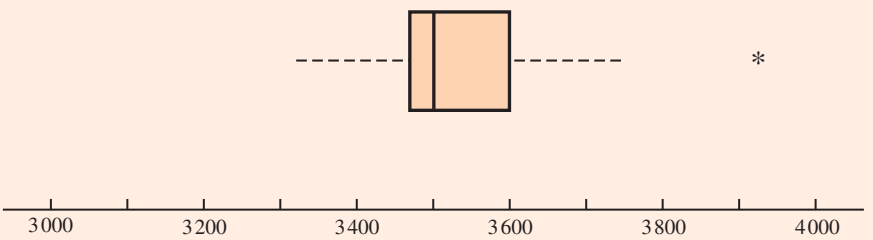


FIGURA 3.6 Diagrama de caja de los datos de los sueldos mensuales iniciales



por lo general en los diagramas de caja. La figura 3.6 muestra la apariencia usual de este tipo de diagrama para los datos de los sueldos.

WEB archivo
MajorSalary

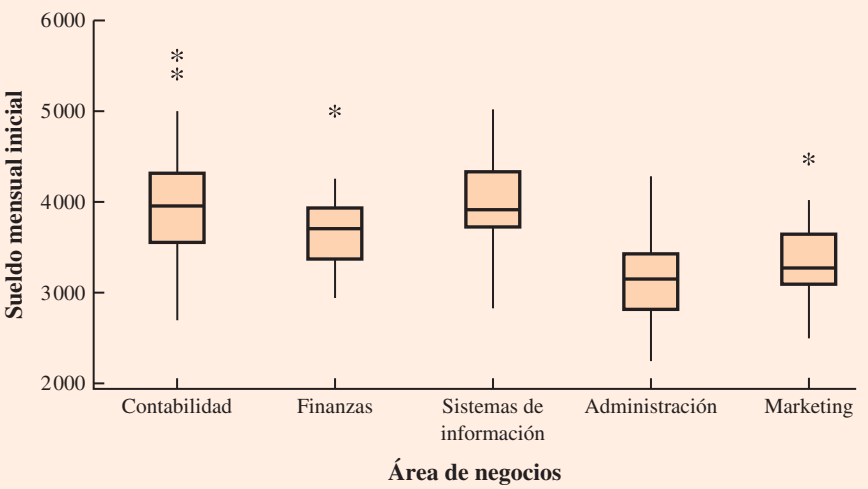
Con la finalidad de comparar los sueldos mensuales iniciales de los licenciados en administración de empresas por área de especialización, se seleccionó una muestra de 111 licenciados recién graduados. Se registraron el campo de especialización y el sueldo mensual inicial de cada profesional. La figura 3.7 muestra los diagramas de caja de Minitab para contabilidad, finanzas, sistemas de información, administración y marketing. Observe que el área de especialización aparece en el eje horizontal, y cada diagrama de caja en el eje vertical por encima del área correspondiente. Mostrar los diagramas de caja de esta manera es una técnica gráfica excelente para hacer comparaciones entre dos o más grupos.

¿Qué observaciones puede hacer acerca de los sueldos iniciales por área de especialización usando los diagramas de caja de la figura 3.7? En específico se observa lo siguiente.

- Los sueldos más altos corresponden a contabilidad; los sueldos más bajos corresponden a administración y marketing.
- Con base en las medianas, la de los sueldos de contabilidad y sistemas de información es similar y mayor. Le sigue finanzas, y administración y contabilidad muestran sueldos con una mediana inferior.
- Existen observaciones atípicas de sueldos altos para las áreas de contabilidad, finanzas y marketing.
- Los sueldos en el área de finanzas parecen tener menos variación, mientras que en contabilidad parecen tener la mayor variación.

Tal vez pueda ver otras interpretaciones basadas en estos diagramas de caja.

FIGURA 3.7 Diagramas de cuadro de Minitab de los sueldos mensuales iniciales por área de especialización



NOTAS Y COMENTARIOS

1. Una ventaja de los procedimientos del análisis exploratorio de datos estriba en que son fáciles de usar, ya que requieren pocos cálculos numéricos. Sencillamente los valores de datos se clasifican en orden ascendente y se identifica el resumen de cinco números. Entonces puede trazarse el diagrama de caja. No es necesario calcular la media y la desviación estándar de los datos.
2. En el apéndice 3.1 se explica cómo elaborar un diagrama de caja de los datos de los sueldos iniciales usando Minitab. El diagrama obtenido se parece al de la figura 3.6, pero girado hacia un lado.

Ejercicios

Métodos

36. Considere una muestra con los datos 27, 25, 20, 15, 30, 34, 28 y 25. Proporcione el resumen de cinco números de los datos.
37. Elabore el diagrama de caja de los datos del ejercicio 36.

AUTO evaluación

38. Muestre el resumen de cinco números y el diagrama de caja de los datos siguientes: 5, 15, 18, 10, 12, 16, 10, 6.
39. Un conjunto de datos tiene un primer cuartil de 42 y un tercer cuartil de 50. Calcule los límites inferior y superior del diagrama de caja correspondiente. ¿Un valor de datos de 65 debe considerarse una observación atípica?

Aplicaciones

40. Naples, Florida, celebra un medio maratón (carrera de 13.1 millas) en enero de cada año. El evento atrae a corredores de todo Estados Unidos y de otras partes del mundo. En enero de 2009 entraron 22 hombres (Men) y 31 mujeres (Women) en la clase de edades de 19 a 24 años. Los tiempos de llegada a la meta en minutos se listan enseguida (*Naples Daily News,* 19 de enero de 2009). Los tiempos se muestran en orden de llegada (Finish).

WEB archivo

Runners

Finish	Men	Women	Finish	Men	Women	Finish	Men	Women
1	65.30	109.03	11	109.05	123.88	21	143.83	136.75
2	66.27	111.22	12	110.23	125.78	22	148.70	138.20
3	66.52	111.65	13	112.90	129.52	23		139.00
4	66.85	111.93	14	113.52	129.87	24		147.18
5	70.87	114.38	15	120.95	130.72	25		147.35
6	87.18	118.33	16	127.98	131.67	26		147.50
7	96.45	121.25	17	128.40	132.03	27		147.75
8	98.52	122.08	18	130.90	133.20	28		153.88
9	100.52	122.48	19	131.80	133.50	29		154.83
10	108.18	122.62	20	138.63	136.57	30		189.27
						31		189.28

a) George Towett, de Marietta, Georgia, llegó en primer lugar de los hombres y Lauren Wald, de Gainesville, Florida, llegó en primer lugar de las mujeres. Compare los tiempos de llegada de los primeros lugares para ambos grupos. Si los 53 corredores hombres y mujeres hubieran competido como un grupo, ¿en qué lugar habría terminado Lauren?

b) ¿Cuál es el tiempo medio para los corredores hombres y mujeres? Compare a los corredores y a las corredoras con base en la mediana de sus tiempos.

c) Proporcione un resumen de cinco números tanto de los hombres como de las mujeres.

d) ¿Hay observaciones atípicas en alguno de los dos grupos?

e) Muestre los diagramas de caja para los dos grupos. ¿Quiénes tienen la mayor variación en los tiempos de llegada: los hombres o las mujeres? Explique.

AUTOevaluación

41. A continuación se proporcionan las ventas anuales, en millones de dólares, de 21 compañías farmacéuticas.

8 408	1 374	1 872	8 879	2 459	11 413
608	14 138	6 452	1 850	2 818	1 356
10 498	7 478	4 019	4 341	739	2 127
3 653	5 794	8 305			

a) Proporcione un resumen de cinco números.
b) Calcule los límites inferior y superior.
c) ¿Los datos contienen observaciones atípicas?
d) Las ventas de $14 138 millones de Johnson & Johnson son las más altas de la lista. Suponga que cometió un error al introducir los datos (una transposición) y que las ventas se introdujeron como $41 138 millones. ¿El método de detección de observaciones del inciso *c*) identifica este problema y permite corregir errores en la introducción de datos?
e) Muestre un diagrama de caja.

42. *Consumer Reports* proporcionó calificaciones de satisfacción del cliente en general para los servicios de telefonía celular AT&T, Sprint, T-Mobile y Verizon en zonas metropolitanas importantes de todo Estados Unidos. La calificación de cada servicio refleja la satisfacción del cliente considerando una variedad de factores como el costo, los problemas de conectividad, las llamadas suspendidas, la interferencia estática y el soporte técnico. Se utilizó una escala de satisfacción de 0 a 100, en la cual 0 indica una insatisfacción total y 100 una satisfacción total. Las calificaciones para los cuatro servicios de telefonía celular en 20 zonas metropolitanas se muestran en seguida (*Consumer Reports,* enero de 2009).

WEB archivo
CellService

Metropolitan Area	AT&T	Sprint	T-Mobile	Verizon
Atlanta	70	66	71	79
Boston	69	64	74	76
Chicago	71	65	70	77
Dallas	75	65	74	78
Denver	71	67	73	77
Detroit	73	65	77	79
Jacksonville	73	64	75	81
Las Vegas	72	68	74	81
Los Ángeles	66	65	68	78
Miami	68	69	73	80
Minneapolis	68	66	75	77
Philadelphia	72	66	71	78
Phoenix	68	66	76	81
San Antonio	75	65	75	80
San Diego	69	68	72	79
San Francisco	66	69	73	75
Seattle	68	67	74	77
St. Louis	74	66	74	79
Tampa	73	63	73	79
Washington	72	68	71	76

a) Considere T-Mobile primero. ¿Cuál es la mediana de la calificación?
b) Elabore un resumen de cinco números para el servicio de esta empresa.
c) ¿Hay observaciones atípicas para T-Mobile? Explique por qué.
d) Repita los incisos *b*) y *c*) para los otros tres servicios de telefonía celular.

e) Presente los diagramas de caja para los cuatro servicios de telefonía celular en una gráfica. Comente qué indica la comparación de diagramas acerca de los cuatro servicios. ¿Cuál recomendó *Consumer Reports* como el mejor en cuanto a la satisfacción del cliente en general?

43. Los Phillies de Filadelfia triunfaron en la Serie Mundial de beisbol de las grandes ligas de 2008 al derrotar a Mantarrayas de Tampa Bay 4 a 3 (*The Philadelphia Inquirer,* 29 de octubre de 2008). Antes, en la clasificatoria de las grandes ligas de beisbol, los Phillies de Filadelfia ganaron el Campeonato de la Liga Nacional al vencer a Los Dodgers de Los Ángeles, mientras que Mantarrayas de Tampa Bay se llevó el Campeonato de la Liga Americana al derrotar a los Medias Rojas de Boston Red Sox. El archivo MLBSalaries contiene los sueldos de los 28 jugadores de cada uno de estos cuatro equipos (base de datos de sueldos de *USA Today*, octubre de 2008). Los datos, mostrados en miles de dólares, se han ordenado del sueldo mayor al menor para cada equipo.

WEB archivo
MLBSalaries

a) Analice los sueldos para el campeón mundial Phillies de Filadelfia. ¿Cuál es la nómina total del equipo? ¿Cuál es la mediana del sueldo? Proporcione el resumen de cinco números.

b) ¿Hay observaciones atípicas para los Phillies de Filadelfia? De ser así, ¿cuántos y de cuánto son los montos de los sueldos?

c) ¿Cuál es la nómina total de cada uno de los otros tres equipos? Elabore el resumen de cinco números para cada equipo e identifique cualesquiera observaciones atípicas.

d) Muestre los diagramas de caja de los sueldos para los cuatro equipos. ¿Cuáles son sus interpretaciones? De estos cuatro equipos, ¿parece que el equipo con sueldos más altos ganó los campeonatos de la liga y la Serie Mundial?

WEB archivo
Mutual

44. Un listado de 46 fondos de inversión y su rendimiento porcentual total de 12 meses se muestra en la tabla 3.5 (*Smart Money,* febrero de 2004).

a) ¿Cuáles son la media y la mediana de los porcentajes de rendimiento para estos fondos de inversión?

b) ¿Cuáles son el primer y el tercer cuartiles?

c) Proporcione un resumen de cinco números.

d) ¿Los datos contienen alguna observación atípica? Muestre un diagrama de caja.

TABLA 3.5 Rendimiento de 12 meses para fondos de inversión

Mutual Fund	Return (%)	Mutual Fund	Return (%)
Alger Capital Appreciation	23.5	Nations Small Company	21.4
Alger LargeCap Growth	22.8	Nations SmallCap Index	24.5
Alger MidCap Growth	38.3	Nations Strategic Growth	10.4
Alger SmallCap	41.3	Nations Value Inv	10.8
AllianceBernstein Technology	40.6	One Group Diversified Equity	10.0
Federated American Leaders	15.6	One Group Diversified Int'l	10.9
Federated Capital Appreciation	12.4	One Group Diversified Mid Cap	15.1
Federated Equity-Income	11.5	One Group Equity Income	6.6
Federated Kaufmann	33.3	One Group Int'l Equity Index	13.2
Federated Max-Cap Index	16.0	One Group Large Cap Growth	13.6
Federated Stock	16.9	One Group Large Cap Value	12.8
Janus Adviser Int'l Growth	10.3	One Group Mid Cap Growth	18.7
Janus Adviser Worldwide	3.4	One Group Mid Cap Value	11.4
Janus Enterprise	24.2	One Group Small Cap Growth	23.6
Janus High-Yield	12.1	PBHG Growth	27.3
Janus Mercury	20.6	Putnam Europe Equity	20.4
Janus Overseas	11.9	Putnam Int'l Capital Opportunity	36.6
Janus Worldwide	4.1	Putnam International Equity	21.5
Nations Convertible Securities	13.6	Putnam Int'l New Opportunity	26.3
Nations Int'l Equity	10.7	Strong Advisor Mid Cap Growth	23.7
Nations LargeCap Enhd. Core	13.2	Strong Growth 20	11.7
Nations LargeCap Index	13.5	Strong Growth Inv	23.2
Nation MidCap Index	19.5	Strong Large Cap Growth	14.5

3.5 Medidas de asociación entre dos variables

Hasta ahora hemos examinado los métodos numéricos que resumen los datos de *una variable a la vez*. Un gerente o quien toma decisiones se interesa con frecuencia en la *relación entre dos variables*. En esta sección se presentan la covarianza y la correlación como medidas descriptivas de la relación entre dos variables.

Para empezar, reconsidere la aplicación referente a una tienda de estéreos y equipos de sonido en San Francisco que se presentó en la sección 2.4. El gerente del establecimiento quiere determinar la relación entre el número de comerciales de televisión transmitidos el fin de semana y las ventas en la tienda durante la semana siguiente. Los datos muestrales con las ventas expresadas en cientos de dólares se proporcionan en la tabla 3.6. Ésta registra 10 observaciones ($n = 10$), una para cada semana. El diagrama de dispersión de la figura 3.8 indica una relación positiva, con las ventas más altas (y) asociadas con un número mayor de comerciales (x). De hecho, el diagrama de dispersión sugiere que se podría usar una línea recta como una aproximación de la relación. En el análisis siguiente se introduce la **covarianza** como una medida descriptiva de la asociación lineal entre dos variables.

Covarianza

Para una muestra de tamaño n con las observaciones (x_1, y_1), (x_2, y_2), etc., la covarianza muestral se define como sigue.

COVARIANZA MUESTRAL

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \qquad (3.10)$$

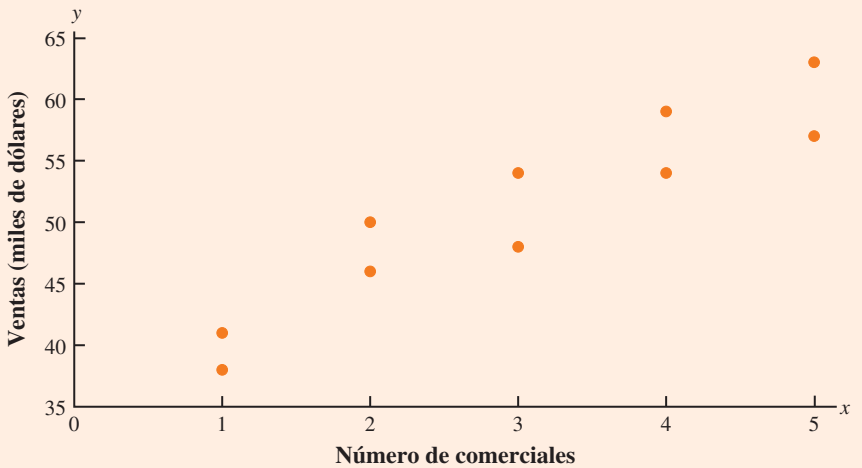
Esta fórmula empareja cada x_i con una y_i. Luego se suman los productos obtenidos al multiplicar la desviación de cada x_i de su media muestral $\bar{x}$ por la desviación de la y_i correspondiente de su media muestral $\bar{y}$; esta suma se divide entonces por $n - 1$.

TABLA 3.6 Datos muestrales para la tienda de estéreos y equipos de sonido

WEB archivo
Stereo

Week	Number of Commercials x	Sales Volume ($100s) y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

FIGURA 3.8 Diagrama para la tienda de estéreos y equipos de sonido



Para medir la solidez de una relación lineal entre el número de comerciales (Number of Commercials) x y el volumen de ventas (Sales Volume) y en el problema de la tienda de estéreos y equipos de sonido, use la ecuación (3.10) a efecto de calcular la covarianza muestral. La tabla 3.7 presenta el cálculo de $\Sigma(x_i - \bar{x})(y_i - \bar{y})$. Observe que $\bar{x} = 30/10 = 3$, y $\bar{y} = 510/10 = 51$. Usando la ecuación (3.10) se obtiene una covarianza muestral de

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{9} = 11$$

TABLA 3.7 Cálculos de la covarianza muestral

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	2	50	−1	−1	1
	5	57	2	6	12
	1	41	−2	−10	20
	3	54	0	3	0
	4	54	1	3	3
	1	38	−2	−13	26
	5	63	2	12	24
	3	48	0	−3	0
	4	59	1	8	8
	2	46	−1	−5	5
Totales	30	510	0	0	99

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{10 - 1} = 11$$

La fórmula para calcular la covarianza de una población de tamaño N es similar a la ecuación (3.10), pero se usa una notación diferente para indicar que se está trabajando con toda la población.

COVARIANZA POBLACIONAL

$$\sigma_{xy} = \frac{\Sigma(x_i - \mu_x)(y_i - \mu_y)}{N} \qquad (3.11)$$

En la ecuación (3.11) la notación μ_x denota la media poblacional de la variable x, y μ_y denota la media poblacional de la variable y. La covarianza poblacional σ_{xy} se define para una población de tamaño N.

Interpretación de la covarianza

Para ayudar en la interpretación de la covarianza muestral, considere la figura 3.9; es igual al diagrama de dispersión de la figura 3.7, con una línea punteada vertical en $\bar{x} = 3$ y una línea punteada horizontal en $\bar{y} = 51$. Las líneas dividen la gráfica en cuatro cuadrantes. Los puntos del cuadrante I corresponden a x_i mayor que $\bar{x}$ y y_i mayor que $\bar{y}$; los puntos del cuadrante II corresponden a x_i menor que $\bar{x}$ y y_i menor que $\bar{y}$, etc. Por tanto, el valor de $(x_i - \bar{x})(y_i - \bar{y})$ debe ser positivo para los puntos del cuadrante I, negativo para los del cuadrante II, positivo para los del cuadrante III, y negativo para los puntos del cuadrante IV.

La covarianza es una medida de la asociación lineal entre dos variables.

Si el valor de s_{xy} es positivo, los puntos con la mayor influencia en s_{xy} deben estar en los cuadrantes I y III. Por ende, un valor positivo para s_{xy} indica una asociación lineal positiva entre x y y; es decir, a medida que el valor de x aumenta, el valor de y también. Si el valor de s_{xy} es negativo, no obstante, los puntos con la mayor influencia en s_{xy} están en los cuadrantes II y IV. Por ende, un valor negativo para s_{xy} indica una asociación lineal negativa entre x y y; es decir, a medida que el valor de x aumenta, el valor de y disminuye. Por último, si los puntos están distribuidos de manera uniforme en los cuatro cuadrantes, el valor de s_{xy} será cercano a cero, lo que indica que no existe una asociación lineal entre x y y. En la figura 3.10 se aprecian los valores de s_{xy} que se expresan con tres tipos distintos de diagramas de dispersión.

FIGURA 3.9 Diagrama de dispersión particionado para la tienda de estéreos y equipos de sonido

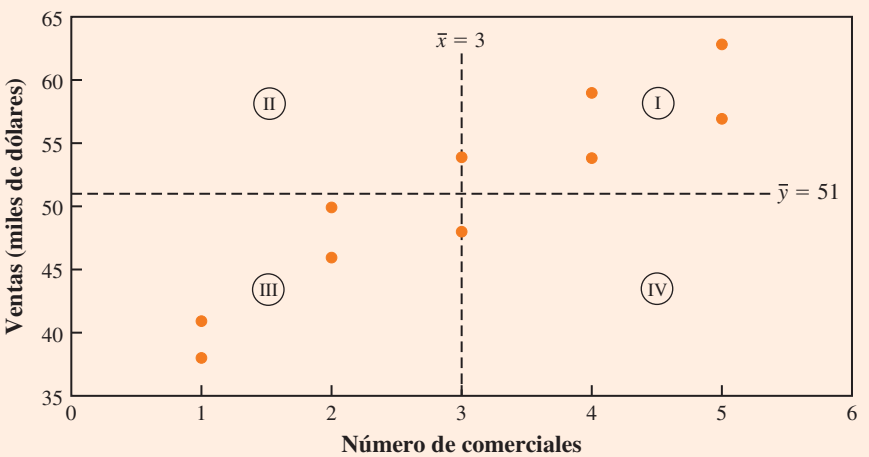
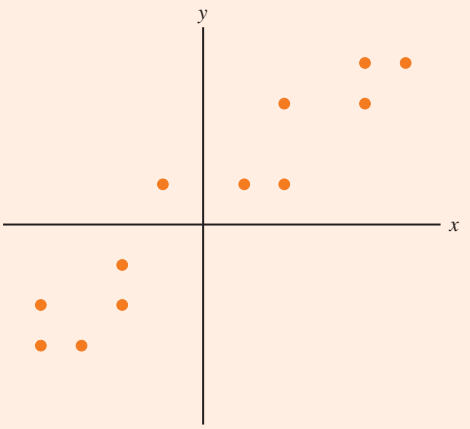
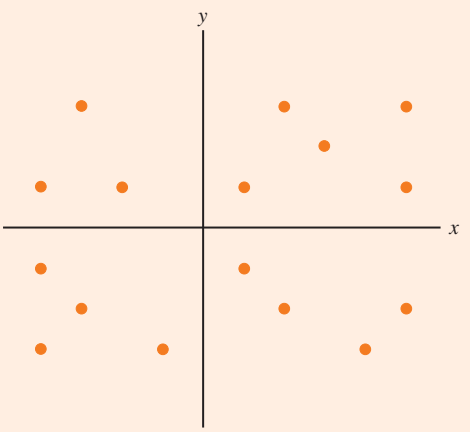


FIGURA 3.10 Interpretación de la covarianza muestral

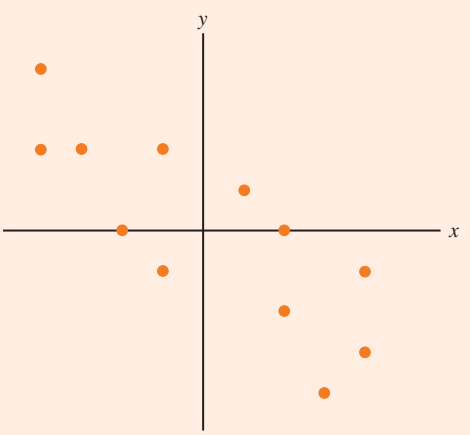
s_{xy} positiva:
(x y y se relacionan linealmente de manera positiva)



s_{xy} aproximadamente 0:
(x y y no se relacionan de manera lineal)



s_{xy} negativa:
(x y y se relacionan linealmente de manera negativa)



Observe de nuevo la figura 3.9. El diagrama de dispersión para la tienda de estéreos y equipos de sonido sigue el patrón del panel superior de la figura 3.10. Como es de esperarse, el valor de la covarianza muestral indica una relación lineal positiva en la que $s_{xy} = 11$.

A partir del análisis anterior, podría parecer que un valor positivo grande para la covarianza indica una relación lineal positiva sólida, y un valor negativo grande indica una relación lineal negativa sólida. Sin embargo, un problema con la covarianza como medida de la solidez de una relación lineal estriba en que su valor depende de las unidades de medida para x y y. Por ejemplo, suponga que estamos interesados en la relación entre la estatura x y el peso y de las personas. Desde luego, la solidez de la relación debe ser la misma, ya sea que la estatura se mida en pies o pulgadas. Sin embargo, la medición en pulgadas no da valores numéricos mucho mayores para $(x_i - \bar{x})$ que cuando la estatura se mide en pies. Por tanto, con la altura medida en pulgadas se obtendría un valor mayor para el numerador $\Sigma(x_i - \bar{x})(y_i - \bar{y})$ en la ecuación (3.10) —y por consiguiente una covarianza mayor—, cuando de hecho la relación no cambia. Una medida de la relación entre dos variables que no se ve afectada por las unidades de medición para x y y es el **coeficiente de correlación**.

Coeficiente de correlación

Para los datos muestrales, el coeficiente de correlación del producto-momento de Pearson se define como se indica a continuación.

COEFICIENTE DE CORRELACIÓN DEL PRODUCTO-MOMENTO DE PEARSON: DATOS MUESTRALES

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{3.12}$$

donde

r_{xy} = coeficiente de correlación muestral
s_{xy} = covarianza muestral
s_x = desviación estándar muestral de x
s_y = desviación estándar muestral de y

La ecuación (3.12) indica que el coeficiente de correlación del producto-momento de Pearson para los datos muestrales (conocido comúnmente de manera más simple como *coeficiente de correlación muestral*) se calcula al dividir la covarianza muestral entre el producto de la desviación estándar muestral de x y la desviación estándar muestral de y.

A continuación se calcula el coeficiente de correlación muestral para la tienda de estéreos y equipos de sonido. Usando los datos de la tabla 3.7 se pueden estimar las desviaciones estándar muestrales para las dos variables:

$$s_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{20}{9}} = 1.49$$

$$s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{566}{9}} = 7.93$$

Ahora, debido a que $s_{xy} = 11$, el coeficiente de correlación muestral es igual a

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{11}{(1.49)(7.93)} = 0.93$$

La fórmula para calcular el coeficiente de correlación de una población, denotado por la letra griega ρ_{xy} (*ro*), se presenta a continuación.

El coeficiente de correlación muestral r_{xy} es el estimador del coeficiente de correlación poblacional ρ_{xy}.

COEFICIENTE DE CORRELACIÓN DEL PRODUCTO-MOMENTO DE PEARSON: DATOS POBLACIONALES

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{3.13}$$

donde

ρ_{xy} = coeficiente de correlación poblacional
σ_{xy} = covarianza poblacional
σ_x = desviación estándar poblacional de x
σ_y = desviación estándar poblacional de y

El coeficiente de correlación muestral r_{xy} proporciona una estimación del coeficiente de correlación poblacional ρ_{xy}.

Interpretación del coeficiente de correlación

Primero se considerará un ejemplo sencillo que ilustra el concepto de una relación lineal positiva perfecta. El diagrama de dispersión de la figura 3.11 representa la relación entre x y y con base en los datos muestrales siguientes.

x_i	y_i
5	10
10	30
15	50

FIGURA 3.11 Diagrama de dispersión que representa una relación lineal positiva perfecta

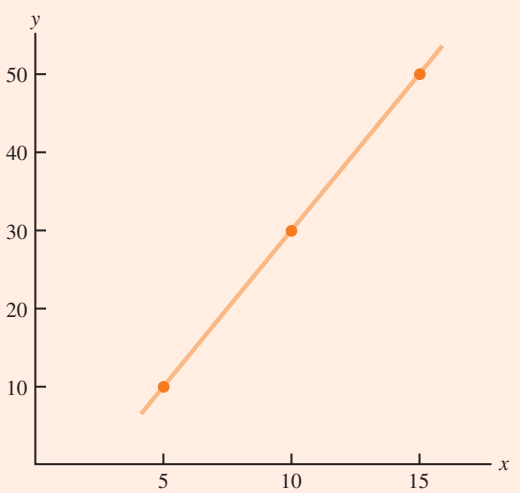


TABLA 3.8 Cálculos utilizados para obtener el coeficiente de correlación muestral

	x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
	5	10	−5	25	−20	400	100
	10	30	0	0	0	0	0
	15	50	5	25	20	400	100
Totales	30	90	0	50	0	800	200
	$\bar{x} = 10$	$\bar{y} = 30$					

La línea recta trazada a través de cada uno de los tres puntos muestra una relación lineal perfecta entre x y y. Con el fin de aplicar la ecuación (3.12) para calcular la correlación muestral, primero se calculan s_{xy}, s_x y s_y. Algunos cálculos se presentan en la tabla 3.8. Con los resultados de esta tabla encontramos

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{200}{2} = 100$$

$$s_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{50}{2}} = 5$$

$$s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{800}{2}} = 20$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{100}{5(20)} = 1$$

El coeficiente de correlación varía de −1 a +1. Los valores cercanos a −1 o a + 1 indican una relación lineal sólida. Entre más se acerque la correlación a cero, más débil es la relación.

Por tanto, se aprecia que el valor del coeficiente de correlación muestral es 1.

En general, se puede demostrar que si todos los puntos de un conjunto de datos caen en una línea recta inclinada con pendiente positiva, el valor del coeficiente de correlación muestral es +1; es decir, un coeficiente de correlación muestral de +1 corresponde a una relación lineal positiva perfecta entre x y y. Por otra parte, si los puntos del conjunto de datos caen en una recta con pendiente negativa, el valor del coeficiente de correlación muestral es −1; es decir, corresponde a una relación lineal negativa perfecta entre x y y.

Suponga ahora que cierto conjunto de datos indica una relación lineal positiva entre x y y pero la relación no es perfecta. El valor de r_{xy} será menor que 1, lo que indica que los puntos en el diagrama de dispersión no estarán todos sobre una línea recta. A medida que los puntos se desvían más y más de una relación lineal positiva perfecta, el valor de r_{xy} se vuelve cada vez más y más pequeño. Cuando éste es igual a cero, indica que no existe una relación lineal entre x y y, y los valores de r_{xy} cercanos a cero indican una relación lineal débil.

Para los datos de la tienda de estéreos y equipos de sonido, $r_{xy} = 0.93$. Por consiguiente, se concluye que existe una relación lineal positiva sólida entre el número de comerciales y las ventas. De manera más específica, un aumento en el número de comerciales se asocia con un incremento en las ventas.

En resumen, se observa que la correlación proporciona una medida de asociación lineal y no necesariamente de causalidad. Una correlación alta entre dos variables no significa que los cambios en una variable ocasionarán cambios en la otra. Por ejemplo, podemos encontrar que la calificación de calidad y el precio típico de la comida en los restaurantes se correlacionan de manera positiva. Sin embargo, un simple incremento en el precio de la comida no causará que la calificación de la calidad aumente.

Ejercicios

Métodos

AUTOevaluación

45. A continuación se presentan cinco observaciones tomadas para dos variables.

x_i	4	6	11	3	16
y_i	50	50	40	60	30

a) Desarrolle un diagrama de dispersión con x en el eje horizontal.
b) ¿Qué indica el diagrama de dispersión elaborado en el inciso *a*) respecto de la relación entre las dos variables?
c) Calcule e interprete la covarianza muestral.
d) Estime e interprete el coeficiente de correlación muestral.

46. A continuación se presentan cinco observaciones tomadas para dos variables.

x_i	6	11	15	21	27
y_i	6	9	6	17	12

a) Elabore un diagrama de dispersión con estos datos.
b) ¿Qué indica el diagrama de dispersión acerca de la relación entre x y y?
c) Calcule e interprete la covarianza muestral.
d) Determine e interprete el coeficiente de correlación muestral.

Aplicaciones

47. Nielsen Media Research proporciona dos medidas de la audiencia televisiva: el rating, que es el porcentaje de hogares que cuenta con un aparato y está viendo un programa, y el share, que es el porcentaje de hogares que tiene el equipo encendido cuyos miembros están viendo un programa determinado. Las cifras siguientes muestran los datos de las calificaciones y las cuotas de Nielsen de la Serie Mundial de Beisbol de las Grandes Ligas durante un periodo de nueve años (Associated Press, 27 de octubre de 2003).

Rating	19	17	17	14	16	12	15	12	13
Share	32	28	29	24	26	20	24	20	22

a) Elabore un diagrama de dispersión con el rating en el eje horizontal.
b) ¿Cuál es la relación entre rating y share? Explique por qué.
c) Calcule e interprete la covarianza muestral.
d) Calcule el coeficiente de correlación muestral. ¿Qué indica este valor acerca de la relación entre rating y share?

48. Un estudio de un departamento de transporte sobre la velocidad de manejo y las millas por galón para automóviles de tamaño mediano dio como resultado los datos siguientes.

Velocidad (millas por hora)	30	50	40	55	30	25	60	25	50	55
Millas por galón	28	25	25	23	30	32	21	35	26	25

Calcule e interprete el coeficiente de correlación muestral.

49. A principios de 2009 el declive económico ocasionó la pérdida de empleos y un incremento en los préstamos morosos para vivienda. La tasa nacional de desempleo fue de 6.5% y el porcentaje de préstamos morosos de 6.12% (*The Wall Street Journal,* 27 de enero de 2009). En la proyección de hacia dónde se dirigía el mercado de bienes raíces el siguiente año, los economistas estudiaron la relación entre la tasa de desempleo y el porcentaje de préstamos morosos. La expectativa era que si la primera seguía en aumento, habría también un incremento en el porcentaje de préstamos con deudores morosos. Los datos siguientes muestran la tasa de

desempleo y el porcentaje de préstamos morosos para 27 de los principales mercados de bienes raíces.

WEB archivo
Housing

Metro Area	Jobless Rate (%)	Delinquent Loan (%)	Metro Area	Jobless Rate (%)	Delinquent Loan (%)
Atlanta	7.1	7.02	Nueva York	6.2	5.78
Boston	5.2	5.31	Orange County	6.3	6.08
Charlotte	7.8	5.38	Orlando	7.0	10.05
Chicago	7.8	5.40	Philadelphia	6.2	4.75
Dallas	5.8	5.00	Phoenix	5.5	7.22
Denver	5.8	4.07	Portland	6.5	3.79
Detroit	9.3	6.53	Raleigh	6.0	3.62
Houston	5.7	5.57	Sacramento	8.3	9.24
Jacksonville	7.3	6.99	St. Louis	7.5	4.40
Las Vegas	7.6	11.12	San Diego	7.1	6.91
Los Ángeles	8.2	7.56	San Francisco	6.8	5.57
Miami	7.1	12.11	Seattle	5.5	3.87
Minneapolis	6.3	4.39	Tampa	7.5	8.42
Nashville	6.6	4.78			

a) Calcule el coeficiente de correlación. ¿Existe una correlación positiva entre la tasa de desempleo (Jobless Rate) y el porcentaje de préstamos de vivienda morosos (Delinquent Loan)? ¿Cuál es su interpretación?

b) Muestre un diagrama de dispersión de la relación entre la tasa de desempleo y el porcentaje de préstamos de vivienda morosos.

50. El promedio industrial Dow Jones (DJIA) y el índice 500 de Standard & Poor's (S&P 500) miden el desempeño del mercado de valores. El DJIA se basa en el precio de las acciones de 30 empresas grandes; el S&P 500, en el precio de las acciones de 500 empresas. Si tanto el DJIA como el S&P 500 miden el desempeño del mercado de valores, ¿cómo se correlacionan? Los datos siguientes ilustran el incremento o el decremento porcentual diario en el DJIA y el S&P 500 para una muestra de nueve días durante un periodo de tres meses (*The Wall Street Journal,* 15 de enero a 10 de marzo de 2006).

WEB archivo
StockMarket

DJIA	0.20	0.82	−0.99	0.04	−0.24	1.01	0.30	0.55	−0.25
S&P 500	0.24	0.19	−0.91	0.08	−0.33	0.87	0.36	0.83	−0.16

a) Elabore un diagrama de dispersión.

b) Calcule el coeficiente de correlación muestral para estos datos.

c) Comente la asociación entre el DJIA y el S&P 500. ¿Necesita revisarlos antes de darse una idea general sobre el desempeño diario del mercado de valores?

51. Las temperaturas diarias altas (High) y bajas (Low) para 14 ciudades de todo el mundo se muestran en el siguiente cuadro (The Weather Channel, 22 de abril de 2009).

WEB archivo
WorldTemp

City	High	Low	City	High	Low
Athens	68	50	London	67	45
Beijing	70	49	Moscow	44	29
Berlin	65	44	Paris	69	44
Cairo	96	64	Rio de Janeiro	76	69
Dublin	57	46	Rome	69	51
Geneva	70	45	Tokyo	70	58
Hong Kong	80	73	Toronto	44	39

a) ¿Cuál es la media muestral de la temperatura alta?
b) ¿Cuál es la media muestral de la temperatura baja?
c) ¿Cuál es la correlación entre las temperaturas alta y baja? Comente.

3.6 Media ponderada y trabajo con datos agrupados

En la sección 3.1 se presentó la media como una de las medidas más importantes de ubicación central. La fórmula para la media de una muestra con n observaciones se vuelve a establecer como sigue.

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad (3.14)$$

En esta fórmula, cada x_i recibe igual importancia o peso. Aunque esta práctica es la más común, en algunos casos la media se calcula confiriendo a cada observación un peso que refleje su importancia. Una media calculada de esta manera se conoce como **media ponderada**.

Media ponderada

La media ponderada se calcula como sigue.

MEDIA PONDERADA

$$\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i} \qquad (3.15)$$

donde

x_i = valor de observación i
w_i = peso de la observación i

Cuando los datos provienen de una muestra, la ecuación (3.15) proporciona la media muestral ponderada. Cuando son de una población, μ reemplaza a $\bar{x}$ y la misma ecuación proporciona la media poblacional ponderada.

Como ejemplo de la necesidad de una media ponderada, considere la muestra siguiente de cinco compras de una materia prima durante los tres meses pasados.

Compra	Costo por libra ($)	Número de libras
1	3.00	1 200
2	3.40	500
3	2.80	2 750
4	2.90	1 000
5	3.25	800

Observe que el costo por libra varía de $2.80 a $3.40, y la cantidad comprada varía de 500 a 2 750 libras. Suponga que un gerente solicitó información sobre el costo medio por libra de la materia prima. Debido a que las cantidades ordenadas varían, se debe usar la fórmula para una media ponderada. Los cinco valores de datos del costo por libra son $x_1 = 3.00$; $x_2 = 3.40$; $x_3 = 2.80$; $x_4 = 2.90$, y $x_5 = 3.25$. El costo medio ponderado por libra se obtuvo al ponderar

cada costo por su cantidad correspondiente. Para este ejemplo, los pesos son $w_1 = 1\,200$; $w_2 = 500$; $w_3 = 2\,750$; $w_4 = 1\,000$, y $w_5 = 800$. Con base en la ecuación (3.15), la media ponderada se calculó como sigue.

$$\bar{x} = \frac{1\,200(3.00) + 500(3.40) + 2\,750(2.80) + 1\,000(2.90) + 800(3.25)}{1\,200 + 500 + 2\,750 + 1\,000 + 800}$$

$$= \frac{18\,500}{6\,250} = 2.96$$

Por tanto, el cálculo de la media ponderada indica que el costo medio por libra para la materia prima es \$2.96. Observe que utilizando la ecuación (3.14) en vez de la fórmula de la media ponderada se habrían obtenido resultados erróneos. En este caso, la media de los cinco valores del costo por libra es (3.00 + 3.40 + 2.80 + 2.90 + 3.25)/5 = 15.35/5 = \$3.07, el cual exagera el costo medio real por libra adquirida.

El cálculo de un promedio escolar es un buen ejemplo del uso de la media ponderada.

La opción de los pesos para el cálculo de una media ponderada en particular depende de la aplicación. Un ejemplo muy conocido para los estudiantes universitarios es el cálculo de un promedio escolar. En éste, los valores de datos manejados son por lo general 4 para una calificación A; 3 para una calificación B; 2 para una calificación C; 1 para una calificación D, y 0 para una calificación F. Los pesos son el número de horas de los créditos ganados por cada calificación. El ejercicio 54 al final de esta sección proporciona un ejemplo de este cálculo de la media ponderada. En otros cálculos, las cantidades como las libras, los dólares o el volumen suelen usarse como pesos. Sea como fuere, cuando las observaciones varían en importancia, el analista debe elegir el peso que mejor refleje la importancia de cada observación en la determinación de la media.

Datos agrupados

En la mayoría de los casos, las medidas de posición y variabilidad se calculan con valores de datos individuales. No obstante, los datos en ocasiones están disponibles sólo en forma agrupada o en forma de distribución de frecuencia. En el análisis siguiente se explica cómo usar la fórmula de la media ponderada para obtener aproximaciones de la media, la varianza y la desviación estándar para **datos agrupados**.

En la sección 2.2 se proporcionó una distribución de frecuencia del tiempo en días requerido para completar las auditorías de fin de año de la firma de contabilidad pública Sanderson and Clifford. La distribución de frecuencia de la duración de las auditorías se ilustra en la tabla 3.9. Con base en esta distribución, ¿cuál es la media muestral de la duración de las auditorías?

Para calcular la media usando sólo los datos agrupados, el punto medio de cada clase se trata como si fuera representativo de los elementos de la clase. Sea M_i el punto medio para la clase i, y f_i la frecuencia de la clase i. La fórmula de la media ponderada (3.15) se utiliza entonces con los valores de datos denotada como M_i y los pesos dados por las frecuencias f_i. En este caso,

TABLA 3.9 Distribución de frecuencia de la duración de la auditoría

Duración de la auditoría (días)	Frecuencia
10–14	4
15–19	8
20–24	5
25–29	2
30–34	1
Total	20

el denominador de la ecuación es la suma de las frecuencias, la cual es el tamaño muestral n. Es decir, $\Sigma f_i = n$. Por tanto, la ecuación para la media muestral de los datos agrupados es la siguiente.

MEDIA MUESTRAL PARA DATOS AGRUPADOS

$$\bar{x} = \frac{\Sigma f_i M_i}{n} \tag{3.16}$$

donde

M_i = punto medio para la clase i
f_i = frecuencia para la clase i
n = tamaño muestral

Con los puntos medios de clase, M_i, a medio camino entre los límites de clase, la primera de 10–14 en la tabla 3.9 tiene un punto medio en $(10 + 14)/2 = 12$. Los cinco puntos medios de clase y el cálculo de la media ponderada para los datos de duración de la auditoría se resumen en la tabla 3.10. Como puede verse, la media muestral de la duración de la auditoría es de 19 días.

Para calcular la varianza de datos agrupados se usa una versión ligeramente alterada de la fórmula para la varianza proporcionada en la ecuación (3.5). En esta ecuación las desviaciones cuadradas de los datos con respecto a la media muestral $\bar{x}$ se escribieron como $(x_i - \bar{x})^2$. Sin embargo, con los datos agrupados, los valores no se conocen. En este caso, el punto medio de la clase, M_i, se trata como si fuera representativo de los x_i valores en la clase correspondiente. Por tanto, las desviaciones cuadradas respecto de la media muestral, $(x_i - \bar{x})^2$, se remplazan por $(M_i - \bar{x})^2$. Así, del mismo modo que con los cálculos de la media muestral para los datos agrupados, se pesa cada valor por la frecuencia de la clase, f_i. La suma de las desviaciones cuadradas con respecto a la media para todos los datos se aproxima por medio de $\Sigma f_i(M_i - \bar{x})^2$. El término $n - 1$ en vez de n aparece en el denominador con el fin de hacer de la varianza muestral la estimación de la varianza poblacional. De ahí que la fórmula siguiente se use con objeto de obtener la varianza muestral para los datos agrupados.

VARIANZA MUESTRAL PARA DATOS AGRUPADOS

$$s^2 = \frac{\Sigma f_i(M_i - \bar{x})^2}{n - 1} \tag{3.17}$$

TABLA 3.10 Cálculo de la media muestral de la duración de la auditoría para los datos agrupados

Duración de la auditoría (días)	Punto medio de la clase (M_i)	Frecuencia (f_i)	$f_i M_i$
10–14	12	4	48
15–19	17	8	136
20–24	22	5	110
25–29	27	2	54
30–34	32	1	32
		20	380

Media muestral $\bar{x} = \frac{\Sigma f_i M_i}{n} = \frac{380}{20} = 19$ días

TABLA 3.11 Cálculo de la varianza muestral de la duración de la auditoría para los datos agrupados (media muestral $\bar{x} = 19$)

Duración de la auditoría (días)	Punto medio de clase (M_i)	Frecuencia (f_i)	Desviación ($M_i - \bar{x}$)	Desviación cuadrada $(M_i - \bar{x})^2$	$f_i(M_i - \bar{x})^2$
10–14	12	4	−7	49	196
15–19	17	8	−2	4	32
20–24	22	5	3	9	45
25–29	27	2	8	64	128
30–34	32	1	13	169	169
		20			570
					$\Sigma f_i(M_i - \bar{x})^2$

$$\text{Varianza muestral } s^2 = \frac{\Sigma f_i(M_i - \bar{x})^2}{n - 1} = \frac{570}{19} = 30$$

El cálculo de la varianza muestral para la duración de la auditoría con base en los datos agrupados se ilustra en la tabla 3.11. La varianza muestral es 30.

La desviación estándar para los datos agrupados es sencillamente la raíz cuadrada de la varianza para tales datos. Para los datos de duración de la auditoría, la desviación estándar muestral es $s = \sqrt{30} = 5.48$.

Antes de concluir con esta sección sobre el cálculo de las medidas de posición y dispersión para los datos agrupados, observe que las fórmulas (3.16) y (3.17) son para una muestra. Las medidas para la población se calculan de modo parecido. Las fórmulas de los datos agrupados para una media y varianza poblacionales se presentan a continuación.

MEDIA POBLACIONAL PARA DATOS AGRUPADOS

$$\mu = \frac{\Sigma f_i M_i}{N} \tag{3.18}$$

VARIANZA POBLACIONAL PARA DATOS AGRUPADOS

$$\sigma^2 = \frac{\Sigma f_i(M_i - \mu)^2}{N} \tag{3.19}$$

NOTAS Y COMENTARIOS

En el cálculo de la estadística descriptiva para los datos agrupados, los puntos medios de las clases se utilizan para aproximar los valores de datos de cada clase. Como resultado, la estadística descriptiva para los datos agrupados se aproxima a la estadística descriptiva que resultaría directamente del uso de los datos originales. Por consiguiente, siempre que sea posible es recomendable calcular los estadísticos descriptivos a partir de los datos originales en vez de hacerlo a partir de los datos agrupados.

Ejercicios

Métodos

52. Considere los datos siguientes y sus pesos correspondientes.

x_i	Peso (w_i)
3.2	6
2.0	3
2.5	2
5.0	8

a) Calcule la media ponderada.
b) Calcule la media muestral de los cuatro valores de datos sin ponderar. Observe la diferencia en los resultados proporcionados por los dos cálculos.

AUTO evaluación

53. Considere los datos muestrales en la frecuencia de distribución siguiente.

Clase	Punto medio	Frecuencia
3–7	5	4
8–12	10	7
13–17	15	9
18–22	20	5

a) Calcule la media muestral.
b) Calcule la varianza muestral y la desviación estándar muestral.

Aplicaciones

AUTO evaluación

54. El promedio de calificaciones para los estudiantes universitarios se basa en el cálculo de una media ponderada. Para la mayoría de los estudiantes, las calificaciones se proporcionan con los valores de datos siguientes: A (4), B (3), C (2), D (1) y F (0). Después de 60 horas de clase de estudios superiores, un alumno de la universidad estatal obtuvo 9 horas de clase de A, 15 de clase de B, 33 de clase de C y 3 horas de clase de D.
a) Calcule el promedio de calificaciones del estudiante.
b) Los alumnos de la universidad estatal deben mantener un promedio de calificaciones de 2.5 para sus primeras 60 horas de clases de estudios superiores con el fin de ser admitidos en el colegio de administración. ¿Este estudiante será admitido?

55. Morningstar da seguimiento al rendimiento total de un número grande de fondos de inversión. La tabla siguiente registra el rendimiento total y el número de fondos para cuatro categorías de fondos de inversión (*Morningstar Funds500*, 2008).

Tipo de fondo	Número de fondos	Rendimiento total (%)
Capital nacional	9 191	4.65
Capital internacional	2 621	18.15
Capital especializado	1 419	11.36
Híbridos	2 900	6.75

a) Usando el número de fondos como pesos, calcule el rendimiento total promedio ponderado para los fondos de inversión cubiertos por Morningstar.
b) ¿Hay alguna dificultad asociada con el uso del "número de fondos" como pesos en el cálculo del rendimiento total promedio ponderado para Morningstar en el inciso *a*)? Comente. ¿Qué más podría usarse para los pesos?
c) Suponga que invirtió $10 000 en fondos de inversión a principios de 2007 y que diversificó la inversión al colocar $2 000 en fondos de capital nacional, $4 000 en fondos de capital

internacional, $3 000 en fondos de capital especializado y $1 000 en fondos híbridos. ¿Cuál es el rendimiento esperado sobre el portafolio?

56. Con base en una encuesta de 425 programas de la maestría en administración de empresas, el informe de *U.S. News & World Report* calificó el programa de la Escuela de Negocios de la Universidad Kelley de Indiana como el 20o. mejor del país (*America's Best Graduate Schools,* 2009). La calificación se basó en parte en encuestas a decanos de la escuela de negocios y a reclutadores corporativos. Se solicitó a todos los consultados que evaluaran la calidad académica general del programa de maestría en una escala de 1 "marginal" a 5 "sobresaliente". Use la muestra de respuestas listada abajo para calcular la calificación media ponderada de los decanos de la escuela de negocios y los reclutadores corporativos. Comente.

Evaluación de la calidad	Decanos de la escuela de negocios	Reclutadores corporativos
5	44	31
4	66	34
3	60	43
2	10	12
1	0	0

57. La distribución de frecuencia siguiente muestra el precio por acción de las 30 empresas del promedio industrial Dow Jones (*Barron's,* 2 de febrero de 2009).

Precio por acción	Número de empresas
$ 0–9	4
$10–19	5
$20–29	7
$30–39	3
$40–49	4
$50–59	4
$60–69	0
$70–79	2
$80–89	0
$90–99	1

a) Calcule el precio medio por acción y la desviación estándar del precio por acción para las empresas del promedio industrial Dow Jones.

b) El 16 de enero de 2006, el precio medio por acción era de $45.83 y la desviación estándar de $18.14. Comente los cambios ocurridos en el precio por acción durante el periodo de tres años.

Resumen

En este capítulo se introdujeron varios estadísticos descriptivos que se utilizan para resumir la posición, la variabilidad y la forma de una distribución de datos. A diferencia de los procedimientos tabulares y gráficos del capítulo 2, las medidas en este capítulo resumen los datos en términos de valores numéricos. Cuando los valores numéricos se obtienen de una muestra, se les llama *estadísticos muestrales*; cuando se obtienen de una población se llaman *parámetros poblacionales*. En seguida se presenta parte de la notación utilizada para ambos conceptos.

En la inferencia estadística, la estadística muestral se conoce como estimador puntual del parámetro poblacional.

	Estadístico muestral	Parámetro poblacional
Media	$\bar{x}$	μ
Varianza	s^2	σ^2
Desviación estándar	s	σ
Covarianza	s_{xy}	σ_{xy}
Correlación	r_{xy}	ρ_{xy}

Se definieron la media, la mediana y la moda como medidas de la posición central. Luego se utilizó el concepto de percentiles para describir otras posiciones en el conjunto de datos. A continuación se presentaron el rango, el rango intercuartílico, la varianza, la desviación estándar y el coeficiente de variación como medidas de variabilidad o dispersión. Nuestra medida principal de la forma de una distribución de datos fue el sesgo. Los valores negativos indican una distribución de datos sesgada a la izquierda; los valores positivos indican una distribución de datos sesgada a la derecha. Luego se describió cómo se usan la media y la desviación estándar al aplicar el teorema de Chebyshev y la regla empírica para proporcionar más información sobre la distribución de los datos e identificar observaciones atípicas.

En la sección 3.4 se muestra cómo elaborar un resumen de cinco números y un diagrama de caja para proporcionar información simultánea sobre la ubicación, la variabilidad y la forma de la distribución. En la sección 3.5 se introdujeron la covarianza y el coeficiente de correlación como medidas de asociación entre dos variables. En la sección final se explicó cómo calcular una media ponderada, así como la media, la varianza y la desviación estándar para datos agrupados.

Los estadísticos descriptivos estudiados pueden obtenerse por medio de software para estadística y hojas de cálculo. En los apéndices del capítulo se explica cómo se usan Minitab, Excel y StatTools para elaborar los estadísticos descriptivos que se trabajaron en este capítulo.

Glosario

Coeficiente de correlación Medida de la asociación lineal entre dos variables que toma los valores entre -1 y $+1$. Los valores cercanos a $+1$ indican una relación lineal positiva sólida; los valores cercanos a -1 indican una relación lineal negativa sólida, y los valores cercanos a cero, la falta de una relación lineal.

Coeficiente de variación Medida de variabilidad relativa calculada al dividir la desviación estándar entre la media y multiplicar por 100.

Covarianza Medida de la asociación lineal entre dos variables. Los valores positivos indican una relación positiva; los valores negativos indican una relación negativa.

Cuartiles Los percentiles 25, 50 y 75, conocidos como primer cuartil, segundo cuartil (mediana) y tercer cuartil, respectivamente. Los cuartiles se usan para dividir un conjunto de datos en cuatro partes, con cada parte conteniendo aproximadamente 25% de los datos.

Datos agrupados Datos disponibles en intervalos de clase según se resumen por una distribución de frecuencia. Los valores individuales de los datos originales no están disponibles.

Desviación estándar Medida de variabilidad calculada al tomar la raíz cuadrada positiva de la varianza.

Diagrama de caja Resumen gráfico de los datos basado en un resumen de cinco números.

Estadístico muestral Valor numérico usado como medida de resumen para una muestra (por ejemplo, la media muestral, $\bar{x}$, la varianza muestral, s^2, y la desviación estándar de la muestra, s).

Estimador puntual Los estadísticos muestrales, como $\bar{x}$, s^2 y s, cuando se utilizan para estimar el parámetro poblacional correspondiente.

Media Medida de la ubicación central calculada al resumir los valores de datos y dividir entre el número de observaciones.

Media ponderada La media obtenida al asignar a cada observación un peso que refleje su importancia.

Mediana Medida de la posición central proporcionada por el valor de en medio cuando los datos se acomodan en orden ascendente.

Moda Medida de la posición, definida como el valor que ocurre con mayor frecuencia.

Observación atípica Valor de datos inusualmente pequeño o inusualmente grande.

Parámetro poblacional Valor numérico utilizado como una medida de resumen para una población (por ejemplo, la media poblacional, μ, la varianza poblaciónal, σ^2, y la desviación estándar de la población, σ).

Percentil Valor tal que por lo menos p por ciento de las observaciones es menor o igual que este valor, y como mínimo $(100 - p)$ por ciento de las observaciones son mayores o iguales que este valor. El percentil 50 es la mediana.

Rango Medida de la variabilidad definida para ser el valor mayor menos el valor menor.

Rango intercuartílico (RIC) Medida de variabilidad definida como la diferencia entre el tercer y el primer cuartiles.

Regla empírica Se usa para calcular el porcentaje de valores de datos que deben estar dentro de una, dos y tres desviaciones estándar de la media para los datos que exhiben una distribución con forma de campana.

Resumen de cinco números Técnica de análisis exploratorio de datos que usa cinco números para resumir los datos: valor menor, primer cuartil, mediana, tercer cuartil y valor más grande.

Sesgo Medida de la forma de una distribución de datos. Los datos sesgados a la izquierda dan como resultado un sesgo negativo; una distribución de datos simétrica genera un sesgo de cero, y los datos sesgados a la derecha producen un sesgo positivo.

Teoremoa de Chebyshev Se utiliza para hacer enunciados sobre la proporción de los valores de datos que deben estar dentro de un número especificado de desviaciones estándar de la media.

valor z Valor calculado al dividir la desviación con respecto a la media $(x_i - \bar{x})$ entre la desviación estándar s. Una puntuación z se conoce como un valor estandarizado y denota el número de desviaciones estándar x_i a partir de la media.

Varianza Medida de variabilidad basada en las desviaciones cuadradas de los valores de datos con respecto a la media.

Fórmulas clave

Media muestral

$$\bar{x} = \frac{\Sigma x_i}{n} \tag{3.1}$$

Media poblacional

$$\mu = \frac{\Sigma x_i}{N} \tag{3.2}$$

Rango intercuartílico

$$\text{RIC} = Q_3 - Q_1 \tag{3.3}$$

Varianza poblacional

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N} \tag{3.4}$$

Varianza muestral

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} \tag{3.5}$$

Desviación estándar

$$\text{Desviación estándar muestral} = s = \sqrt{s^2} \tag{3.6}$$

$$\text{Desviación estándar poblacional} = \sigma = \sqrt{\sigma^2} \tag{3.7}$$

Coeficiente de variación

$$\left(\frac{\text{desviación estándar}}{\text{media}} \times = 100\right)\% \tag{3.8}$$

Valor z

$$z_i = \frac{x_i - \bar{x}}{s} \tag{3.9}$$

Covarianza muestral

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \tag{3.10}$$

Covarianza poblacional

$$\sigma_{xy} = \frac{\Sigma(x_i - \mu_x)(y_i - \mu_y)}{N} \tag{3.11}$$

Coeficiente de correlación del producto-momento de Pearson: datos muestrales

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{3.12}$$

Coeficiente de correlación del producto-momento de Pearson: datos poblacionales

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{3.13}$$

Media ponderada

$$\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i} \tag{3.15}$$

Media muestral para datos agrupados

$$\bar{x} = \frac{\Sigma f_i M_i}{n} \tag{3.16}$$

Varianza muestral para datos agrupados

$$s^2 = \frac{\Sigma f_i (M_i - \bar{x})^2}{n - 1} \tag{3.17}$$

Media poblacional para datos agrupados

$$\mu = \frac{\Sigma f_i M_i}{N} \tag{3.18}$$

Varianza poblacional para datos agrupados

$$\sigma^2 = \frac{\Sigma f_i (M_i - \mu)^2}{N} \tag{3.19}$$

Ejercicios complementarios

58. Según la encuesta del gasto anual de los consumidores, el promedio mensual del cargo a la tarjeta de crédito Visa del Bank of America fue de $1 838 (*U.S. Airways Attaché Magazine,* diciembre de 2003). Una muestra de cargos mensuales a tarjetas de crédito proporciona los datos siguientes.

WEB archivo
Visa

236	1 710	1 351	825	7 450
316	4 135	1 333	1 584	387
991	3 396	170	1 428	1 688

a) Calcule la media y la mediana.
b) Estime el primer y tercer cuartiles.
c) Calcule el rango y el rango intercuartílico.
d) Determine la varianza y la desviación estándar.
e) La medida del sesgo para estos datos es 2.12. Comente la forma de esta distribución. ¿Es la que usted esperaría? ¿Por qué?
f) ¿Los datos contienen observaciones atípicas?

59. La Oficina del Censo de Estados Unidos (U.S. Census Bureau) proporciona estadísticas sobre la vida familiar en este país, incluyendo la edad en el primer matrimonio, el estado marital actual y el tamaño de la vivienda (sitio web U.S. Census Bureau, 20 de marzo de 2006). Los datos siguientes muestran la edad en el primer matrimonio para una muestra de hombres (Men) y una de mujeres (Women).

WEB archivo
Ages

Hombres	26	23	28	25	27	30	26	35	28
	21	24	27	29	30	27	32	27	25
Mujeres	20	28	23	30	24	29	26	25	
	22	22	25	23	27	26	19		

a) Determine la edad media en la época del primer matrimonio para hombres y mujeres.
b) Calcule el primer y tercer cuartiles para ambos grupos.
c) Hace 25 años la edad media en la época del primer matrimonio era de 25 para los hombres y 22 para las mujeres. ¿Qué elementos proporciona esta información para comprender la decisión de cuándo casarse entre la gente joven en la actualidad?

60. El rendimiento del dividendo es el dividendo anual por acción que una empresa paga, dividido entre el precio por acción actual de mercado expresado como porcentaje. Una muestra de 10 empresas grandes proporciona los siguientes datos de rendimiento del dividendo (*The Wall Street Journal,* 16 de enero de 2004).

Empresa	Rendimiento %	Empresa	Rendimiento %
Altria Group	5.0	General Motors	3.7
American Express	0.8	JPMorgan Chase	3.5
Caterpillar	1.8	McDonald's	1.6
Eastman Kodak	1.9	United Technology	1.5
ExxonMobil	2.5	Wal-Mart Stores	0.7

a) ¿Cuáles son la media y la mediana de los rendimientos?
b) ¿Cuáles son la varianza y la desviación estándar?
c) ¿Qué empresa proporciona el rendimiento del dividendo más alto?
d) ¿Cuál es el valor z para McDonald's? Interprete este valor z.
e) ¿Cuál es el valor z para General Motors? Interprételo.
f) Con base en la puntuación z, ¿los datos contienen alguna observación atípica?

61. El Departamento de Educación de Estados Unidos informa que alrededor de 50% de todos los estudiantes universitarios usa un préstamo estudiantil para ayudarse a cubrir los gastos escolares (National Center for Educational Studies, enero de 2006). En la siguiente lista se observa una muestra de alumnos que se graduaron con una deuda de préstamos estudiantil. Los datos, en miles de dólares, registran montos típicos de deuda después de la graduación.

10.1 14.8 5.0 10.2 12.4 12.2 2.0 11.5 17.8 4.0

a) Para aquellos alumnos que usan un préstamo estudiantil, ¿cuál es la deuda media después de la graduación?
b) ¿Cuál es la varianza? ¿La desviación estándar?

62. Los propietarios de pequeñas empresas con frecuencia acuden a compañías de servicios externos para manejar la nómina de sus empleados. Esto se debe a que las pequeñas empresas se enfrentan a regulaciones fiscales complicadas, y las multas por errores en las declaraciones fiscales son costosas. Según el Internal Revenue Service, 26% de todas las devoluciones de impuestos de empleo de las pequeñas empresas contenía errores que dieron como resultado una multa fiscal al propietario (*The Wall Street Journal,* 30 de enero de 2006). La multa fiscal para una muestra de 20 pequeñas empresas se presenta a continuación.

WEB archivo
Penalty

820	270	450	1 010	890	700	1 350	350	300	1 200
390	730	2 040	230	640	350	420	270	370	620

a) ¿Cuál es la multa fiscal media para las devoluciones de impuestos sobre nómina llenadas de manera inapropiada?
b) ¿Cuál es la desviación estándar?
c) ¿La multa más alta de $2 040 es una observación atípica?
d) ¿Cuáles son algunas ventajas de contratar una empresa de servicios de nómina para el propietario de una pequeña empresa a efecto de que maneje los servicios de nómina de los empleados, incluidas las devoluciones de impuestos de empleo?

63. El transporte público y el automóvil son dos medios que un empleado puede usar para ir al trabajo cada día. Las muestras de los tiempos registrados para cada método se listan enseguida; los tiempos se proporcionan en minutos.

Transporte público	28	29	32	37	33	25	29	32	41	34
Automóvil	29	31	33	32	34	30	31	32	35	33

a) Calcule el tiempo de la media muestral para ir al trabajo en cada medio.
b) Calcule la desviación estándar muestral para cada método.
c) Con base en sus resultados de los incisos *a*) y *b*), ¿cuál método de transporte debe preferirse? Explique por qué.
d) Elabore un diagrama de caja para cada método. ¿Una comparación de los diagramas apoya su conclusión del inciso *c*)?

64. La Asociación Nacional de Agentes Inmobiliarios (National Association of Realtors) informó el precio medio de la vivienda en Estados Unidos y el incremento en éste durante un periodo de cinco años (*The Wall Street Journal,* 16 de enero de 2006). Utilice los precios de la muestra de viviendas listados aquí para responder las preguntas siguientes.

WEB archivo
Homes

995.9	48.8	175.0	263.5	298.0	218.9	209.0
628.3	111.0	212.9	92.6	2 325.0	958.0	212.5

a) ¿Cuál es el la mediana del precio de la muestra de vivienda?
b) En enero de 2001, la Asociación Nacional de Agentes Inmobiliarios informó que en Estados Unidos la mediana del precio de la vivienda fue de $139 300. ¿Cuál fue el incremento del porcentaje en la mediana del precio durante el periodo de cinco años?
c) ¿Cuál es el primer y el tercer cuartil para los datos de la muestra?
d) Proporcione un resumen de cinco números para los precios de la vivienda.
e) ¿Los datos contienen algunas observaciones atípicas.
f) ¿Cuál es el precio medio de la vivienda para la muestra? ¿Por qué la Asociación Nacional de Agentes Inmobiliarios prefiere usar la mediana del precio de las casas en su informe?

65. La Encuesta de la Comunidad Estadounidense de la Oficina del Censo de Estados Unidos dio a conocer el porcentaje de niños menores de 18 años que había vivido por debajo del nivel de pobreza durante los 12 meses anteriores (sitio web del U.S. Census Bureau, agosto de 2008). Las regiones de Estados Unidos noreste (NE), sureste (SE), oeste medio (MW), suroeste (SW) oeste (W) y el porcentaje de niños menores de 18 años que había vivido por debajo del nivel de pobreza se listan para cada estado.

WEB archivo
PovertyLevel

State	Region	Poverty %	State	Region	Poverty %
Alabama	SE	23.0	Montana	W	17.3
Alaska	W	15.1	Nebraska	MW	14.4
Arizona	SW	19.5	Nevada	W	13.9
Arkansas	SE	24.3	New Hampshire	NE	9.6
California	W	18.1	New Jersey	NE	11.8
Colorado	W	15.7	New Mexico	SW	25.6
Connecticut	NE	11.0	New York	NE	20.0
Delaware	NE	15.8	North Carolina	SE	20.2
Florida	SE	17.5	North Dakota	MW	13.0
Georgia	SE	20.2	Ohio	MW	18.7
Hawaii	W	11.4	Oklahoma	SW	24.3
Idaho	W	15.1	Oregon	W	16.8
Illinois	MW	17.1	Pennsylvania	NE	16.9
Indiana	MW	17.9	Rhode Island	NE	15.1
Iowa	MW	13.7	South Carolina	SE	22.1
Kansas	MW	15.6	South Dakota	MW	16.8
Kentucky	SE	22.8	Tennessee	SE	22.7
Louisiana	SE	27.8	Texas	SW	23.9
Maine	NE	17.6	Utah	W	11.9
Maryland	NE	9.7	Vermont	NE	13.2
Massachusetts	NE	12.4	Virginia	SE	12.2
Michigan	MW	18.3	Washington	W	15.4
Minnesota	MW	12.2	West Virginia	SE	25.2
Mississippi	SE	29.5	Wisconsin	MW	14.9
Missouri	MW	18.6	Wyoming	W	12.0

a) ¿Cuál es la mediana del porcentaje del nivel de pobreza (Poverty) para los 50 estados?
b) ¿Cuáles son el primer y el tercer cuartiles? ¿Cuál es su interpretación de los cuartiles?
c) Muestre un diagrama de caja para los datos. Interprete el diagrama respecto de lo que indica acerca del nivel de pobreza para los niños de Estados Unidos. ¿Algún estado (State) se considera una observación atípica? Comente.
d) Identifique los estados en el cuartil inferior. ¿Cuál es su interpretación de este grupo y qué región o regiones se representan en este cuartil?

66. La revista *Travel + Leisure* presentó su lista anual de los 500 mejores hoteles del mundo (*Travel + Leisure,* enero de 2009). La revista proporciona una calificación para cada hotel junto con una breve descripción que incluye su tamaño, servicios y costo por noche en habitación doble. Una muestra de 12 de los hoteles de más alta calificación en Estados Unidos se presenta a continuación.

WEB archivo
Travel

Hotel	Location	Rooms	Cost/Night
Boulders Resort & Spa	Phoenix, AZ	220	499
Disney's Wilderness Lodge	Orlando, FL	727	340
Four Seasons Hotel Beverly Hills	Los Ángeles, CA	285	585
Four Seasons Hotel	Boston, MA	273	495
Hay-Adams	Washington, DC	145	495
Inn on Biltmore Estate	Asheville, NC	213	279
Loews Ventana Canyon Resort	Phoenix, AZ	398	279
Mauna Lani Bay Hotel	Isla de Hawaii	343	455
Montage Laguna Beach	Laguna Beach, CA	250	595
Sofitel Water Tower	Chicago, IL	414	367
St. Regis Monarch Beach	Dana Point, CA	400	675
The Broadmoor	Colorado Springs, CO	700	420

a) ¿Cuál es el número medio de habitaciones (Rooms)?
b) ¿Cuál es el costo medio por noche (Cost/Night) para una habitación doble?

c) Elabore un diagrama de dispersión con el número de habitaciones en el eje horizontal y el costo por noche en el eje vertical. ¿Parece haber una relación entre el número de habitaciones y el costo por noche? Comente.

d) ¿Cuál es el coeficiente de correlación muestral? ¿Qué le dice sobre la relación entre el número de habitaciones y el costo por noche para una habitación doble? ¿Esto le parece razonable? Comente.

67. Morningstar da seguimiento al rendimiento de un gran número de empresas y publica una evaluación de cada una. Junto con una variedad de datos financieros, Morningstar incluye una estimación del valor justo (Fair Value) para el precio que debe pagarse por una acción de las acciones comunes de la empresa. Los datos para 30 empresas se encuentran en el archivo llamado *FairValue*. Los datos incluyen la estimación del precio justo por acción de las acciones comunes, el precio por acción más reciente y la utilidad por acción para la empresa (*Morningstar Stocks500,* 2008).

WEB archivo
FairValue

a) Elabore un diagrama de dispersión para los datos del precio justo y del precio por acción, con este último sobre el eje horizontal. ¿Cuál es el coeficiente de correlación muestral y qué puede decir acerca de la relación entre las variables?

b) Desarrolle un diagrama de dispersión para los datos del precio justo y del precio por acción con este último sobre el eje horizontal. ¿Cuál es el coeficiente de correlación muestral y qué puede decir acerca de la relación entre las variables?

68. ¿El registro de un equipo de béisbol de ligas mayores durante el entrenamiento de primavera indica cómo jugará durante la temporada regular? En los últimos seis años el coeficiente de correlación entre el porcentaje de victorias de un equipo en el entrenamiento de primavera y su porcentaje de triunfos en la temporada regular es de 0.18 (*The Wall Street Journal,* 30 de marzo de 2009). Enseguida se listan los porcentajes de victorias para los 14 equipos de la Liga Americana durante la temporada 2008.

WEB archivo
SpringTraining

Team	Spring Training	Regular Season	Team	Spring Training	Regular Season
Baltimore Orioles	0.407	0.422	Minnesota Twins	0.500	0.540
Boston Red Sox	0.429	0.586	New York Yankees	0.577	0.549
Chicago White Sox	0.417	0.546	Oakland A's	0.692	0.466
Cleveland Indians	0.569	0.500	Seattle Mariners	0.500	0.377
Detroit Tigers	0.569	0.457	Tampa Bay Rays	0.731	0.599
Kansas City Royals	0.533	0.463	Texas Rangers	0.643	0.488
Los Ángeles Angels	0.724	0.617	Toronto Blue Jays	0.448	0.531

a) ¿Cuál es el coeficiente de correlación entre los porcentajes de victoria del entrenamiento de primavera (Spring Training) y de la temporada regular (Regular Season)?

b) ¿Qué indica su conclusión acerca del registro de un equipo durante el entrenamiento de primavera sobre cómo jugará durante la temporada regular? ¿Cuáles son algunas razones para que esto ocurra? Comente.

69. Los días para el vencimiento de una muestra de cinco fondos del mercado de dinero se listan enseguida junto con los montos en dólares de las cantidades invertidas en los fondos. Utilice la media ponderada para determinar el número medio de días para el vencimiento de los dólares invertidos en estos cinco fondos del mercado de dinero.

Días para el vencimiento	Valor monetario (millones)
20	20
12	30
7	10
5	15
6	10

70. La velocidad de los automóviles que viajan por una autopista con un límite de velocidad establecido de 55 millas por hora se comprueba mediante un sistema de radar de la policía estatal. A continuación se presenta una distribución de frecuencia de las velocidades.

Velocidad (millas por hora)	Frecuencia
45–49	10
50–54	40
55–59	150
60–64	175
65–69	75
70–74	15
75–79	10
Total	475

a) ¿Cuál es la velocidad media de los automóviles que viajan en esta autopista?
b) Calcule la varianza y la desviación estándar.

Caso a resolver 1 Pelican Stores

Pelican Stores, una división de National Clothing, es una cadena de tiendas de ropa para dama que opera en todo Estados Unidos. La cadena lanzó recientemente una promoción en la que se enviaron cupones de descuento a los clientes de otras tiendas de National Clothing. Los datos recabados de una muestra de 100 transacciones de tarjetas de crédito en Pelican Stores durante un día, mientras la promoción estuvo vigente, se encuentran en el archivo llamado *PelicanStores*. La tabla 3.12 presenta una parte del conjunto de datos. El método de pago proprietary card se refiere a los cargos realizados usando una tarjeta de National Clothing. A los clientes (Customer)

WEB archivo
PelicanStores

TABLA 3.12 Muestra de 100 compras con tarjeta de crédito en Pelican Stores

Customer	Type of Customer	Items	Net Sales	Method of Payment	Gender	Marital Status	Age
1	Regular	1	39.50	Discover	Male	Married	32
2	Promotional	1	102.40	Proprietary card	Female	Married	36
3	Regular	1	22.50	Proprietary card	Female	Married	32
4	Promotional	5	100.40	Proprietary card	Female	Married	28
5	Regular	2	54.00	MasterCard	Female	Married	34
6	Regular	1	44.50	MasterCard	Female	Married	44
7	Promotional	2	78.00	Proprietary card	Female	Married	30
8	Regular	1	22.50	Visa	Female	Married	40
9	Promotional	2	56.52	Proprietary card	Female	Married	46
10	Regular	1	44.50	Proprietary card	Female	Married	36
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
96	Regular	1	39.50	MasterCard	Female	Married	44
97	Promotional	9	253.00	Proprietary card	Female	Married	30
98	Promotional	10	287.59	Proprietary card	Female	Married	52
99	Promotional	2	47.60	Proprietary card	Female	Married	30
100	Promotional	1	28.44	Proprietary card	Female	Married	44

que efectuaron una compra utilizando un cupón de descuento se les llama *clientes de promoción* y a los que compraron, pero no usaron un cupón de descuento se les denomina *clientes regulares*. Dado que los cupones promocionales no se enviaron a los compradores regulares de Pelican Stores, la gerencia considera las ventas realizadas a personas que presentaron los cupones de promoción como ventas que de lo contrario no se hubieran hecho. Por supuesto, Pelican también espera que los clientes de promoción sigan comprando en sus tiendas.

La mayoría de las variables mostradas en la tabla 3.12 se explican por sí mismas, pero dos requieren una aclaración.

Artículos (Items)	Número total de artículos adquiridos.
Ventas netas (Net Sales)	Monto total ($) cargado a la tarjeta de crédito.

A la gerencia de Pelican le gustaría usar estos datos muestrales para enterarse de su base de clientes y evaluar la promoción de los cupones de descuento.

Informe gerencial

Utilice los métodos tabular y gráfico de la estadística descriptiva para resumir los datos y comente sus hallazgos. Como mínimo, su informe debe incluir lo siguiente:

1. Estadísticos descriptivos sobre las ventas netas y sobre las ventas netas por varias clasificaciones de clientes.
2. Estadísticos descriptivos concernientes a la relación entre la edad (Age) y las ventas netas.

Caso a resolver 2 Industria del cine

La industria estadounidense del cine es un negocio competitivo. Más de 50 estudios producen un total de 300 a 400 películas nuevas cada año (Motion Pictures), y el éxito financiero de cada una varía considerablemente. Las ventas brutas del fin de semana de estreno (Opening Gross Sales), las ventas brutas totales (Total Gross Sales), el número de cines (Number of Theaters) donde la película se exhibe y el número de semanas que ésta estuvo entre las primeras 60 (Weeks in Top 60) en ventas brutas son variables comunes utilizadas para medir el éxito de un título. Los datos recabados de una muestra de 100 filmes producidos en 2005 se incluyen en el archivo llamado *Movies*. La tabla 3.13 muestra los datos de las primeras 10 películas de este archivo.

WEB archivo
Movies

TABLA 3.13 Datos del desempeño de 10 películas

Motion Picture	Opening Gross Sales ($millions)	Total Gross Sales ($millions)	Number of Theaters	Weeks in Top 60
Coach Carter	29.17	67.25	2574	16
Ladies in Lavender	0.15	6.65	119	22
Batman Begins	48.75	205.28	3858	18
Unleashed	10.90	24.47	1962	8
Pretty Persuasion	0.06	0.23	24	4
Fever Pitch	12.40	42.01	3275	14
Harry Potter and the Goblet of Fire	102.69	287.18	3858	13
Monster-in-Law	23.11	82.89	3424	16
White Noise	24.11	55.85	2279	7
Mr. and Mrs. Smith	50.34	186.22	3451	21

Informe gerencial

Utilice los métodos numéricos de la estadística descriptiva presentados en este capítulo para saber cómo estas variables contribuyen al éxito de una película. Incluya lo siguiente en su informe.

1. Los estadísticos descriptivos de cada una de las cuatro variables junto con un análisis de lo que cada estadístico descriptivo indica sobre la industria del cine.
2. ¿Qué películas, si las hay, deben considerarse observaciones atípicas de alto desempeño? Explique por qué.
3. La estadística descriptiva muestra la relación entre las ventas brutas totales y cada una de las otras variables. Comente.

Caso a resolver 3 Escuelas de negocios de Asia-Pacífico

WEB archivo
Asian

La consecución de un título de posgrado en los negocios es ahora internacional. Una encuesta muestra que cada vez más asiáticos eligen la ruta de la maestría en administración de empresas (MBA) para lograr el éxito corporativo. Como resultado, el número de solicitantes para los cursos de MBA en escuelas de Asia-Pacífico sigue aumentando.

En toda la región, miles de asiáticos muestran una creciente voluntad de dejar de lado temporalmente su carrera y pasar dos años en la búsqueda de un título de negocios teórico. Los cursos en estas escuelas son notoriamente difíciles e incluyen economía, banca, marketing, ciencias del comportamiento, relaciones laborales, toma de decisiones, pensamiento estratégico, derecho de los negocios, y mucho más. El conjunto de datos de la tabla 3.14 muestra algunas características de las principales escuelas de negocios de Asia-Pacífico.

Informe gerencial

Use los métodos de la estadística descriptiva para resumir los datos de la tabla 3.14. Comente sus hallazgos.

1. Incluya un resumen para cada variable del conjunto de datos. Comente e interprete con base en los máximos y los mínimos, así como los medios y las proporciones apropiados. ¿Qué elementos de comprensión nuevos proporcionan estos estadísticos descriptivos respecto de las escuelas de negocios de Asia-Pacífico?
2. Resuma los datos para comparar lo siguiente:
   a) Cualquier diferencia entre los costos de clases locales y en el extranjero.
   b) Alguna diferencia entre los sueldos iniciales medios para las escuelas que requieren y no requieren experiencia laboral.
   c) Cualquier diferencia entre los sueldos iniciales para escuelas que requieren y no requieren pruebas de inglés.
3. ¿Los sueldos iniciales parecen estar relacionados con las clases?
4. Presente resúmenes gráficos y numéricos adicionales que sean benéficos para comunicar los datos de la tabla 3.14 a otras personas.

Caso a resolver 4 Transacciones del sitio web de Heavenly Chocolates

Heavenly Chocolates fabrica y vende productos de chocolate de calidad en su planta y tienda minorista ubicada en Saratoga Springs, Nueva York. Hace dos años la empresa desarrolló un sitio web y comenzó a vender sus productos por Internet. Las ventas electrónicas han excedido las expectativas de la empresa y la gerencia ahora está considerando estrategias para incrementarlas aún más. Para saber más sobre los clientes del sitio web, se seleccionó una muestra de 50 transacciones de Heavenly Chocolate de las ventas del mes anterior. Datos que ilustran

TABLA 3.14 Datos de 25 escuelas de negocios de Asia-Pacífico

Escuela de negocios	Inscripción de tiempo completo	Estudiantes por facultad	Clases locales ($)	Clases en el extranjero ($)	Edad	Extranjero%	GMAT	Examen de inglés	Experiencia de trabajo	Sueldo inicial ($)
Melbourne Business School	200	5	24 420	29 600	28	47	Sí	No	Sí	71 400
University of New South Wales (Sydney)	228	4	19 993	32 582	29	28	Sí	No	Sí	65 200
Indian Institute of Management (Ahmedabad)	392	5	4 300	4 300	22	0	No	No	No	7 100
Chinese University of Hong Kong	90	5	11 140	11 140	29	10	Sí	No	No	31 000
International University of Japan (Niigata)	126	4	33 060	33 060	28	60	Sí	Sí	No	87 000
Asian Institute of Management (Manila)	389	5	7 562	9 000	25	50	Sí	No	Sí	22 800
Indian Institute of Management (Bangalore)	380	5	3 935	16 000	23	1	Sí	No	No	7 500
National University of Singapore	147	6	6 146	7 170	29	51	Sí	Sí	Sí	43 300
Indian Institute of Management (Calcutta)	463	8	2 880	16 000	23	0	No	No	No	7 400
Australian National University (Canberra)	42	2	20 300	20 300	30	80	Sí	Sí	Sí	46 600
Nanyang Technological University (Singapore)	50	5	8 500	8 500	32	20	Sí	No	Sí	49 300
University of Queensland (Brisbane)	138	17	16 000	22 800	32	26	No	No	Sí	49 600
Hong Kong University of Science and Technology	60	2	11 513	11 513	26	37	Sí	No	Sí	34 000
Macquarie Graduate School of Management (Sydney)	12	8	17 172	19 778	34	27	No	No	Sí	60 100
Chulalongkorn University (Bangkok)	200	7	17 355	17 355	25	6	Sí	No	Sí	17 600
Monash Mt. Eliza Business School (Melbourne)	350	13	16 200	22 500	30	30	Sí	Sí	Sí	52 500
Asian Institute of Management (Bangkok)	300	10	18 200	18 200	29	90	No	Sí	Sí	25 000
University of Adelaide	20	19	16 426	23 100	30	10	No	No	Sí	66 000
Massey University (Palmerston North, New Zealand)	30	15	13 106	21 625	37	35	No	Sí	Sí	41 400
Royal Melbourne Institute of Technology Business Graduate School	30	7	13 880	17 765	32	30	No	Sí	Sí	48 900
Jamnalal Bajaj Institute of Management Studies (Mumbai)	240	9	1 000	1 000	24	0	No	No	Sí	7 000
Curtin Institute of Technology (Perth)	98	15	9 475	19 097	29	43	Sí	No	Sí	55 000
Lahore University of Management Sciences	70	14	11 250	26 300	23	2.5	No	No	No	7 500
University Sains Malaysia (Penang)	30	5	2 260	2 260	32	15	No	Sí	Sí	16 000
De La Salle University (Manila)	44	17	3 300	3 600	28	3.5	Sí	No	Sí	13 100

TABLA 3.15 Muestra de 50 transacciones del sitio web de Heavenly Chocolates

WEB archivo
Shoppers

Customer	Day	Browser	Time (min)	Pages Viewed	Amount Spent ($)
1	Mon	Internet Explorer	12.0	4	54.52
2	Wed	Other	19.5	6	94.90
3	Mon	Internet Explorer	8.5	4	26.68
4	Tue	Firefox	11.4	2	44.73
5	Wed	Internet Explorer	11.3	4	66.27
6	Sat	Firefox	10.5	6	67.80
7	Sun	Internet Explorer	11.4	2	36.04
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
48	Fri	Internet Explorer	9.7	5	103.15
49	Mon	Other	7.3	6	52.15
50	Fri	Internet Explorer	13.4	3	98.75

el día de la semana (Day) en que se realizó cada transacción, el tipo de explorador (Browser) usado por el cliente, el tiempo invertido en el sitio web (Time), el número de páginas visitadas (Pages Viewed,) y la cantidad gastada (Amount Spent) por cada uno de los 50 clientes están contenidos en el archivo llamado *Shoppers*. Una porción de los datos se muestra en la tabla 3.15.

A Heavenly Chocolates le gustaría usar los datos de la muestra para determinar si los compradores en línea que pasaron más tiempo y vieron más páginas también gastaron más dinero durante su visita al sitio web. A la empresa también le gustaría investigar el efecto que el día de la semana y el tipo de explorador tienen sobre las ventas.

Informe gerencial

Use los métodos de la estadística descriptiva para saber más acerca de los clientes que visitan el sitio web de Heavenly Chocolates. Incluya lo siguiente en su informe.

1. Resúmenes gráficos y numéricos para el tiempo que el comprador pasa en el sitio web, el número de páginas visitadas y la cantidad media gastada por transacción. Comente los datos que obtuvo acerca de los compradores en línea de Heavenly Chocolates a partir de estos resúmenes numéricos.
2. Resuma la frecuencia, los dólares totales y la cantidad media gastados por transacción para cada día de la semana. ¿Qué observaciones puede usted hacer sobre el negocio de Heavenly Chocolates con base en el día de la semana? Comente.
3. Resuma la frecuencia, los dólares totales y la cantidad media gastados por transacción para cada tipo de navegador. ¿Qué observaciones puede hacer acerca del negocio con base en el tipo de explorador? Comente.
4. Elabore un diagrama de dispersión y calcule el coeficiente de correlación muestral para explorar la relación entre el tiempo invertido en el sitio web y la cantidad gastada. Utilice el eje horizontal para el tiempo invertido. Comente.
5. Prepare un diagrama de dispersión y calcule el coeficiente de correlación muestral para explorar la relación entre el número de páginas visitadas y la cantidad gastada. Utilice el eje horizontal para el número de páginas web consultadas. Comente.
6. Elabore un diagrama de dispersión y calcule el coeficiente de correlación muestral para explorar la relación entre el tiempo pasado en el sitio web y el número de páginas visitadas. Use el eje horizontal para representar el número de páginas visitadas. Comente.

Apéndice 3.1 Estadística descriptiva usando Minitab

En este apéndice se describe cómo se usa Minitab para calcular una variedad de estadísticos descriptivos y desplegar diagramas de caja. Luego se explica su uso para obtener las medidas de covarianza y de correlación para dos variables.

Estadística descriptiva

La tabla 3.1 proporcionó los sueldos iniciales de 12 licenciados en administración de empresas recién graduados de la escuela de negocios. Estos datos están disponibles en el archivo Start-Salary. La figura 3.12 muestra la estadística descriptiva de los datos de los sueldos iniciales obtenidos con Minitab. Las definiciones de los encabezados se muestran en seguida.

N	Número de valores de datos
N*	Número de valores de datos faltantes
Mean	Media
SE Mean	Error estándar de la media
StDev	Desviación estándar
Minimum	Valor de datos mínimo
Q1	Primer cuartil
Median	Mediana
Q3	Tercer cuartil
Maximum	Valor de datos máximo

La etiqueta SE Mean se refiere al *error estándar de la media.* Se calcula dividiendo la desviación estándar entre la raíz cuadrada de *N*. La interpretación y el uso de esta medida se estudian en el capítulo 7, cuando se presentan los temas de muestreo y distribuciones del muestreo.

Aunque las medidas numéricas del rango, el rango intercuartílico, la varianza y el coeficiente de variación no aparecen en el resultado de Minitab, estos valores se calculan fácilmente a partir de los resultados de la figura 3.12 como sigue.

$$\text{Rango} = \text{máximo} - \text{mínimo}$$
$$\text{RIC} = Q_3 - Q_1$$
$$\text{Varianza} = (\text{StDev})^2$$
$$\text{Coeficiente de variación} = (\text{StDev/Mean}) \times 100$$

Por último, observe que los cuartiles de Minitab $Q_1 = 3\,457.5$ y $Q_3 = 3\,625$ son ligeramente diferentes de los cuartiles $Q_1 = 3\,465$ y $Q_3 = 3\,600$ calculados en la sección 3.1. Las distintas convenciones* que se usaron para identificar los cuartiles explican esta variación. Por consiguiente, los valores Q_1 y Q_3 proporcionados por una convención tal vez no sean idénticos a los derivados de otra convención. No obstante, cualesquiera diferencias tienden a ser insignificantes

FIGURA 3.12 Estadísticos descriptivos proporcionados por Minitab

N	N*	Mean	SE Mean	StDev
12	0	3540.0	47.8	165.7
Minimum	**Q1**	**Median**	**Q3**	**Maximum**
3310.0	3457.5	3505.0	3625.0	3925.0

* Con las n observaciones arregladas en orden ascendente (del valor menor al valor mayor), Minitab usa las posiciones dadas por $(n + 1)/4$ y $3(n + 1)/4$ para ubicar a Q_1 y Q_3, respectivamente. Cuando una posición es fraccional, Minitab hace una interpolación entre los dos valores de datos ordenados adyacentes para determinar el cuartil correspondiente.

y los resultados proporcionados no deben inducir al usuario a errores al hacer las interpretaciones usuales asociadas con los cuartiles.

Enseguida se explicará cómo se generan los estadísticos de la figura 3.12. Los datos de los sueldos iniciales están en la columna C2 de la hoja de trabajo de StartSalary. Los pasos siguientes guían para generar los estadísticos descriptivos.

WEB archivo
StartSalary

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Basic Statistics**.
Paso 3. Elija **Display Descriptive Statistics**.
Paso 4. Cuando el cuadro de diálogo Display Descriptive Statistics aparezca:
Introduzca C2 en el cuadro **Variables**.
Haga clic en **OK**.

Diagrama de caja

Los pasos siguientes usan el archivo StartSalary para generar el diagrama de caja sobre los datos de los sueldos iniciales.

Paso 1. Seleccione el menú **Graph**.
Paso 2. Elija **Boxplot**.
Paso 3. Seleccione **Simple** y haga clic en **OK**.
Paso 4. Cuando aparezca el cuadro de diálogo Boxplot-One Y, Simple:
Introduzca C2 en el cuadro **Graph variables**.
Haga clic en **OK**.

Covarianza y correlación

WEB archivo
Stereo

La tabla 3.6 proporciona el número de comerciales y el volumen de ventas de una tienda de estéreos y equipos de sonido. Estos datos están disponibles en el archivo *Stereo*; el número de comerciales se encuentra en la columna C2 y el volumen de ventas en la columna C3. Los pasos siguientes muestran cómo se usa Minitab para calcular la covarianza de las dos variables.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Basic Statistics**.
Paso 3. Elija **Covariance**.
Paso 4. Cuando el cuadro de diálogo Covariance aparezca:
Introduzca C2 C3 en el cuadro **Variables**.
Haga clic en **OK**.

Para obtener el coeficiente de correlación del número de comerciales y el volumen de ventas sólo es necesario realizar un cambio en el procedimiento anterior. En el paso 3 elija la opción **Correlation**.

Apéndice 3.2 Estadística descriptiva usando Excel

Excel se puede utilizar para generar los estadísticos descriptivos de este capítulo. En este apéndice se explica cómo se usa para obtener varias medidas de posición y variabilidad para una sola variable, así como la covarianza y el coeficiente de correlación como medidas de asociación entre dos variables.

Uso de las funciones de Excel

Excel proporciona funciones para calcular la media, la mediana, la moda, la varianza muestral y la desviación estándar de la muestra. El uso de estas funciones se explica mediante el cálculo

FIGURA 3.13 Uso de las funciones de Excel para calcular la media, mediana, moda y desviación estándar

	A	B	C	D	E	F
1	**Graduate**	**Starting Salary**		**Mean**	=AVERAGE(B2:B13)	
2	1	3450		**Median**	=MEDIAN(B2:B13)	
3	2	3550		**Mode**	=MODE(B2:B13)	
4	3	3650		**Variance**	=VAR(B2:B13)	
5	4	3480		**Standard Deviation**	=STDEV(B2:B13)	
6	5	3355				
7	6	3310				
8	7	3490				
9	8	3730				
10	9	3540				
11	10	3925				
12	11	3520				
13	12	3480				
14						

	A	B	C	D	E	F
1	**Graduate**	**Starting Salary**		**Mean**	3540	
2	1	3450		**Median**	3505	
3	2	3550		**Mode**	3480	
4	3	3650		**Variance**	27440.91	
5	4	3480		**Standard Deviation**	165.65	
6	5	3355				
7	6	3310				
8	7	3490				
9	8	3730				
10	9	3540				
11	10	3925				
12	11	3520				
13	12	3480				
14						

WEB archivo
StartSalary

de la media, la mediana, la varianza muestral y la desviación estándar muestral de los datos de los sueldos iniciales de la tabla 3.1. Vuelva a observar la figura 3.13 mientras se describen los pasos involucrados. Los datos se introducen en la columna B.

La función AVERAGE de Excel se usa para calcular la media al introducir la fórmula siguiente en la celda E1.

=AVERAGE(B2:B13)

De modo parecido, las fórmulas =MEDIAN(B2:B13), =MODE(B2:B13), =VAR(B2:B13) y =STDEV(B2:B13) se introducen en las celdas E2:E5, respectivamente, para calcular la mediana, la moda, la varianza y la desviación estándar. La hoja de trabajo en segundo plano muestra que los valores estimados con las funciones de Excel son los mismos que aquellos calculados antes en el capítulo.

Excel proporciona también funciones para calcular la covarianza y el coeficiente de correlación. Debe tener cuidado cuando las use debido a que la función de covarianza trata los datos como una población y la función de correlación los trata como una muestra. Por tanto, el resultado obtenido usando la función de covarianza de Excel debe ajustarse para proporcionar la covarianza muestral. Enseguida se explica cómo usar estas funciones para calcular la covarianza muestral y el coeficiente de correlación muestral para los datos de la tienda de estéreos y equipos de sonido de la tabla 3.7. Vuelva a observar la figura 3.14 mientras se presentan los pasos involucrados.

WEB archivo
Stereo

La función de covarianza de Excel, COVAR, sirve para calcular la covarianza poblacional al introducir la fórmula siguiente en la celda F1.

=COVAR(B2:B11,C2:C11)

De manera similar, la fórmula =CORREL(B2:B11,C2:C11) se introduce en la celda F2 para calcular el coeficiente de correlación muestral. La hoja de trabajo en segundo plano muestra los

FIGURA 3.14 Uso de las funciones de Excel para calcular la covarianza y la correlación

	A	B	C	D	E	F	G
1	Week	Commercials	Sales		Population Covariance	=COVAR(B2:B11,C2:C11)	
2	1	2	50		Sample Correlation	=CORREL(B2:B11,C2:C11)	
3	2	5	57				
4	3	1	41				
5	4	3	54				
6	5	4	54				
7	6	1	38				
8	7	5	63				
9	8	3	48				
10	9	4	59				
11	10	2	46				
12							

	A	B	C	D	E	F	G
1	Week	Commercials	Sales		Population Covariance	9.90	
2	1	2	50		Sample Correlation	0.93	
3	2	5	57				
4	3	1	41				
5	4	3	54				
6	5	4	54				
7	6	1	38				
8	7	5	63				
9	8	3	48				
10	9	4	59				
11	10	2	46				
12							

valores estimados usando las funciones de Excel. Observe que el valor del coeficiente de correlación muestral (0.93) es el mismo que se calculó usando la ecuación (3.12). Sin embargo, el resultado proporcionado por la función COVAR de Excel, 9.9, se obtuvo al tratar los datos como una población. Por tanto, este resultado debe ajustarse para obtener la covarianza muestral. El ajuste es muy sencillo. Primero note que la fórmula de la covarianza poblacional, la ecuación (3.11), requiere que se divida entre el número total de observaciones en el conjunto de datos, pero la fórmula para la covarianza muestral, la ecuación (3.10), requiere que se divida entre el número total de observaciones menos 1. Por tanto, para usar el resultado de Excel de 9.9 a efecto de calcular la covarianza muestral, sencillamente se multiplica 9.9 por $n/(n-1)$. Como $n = 10$, se obtiene

$$s_{xy} = \left(\frac{10}{9}\right)9.9 = 11$$

Por tanto, la covarianza muestral de los datos de la tienda de estéreos y equipos de sonido es 11.

Uso de la herramienta Descriptive Statistics de Excel

WEB archivo
StartSalary

Como ya se demostró, Excel proporciona funciones estadísticas para calcular los estadísticos descriptivos de un conjunto de datos. Estas funciones se usan para determinar un estadístico a la vez (por ejemplo, la media, la varianza, etc.). Excel también cuenta con una variedad de herramientas para análisis de datos. Una de estas herramientas, llamada Descriptive Statistics, permite al usuario calcular una variedad de estadísticos descriptivos en una sola operación. Enseguida se explica cómo usar esta herramienta para calcular los estadísticos descriptivos de los datos de los sueldos iniciales de la tabla 3.1.

Paso 1. Haga clic en la ficha **Data** de la cinta de opciones.
Paso 2. En el grupo **Analysis** haga clic en **Data Analysis**.
Paso 3. Cuando el cuadro de diálogo Data Analysis aparezca:
Elija **Descriptive Statistics**.
Haga clic en **OK**.

FIGURA 3.15 Resultado de la herramienta Descriptive Statistics de Excel

	A	B	C	D	E	F
1	**Graduate**	**Starting Salary**		***Starting Salary***		
2	1	3450				
3	2	3550		**Mean**	3540	
4	3	3650		Standard Error	47.82	
5	4	3480		**Median**	3505	
6	5	3355		**Mode**	3480	
7	6	3310		**Standard Deviation**	165.65	
8	7	3490		**Sample Variance**	27440.91	
9	8	3730		Kurtosis	1.7189	
10	9	3540		**Skewness**	1.0911	
11	10	3925		**Range**	615	
12	11	3520		**Minimum**	3310	
13	12	3480		**Maximum**	3925	
14				**Sum**	42480	
15				**Count**	12	
16						

Paso 4. Cuando el cuadro de diálogo Descriptive Statistics aparezca:
Introduzca B1:B13 en el cuadro **Input Range**.
Seleccione **Grouped By Columns**.
Elija **Labels in First Row**.
Seleccione **Output Range**.
Introduzca D1 en el cuadro **Output Range** (para identificar la esquina superior izquierda de la sección de la hoja de trabajo donde aparecerá el estadístico descriptivo).
Seleccione **Summary statistics**.
Haga clic en **OK**.

Las celdas D1:E15 de la figura 3.15 muestran los estadísticos descriptivos proporcionados por Excel. Las entradas en negritas son los estadísticos descriptivos que se cubren en este capítulo. Los que no están en negritas se cubren después en el libro o se estudian en libros más avanzados.

Apéndice 3.3 Estadística descriptiva usando StatTools

En este apéndice se describe cómo se utiliza StatTools para calcular una variedad de estadísticos descriptivos y desplegar diagramas de caja. Luego se muestra cómo se usa StatTools para obtener las medidas de covarianza y correlación para dos variables.

Estadística descriptiva

WEB archivo
StartSalary

Los datos de los sueldos iniciales de la tabla 3.1 son útiles para ilustrar. Primero se verá el uso de Data Set Manager con el fin de crear un conjunto de datos StatTools para estos datos usando el procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes generarán una variedad de estadísticos descriptivos.

Paso 1. Haga clic en la ficha **StatTools** de la cinta de opciones.
Paso 2. En **Analyses Group,** haga clic en **Summary Statistics**.
Paso 3. Elija la opción **One-Variable Summary**.

Paso 4. Cuando el cuadro de diálogo One-Variable Summary Statistics se abra:
En la sección **Variables** seleccione **Starting Salary**.
Haga clic en **OK**.

Aparecerá una variedad de estadísticos descriptivos.

Diagramas de caja

Los datos de los sueldos iniciales de la tabla 3.1 se usan para ilustrar. Primero se utiliza Data Set Manager con el fin crear un conjunto de datos StatTools para estos datos mediante el procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes crearán un diagrama de caja para estos datos.

WEB archivo
StartSalary

Paso 1. Haga clic en la ficha **StatTools** en la cinta de opciones.
Paso 2. En **Analyses Group**, haga clic en **Summary Graphs**.
Paso 3. Elija la opción **Box-Whisker Plot**.
Paso 4. Cuando el cuadro de diálogo StatTools–Box–Whisker Plot aparezca:
En la sección **Variables** seleccione **Starting Salary**.
Haga clic en **OK**.

El símbolo □ se usa para identificar una observación atípica, y x para identificar la media.

Covarianza y correlación

Utilizamos los datos de la tienda de estéreos y equipos de sonido de la tabla 3.7 para demostrar el cálculo de la covarianza muestral y el coeficiente de correlación muestral. Primero se usa Data Set Manager con el fin crear un conjunto de datos StatTools para estos datos por medio del procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes proporcionarán la covarianza muestral y el coeficiente de correlación muestral.

WEB archivo
Stereo

Paso 1. Haga clic en la ficha **StatTools** en la cinta de opciones.
Paso 2. En **Analyses Group,** haga clic en **Summary Statistics**.
Paso 3. Elija la opción **Correlation and Covariance**.
Paso 4. Cuando el cuadro de diálogo StatTools–Correlation and Covariance aparezca:
En la sección **Variables**:
Elija **No. of Commercials**.
Seleccione **Sales Volume**.
En la sección **Tables to Create**:
Seleccione **Table of Correlations**.
Elija **Table of Covariances**.
En la sección **Table Structure**, seleccione **Symmetric**.
Haga clic en **OK**.

Una tabla muestra el coeficiente de correlación y la covarianza aparecerá.

CAPÍTULO 4

Introducción a la probabilidad

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: OCEANWIDE SEAFOOD

4.1 EXPERIMENTOS, REGLAS DE CONTEO Y ASIGNACIÓN DE PROBABILIDADES
Reglas de conteo, combinaciones y permutaciones
Asignación de probabilidades
Probabilidades para el proyecto de KP&L

4.2 EVENTOS Y SUS PROBABILIDADES

4.3 ALGUNAS RELACIONES BÁSICAS DE PROBABILIDAD
Complemento de un evento
Ley de la adición

4.4 PROBABILIDAD CONDICIONAL
Eventos independientes
Ley de la multiplicación

4.5 TEOREMA DE BAYES
Método tabular

ESTADÍSTICA *en* LA PRÁCTICA

OCEANWIDE SEAFOOD*

SPRINGBORO, OHIO

Oceanwide Seafood es el principal proveedor de pescado y mariscos de calidad del suroeste de Ohio. La empresa vende más de 90 variedades de mariscos frescos y congelados de todo el mundo y prepara cortes especiales según las especificaciones de sus clientes, que incluyen los principales restaurantes y minoristas de alimentos en Ohio, Kentucky e Indiana. La empresa, fundada en 2005, ha logrado tener éxito al proporcionar un excelente servicio al cliente y mariscos de calidad excepcional.

La probabilidad y la información estadística se utilizan para la toma de decisiones operativas y de marketing. Por ejemplo, para seguir la pista del crecimiento de la empresa y establecer los futuros niveles meta de ventas, se utiliza una serie de tiempo que muestra las ventas mensuales. Estadísticos como el tamaño medio de los pedidos del cliente y el número medio de días que tarda en hacer los pagos ayudan a identificar a los mejores clientes de la empresa, así como a proporcionar puntos de referencia para el manejo de los problemas de las cuentas por cobrar. Además, los datos sobre los niveles mensuales de inventario se usan en el análisis de la utilidad de operación y las tendencias en las ventas de productos.

El análisis de probabilidad ha ayudado a Oceanwide a determinar precios razonables y rentables para sus productos. Por ejemplo, cuando recibe un pescado entero fresco de uno de sus proveedores, éste se procesa y corta para cumplir con los pedidos de cada cliente. Un atún entero fresco de 100 libras conservado en hielo podría costarle a Oceanwide \$500. A simple vista, el costo para la empresa parece ser \$500/100 = \$5 por libra. Sin embargo, debido a la pérdida en la operación de procesamiento y corte, un atún entero de 100 libras no proporcionará 100 libras de producto terminado. Si la operación de procesamiento y corte produce 75% del atún entero, el número de libras de producto terminado disponible para vender a los clientes sería 0.75(100) = 75 libras, no 100 libras. En este caso, el costo real del atún para la empresa sería \$500/75 = \$6.67 por libra. Por tanto, Oceanwide necesitaría determinar un costo de \$6.67 por libra para que el precio que fija a sus clientes sea rentable.



El atún de aleta azul se envía a Oceanwide Seafood casi todos los días. © Gregor Kervina, 2009/Fotografía usada con autorización de Shutterstock.com.

Para ayudar a determinar el porcentaje del rendimiento probable del procesamiento y corte de atún entero, se recabaron datos sobre el rendimiento de una muestra del producto entero. La variable y denota el porcentaje de rendimiento del producto. Utilizando los datos, Oceanwide pudo determinar que 5% de las veces dicho rendimiento fue por lo menos de 90%. En la notación de probabilidad condicional, ésta se escribe $P(Y \geq 90\% \mid \text{atún}) = 0.05$; es decir, la probabilidad de que el rendimiento sea por lo menos de 90%, teniendo en cuenta que el pescado es un atún, es 0.05. Si Oceanwide estableció el precio de venta del producto sobre la base de un rendimiento de 90%, la empresa obtendrá un rendimiento menor al esperado 95% de las veces. Como resultado, estaría subestimando su costo por libra y también el precio para sus clientes. Otra información de probabilidad condicional para otros porcentajes de rendimiento ayudaron a la gerencia a seleccionar un rendimiento de 70% como base para determinar el costo del atún y el precio que fija para sus clientes. Probabilidades condicionales parecidas sobre otros productos del mar permitieron establecer porcentajes para fijar precios por rendimiento para cada tipo de producto del mar. En este capítulo usted aprenderá a calcular e interpretar las probabilidades condicionales y otras más que son útiles en el proceso de toma de decisiones.

* Los autores agradecen a Dale Hartlage, presidente de Oceanwide Seafood Company, por proporcionar este artículo para la sección Estadística en la práctica.

Los gerentes o administradores suelen basar sus decisiones en un análisis de incertidumbre como los siguientes:

1. ¿Qué posibilidades hay de que las ventas disminuyan si los precios aumentan?
2. ¿Cuál es la probabilidad de que un nuevo método de ensamble mejore la productividad?
3. ¿Qué tan probable es que este proyecto se complete a tiempo?
4. ¿Qué posibilidad hay de que una nueva inversión sea rentable?

Algunos de los primeros trabajos sobre probabilidad tuvieron su origen en una serie de cartas entre Pierre de Fermat y Blaise Pascal en la década de 1650.

La **probabilidad** es una medida numérica de la posibilidad de que un evento ocurra. Por tanto, se utiliza como una medida del grado de incertidumbre asociado con cada uno de los cuatro eventos previamente listados. Si las probabilidades están disponibles, se puede determinar la posibilidad de ocurrencia de cada evento.

Los valores de probabilidad siempre se asignan en una escala de 0 a 1. Una probabilidad cercana a 0 indica que es poco probable que un evento ocurra, una probabilidad cercana a 1 indica que es casi seguro que un evento se produzca. Otras probabilidades entre 0 y 1 representan grados de posibilidad de que un evento ocurra. Por ejemplo, si se considera el evento "lluvia para mañana", se entiende que cuando el informe del clima indica "una probabilidad de lluvia casi nula", significa que la posibilidad de lluvia es muy baja. Sin embargo, si se informa una probabilidad de 0.90 de que llueva, es probable que llueva. Una medida de 0.50 indica que la probabilidad de que llueva es igual a la de que no llueva. La figura 4.1 representa el punto de vista de la probabilidad como una medida numérica de la posibilidad de que un evento ocurra.

4.1 Experimentos, reglas de conteo y asignación de probabilidades

En el estudio de la probabilidad, un **experimento** se define como un proceso que genera resultados bien definidos. En cada repetición ocurre uno y sólo uno de los resultados posibles del experimento. En seguida se listan varios ejemplos de experimentos y sus resultados correspondientes.

Experimento	Resultados del experimento
Lanzar una moneda	Cara, cruz
Seleccionar una parte para inspeccionarla	Defectuosa, sin defectos
Hacer una llamada de ventas	Comprar, no comprar
Arrojar un dado	1, 2, 3, 4, 5, 6
Jugar un partido de futbol americano	Ganar, perder, empatar

Cuando se especifican todos los resultados posibles del experimento, el **espacio muestral** de éste queda definido.

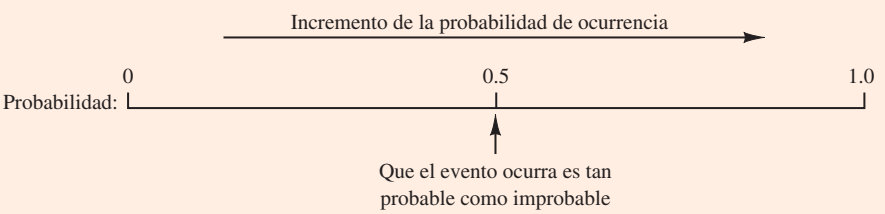
ESPACIO MUESTRAL

El espacio muestral es el conjunto de todos los resultados del experimento.

A los resultados del experimento también se les llama puntos de la muestra.

Un resultado experimental también se conoce como **punto de la muestra** para identificarlo como un elemento del espacio muestral.

FIGURA 4.1 La probabilidad como una medida numérica de la posibilidad de que un evento ocurra



Considere el primer experimento de la tabla anterior, es decir, el lanzamiento de una moneda. La cara que cae hacia arriba, ya sea cara o cruz, determina los resultados del experimento (puntos de la muestra). Si S denota el espacio muestral, se utiliza la siguiente notación para describirlo.

$$S = \{\text{cara, cruz}\}$$

El espacio muestral para el segundo experimento de la tabla, en el que se selecciona una parte para inspeccionarla, se describe como sigue:

$$S = \{\text{defectuosa, sin defectos}\}$$

Los dos ejemplos que se acaban de describir tienen dos resultados del experimento (puntos de la muestra). Sin embargo, suponga que se considera el cuarto caso listado en la tabla: el tiro de un dado. Los resultados del experimento posibles, que se definen como el número de puntos que tiene la cara superior del dado, son los seis puntos del espacio muestral de este experimento.

$$S = \{1, 2, 3, 4, 5, 6\}$$

Reglas de conteo, combinaciones y permutaciones

La identificación y el conteo de los resultados del experimento es un paso necesario en la asignación de probabilidades. Ahora se estudiarán tres reglas de conteo útiles.

Experimentos de pasos múltiples La primera regla de conteo se aplica a los experimentos de pasos múltiples. Considere un experimento que consiste en lanzar dos monedas. Los resultados se definen en función del patrón de caras y cruces que muestra la cara superior de las dos monedas. ¿Cuántos resultados son posibles para este experimento? El ejemplo de lanzar dos monedas se considera un experimento de dos pasos en el cual el paso 1 es el lanzamiento de la primera moneda y el paso 2 el lanzamiento de la segunda. Si se utiliza H para denotar una cara y T para una cruz, (H, H) indica el resultado experimental en el que hay una cara en la primera moneda y una cara en la segunda. Siguiendo esta notación, el espacio muestral (S) para este experimento se describe como sigue:

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

Por tanto, hay cuatro resultados experimentales posibles. En este caso, es fácil listarlos todos.

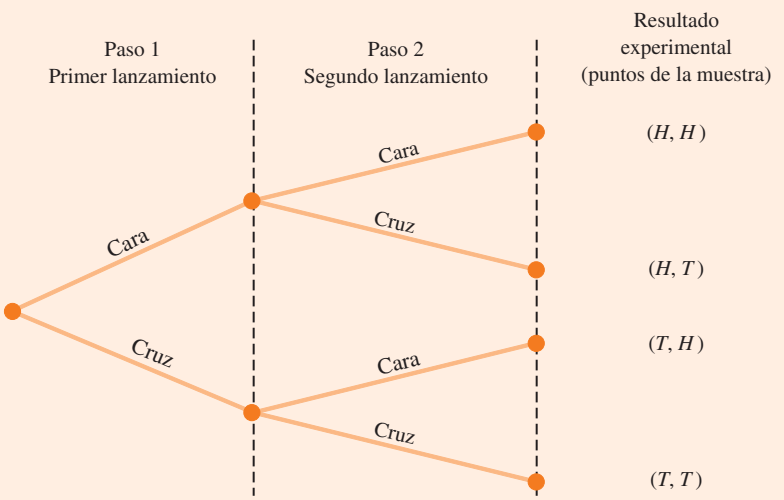
La regla de conteo para experimentos de pasos múltiples permite determinar el número de resultados del experimento sin listarlos.

REGLA DE CONTEO PARA EXPERIMENTOS DE PASOS MÚLTIPLES

Si un experimento se describe como una secuencia de k pasos con n_1 resultados posibles en el primer paso, n_2 resultados posibles en el segundo paso, y así sucesivamente, el número total de resultados del experimento está dado por $(n_1)\,(n_2)\,\ldots\,(n_k)$.

Si se considera el experimento del lanzamiento de dos monedas como la secuencia de lanzar primero una moneda $(n_1 = 2)$ y luego la otra $(n_2 = 2)$, al aplicar la regla de conteo puede verse que $(2)(2) = 4$, por lo que hay cuatro resultados experimentales distintos posibles. Como se mostró, estos resultados son $S = \{(H, H), (H, T), (T, H), (T, T)\}$. El número de resultados en un experimento que consiste en lanzar seis monedas es $(2)(2)(2)(2)(2)(2) = 64$.

FIGURA 4.2 Diagrama de árbol para el experimento del lanzamiento de dos monedas



Sin el diagrama de árbol, podría pensarse que hay sólo tres resultados experimentales posibles para dos lanzamientos de una moneda: 0 caras, 1 cara y 2 caras.

Un **diagrama de árbol** es una representación gráfica que ayuda a visualizar un experimento de pasos múltiples. La figura 4.2 muestra un diagrama de árbol para el experimento del lanzamiento de dos monedas. La secuencia de pasos va de izquierda a derecha a través del árbol. El paso 1 corresponde al lanzamiento de la primera moneda y el paso 2, al lanzamiento de la segunda. En cada paso, los dos resultados posibles son cara o cruz. Observe que a cada resultado posible del paso 1 le corresponden las dos ramas de los dos resultados posibles del paso 2. Cada uno de los puntos en el extremo derecho del árbol representa un resultado experimental. Cada trayectoria que recorre por el árbol desde el nodo que está en el extremo izquierdo hasta uno de los nodos en el extremo derecho es una secuencia única de resultados.

Ahora se explicará cómo se utilizan la reglas de conteo para experimentos de pasos múltiples mediante el análisis de un proyecto de expansión de Kentucky Power & Light Company (KP&L), el cual tiene la finalidad de incrementar la capacidad de generación de una de sus plantas en el norte de Kentucky. El proyecto está dividido en dos etapas o pasos secuenciales: etapa 1 (diseño) y etapa 2 (construcción). Aun cuando cada una se programará y controlará lo más detalladamente posible, la gerencia no puede predecir el tiempo exacto requerido para completar cada etapa. Un análisis de proyectos de construcción similares reveló que la duración posible de la etapa de diseño sería de 2, 3 o 4 meses y la duración probable de la fase de construcción sería de 6, 7 u 8 meses. Además, debido a la necesidad apremiante de tener más electricidad, la gerencia fijó una meta de 10 meses para completar todo el proyecto.

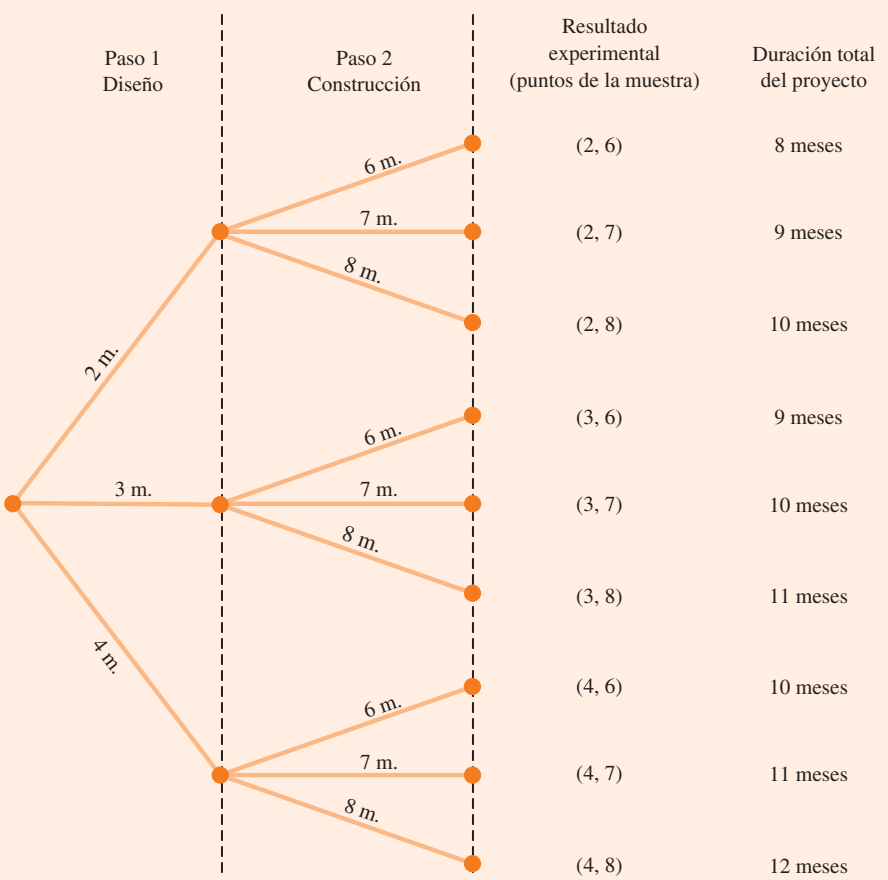
Como este proyecto tiene tres tiempos de terminación posibles para la etapa de diseño (paso 1) y tres tiempos de terminación posibles para la de construcción (paso 2), se aplica la regla de conteo para los experimentos de pasos múltiples para determinar un total de $(3)(3) = 9$ resultados del experimento. Para describir dichos resultados se utiliza una notación de dos números; por ejemplo, (2, 6) indica que la etapa de diseño se completará en 2 meses y la de construcción en 6. Este resultado experimental implica un total de $2 + 6 = 8$ meses para completar todo el plan. La tabla 4.1 resume los nueve resultados del experimento del problema de KP&L. El diagrama de árbol de la figura 4.3 muestra cómo ocurren los nueve resultados (puntos de la muestra).

La regla de conteo y el diagrama de árbol ayudan al gerente de proyectos a identificar los resultados del experimento y a determinar la duración posible del proyecto. A partir de la

TABLA 4.1 Resultados del experimento (puntos de la muestra) del proyecto de KP&L

Duración (meses)			
Etapa 1 Diseño	**Etapa 2 Construcción**	**Notación para resultados del experimento**	**Duración total del proyecto (meses)**
2	6	(2, 6)	8
2	7	(2, 7)	9
2	8	(2, 8)	10
3	6	(3, 6)	9
3	7	(3, 7)	10
3	8	(3, 8)	11
4	6	(4, 6)	10
4	7	(4, 7)	11
4	8	(4, 8)	12

FIGURA 4.3 Diagrama de árbol del proyecto de KP&L



información de la figura 4.3 se ve que éste durará de 8 a 12 meses, y que seis de los nueve resultados del experimento tienen la duración deseada de 10 meses o menos. Aun cuando la identificación de los resultados del experimento puede parecer útil, es necesario considerar cómo se asignan los valores de probabilidad a dichos resultados antes de evaluar la probabilidad de que el proyecto se complete dentro de los 10 meses deseados.

Combinaciones Una segunda regla de conteo útil permite contar el número de resultados cuando el experimento consiste en la selección de n objetos de un conjunto (generalmente mayor) de N objetos. Ésta se conoce como *regla de conteo para combinaciones*.

REGLA DE CONTEO PARA COMBINACIONES

El número de combinaciones de N objetos tomados n a la vez es

$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!} \qquad (4.1)$$

donde

$$N! = N(N-1)(N-2)\cdots(2)(1)$$
$$n! = n(n-1)(n-2)\cdots(2)(1)$$

y, por definición,

$$0! = 1$$

En el muestreo de una población finita de tamaño N, la regla de conteo para combinaciones ayuda a determinar el número de muestras diferentes de tamaño n que pueden seleccionarse.

La notación ! significa *factorial;* por ejemplo, 5 factorial es 5! = (5)(4)(3)(2)(1) = 120.

Como ejemplo del uso de la regla de conteo para combinaciones, considere un procedimiento de control de calidad en el cual un inspector selecciona al azar de dos a cinco partes para buscar defectos. En un grupo de cinco partes, ¿cuántas combinaciones de dos partes pueden seleccionarse? La regla de conteo de la ecuación (4.1) muestra que con $N = 5$ y $n = 2$; tenemos

$$C_2^5 = \binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{(5)(4)(3)(2)(1)}{(2)(1)(3)(2)(1)} = \frac{120}{12} = 10$$

Por tanto, 10 resultados son posibles para el experimento de selección de dos partes al azar de un grupo de cinco. Si las cinco partes se etiquetan como A, B, C, D y E, las 10 combinaciones o resultados del experimento son AB, AC, AD, AE, BC, BD, BE, CD, CE y DE.

Como otro ejemplo, considere el sistema de lotería de Florida que utiliza la selección al azar de seis enteros de un grupo de 53 para determinar al ganador de la semana. La regla de conteo para combinaciones, la ecuación (4.1), se utiliza para determinar el número de maneras en que seis enteros diferentes pueden seleccionarse de un grupo de 53.

$$\binom{53}{6} = \frac{53!}{6!(53-6)!} = \frac{53!}{6!47!} = \frac{(53)(52)(51)(50)(49)(48)}{(6)(5)(4)(3)(2)(1)} = 22\,957\,480$$

La regla de conteo para combinaciones muestra que el evento de ganar la lotería es muy poco probable.

La regla de conteo para combinaciones establece que casi 23 millones de resultados experimentales son posibles en el sorteo de la lotería. Una persona que compra un billete tiene 1 oportunidad en 22 957 480 de ganar.

Permutaciones Una tercera regla de conteo que en ocasiones es útil es la regla de conteo para permutaciones. Ésta permite que una persona calcule el número de resultados experimentales cuando se seleccionan n objetos de un conjunto de N objetos y el orden de selección es

importante. Los mismos n objetos seleccionados en un orden distinto se consideran un resultado experimental diferente.

REGLA DE CONTEO PARA PERMUTACIONES

El número de permutaciones de N objetos tomados n a la vez está dado por

$$P_n^N = n!\binom{N}{n} = \frac{N!}{(N-n)!} \tag{4.2}$$

La regla de conteo para permutaciones se relaciona estrechamente con la regla de conteo para combinaciones; sin embargo, un experimento produce más permutaciones que combinaciones para el mismo número de objetos debido a que cada selección de n objetos se ordena de $n!$ maneras distintas.

Como ejemplo, considere de nuevo el proceso de control de calidad en el que un inspector selecciona dos de cinco partes distintas para inspeccionarlas en busca de defectos. ¿Cuántas permutaciones pueden seleccionarse? La regla de conteo de la ecuación (4.2) muestra que con $N = 5$ y $n = 2$ se tiene

$$P_2^5 = \frac{5!}{(5-2)!} = \frac{5!}{3!} = \frac{(5)(4)(3)(2)(1)}{(3)(2)(1)} = \frac{120}{6} = 20$$

Por tanto, hay 20 resultados posibles para el experimento de seleccionar dos partes al azar de un grupo de cinco cuando se toma en cuenta el orden de selección. Si las partes se etiquetan como A, B, C, D y E, las 20 permutaciones son AB, BA, AC, CA, AD, DA, AE, EA, BC, CB, BD, DB, BE, EB, CD, DC, CE, EC, DE y ED.

Asignación de probabilidades

Ahora se explicará cómo asignar las probabilidades a los resultados del experimento. Los enfoques de tres pasos más usuales son el método clásico, el de frecuencia relativa y el subjetivo. Sea cual fuere el método empleado, se deben cumplir dos **requisitos básicos para la asignación de probabilidades**.

REQUISITOS BÁSICOS PARA LA ASIGNACIÓN DE PROBABILIDADES

1. La probabilidad asignada a cada resultado experimental debe estar entre 0 y 1, inclusive. Si E_i denota el i-ésimo resultado del experimento y $P(E_i)$ su probabilidad, entonces este requisito se escribe como

$$0 \leq P(E_i) \leq 1 \text{ para toda } i \tag{4.3}$$

2. La suma de las probabilidades para todos los resultados del experimento debe ser igual a 1. Para n resultados, este requisito se escribe como

$$P(E_1) + \mathrm{P(E_2)} + \cdots + P(E_n) = 1 \tag{4.4}$$

El **método clásico** de asignación de probabilidades es apropiado cuando todos los resultados del experimento son igualmente probables. Si n resultados son posibles, una probabilidad de $1/n$ se asigna a cada resultado experimental. Cuando se utiliza este método, los dos requisitos básicos para la asignación de probabilidades se cumplen de manera automática.

Como ejemplo, considere el experimento del lanzamiento de una moneda sin truco; los dos resultados, es decir, cara y cruz, son igualmente probables. Dado que uno de los dos resultados igualmente probables es una cara, la probabilidad de observar una cara es 1/2, o 0.50. Asimismo, la probabilidad de observar una cruz también es 1/2 o 0.50.

En otro ejemplo, considere el experimento de arrojar un dado. Parecería razonable concluir que los seis resultados posibles son igualmente probables y, por consiguiente, a cada resultado se le asigna una probabilidad de 1/6. Si $P(1)$ denota la probabilidad de que un punto aparezca en la cara superior del dado, entonces $P(1) = 1/6$. De igual manera, $P(2) = 1/6$, $P(3) = 1/6$, $P(4) = 1/6$, $P(5) = 1/6$ y $P(6) = 1/6$. Observe que estas probabilidades satisfacen los dos requisitos básicos de las ecuaciones (4.3) y (4.4), ya que cada una es mayor o igual que cero y suman 1.0.

El **método de frecuencia relativa** de asignación de probabilidades es apropiado cuando los datos están disponibles para estimar la proporción del tiempo en que ocurrirá el resultado si el experimento se repite un gran número de veces. Como ejemplo considere un estudio de los tiempos de espera en el departamento de rayos X para un hospital local. Un empleado registró el número de pacientes que esperan el servicio a las 9:00 a.m. durante 20 días sucesivos y obtuvo los resultados siguientes.

Número de pacientes que esperan	Número de días que el resultado ocurrió
0	2
1	5
2	6
3	4
4	3
Total	20

Estos datos arrojan que en 2 de los 20 días, cero pacientes esperaban por el servicio; en 5 de los días, un paciente esperaba por el servicio, etc. Utilizando el método de la frecuencia relativa, se asignaría una probabilidad de 2/20 = 0.10 para el resultado experimental de cero pacientes esperando; 5/20 = 0.25 para un paciente que espera; 6/20 = 0.30 para dos pacientes; 4/20 = 0.20 para tres sujetos y 3/20 = 0.15 para cuatro. Al igual que con el método clásico, el uso del método de la frecuencia relativa cumple automáticamente con los dos requisitos básicos de las ecuaciones (4.3) y (4.4).

El **método subjetivo** de asignación de probabilidades es más apropiado cuando no se puede asumir en forma realista que los resultados del experimento son igualmente probables y cuando se dispone de pocos datos relevantes. Cuando el método subjetivo se utiliza para asignar probabilidades a los resultados del experimento, es posible usar cualquier información disponible, como nuestra experiencia o intuición. Después de considerar toda la información disponible, un valor de probabilidad que expresa nuestro *grado de creencia* (en una escala de 0 a 1) de que el resultado experimental ocurrirá se especifica. Debido a que la probabilidad subjetiva expresa el grado de creencia de una persona, es personal. Utilizando este método, se puede esperar que distintas personas asignen probabilidades diferentes al mismo resultado experimental.

El método subjetivo exige un cuidado especial para asegurar que los dos requisitos básicos de las ecuaciones (4.3) y (4.4) se satisfagan. Sin considerar el grado de creencia de una persona, el valor de la probabilidad asignada a cada resultado experimental debe ser de entre 0 y 1, inclusive, y la suma de todas las probabilidades para los resultados experimentales debe ser igual a 1.0.

Considere el caso en el que Tom y Judy Elsbernd hacen una oferta para comprar una casa. Hay dos resultados posibles:

E_1 = su oferta es aceptada

E_2 = su oferta es rechazada

Judy cree que la probabilidad de que su oferta sea aceptada es de 0.8; por tanto, establecería $P(E_1) = 0.8$ y $P(E_2) = 0.2$. Tom, no obstante, cree que la probabilidad de que su oferta se acepte es de 0.6; por consiguiente, establecería $P(E_1) = 0.6$ y $P(E_2) = 0.4$. Note que la estimación de la probabilidad para E_1 de Tom refleja un pesimismo mayor de que su oferta será aceptada.

El teorema de Bayes (vea la sección 4.5) proporciona un medio para combinar de manera subjetiva determinadas probabilidades previas con las probabilidades obtenidas por otros medios para lograr las probabilidades revisadas, o posteriores.

Tanto las probabilidades asignadas de Judy como las de Tom satisfacen los dos requisitos básicos. El hecho de que sus estimaciones sean diferentes recalca la naturaleza personal del método subjetivo.

Aun cuando en las situaciones de negocios puede aplicarse ya sea el método clásico o el método de frecuencia relativa, los gerentes tal vez quieran proporcionar estimaciones de probabilidad subjetivas. En estos casos, las mejores estimaciones con frecuencia se obtienen al combinar las estimaciones de los métodos clásico y de frecuencia relativa con las de probabilidad subjetivas.

Probabilidades para el proyecto de KP&L

Para realizar otro análisis sobre el proyecto de KP&L, se deben desarrollar las probabilidades de cada uno de los nueve resultados del experimento listados en la tabla 4.1. Sobre la base de la experiencia y el juicio, la gerencia concluyó que los resultados del experimento no eran igualmente probables. Por consiguiente, no podría utilizarse el método clásico de asignación de probabilidades. La gerencia decidió, por tanto, efectuar un estudio de los tiempos de terminación de proyectos similares realizados por KP&L durante los tres años pasados. Los resultados de un análisis de 40 proyectos se resumen en la tabla 4.2.

Después de revisar los resultados del estudio, la gerencia optó por emplear el método de frecuencia relativa de asignación de probabilidades. Podría haber proporcionado estimaciones de probabilidad subjetivas, pero pensó que el proyecto actual era muy parecido a los 40 anteriores. Así, el método de frecuencia relativa se consideró el mejor.

Al usar los datos de la tabla 4.2 para calcular las probabilidades, se observa que el resultado (2, 6) —la etapa 1 completada en 2 meses y la etapa 2 completada en 6 meses— ocurrió seis veces en los 40 proyectos. El método de frecuencia relativa se utiliza para asignar una probabilidad de $6/40 = 0.15$ a este resultado. Asimismo, el resultado (2, 7) también ocurrió en seis de los 40 proyectos, proporcionando una probabilidad de $6/40 = 0.15$. Si se continúa de esta manera, se obtienen las asignaciones de probabilidad para los puntos de la muestra del proyecto de KP&L presentados en la tabla 4.3. Observe que $P(2, 6)$ representa la probabilidad del punto de muestreo (2, 6); $P(2, 7)$ la del punto de muestreo (2, 7), etcétera.

TABLA 4.2 Resultados de terminación de 40 proyectos de KP&L

Duración (meses)			
Etapa 1 Diseño	Etapa 2 Construcción	Punto de muestreo	Número de proyectos anteriores con estos tiempos de terminación
2	6	(2, 6)	6
2	7	(2, 7)	6
2	8	(2, 8)	2
3	6	(3, 6)	4
3	7	(3, 7)	8
3	8	(3, 8)	2
4	6	(4, 6)	2
4	7	(4, 7)	4
4	8	(4, 8)	6
		Total	40

TABLA 4.3 Asignaciones de probabilidad para el proyecto de KP&L con base en el método de frecuencia relativa

Punto de la muestreo	Duración del proyecto	Probabilidad del punto de muestreo
(2, 6)	8 meses	$P(2, 6) = 6/40 = 0.15$
(2, 7)	9 meses	$P(2, 7) = 6/40 = 0.15$
(2, 8)	10 meses	$P(2, 8) = 2/40 = 0.05$
(3, 6)	9 meses	$P(3, 6) = 4/40 = 0.10$
(3, 7)	10 meses	$P(3, 7) = 8/40 = 0.20$
(3, 8)	11 meses	$P(3, 8) = 2/40 = 0.05$
(4, 6)	10 meses	$P(4, 6) = 2/40 = 0.05$
(4, 7)	11 meses	$P(4, 7) = 4/40 = 0.10$
(4, 8)	12 meses	$P(4, 8) = 6/40 = 0.15$
		Total 1.00

NOTAS Y COMENTARIOS

1. En estadística, la noción de experimento difiere de alguna manera de la que se maneja en las ciencias físicas. En éstas, los investigadores realizan con frecuencia un experimento en un laboratorio o en un entorno controlado con el fin de aprender sobre la causa y el efecto. En los experimentos estadísticos, la probabilidad determina los resultados. Aun cuando el experimento se repite exactamente de la misma manera, puede ocurrir un resultado muy diferente. Debido a esta influencia de la probabilidad del resultado, los experimentos de estadística a veces se denominan *experimentos aleatorios.*
2. Cuando se obtiene una muestra al azar de una población de tamaño N sin remplazarla, se utiliza la regla de conteo para combinaciones con el fin de encontrar el número de muestras diferentes de tamaño n que pueden seleccionarse.

Ejercicios

Métodos

1. Un experimento consta de tres pasos con tres resultados posibles para el primer paso, dos resultados posibles para el segundo y cuatro para el tercero. ¿Cuántos resultados experimentales existen para todo el experimento?

AUTOevaluación

2. ¿De cuántas maneras pueden seleccionarse tres elementos de un grupo de seis? Utilice las letras A, B, C, D, E y F para identificar los elementos y elabore una lista cada una de las distintas combinaciones de tres elementos.
3. ¿Cuántas permutaciones de tres elementos pueden seleccionarse de un grupo de seis? Utilice las letras A, B, C, D, E y F para identificar los elementos y elabore una lista de cada una de las permutaciones de B, D y F.
4. Considere el experimento de lanzar una moneda tres veces.
   a) Elabore un diagrama de árbol para el experimento.
   b) Prepare una lista de los resultados del experimento.
   c) ¿Cuál es la probabilidad para cada resultado experimental?
5. Suponga que un experimento tiene cinco resultados igualmente probables: E_1, E_2, E_3, E_4, E_5. Asigne probabilidades a cada resultado y muestre que se cumplen los requisitos de las ecuaciones (4.3) y (4.4). ¿Qué método utilizó?

AUTOevaluación

6. Un experimento con tres resultados se repitió 50 veces y mostró que E_1 ocurrió 20 veces, E_2 13 veces y E_3 17 veces. Asigne probabilidades a los resultados. ¿Qué método usó?
7. Alguien que toma decisiones asignó de manera subjetiva las probabilidades siguientes a los cuatro resultados de un experimento: $P(E_1) = 0.10$, $P(E_2) = 0.15$, $P(E_3) = 0.40$ y $P(E_4) = 0.20$. ¿Son válidas estas asignaciones de probabilidad? Explique por qué.

Aplicaciones

8. En la ciudad de Mildford, las aplicaciones para los cambios de zonificación pasan por un proceso de dos pasos: una revisión de la comisión de planeación y una decisión final del consejo ciudadano. En el paso 1 la comisión de planeación revisa el cambio de zona solicitado y hace una recomendación positiva o negativa respecto de ese cambio. En el paso 2 el consejo ciudadano revisa la recomendación y luego vota para aprobar o desaprobar el cambio de zona. Suponga que el desarrollador de un complejo de departamentos presenta una solicitud para un cambio de zona. Considere el proceso de aplicación como un experimento.
   a) ¿Cuántos puntos de la muestra hay para este experimento? Lístelos.
   b) Construya un diagrama de árbol para el experimento.

AUTOevaluación

9. El muestreo aleatorio simple utiliza una muestra de tamaño n de una población de tamaño N para obtener datos que se pueden usar para hacer inferencias sobre las características de una población. Suponga que de una población de 50 cuentas bancarias se quiere tomar una muestra al azar de cuatro cuentas con el fin de aprender acerca de la población. ¿Cuántas muestras al azar diferentes de las cuatro cuentas son posibles?

AUTOevaluación

10. En Estados Unidos, muchos estudiantes han acumulado una deuda cuando se gradúan de la universidad. En la tabla siguiente se muestra el porcentaje de graduados que al terminar han acumulado una deuda y el monto medio de ésta para los estudiantes de cuatro universidades y cuatro colegios de arte (*U.S. News and World Report, America's Best Colleges,* 2008).

Universidad	% con deuda	Monto ($)	Colegio	% con deuda	Monto ($)
Pace	72	32 980	Wartburg	83	28 758
Iowa State	69	32 130	Morehouse	94	27 000
Massachusetts	55	11 227	Wellesley	55	10 206
SUNY-Albany	64	11 856	Wofford	49	11 012

   a) Si usted elige al azar a un graduado de Morehouse College, ¿cuál es la probabilidad de que este estudiante se haya graduado con una deuda?
   b) Si escoge una de estas ocho instituciones para un estudio de seguimiento sobre los préstamos a estudiantes, ¿cuál es la probabilidad de que considere una institución que tenga más de 60% graduados con deuda?
   c) Si usted elige una de estas ocho instituciones para un estudio de seguimiento sobre los préstamos a estudiantes, ¿cuál es la probabilidad de que escoja una institución donde la deuda media de los graduados endeudados sea de más de $30 000?
   d) ¿Cuál es la probabilidad de que un graduado de Pace University no tenga deuda?
   e) Para los graduados de Pace University con deuda, el monto medio de ésta es de $32 980. Considerando a todos los graduados de dicha universidad, ¿cuál es la deuda media por sujeto?

11. La National Highway Traffic Safety Administration (NHTSA) realizó un estudio para enterarse de cómo usan los cinturones de seguridad los conductores de todo el país (Associated Press, 25 de agosto de 2003). Los datos muestrales consistentes con la encuesta de la NHSTA son los siguientes.

	¿El conductor usa cinturón de seguridad?	
Región	Sí	No
Noreste	148	52
Oeste medio	162	54
Sur	296	74
Oeste	252	48
Total	858	228

a) Para Estados Unidos, ¿cuál es la probabilidad de que un conductor use cinturón de seguridad?
b) La probabilidad de uso del cinturón para un conductor estadounidense un año antes fue de 0.75. El jefe de la NHTSA, el Dr. Jeffrey Runge, había esperado una probabilidad de 0.78 en 2003. ¿Se sentiría complacido con los resultados de la encuesta de 2003?
c) ¿Cuál es la probabilidad del uso del cinturón de seguridad por región del país? ¿En qué región se usa más?
d) ¿Qué proporción de los conductores de la muestra proviene de cada región del país? ¿Qué región tuvo la mayoría de conductores selecionados? ¿Cuál tuvo la segunda mayoría?
e) Suponiendo que el número total de conductores de cada región es el mismo, ¿ve usted alguna razón por la cual la estimación de probabilidad del inciso *a*) podría ser demasiado alta? Explique.

12. La lotería Powerball se juega dos veces a la semana en 28 estados, las Islas Vírgenes y el distrito de Columbia. Para jugarla, un participante debe comprar un boleto y luego seleccionar cinco dígitos de los números de 1 al 55 y un número de Powerball de los dígitos 1 al 42. Para determinar los números ganadores para cada juego, los oficiales de la lotería extrajeron cinco bolas blancas de una urna con 55 bolas blancas y una bola roja de una urna con 42 bolas rojas. Para ganar la lotería, los números de un participante deben coincidir con los de las cinco bolas blancas en cualquier orden y con el número de la bola Powerball roja. Ocho colaboradores de la planta ConAgra Foods en Lincoln, Nebraska, reclamaron el premio mayor récord de $365 millones el 18 de febrero de 2006, al coincidir los números 15-17-43-44-49 y la bola Powerball número 29. Otros premios en efectivo se otorgan cada vez que el juego se realiza. Por ejemplo, se paga un premio de $200 000 si los cinco números del participante coinciden con los números de las cinco bolas blancas (sitio de Powerball, 19 de marzo de 2006).
a) Calcule el número de formas en que los primeros cinco números pueden ser seleccionados.
b) ¿Cuál es la probabilidad de ganar un premio de $200 000 por coincidir los números de las cinco bolas blancas?
c) ¿Cuál es la probabilidad de ganar el premio mayor Powerball?

13. Una empresa que fabrica pasta dental estudia cinco diseños de empaque diferentes. Suponiendo que un diseño tiene igual probabilidad de ser seleccionado por un consumidor como cualquier otro, ¿qué probabilidad de selección asignaría a cada uno de los diseños de empaque? En un experimento real se pidió a 100 consumidores que seleccionaran el diseño de su preferencia. Se obtuvieron los datos siguientes. ¿Los datos confirman la creencia de que un diseño tiene la misma probabilidad de ser seleccionado que otro? Explique por qué.

Diseño	Número de veces preferido
1	5
2	15
3	30
4	40
5	10

4.2 Eventos y sus probabilidades

En la introducción de este capítulo se usó el término *evento* de manera muy parecida a como se utiliza en el lenguaje cotidiano. Luego, en la sección 4.1 se presentó el concepto de experimento y los resultados del experimento o puntos de la muestra correspondientes. Los puntos de la muestra y los eventos proporcionan la base del estudio de la probabilidad. Por consiguiente, ahora un **evento** se define de manera formal en relación con los puntos de la muestra. Esta definición es la base para determinar la probabilidad de un evento.

EVENTO

Un evento es una colección de puntos de la muestra.

Como ejemplo, retome el proyecto de KP&L y suponga que el gerente está interesado en el evento de que el proyecto completo se termine en 10 meses o menos. Al observar la tabla 4.3 se ve que seis puntos de la muestra —(2, 6), (2, 7), (2, 8), (3, 6), (3, 7) y (4, 6)— proporcionan una duración de 10 meses o menos. C denota el evento de que el proyecto dure 10 meses o menos; escribimos

$$C = \{(2, 6), (2, 7), (2, 8), (3, 6), (3, 7), (4, 6)\}$$

Se dice que el evento C ocurre si *cualquiera* de estos seis puntos de la muestra aparece como el resultado experimental.

Otros eventos que podrían ser de interés para la gerencia de KP&L son los siguientes.

L = El evento de que el proyecto se complete en *menos* de 10 meses

M = El evento de que el proyecto se complete en *más* de 10 meses

Con ayuda de la información de la tabla 4.3, vemos que estos eventos constan de los puntos de la muestra siguientes:

$$L = \{(2, 6), (2, 7), (3, 6)\}$$
$$M = \{(3, 8), (4, 7), (4, 8)\}$$

Una variedad de eventos adicionales puede definirse para el proyecto de KP&L, pero en cada caso el evento debe identificarse como una colección de puntos de la muestra para el experimento.

Dadas las probabilidades de los puntos de la muestra mostrados en la tabla 4.3, podemos utilizar la definición siguiente para calcular la probabilidad de cualquier evento que la gerencia de KP&L podría desear considerar.

PROBABILIDAD DE UN EVENTO

La probabilidad de cualquier evento es igual a la suma de las probabilidades de los puntos de la muestra del evento.

Con ayuda de esta definición, se calcula la probabilidad de un evento particular al sumar las probabilidades de los puntos de la muestra (resultados del experimento) que conforman el evento. Ahora se puede calcular la probabilidad de que el proyecto tarde en completarse 10 meses o menos. Debido a que este evento está dado por $C = \{(2, 6), (2, 7), (2, 8), (3, 6), (3, 7), (4, 6)\}$, la probabilidad del evento C, denotada $P(C)$, está dada por

$$P(C) = P(2, 6) + P(2, 7) + P(2, 8) + P(3, 6) + P(3, 7) + P(4, 6)$$

Revisando las probabilidades de los puntos de la muestra de la tabla 4.3 tenemos

$$P(C) = 0.15 + 0.15 + 0.05 + 0.10 + 0.20 + 0.05 = 0.70$$

De modo parecido, debido a que el evento de que el proyecto se complete en menos de 10 meses está dado por $L = \{(2, 6), (2, 7), (3, 6)\}$, la probabilidad de este evento está determinada por

$$\begin{aligned} P(L) &= P(2, 6) + P(2, 7) + P(3, 6) \\ &= 0.15 + 0.15 + 0.10 = 0.40 \end{aligned}$$

Por último, para el evento de que el proyecto se termine en más de 10 meses, tenemos $M = \{(3, 8), (4, 7), (4, 8)\}$, y por tanto

$$\begin{aligned} P(M) &= P(3, 8) + P(4, 7) + P(4, 8) \\ &= 0.05 + 0.10 + 0.15 = 0.30 \end{aligned}$$

Al utilizar estos resultados de la probabilidad, ahora es posible decir a la gerencia de KP&L que hay una probabilidad de 0.70 de que el proyecto se complete en 10 meses o menos, una probabilidad de 0.40 de que se complete en menos de 10 meses y una probabilidad de 0.30 de que concluya en más de 10 meses. Este procedimiento de cálculo de las probabilidades del evento puede repetirse para cualquier evento de interés para la gerencia de KP&L.

En cualquier momento se pueden identificar todos los puntos de la muestra de un experimento y asignar probabilidades a cada uno, y podemos calcular la probabilidad de un evento utilizando la definición. No obstante, en muchos experimentos un número grande de puntos de la muestra hace muy engorrosa, si no es que imposible, la identificación de estos puntos, así como la determinación de sus probabilidades asociadas. En las secciones restantes de este capítulo se presentan algunas relaciones de probabilidad básicas que se usan para calcular la probabilidad de un evento sin conocimiento de todas las probabilidades de los puntos de la muestra.

NOTAS Y COMENTARIOS

1. El espacio muestral, S, es un evento. Debido a que contiene todos los resultados del experimento, tiene una probabilidad de 1; es decir, $P(S) = 1$.
2. Cuando se utiliza el método clásico para asignar probabilidades, el supuesto es que los resultados del experimento son igualmente probables. En estos casos, la probabilidad de un evento se calcula contando el número de resultados del experimento en el evento y dividiendo el resultado entre el número total de resultados del experimento.

Ejercicios

Métodos

14. Un experimento tiene cuatro resultados igualmente probables: E_1, E_2, E_3 y E_4.
    a) ¿Cuál es la probabilidad de que E_2 ocurra?
    b) ¿Cuál es la probabilidad de que cualesquiera de los dos resultados ocurran (por ejemplo, E_1 o E_3)?
    c) ¿Cuál es la probabilidad de que cualesquiera de los tres resultados ocurran (por ejemplo, E_1 o E_2 o E_4)?

AUTO evaluación

15. Considere el experimento de seleccionar una carta de una baraja de 52 cartas. Cada carta corresponde a un punto muestral con una probabilidad de 1/52.
    a) Elabore una lista de los puntos de la muestra en el evento de seleccionar un as.
    b) Liste los puntos de la muestra en el evento de elegir una carta de bastos.
    c) Elabore una lista de los puntos de la muestra en el evento de seleccionar una figura (jota, reina o rey).
    d) Calcule las probabilidades asociadas con cada uno de los eventos de los incisos *a*), *b*) y *c*).

16. Considere el experimento de arrojar un par de dados. Suponga que le interesa la suma de los valores de las caras mostradas en el dado.
    a) ¿Cuántos puntos de la muestra son posibles? (*Sugerencia:* utilice la regla de conteo para los experimentos de pasos múltiples.)
    b) Elabore una lista de los puntos de la muestra.
    c) ¿Cuál es la probabilidad de obtener un valor de 7?
    d) ¿Cuál es la probabilidad de obtener un valor de 9 o mayor?
    e) Debido a que cada tiro tiene seis valores pares de eventos posibles (2, 4, 6, 8, 10 y 12) y sólo cinco valores impares posibles (3, 5, 7, 9 y 11), el dado debe mostrar más a menudo valores pares que impares. ¿Está usted de acuerdo con este enunciado? Explique.
    f) ¿Qué método utilizó para asignar las probabilidades requeridas?

Aplicaciones

AUTOevaluación

17. Revise los puntos de la muestra de KP&L y las probabilidades de los puntos de la muestra de las tablas 4.2 y 4.3.
    a) La etapa de diseño (etapa 1) rebasará el presupuesto si tarda 4 meses en completarse. Elabore una lista de los puntos de la muestra en el evento de que la etapa de diseño sobrepase el presupuesto.
    b) ¿Cuál es la probabilidad de que la etapa de diseño rebase el presupuesto?
    c) La etapa de construcción (etapa 2) rebasará el gasto presupuestado si tarda 8 meses en completarse. Elabore una lista de los puntos de la muestra en el evento de que la etapa de construcción sobrepase el presupuesto.
    d) ¿Cuál es la probabilidad de que la fase de construcción rebase el presupuesto?
    e) ¿Cuál es la probabilidad de que ambas etapas lo sobrepasen?

18. Para investigar con qué frecuencia las familias suelen comer en casa, Harris Interactive encuestó a 496 adultos que vivían con niños menores de 18 años (*USA Today,* 3 de enero de 2007). Los resultados de la encuesta se muestran en la tabla siguiente.

Número de comidas familiares por semana	Número de respuestas a la encuesta
0	11
1	11
2	30
3	36
4	36
5	119
6	114
7 o más	139

Para una familia seleccionada al azar con niños menores de 18 años, calcule lo siguiente:
    a) La probabilidad de que la familia no coma en casa durante la semana.
    b) La probabilidad de que la familia coma por lo menos cuatro veces en casa durante la semana.
    c) La probabilidad de que la familia coma dos o menos veces en casa durante la semana.

19. La National Sporting Goods Association realizó una encuesta a personas de 7 años de edad o mayores acerca de su participación en actividades deportivas (*Statistical Abstract of the United States*, 2002). La población total en este grupo de edades se reportó en 248.5 millones, con 120.9 millones de hombres y 127.6 millones de mujeres. El número de participantes para las cinco actividades deportivas principales se muestra enseguida.

	Participantes (millones)	
Actividad	Hombre	Mujer
Ciclismo	22.2	21.0
Acampar	25.6	24.3
Ejercitarse caminando	28.7	57.7
Ejercitarse con equipo	20.4	24.4
Nadar	26.4	34.4

    a) Para una mujer seleccionada al azar, estime la probabilidad de participación en cada una de las actividades deportivas.
    b) Para un hombre seleccionado al azar, calcule la probabilidad de participación en cada una de las actividades deportivas.
    c) Para una persona seleccionada al azar, ¿cuál es la probabilidad de que se ejercite caminando?
    d) Suponga que acaba de ver a una persona que se ejercita caminando. ¿Cuál es la probabilidad de que se trate de una mujer? ¿Cuál es la probabilidad de que sea hombre?

20. La revista *Fortune* publica una lista anual de las 500 empresas más grandes de Estados Unidos. Los datos siguientes muestran los cinco estados con el número más grande de empresas *Fortune 500* (*The New York Times Almanac,* 2006).

Estado	Número de empresas
Nueva York	54
California	52
Texas	48
Illinois	33
Ohio	30

Suponga que una empresa *Fortune 500* es elegida al azar para un cuestionario de seguimiento. ¿Cuáles son las probabilidades de los eventos siguientes?

a) Sea N el evento de que las oficinas corporativas de la empresa tienen su sede en Nueva York. Calcule $P(N)$.

b) Sea T el evento de que las oficinas corporativas de la empresa tienen su sede en Texas. Calcule $P(T)$.

c) Sea B el evento de que la sede de las oficinas corporativas de la empresa está en estos cinco estados. Calcule $P(B)$.

21. La población adulta estadounidense por edad es la siguiente (*The World Almanac,* 2009). Los datos se proporcionan en millones de personas.

Edad	Número
18 a 24	29.8
25 a 34	40.0
35 a 44	43.4
45 a 54	43.9
55 a 64	32.7
65 y más	37.8

Suponga que una persona de esta población será elegida al azar.

a) ¿Cuál es la probabilidad de que la persona tenga de 18 a 24 años?

b) ¿Cuál es la probabilidad de que tenga de 18 a 34 años?

c) ¿Cuál es la probabilidad de que tenga 45 años?

4.3 Algunas relaciones básicas de probabilidad

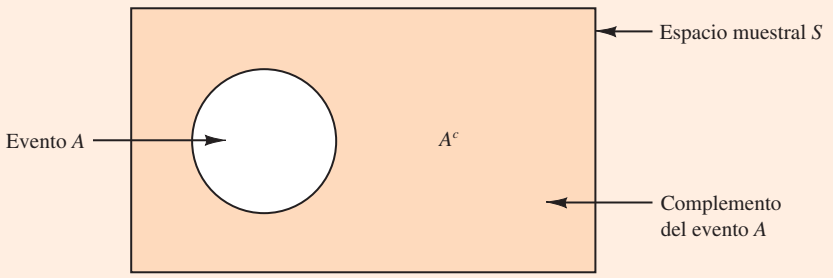
Complemento de un evento

Dado un evento A, el **complemento de A** se define como el evento que consta de todos los puntos de la muestra que *no* están en A. El complemento de A se denota por medio de A^c. La figura 4.4 es un diagrama, conocido como **diagrama de Venn**, el cual ilustra el concepto de complemento. El área rectangular representa el espacio muestral para el experimento y como tal contiene todos los puntos de la muestra posibles. El círculo representa el evento A y contiene sólo los puntos de la muestra que pertenecen a A. La región sombreada del rectángulo contiene todos los puntos de la muestra que no están en el evento A y es por definición el complemento de A.

En cualquier probabilidad de aplicación debe ocurrir cualquier evento A o su complemento A^c. Por consiguiente, tenemos

$$P(A) = P(A^c) = 1$$

FIGURA 4.4 El complemento del evento A está sombreado



Al calcular $P(A)$, se obtiene el resultado siguiente.

CÁLCULO DE LA PROBABILIDAD UTILIZANDO EL COMPLEMENTO

$$P(A) = 1 - P(A^c) \tag{4.5}$$

La ecuación (4.5) muestra la probabilidad de que un evento A se calcule fácilmente si se conoce la probabilidad de su complemento, $P(A^c)$.

Como ejemplo, considere el caso de un gerente de ventas quien, después de revisar los informes de ventas, establece que 80% de los contactos de clientes nuevos no generan ninguna venta. Al hacer que A denote el evento de que se realiza una venta y A^c denote el evento de que no se realice, el gerente establece que $P(A^c) = 0.80$. Utilizando la ecuación (4.5), vemos que

$$P(A) = 1 - P(A^c) = 1 - 0.80 = 0.20$$

Se puede concluir que un contacto de un cliente nuevo tiene una probabilidad de 0.20 de generar una venta.

En otro ejemplo, un agente de compras establece una probabilidad de 0.90 de que un proveedor envíe mercancía sin partes defectuosas. Utilizando el complemento, se puede concluir que hay una probabilidad de $1 - 0.90 = 0.10$ de que la mercancía contenga partes defectuosas.

Ley de la adición

La ley de la adición es útil cuando interesa conocer la probabilidad de que ocurra por lo menos uno de dos eventos. Es decir, con los eventos A y B nos interesa conocer la probabilidad de que ocurra el evento A o el evento B, o ambos.

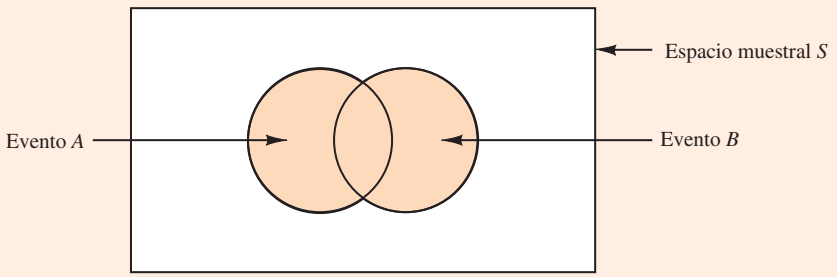
Antes de presentar la ley de la adición, debemos estudiar dos conceptos relacionados con la combinación de eventos: la *unión* de eventos y la *intersección* de eventos. Dados dos eventos A y B, la **unión de A y B** se define como sigue.

UNIÓN DE DOS EVENTOS

La *unión* de A y B es el evento que contiene *todos* los puntos de la muestra que pertenecen a A *o* B *o ambos*. La unión se denota mediante $A \cup B$.

El diagrama de Venn de la figura 4.5 representa la unión de los eventos A y B. Observe que los dos círculos contienen todos los puntos de la muestra del evento A, así como todos los puntos

FIGURA 4.5 Unión de los eventos A y B sombreada



de la muestra del evento B. El hecho de que los círculos se traslapen indica que algunos puntos de la muestra están contenidos tanto en A como en B.

A continuación se presenta la definición de **intersección de A y B**.

INTERSECCIÓN DE DOS EVENTOS

Dados dos eventos A y B, la *intersección* de A y B es el evento que contiene los puntos de la muestra que pertenecen a *tanto a A como a B*. La intersección se denota por medio de $A \cap B$.

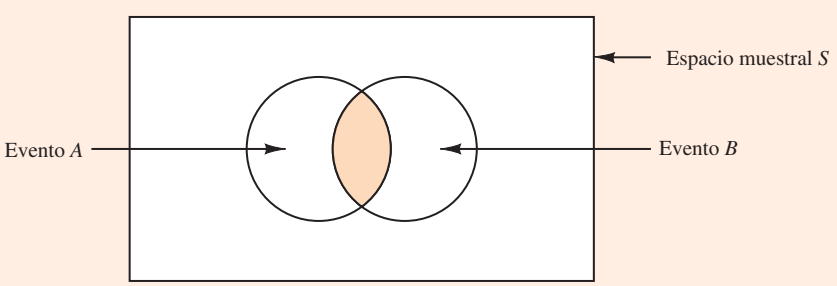
El diagrama de Venn que representa la intersección de los eventos A y B se muestra en la figura 4.6. El área donde los dos círculos se traslapan es la intersección; contiene los puntos de la muestra que están tanto en A como en B.

Ahora se estudiará la ley de la adición. La **ley de la adición** proporciona una manera de calcular la probabilidad de que ocurra el evento A o el evento B o ambos. En otras palabras, la ley de la adición se utiliza para calcular la probabilidad de la unión de dos eventos. La ley de la adición se escribe como sigue.

LEY DE LA ADICIÓN

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad \textbf{(4.6)}$$

FIGURA 4.6 Intersección de los eventos A y B sombreada



Para entender de manera intuitiva la ley de la adición, considere que los dos primeros términos de la ley, $P(A) + P(B)$, representan todos los puntos de la muestra en $A \cup B$. Sin embargo, debido a que los puntos de la muestra en la intersección $A \cap B$ están en A y en B, cuando se calcula $P(A) + P(B)$, en realidad se están contando dos veces cada uno de los puntos de la muestra en $A \cap B$. Este conteo excesivo se corrige al restar $P(A \cap B)$.

Como ejemplo de una aplicación de la ley de la adición, considere el caso de una pequeña planta de ensamble con 50 empleados. Se espera que cada trabajador complete las asignaciones de trabajo a tiempo y de tal manera que el producto ensamblado apruebe la inspección final. De vez en cuando, algunos trabajadores no cumplen con los estándares de desempeño, ya que terminan la tarea con atraso o ensamblan un producto defectuoso. Al final del periodo de evaluación del desempeño, el gerente de producción encontró que 5 de los 50 trabajadores terminaron el trabajo con atraso, 6 de los 50 ensamblaron un producto defectuoso y 2 de los 50 terminaron con atraso *y* ensamblaron un producto defectuoso.

Sean

$$L = \text{evento de que el trabajo se termine con atraso}$$
$$D = \text{evento de que el producto ensamblado esté defectuoso}$$

La información de la frecuencia relativa conduce a las probabilidades siguientes.

$$P(L) = \frac{5}{50} = 0.10$$

$$P(D) = \frac{6}{50} = 0.12$$

$$P(L \cap D) = \frac{2}{50} = 0.04$$

Después de revisar los datos de desempeño, el gerente de producción decidió asignar una calificación baja a cualquier empleado cuyo trabajo estuviera atrasado o defectuoso, por lo que el evento de interés es $L \cup D$. ¿Cuál es la probabilidad de que el gerente asigne una calificación de bajo desempeño a un empleado?

Note que la pregunta de probabilidad trata de la unión de dos eventos. En concreto, se desea conocer $P(L \cup D)$. Mediante la ecuación (4.6) tenemos

$$P(L \cup D) = P(L) + P(D) - P(L \cap D)$$

Al conocer los valores de las tres probabilidades en el lado derecho de esta expresión, se puede escribir

$$P(L \cup D) = 0.10 + 0.12 - 0.04 = 0.18$$

Este cálculo indica que hay una probabilidad de 0.18 de que un empleado seleccionado al azar reciba una calificación de bajo desempeño.

En otro ejemplo de la ley de la adición, considere un estudio reciente realizado por el jefe de personal de una importante firma de software. El estudio reveló que 30% de los empleados que dejaron la empresa en un plazo de dos años lo hizo principalmente porque se sentía insatisfecho con su sueldo, 20% se fue porque no estaba satisfecho con el trabajo que se le asignó y 12% indicó insatisfacción tanto con su sueldo como con el trabajo asignado. ¿Cuál es la probabilidad de que un empleado que deja la empresa en un plazo de dos años lo haga

debido a su insatisfacción con el sueldo, a su insatisfacción con el trabajo asignado o a ambas cosas?

Sea

S = evento de que el empleado deje la empresa debido al sueldo

W = evento de que el empleado deje la empresa debido al trabajo asignado

Se tiene $P(S) = 0.30$; $P(W) = 0.20$, y $P(S \cap W) = 0.12$. Utilizando la ecuación (4.6), la ley de la adición, tenemos

$$P(S \cup W) = P(S) + P(W) - P(S \cap W) = 0.30 + 0.20 - 0.12 = 0.38$$

Se obtuvo una probabilidad de 0.38 de que un empleado abandone la empresa por las razones del sueldo o el trabajo asignado.

Antes de concluir nuestro análisis de la ley de la adición, considere un caso especial que surge para los **eventos mutuamente excluyentes**.

EVENTOS MUTUAMENTE EXCLUYENTES

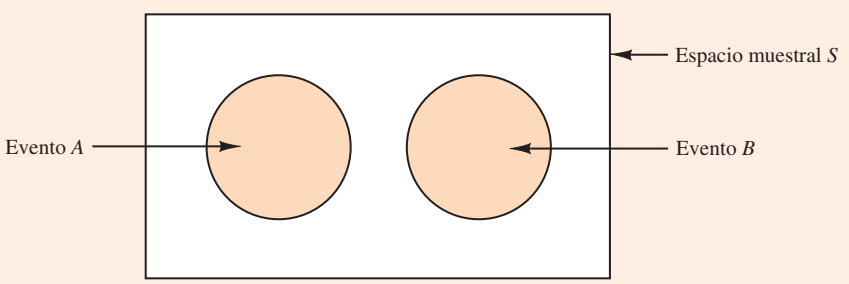
Se dice que dos eventos son mutuamente excluyentes si no tienen puntos de la muestra en común.

Los eventos A y B son mutuamente excluyentes si, cuando ocurre un evento, el otro no puede ocurrir. Por tanto, un requisito para que A y B sean mutuamente excluyentes consiste en que su intersección no debe contener puntos de la muestra. El diagrama de Venn que representa dos eventos mutuamente excluyentes A y B se muestra en la figura 4.7. En este caso $P(A \cap B) = 0$, y la ley de la adición puede escribirse como sigue.

LEY DE LA ADICIÓN PARA EVENTOS MUTUAMENTE EXCLUYENTES

$$P(A \cup B) = P(A) + P(B)$$

FIGURA 4.7 Eventos mutuamente excluyentes



Ejercicios

Métodos

22. Suponga que tiene un espacio muestral con cinco resultados experimentales igualmente probables: E_1, E_2, E_3, E_4 y E_5. Sea

$$A = \{E_1, E_2\}$$
$$B = \{E_3, E_4\}$$
$$C = \{E_2, E_3, E_5\}$$

a) Calcule $P(A)$, $P(B)$ y $P(C)$.
b) Encuentre $P(A \cup B)$. ¿A y B son mutuamente excluyentes?
c) Calcule A^c, C^c, $P(A^c)$ y $P(C^c)$.
d) Defina $A \cup B^c$ y $P(A \cup B^c)$.
e) Calcule $P(B \cup C)$.

AUTOevaluación

23. Suponga que tiene un espacio muestral $S = \{E_1, E_2, E_3, E_4, E_5, E_6, E_7\}$, donde $E_1, E_2, \ldots, E_7$ denotan los puntos de la muestra. Las asignaciones de probabilidad siguientes se aplican: $P(E_1) = 0.05$; $P(E_2) = 0.20$; $P(E_3) = 0.20$; $P(E_4) = 0.25$; $P(E_5) = 0.15$; $P(E_6) = 0.10$, y $P(E_7) = 0.05$. Sean

$$A = \{E_1, E_4, E_6\}$$
$$B = \{E_2, E_4, E_7\}$$
$$C = \{E_2, E_3, E_5, E_7\}$$

a) Calcule $P(A)$, $P(B)$ y $P(C)$.
b) Encuentre $A \cup B$ y $P(A \cup B)$.
c) Calcule $A \cap B$ y $P(A \cap B)$.
d) ¿A y C son mutuamente excluyentes?
e) Calcule B^c y $P(B^c)$.

Aplicaciones

24. Clarkson University encuestó al alumnado para conocer qué pensaba sobre la universidad. Una parte de la encuesta solicitaba a los alumnos que indicaran si su experiencia general en Clarkson estaba por debajo de sus expectativas, cumplía con las mismas o las rebasaba. Los resultados mostraron que 4% de los encuestados no proporcionó respuesta, 26% dijo que su experiencia estaba por debajo de sus expectativas y 65% afirmó que su experiencia cumplía con sus expectativas.
a) Si se elige un estudiante al azar, ¿cuál es la probabilidad de que él diga que su experiencia *rebasó* sus expectativas?
b) Si se escoge un alumno al azar, ¿cuál es la probabilidad de que él diga que su experiencia cumplió o rebasó sus expectativas?

25. La Oficina del Censo de Estados Unidos proporciona datos sobre el número de adultos jóvenes, entre 18 y 24 años, que viven en la casa de sus padres.[1] Sean

M = el evento de que un hombre adulto joven viva en casa de sus padres
F = el evento de que una mujer adulta joven viva en casa de sus padres

Si se seleccionan al azar un hombre adulto joven y una mujer adulta joven, los datos de la Oficina del Censo permiten concluir $P(M) = 0.56$ y $P(F) = 0.42$ (*The World Almanac,* 2006). La probabilidad de que ambos estén viviendo en la casa de sus padres es 0.24.
a) ¿Cuál es la probabilidad de que por lo menos uno de los dos adultos jóvenes seleccionados viva en casa de sus padres?
b) ¿Cuál es la probabilidad de que ambos adultos jóvenes vivan solos (ninguno vive en casa de sus padres)?

[1] Los datos incluyen adultos jóvenes solos que viven en los dormitorios de la universidad, debido a que se supone que regresan a casa de sus padres cuando no hay clases.

26. La información sobre los fondos de inversión proporcionada por Morningstar Investment Research incluye el tipo de fondo, es decir, capital nacional, capital internacional o renta fija y la calificación Morningstar para el fondo. Ésta se expresa con 1 estrella (calificación menor) a 5 estrellas (calificación mayor). Una muestra de 25 fondos de inversión fue seleccionada de *Morningstar Funds500* (2008). Se obtuvieron los conteos siguientes:
    - Dieciséis fondos de inversión eran fondos de capital nacional.
    - Trece fondos de inversión se calificaron con 3 estrellas o menos.
    - Siete de los fondos de capital nacional se calificaron con 4 estrellas.
    - Dos fondos de capital nacional se calificaron con 5 estrellas.

    Suponga que uno de estos 25 fondos de inversión es seleccionado al azar con el fin de conocer más sobre el fondo y su estrategia de inversión.
    a) ¿Cuál es la probabilidad de seleccionar un fondo de capital nacional?
    b) ¿Cuál es la probabilidad de elegir un fondo con una calificación de 4 o 5 estrellas?
    c) ¿Cuál es la probabilidad de seleccionar un fondo de capital nacional *y* que tiene una calificación de 4 o 5 estrellas??
    d) ¿Cuál es la probabilidad de escoger un fondo de capital nacional *o* que tiene una calificación de 4 o 5 estrellas?

27. ¿Qué ligas de basquetbol colegial de la NCAA tienen mayor probabilidad de hacer que un equipo juegue en el partido del campeonato nacional de basquetbol colegial? Durante los últimos 20 años, la Atlantic Coast Conference (ACC) califica primero por tener un equipo en el partido del campeonato 10 veces. La Southeastern Conference (SEC) se clasifica en segundo lugar por tener un equipo en el partido de campeonato 8 veces. Sin embargo, estas dos ligas tuvieron equipos en el partido del campeonato sólo una vez, cuando Arkansas (SEC) derrotó a Duke (ACC) 76-70 en 1994 (sitio web de la NCAA, abril de 2009). Utilice estos datos para estimar las probabilidades siguientes.
    a) ¿Cuál es la probabilidad de que la ACC tenga un equipo en el partido del campeonato?
    b) ¿Cuál es la probabilidad para la SEC?
    c) ¿Cuál es la probabilidad de que la ACC y la SEC tengan ambos equipos en el partido del campeonato?
    d) ¿Cuál es la probabilidad de que por lo menos un equipo de estas dos ligas juegue en el partido del campeonato? Es decir, ¿cuál es la probabilidad de que un equipo de la ACC o la SEC juegue en el campeonato?
    e) ¿Cuál es la probabilidad de que el partido del campeonato no tenga un equipo de una de estas dos ligas?

AUTOevaluación

28. Una encuesta de suscriptores a una revista reveló que 45.8% rentó un automóvil durante los 12 meses anteriores por razones de trabajo, 54% lo rentó en el mismo periodo por razones personales y 30% tanto por razones de trabajo como personales.
    a) ¿Cuál es la probabilidad de que un suscriptor rentara un automóvil durante los 12 meses anteriores por razones de trabajo o personales?
    b) ¿Cuál es la probabilidad de que un suscriptor no rentara un automóvil durante el periodo de referencia por razones de trabajo o personales?

29. Los estudiantes de bachillerato con registros académicos sólidos aplican para las universidades más selectivas de Estados Unidos en números mayores cada año. Debido a que el número de vacantes permanece relativamente estable, algunas universidades rechazan más aspirantes de forma anticipada. La Universidad de Pennsylvania recibió 2 851 solicitudes de admisión anticipadas. De este grupo, aceptó a 1 033 estudiantes, rechazó a 854 en el acto y difirió 964 al grupo de admisión regular para una consideración posterior. En el pasado, la universidad ha admitido a 18% de los estudiantes diferidos que presentó una solicitud de admisión anticipada durante el proceso de admisión regular. Contando tanto a los alumnos aceptados de forma anticipada como durante el proceso de admisión regular, el tamaño total de la generación fue de 2 375 (*USA Today,* 24 de enero de 2001). E, R y D representan los eventos de que un estudiante que solicita la admisión anticipada sea aceptado de forma anticipada, rechazado en el acto o diferido al grupo de admisiones regulares.
    a) Utilice los datos para estimar $P(E)$, $P(R)$ y $P(D)$.
    b) ¿Los eventos E y D son mutuamente excluyentes? Calcule $P(E \cap D)$.

c) Para los 2 375 estudiantes admitidos en la universidad, ¿cuál es la probabilidad de que uno seleccionado al azar sea aceptado durante la admisión anticipada?

d) Suponga que un estudiante presenta una solicitud de ingreso a la universidad para una admisión anticipada. ¿Cuál es la probabilidad de que sea aceptado por una admisión anticipada o sea diferido e ingresado después durante el proceso regular de admisión?

4.4 Probabilidad condicional

La probabilidad de un evento a menudo es influida por el hecho de si otro evento relacionado ha ocurrido ya. Suponga que se tiene un evento A con probabilidad $P(A)$. Si se obtiene nueva información y se aprende que un evento relacionado, denotado por B, ya ocurrió, esta información se puede aprovechar mediante el cálculo de una nueva probabilidad del evento A, a la cual se denomina **probabilidad condicional**, y se escribe $P(A \mid B)$. La notación se utiliza para indicar que se está considerando la probabilidad del evento A *dada* la condición de que B ha ocurrido. De ahí que la notación $P(A \mid B)$ se lea "la probabilidad de A dado B".

Como ejemplo de la aplicación de la probabilidad condicional, considere la situación del estado de ascensos de oficiales hombres y mujeres de una fuerza policiaca metropolitana en el este de Estados Unidos. La policía local está formada por 1 200 oficiales, 960 hombres y 240 mujeres. Durante los últimos dos años fueron ascendidos 324 oficiales de policía. La composición específica de la promoción de hombres y mujeres se muestra en la tabla 4.4.

Después de revisar el registro de ascensos, un comité de mujeres policía planteó un caso de discriminación sobre la base de que 288 oficiales hombres fueron promovidos, en comparación con sólo 36 mujeres. La comandancia argumentó que el número relativamente bajo de ascensos de las oficiales femeninas no se debe a discriminación, sino al hecho de que en la policía hay relativamente pocos miembros que son mujeres. Enseguida se mostrará cómo se utiliza la probabilidad condicional para analizar la acusación de discriminación.

Sean

M = el evento de que un oficial es hombre

W = el evento de que un oficial es mujer

A = el evento de que un oficial es promovido

A^c = el evento de que un oficial no es promovido

La división de los valores de la tabla 4.4 entre el total de 1 200 oficiales permite resumir la información disponible con los valores de probabilidad siguientes.

Una probabilidad de $P(M \cap A) = 288/1\,200 = 0.24$ de que un agente elegido al azar sea hombre *y* sea promovido

Una probabilidad de $P(M \cap A^c) = 672/1\,200 = 0.56$ de que un agente elegido al azar sea hombre *y* no sea promovido

TABLA 4.4 Estado de la promoción de los oficiales de policía durante los dos años anteriores

	Hombres	Mujeres	Total
Promovido(a)	288	36	324
No promovido(a)	672	204	876
Total	960	240	1200

TABLA 4.5 Tabla de probabilidad conjunta para promociones

Las probabilidades conjuntas aparecen en el cuerpo de la tabla

	Hombres (M)	Mujeres (W)	Total
Promovido (A)	0.24	0.03	0.27
No promovido (A^c)	0.56	0.17	0.73
Total	0.80	0.20	1.00

Las probabilidades marginales aparecen en los bordes de la tabla.

Una probabilidad de $P(W \cap A) = 36/1\,200 = 0.03$ de que un oficial elegido al azar sea mujer *y* sea promovida

Una probabilidad de $P(W \cap A^c) = 204/1\,200 = 0.17$ de que un agente elegido al azar sea mujer *y* no sea promovida

Debido a que cada uno de estos valores da la probabilidad de la intersección de dos eventos, las probabilidades se llaman **probabilidades conjuntas**. La tabla 4.5, que proporciona un resumen de la información de probabilidad sobre la situación de la promoción de oficiales de policía, se conoce como *tabla de probabilidad conjunta.*

Los valores en los bordes de esta tabla proporcionan las probabilidades de cada caso por separado. Es decir, $P(M) = 0.80$; $P(W) = 0.20$; $P(A) = 0.27$, y $P(A^c) = 0.73$. Estos datos se refieren a las **probabilidades marginales** debido a su ubicación en los bordes de la tabla de probabilidad conjunta. Observe que las probabilidades marginales se encuentran al sumar las probabilidades conjuntas en la fila o columna correspondiente de la tabla. Por ejemplo, la probabilidad marginal de ser promovido es $P(A) = P(M \cap A) + P(W \cap A) = 0.24 + 0.03 = 0.27$. De las probabilidades marginales, también vemos que 80% de la fuerza policiaca son hombres y 20% mujeres, y que 27% de todos los oficiales fueron promovidos y 73% no fueron promovidos.

Para comenzar, se hará el análisis de probabilidad condicional mediante el cálculo de la probabilidad de que un oficial sea promovido dado que es hombre. En la notación de la probabilidad condicional se trata de determinar $P(A \mid M)$. Para calcularla, primero observe que esta notación simplemente significa que se está considerando la probabilidad del evento A (promoción), dado que la condición designada como el evento M (el oficial es hombre) se sabe que existe. Por tanto $P(A \mid M)$ indica que estamos interesados sólo en el estado de la promoción de los 960 oficiales hombres. Debido a que 288 de estos 960 oficiales fueron ascendidos, la probabilidad de ser promovido, dado que el oficial es hombre, es de $288/960 = 0.30$. En otras palabras, dado su género, ese oficial tenía una probabilidad de 30% de ser promovido en los últimos dos años.

Este procedimiento fue fácil de aplicar debido a que los valores de la tabla 4.4 muestran el número de oficiales en cada categoría. Ahora queremos demostrar cómo las probabilidades condicionales como $P(A \mid M)$ se calculan directamente de las probabilidades de eventos relacionados más que de los datos de la frecuencia de la tabla 4.4.

Hemos mostrado que $P(A \mid M) = 288/960 = 0.30$. Ahora dividamos tanto el numerador como el denominador de esta fracción entre 1 200, el número total de oficiales que participaron en el estudio.

$$P(A \mid M) = \frac{288}{960} = \frac{288/1\,200}{960/1\,200} = \frac{0.24}{0.80} = 0.30$$

Ahora se ve que la probabilidad condicional $P(A \mid M)$ se calcula como 0.24/0.80. Revise la tabla de probabilidad conjunta (tabla 4.5). Tome nota en particular de que 0.24 es la probabilidad

conjunta de A y M; es decir, $P(A \cap M) = 0.24$. También note que 0.80 es la probabilidad marginal de que un oficial elegido al azar es hombre; es decir, $P(M) = 0.80$. Por tanto, la probabilidad condicional $P(A \mid M)$ se calcula como la razón de la probabilidad conjunta $P(A \cap M)$ a la probabilidad marginal $P(M)$.

$$P(A \mid M) = \frac{P(A \cap M)}{P(M)} = \frac{0.24}{0.80} = 0.30$$

El hecho de que las probabilidades condicionales se calculen como la razón de una probabilidad conjunta a una probabilidad marginal proporciona la fórmula general siguiente para los cálculos de la probabilidad condicional para dos eventos A y B.

PROBABILIDAD CONDICIONAL

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{4.7}$$

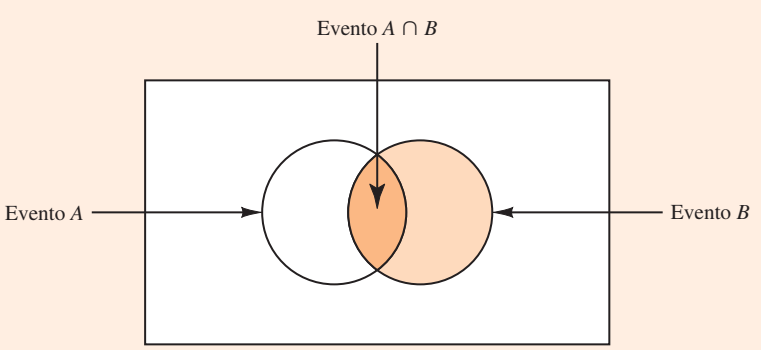
o

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{4.8}$$

El diagrama de Venn de la figura 4.8 es útil en la obtención de una comprensión intuitiva de la probabilidad condicional. El círculo de la derecha muestra que el evento B ha ocurrido; la parte del círculo que se superpone con el evento A denota el evento $(A \cap B)$. Se sabe que una vez que B ha ocurrido, la única manera en que se puede observar también A es que el evento $(A \cap B)$ ocurra. Por tanto, la razón $P(A \cap B)/P(B)$ proporciona la probabilidad condicional de que el evento A tendrá lugar, dado que el evento B ha ocurrido ya.

Retome el problema de discriminación contra las oficiales mujeres. La probabilidad marginal de la fila 1 de la tabla 4.5 muestra que la probabilidad de promoción de un oficial es $P(A) = 0.27$ (con independencia de que sea hombre o mujer). Sin embargo, el problema fundamental en el caso de la discriminación implica las dos probabilidades condicionales $P(A \mid M)$ y $P(A \mid W)$. Es decir, ¿cuál es la probabilidad de una promoción *dado* que el policía es hombre, y cuál es la probabilidad *dado* que el policía es mujer? Si estas dos probabilidades son iguales, un argumento de discriminación no tiene ningún fundamento, porque las posibilidades son iguales para los policías de ambos géneros. Sin embargo, una diferencia en las dos probabilidades condicionales apoyará la posición de que los policías hombres y mujeres son tratados de manera diferente en las decisiones de promoción.

FIGURA 4.8 Probabilidad condicional $P(A \mid B) = P(A \cap B)/P(B)$



Ya determinamos que $P(A \mid M) = 0.30$. Ahora utilizamos los valores de la tabla 4.5 y la relación básica de la probabilidad condicional en la ecuación (4.7) para calcular la probabilidad de que un policía sea promovido dado que es mujer; es decir, $P(A \mid W)$. Con ayuda de la ecuación (4.7), reemplazando W con B obtenemos

$$P(A \mid W) = \frac{P(A \cap W)}{P(W)} = \frac{0.03}{0.20} = 0.15$$

¿A qué conclusión llega? La probabilidad de una promoción, dado que el policía es hombre, es de 0.30, el doble de la probabilidad de 0.15 considerando que el policía es mujer. Aun cuando el uso de la probabilidad condicional no prueba por sí misma que existe discriminación en el caso, los valores de probabilidad condicional apoyan el argumento presentado por los policías hombres.

Eventos independientes

En la ilustración anterior, $P(A) = 0.27$; $P(A \mid M) = 0.30$, y $P(A \mid W) = 0.15$. Vemos que la probabilidad de una promoción (evento A) no ha cambiado ni se ha visto influida por el hecho de que el policía sea hombre o mujer. En particular, debido a que $P(A \mid M) \neq P(A)$, diríamos que los eventos A y M son dependientes. Es decir, la probabilidad del evento A (promoción) se ve alterada o afectada por conocer que el evento M (el policía es hombre) existe. Asimismo, con $P(A \mid W) \neq P(A)$, diríamos que A y W son *eventos dependientes.* No obstante, si la probabilidad del evento A no cambia por la existencia del evento M —es decir, $P(A \mid M) = P(A)$— diríamos que A y M son **eventos independientes**. Esta situación conduce a la definición siguiente de la independencia de dos eventos.

EVENTOS INDEPENDIENTES

Dos eventos A y B son independientes si

$$P(A \mid B) = P(A) \tag{4.9}$$

o

$$P(B \mid A) = P(B) \tag{4.10}$$

De lo contrario, los eventos son dependientes.

Ley de la multiplicación

Mientras que la ley aditiva de la probabilidad se utiliza para calcular la probabilidad de la unión de dos eventos, la ley de la multiplicación se utiliza para calcular la probabilidad de la intersección de dos eventos. Esta última ley se basa en la definición de la probabilidad condicional. Con ayuda de las ecuaciones (4.7) y (4.8) y calculando $P(A \cap B)$, se obtiene la **ley de la multiplicación**.

LEY DE LA MULTIPLICACIÓN

$$P(A \cap B) = P(B)P(A \mid B) \tag{4.11}$$

o

$$P(A \cap B) = P(A)P(B \mid A) \tag{4.12}$$

Para ilustrar el uso de la ley que se comenta, considere un departamento de circulación de periódicos donde se sabe que 84% de las familias en un vecindario en particular se suscribe a la edición diaria del periódico. Si D denota el evento de que una familia se suscribe a la edición diaria, $P(D) = 0.84$. Además, se sabe que la probabilidad de que una familia que ya cuenta

con una suscripción también adquiera la edición dominical (evento S) es de 0.75; es decir, $P(S \mid D) = 0.75$. ¿Cuál es la probabilidad de que una familia se suscriba tanto a las ediciones dominicales como a las ediciones diarias del periódico? Utilizando la ley de la multiplicación, calculamos el $P(S \cap D)$ deseado como

$$P(S \cap D) = P(D)P(S \mid D) = 0.84(0.75) = 0.63$$

Se sabe que 63% de las familias se suscribe tanto a las ediciones dominicales como a las diarias.

Antes de concluir esta sección, considere el caso especial de la ley de la multiplicación cuando los eventos involucrados son independientes. Recuerde que los eventos A y B son independientes siempre que $P(A \mid B) = P(A)$ o $P(B \mid A) = P(B)$. Por consiguiente, utilizando las ecuaciones (4.11) y (4.12) para el caso especial de los eventos independientes, obtenemos la ley de la multiplicación siguiente.

LEY DE LA MULTIPLICACIÓN PARA EVENTOS INDEPENDIENTES

$$P(A \cap B) = P(A)P(B) \qquad (4.13)$$

Para calcular la probabilidad de la intersección de dos eventos independientes, sencillamente se multiplican las probabilidades correspondientes. Observe que la ley de la multiplicación para eventos independientes proporciona otra manera de determinar si A y B son independientes. Es decir, si $P(A \cap B) = P(A)P(B)$, entonces A y B son independientes; si $P(A \cap B) \neq P(A)P(B)$, entonces A y B son dependientes.

Como una aplicación de la ley de la multiplicación para eventos independientes, considere la situación de un gerente de estaciones de servicio que sabe, a partir de su experiencia, que 80% de los clientes usa tarjeta de crédito cuando compra gasolina. ¿Cuál es la probabilidad de que los siguientes dos clientes que compren gasolina usen tarjeta de crédito? Si

A = el evento de que el primer cliente use tarjeta de crédito

B = el evento de que el segundo cliente use tarjeta de crédito

entonces el evento de interés está en $A \cap B$. Debido a que no existe más información, es posible asumir de manera razonable que A y B son eventos independientes. Por tanto,

$$P(A \cap B) = P(A)P(B) = (0.80)(0.80) = 0.64$$

Para resumir esta sección, observe que nuestro interés en la probabilidad condicional está motivado por el hecho de que los eventos con frecuencia están relacionados. En estos casos, se dice que los eventos son dependientes y las fórmulas de la probabilidad condicional en las ecuaciones (4.7) y (4.8) deben usarse para calcular las probabilidades respectivas. Si dos eventos no están relacionados, son independientes; en este caso, la probabilidad de ninguno de ellos se ve afectada por el hecho de que el otro evento ocurra.

NOTAS Y COMENTARIOS

No confunda la noción de eventos mutuamente excluyentes con la de eventos independientes. Dos eventos con probabilidades diferentes de cero no pueden ser mutuamente excluyentes e independientes. Si se sabe que ocurre un evento mutuamente excluyente, el otro no puede ocurrir; por tanto, la probabilidad de que el otro evento ocurra se reduce a cero: son eventos dependientes.

Ejercicios

Métodos

AUTOevaluación

30. Suponga que tenemos dos eventos, A y B, con $P(A) = 0.50$; $P(B) = 0.60$, y $P(A \cap B) = 0.40$.
   a) Calcule $P(A \mid B)$.
   b) Calcule $P(B \mid A)$.
   c) ¿Los eventos A y B son independientes? ¿Por qué?

31. Suponga que tenemos dos eventos, A y B, que son mutuamente excluyentes. Suponga además que sabemos que $P(A) = 0.30$ y $P(B) = 0.40$.
    a) ¿Cuánto es $P(A \cap B)$?
    b) ¿Cuánto es $P(A \mid B)$?
    c) Un alumno de estadística sostiene que los conceptos de eventos mutuamente excluyentes y de eventos independientes en realidad son lo mismo, y que por tanto si los eventos son mutuamente excluyentes, deben ser independientes. ¿Está de acuerdo con esta afirmación? Utilice la información de probabilidad de este problema para argumentar su respuesta.
    d) ¿Qué conclusión general formularía acerca de los eventos mutuamente excluyentes e independientes dados los resultados de este problema?

Aplicaciones

32. La industria automotriz vendió 657 000 vehículos en Estados Unidos durante enero de 2009 (*The Wall Street Journal,* 4 de febrero de 2009). Este volumen se redujo 37% desde enero de 2008 a medida que las condiciones económicas continuaron deteriorándose. Los tres grandes fabricantes de automóviles de Estados Unidos, a saber General Motors, Ford y Chrysler, vendieron 280 500 vehículos, 48% menos desde enero de 2008. Un resumen de las ventas por fabricante y tipo de vehículo vendido se muestra en la tabla siguiente. Los datos están en miles de unidades. Los fabricantes líderes no estadounidenses son Toyota, Honda y Nissan. La categoría camión ligero incluye los modelos pickup, minivan, SUV y crossover.

		Tipo de vehículo	
		Automóvil	**Camión ligero**
Fabricante	**Estadounidense**	87.4	193.1
	No estadounidense	228.5	148.0

    a) Elabore una tabla de probabilidad conjunta para estos datos y utilícela para responder las preguntas restantes.
    b) ¿Cuáles son las probabilidades marginales? ¿Qué le dicen sobre las probabilidades asociadas con el fabricante y el tipo de vehículo vendido?
    c) Si un vehículo fue producido por una de las automotrices estadounidenses, ¿cuál es la probabilidad de que la unidad sea un automóvil? ¿Y de que sea un camión ligero?
    d) Si un vehículo no fue producido por uno de los fabricantes estadounidenses, ¿cuál es la probabilidad de que se trate de un automóvil? ¿Cuál es la probabilidad de que sea un camión ligero?
    e) Si la unidad era un camión ligero, ¿cuál es la probabilidad de que haya sido producido por uno de los fabricantes estadounidenses?
    f) ¿Qué le dice la información de probabilidad sobre las ventas?

AUTO evaluación

33. En una encuesta de estudiantes de maestría se obtuvieron los datos siguientes sobre la primera razón de los "estudiantes para solicitar el ingreso en la escuela en que se matricularon".

		Razón de la solicitud			
		Calidad de la escuela	**Costo o conveniencia de la escuela**	**Otros**	**Totales**
Estado de inscripción	**Tiempo completo**	421	393	76	890
	Tiempo parcial	400	593	46	1 039
	Totales	821	986	122	1 929

    a) Elabore una tabla de probabilidad conjunta para estos datos.
    b) Use las probabilidades marginales de la calidad de la escuela, el costo o la conveniencia de la escuela y otros para comentar la razón más importante para elegirla.

c) Si un estudiante es de tiempo completo, ¿cuál es la probabilidad de que la calidad sea la primera razón para elegir una escuela?

d) Si es de tiempo parcial, ¿cuál es la probabilidad de que la calidad sea la primera razón para elegir una escuela?

e) Sea *A* el evento de que un estudiante asiste de tiempo completo y *B* el evento de que lista la calidad de la escuela como la primera razón para solicitar el ingreso. ¿Los eventos *A* y *B* son independientes? Justifique su respuesta.

34. El Departamento de Transporte de Estados Unidos informó que durante noviembre, 83.4% de los vuelos de Southwest Airlines, 75.1% de los de US Airways y 70.1% de los de JetBlue llegaron a tiempo (*USA Today,* 4 de enero de 2007). Suponga que este desempeño a tiempo es aplicable para los vuelos que arriban a la explanada *A* de Rochester International Airport, y que 40% de las llegadas a la explanada *A* son vuelos de Southwest Airlines, 35% de US Airways y 25% de JetBlue.

a) Elabore una tabla de probabilidad conjunta con tres filas (aerolíneas) y dos columnas (arribo a tiempo frente a llegadas con retraso).

b) Se acaba de anunciar que el vuelo 1424 llegará por la puerta 20 en la explanada *A*. ¿Cuál es la aerolínea más probable para este arribo?

c) ¿Cuál es la probabilidad de que el vuelo 1424 llegue a tiempo?

d) Suponga que se anuncia que el vuelo 1424 llegará tarde. ¿Cuál es la aerolínea más probable para esta llegada? ¿Cuál es la menos probable?

35. Con base en el estudio Ameriprise Financial Money Across Generations, 9 de cada 10 padres con hijos adultos de 20 a 35 años los han apoyado con algún tipo de ayuda financiera que abarca la universidad, un automóvil, la renta, artículos, pagos a la tarjeta de crédito o pagos para casa (*Money,* enero de 2009). La tabla siguiente con los datos muestrales consistentes con el estudio indica el número de veces que los padres han proporcionado ayuda financiera a sus hijos adultos para comprar un automóvil o pagar la renta.

		Pagar renta	
		Sí	**No**
Comprar un automóvil	**Sí**	56	52
	No	14	78

a) Elabore una tabla de probabilidad conjunta y utilícela para responder las preguntas restantes.

b) Con base en las probabilidades marginales sobre comprar un automóvil y pagar la renta, ¿es más probable que los padres apoyen a sus hijos adultos con la compra de un automóvil o el pago de la renta? ¿Cuál es su interpretación de las probabilidades marginales?

c) Si los padres proporcionaron respaldo financiero para comprar un automóvil, ¿cuál es la probabilidad de que apoyaran *con el pago de la renta*?

d) Si los padres no proveyeron ayuda financiera para comprar un automóvil, ¿cuál es la probabilidad de que apoyaran con el pago de la renta?

e) ¿La ayuda económica para comprar un automóvil es independiente de la proporcionada para pagar la renta? Utilice las probabilidades para justificar su respuesta.

f) ¿Cuál es la probabilidad de que los padres proporcionaran ayuda financiera para sus hijos adultos, ya sea para comprar un automóvil o pagar la renta?

36. Jerry Stackhouse de los Mavericks de Dallas de la Asociación Nacional de Basquetbol es el mejor lanzador de tiro libre del equipo, al anotar 89% de sus tiros (sitio web de ESPN, julio de 2008). Suponga que más tarde, en un partido de basquetbol, le cometen una falta o *foul* a Jerry Stackhouse y se le otorgan dos tiros.

a) ¿Cuál es la probabilidad de que anote ambos tiros?

b) ¿Cuál es la probabilidad de que anote por lo menos uno?

c) ¿Cuál es la probabilidad de que falle ambos tiros?

d) Después, en un partido de basquetbol, un equipo comete faltas frecuentes de manera deliberada contra un jugador adversario con el fin de detener el reloj del partido. La estrategia usual es cometer falta deliberadamente contra el peor tirador de tiros libres del otro equipo. Suponga que el centro de los Mavericks de Dallas anota 58% de sus tiros libres. Calcule las probabilidades para el centro como en los incisos *a*), *b*) y *c*), y muestre que cometer faltas contra el centro de los Mavericks de Dallas es una mejor estrategia que cometerlas contra Jerry Stackhouse.

37. Visa Card USA estudió con qué frecuencia los consumidores jóvenes, de 18 a 24 años de edad, usan tarjetas (de débito y crédito) al realizar compras (Associated Press, 16 de enero de 2006). Los resultados del estudio proporcionaron las probabilidades siguientes.

- La probabilidad de que un cliente use una tarjeta cuando realiza una compra es 0.37.
- Dado que el cliente usa una tarjeta, hay una probabilidad de 0.19 de que tenga de 18 a 24 años.
- Dado que el consumidor usa una tarjeta, hay una probabilidad de 0.81 de que tenga más de 24 años.

Los datos de la Oficina del Censo de Estados Unidos muestran que 14% de la población de consumidores tiene de 18 a 24 años.

a) Dado que el cliente tiene entre 18 y 24 años de edad, ¿cuál es la probabilidad de que use una tarjeta?
b) Dado que el consumidor tiene 24 años, ¿cuál es la probabilidad de que utilice una tarjeta?
c) ¿Cuál es la interpretación de las probabilidades en los incisos *a*) y *b*)?
d) ¿Las empresas como Visa, MasterCard y Discover deben otorgar tarjetas al grupo de edades de 18 a 24 años antes de que estos consumidores tengan tiempo para establecer un historial de crédito? Si no es así, ¿por qué? Si es así, ¿qué restricciones a estos grupos de edad podrían mencionar las empresas?

38. Un estudio del consumidor de Morgan Stanley encuestó a hombres y mujeres y les preguntó si preferían beber agua simple embotellada o una bebida rehidratante como Gatorade o el agua Propel Fitness (*The Atlanta Journal-Constitution*, 28 de diciembre de 2005). Suponga que 200 hombres y 200 mujeres participaron en el estudio, y 280 informaron que preferían el agua embotellada simple. Del grupo que prefiere una bebida deportiva, 80 eran hombres y 40 mujeres. Sean

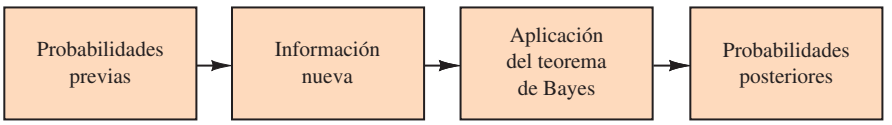
M = el evento de que el consumidor es hombre
W = el evento de que el consumidor es mujer
B = el evento de que el consumidor prefirió agua simple embotellada
S = el evento de que el consumidor prefirió una bebida rehidratante

a) ¿Cuál es la probabilidad de que una persona del estudio prefiriera agua embotellada simple?
b) ¿Cuál es la probabilidad de que prefiriera una bebida rehidratante?
c) ¿Cuáles son las probabilidades condicionales $P(M \mid S)$ y $P(W \mid S)$?
d) ¿Cuáles son las probabilidades conjuntas $P(M \cap S)$ y $P(W \cap S)$?
e) Dado que un consumidor es hombre, ¿cuál es la probabilidad de que prefiera una bebida rehidratante?
f) Dado que un consumidor es mujer, ¿cuál es la probabilidad de que prefiera una bebida rehidratante?
g) ¿La preferencia por una bebida rehidratante es independiente de si el consumidor es hombre o mujer? Explique con ayuda de la información de probabilidad.

4.5 Teorema de Bayes

En el estudio de la probabilidad condicional se indicó que la revisión de las probabilidades cuando se obtiene nueva información es una fase importante del análisis de probabilidad. A menudo comenzamos el análisis con las estimaciones de **probabilidad previa** o inicial para eventos específicos de interés. Por tanto, de fuentes como una muestra, un informe especial o una prueba de productos se obtiene información adicional sobre los eventos. Con esta nueva información actualizamos los valores de probabilidad previos mediante el cálculo de las probabilidades revisadas, conocidas como **probabilidades posteriores**. El **teorema de Bayes** proporciona un medio para efectuar estos cálculos. Los pasos en este proceso de revisión de la probabilidad se muestran en la figura 4.9.

FIGURA 4.9 Revisión de la probabilidad utilizando el teorema de Bayes



Como una aplicación del teorema de Bayes, considere una empresa de manufactura que recibe embarques de refacciones de dos proveedores diferentes. Sea A_1 el evento de que una refacción proviene del proveedor 1, y A_2 el evento de que una refacción proviene del proveedor 2. En la actualidad, 65% de las partes adquiridas por la empresa son del proveedor 1 y el 35% restante son del proveedor 2. De ahí que si una refacción es seleccionada al azar, se le asignarían las probabilidades previas $P(A_1) = 0.65$ y $P(A_2) = 0.35$.

La calidad de las partes adquiridas varía con la fuente de suministro. Los datos históricos sugieren que las calificaciones de calidad de los dos proveedores se muestran en la tabla 4.6. Si G denota el evento de que una refacción está en buen estado y B denota el evento de que una refacción está en mal estado, la información de la tabla 4.6 proporciona los valores de probabilidad condicional siguientes.

$$P(G \mid A_1) = 0.98 \quad P(B \mid A_1) = 0.02$$
$$P(G \mid A_2) = 0.95 \quad P(B \mid A_2) = 0.05$$

El diagrama de árbol de la figura 4.10 representa el proceso de la empresa que recibe una refacción de uno de los dos proveedores y luego descubre que está en buen o mal estado como un experimento de dos pasos. Se ve que los cuatro resultados del experimento son posibles; dos corresponden a la refacción que está en buen estado y dos a la que está en mal estado.

Cada uno de los resultados es la intersección de dos eventos, así que se puede utilizar la regla de la multiplicación para calcular las probabilidades. Por ejemplo,

$$P(A_1, G) = P(A_1 \cap G) = P(A_1)P(\mathrm{G} \mid A_1)$$

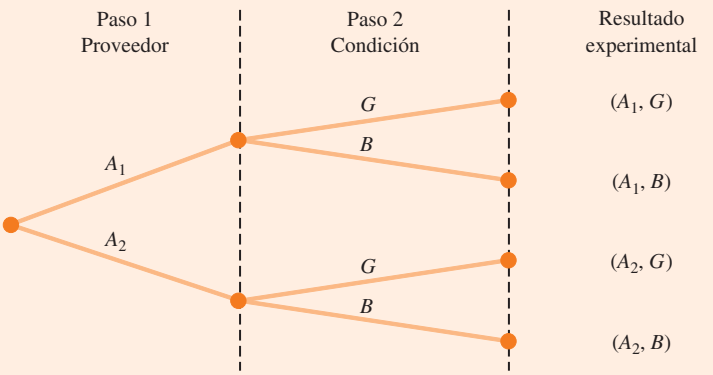
El proceso de calcular estas probabilidades conjuntas puede representarse en lo que se llama un *árbol de probabilidad* (figura 4.11). De izquierda a derecha por el árbol, las probabilidades para cada rama en el paso 1 son previas y las probabilidades para cada rama en el paso 2 son condicionales. Para encontrar las probabilidades de cada resultado del experimento, sencillamente se multiplican las probabilidades de las ramas que conducen al resultado. Cada una de estas probabilidades conjuntas se muestra en la figura 4.11 junto con las probabilidades conocidas para cada rama.

Suponga ahora que las refacciones de los dos proveedores se usan en el proceso de manufactura de la empresa y que una máquina se descompone porque intenta procesar una refacción en mal estado. Dada la información de que la refacción esta defectuosa, ¿cuál es la probabilidad

TABLA 4.6 Niveles históricos de calidad de los dos proveedores

	Porcentaje de refacciones en buen estado	Porcentaje de refacciones en mal estado
Proveedor 1	98	2
Proveedor 2	95	5

FIGURA 4.10 Diagrama de árbol para el ejemplo de los dos proveedores



Nota. El paso 1 ilustra que la refacción llega de uno de dos proveedores, y el paso 2 muestra si la refacción es buena o mala.

de que provenga del proveedor 1 y cuál de que provenga del proveedor 2? Con la información del árbol de probabilidades (figura 4.11), el teorema de Bayes ayuda a responder estas preguntas.

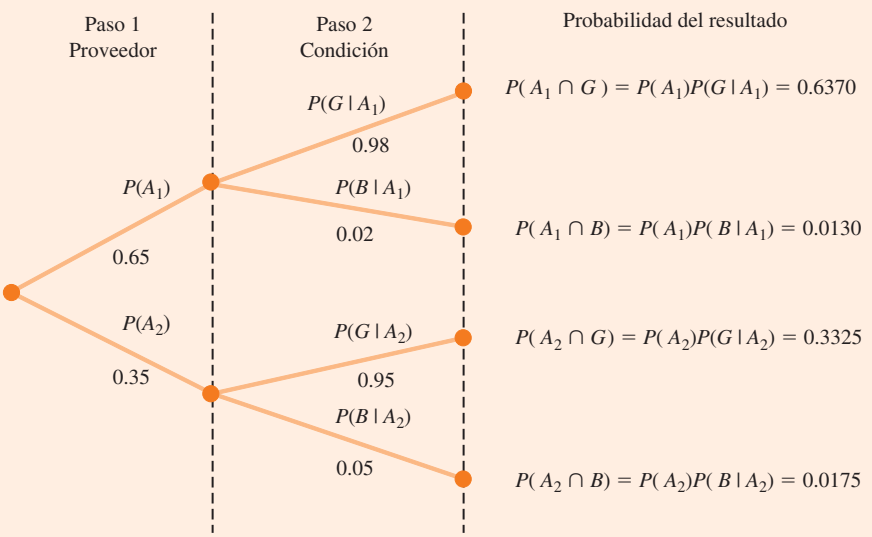
Partiendo de que B denota el evento de que la refacción se encuentra en mal estado, se buscan las probabilidades posteriores $P(A_1 \mid B)$ y $P(A_2 \mid B)$. A partir de la ley de la probabilidad condicional sabemos que

$$P(A_1 \mid B) = \frac{P(A_1 \cap B)}{P(B)} \tag{4.14}$$

Al remitirse al árbol de probabilidad, vemos que

$$P(A_1 \cap B) = P(A_1)P(B \mid A_1) \tag{4.15}$$

FIGURA 4.11 Árbol de probabilidad para el ejemplo de dos proveedores



Para obtener $P(B)$, note que el evento B puede ocurrir sólo de dos maneras: $(A_1 \cap B)$ y $(A_2 \cap B)$. Por tanto, tenemos

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) \qquad \textbf{(4.16)}$$
$$= P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)$$

Al sustituir las ecuaciones (4.15) y (4.16) en la ecuación (4.14) y escribir un resultado parecido para $P(A_2 \mid B)$, se obtiene el teorema de Bayes para el caso de dos eventos.

Al reverendo Thomas Bayes (1702-1761), ministro presbiteriano, se le atribuye el trabajo original que condujo a la versión actual en uso del teorema de Bayes.

TEOREMA DE BAYES (CASO DE DOS EVENTOS)

$$P(A_1 \mid B) = \frac{P(A_1)P(B \mid A_1)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)} \qquad \textbf{(4.17)}$$

$$P(A_2 \mid B) = \frac{P(A_2)P(B \mid A_2)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)} \qquad \textbf{(4.18)}$$

Con ayuda de la ecuación (4.17) y los valores de la probabilidad proporcionados en el ejemplo, tenemos

$$P(A_1 \mid B) = \frac{P(A_1)P(B \mid A_1)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)}$$
$$= \frac{(0.65)(0.02)}{(0.65)(0.02) + (0.35)(0.05)} = \frac{0.0130}{0.0130 + 0.0175}$$
$$= \frac{0.0130}{0.0305} = 0.4262$$

Además, con ayuda de la ecuación (4.18), se obtiene $P(A_2 \mid B)$.

$$P(A_2 \mid B) = \frac{(0.35)(0.05)}{(0.65)(0.02) + (0.35)(0.05)}$$
$$= \frac{0.0175}{0.0130 + 0.0175} = \frac{0.0175}{0.0305} = 0.5738$$

Considere que en esta aplicación se inició con una probabilidad de 0.65 de que una refacción seleccionada al azar fuera del proveedor 1. Sin embargo, dada la información de que la refacción se encuentra en mal estado, la probabilidad de que sea del proveedor 1 baja a 0.4262. De hecho, si la parte se encuentra en mal estado, tiene una posibilidad mayor que 50–50 de provenir del proveedor 2, es decir, $P(A_2 \mid B) = 0.5738$.

El teorema de Bayes es válido cuando los eventos de los que se quiere calcular las probabilidades posteriores son mutuamente excluyentes y su unión es el espacio muestral total.[2] Para el caso de los n eventos mutuamente excluyentes $A_1, A_2, \ldots, A_n$, cuya unión es el espacio muestral entero, el teorema de Bayes se utiliza para calcular cualquier probabilidad posterior $P(A_i \mid B)$ como se muestra aquí.

TEOREMA DE BAYES

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(A_1)P(\mathrm{B} \mid A_1) + P(A_2)P(B \mid A_2) + \cdots + P(A_n)P(B \mid A_n)} \qquad \textbf{(4.19)}$$

[2] Si la unión de los eventos es todo el espacio muestral entero, se dice que los eventos son *colectivamente exhaustivos*.

Con las probabilidades previas $P(A_1), P(A_2), \ldots, P(A_n)$ y las probabilidades condicionales apropiadas $P(B \mid A_1), P(B \mid A_2), \ldots, P(B \mid A_n)$, la ecuación (4.19) se usa para calcular la probabilidad posterior de los eventos $A_1, A_2, \ldots, A_n$.

Método tabular

Un método tabular es útil para efectuar los cálculos del teorema de Bayes. Un método de este tipo se muestra en la tabla 4.7 para el problema del proveedor de refacciones. Los cálculos mostrados allí se realizan con los pasos siguientes.

Paso 1. Prepare las tres columnas siguientes:
Columna 1 - Los eventos mutuamente excluyentes A_i que se desean para las probabilidades posteriores
Columna 2 - Las probabilidades previas $P(A_i)$ para los eventos
Columna 3 - Las probabilidades condicionales $P(B \mid A_i)$ de la nueva información B dada a cada evento

Paso 2. En la columna 4, calcule las probabilidades conjuntas $P(A_i \cap B)$ para cada evento y la nueva información B mediante la ley de la multiplicación. Estas probabilidades conjuntas se calculan multiplicando las probabilidades previas de la columna 2 por las probabilidades condicionales correspondientes de la columna 3, es decir, $P(A_i \cap B) = P(A_i)P(B \mid A_i)$.

Paso 3. Sume las probabilidades conjuntas de la columna 4. La suma es la probabilidad de la nueva información, $P(B)$. Por tanto, en la tabla 4.7 se ve que existe una probabilidad de 0.0130 de que la refacción provenga del proveedor 1 y se encuentre en mal estado, y una probabilidad de 0.0175 de que provenga del proveedor 2 y esté defectuosa. Debido a que estas son las dos únicas formas en que puede obtenerse una refacción en mal estado, la suma 0.0130 + 0.0175 muestra una probabilidad general de 0.0305 de encontrar una refacción en mal estado en los embarques combinados de los dos proveedores.

Paso 4. En la columna 5, calcule las probabilidades posteriores utilizando la relación básica de la probabilidad condicional.

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)}$$

Note que las probabilidades conjuntas $P(A_i \cap B)$ están en la columna 4 y la probabilidad $P(B)$ es la suma de la columna 4.

TABLA 4.7 Método tabular de los cálculos del teorema de Bayes para el problema de los dos proveedores

(1) Eventos A_i	(2) Probabilidades previas $P(A_i)$	(3) Probabilidades condicionales $P(B \mid A_i)$	(4) Probabilidades conjuntas $P(A_i \cap B)$	(5) Probabilidades posteriores $P(A_i \mid B)$
A_1	0.65	0.02	0.0130	0.0130/0.0305 = 0.4262
A_2	0.35	0.05	0.0175	0.0175/0.0305 = 0.5738
	1.00		$P(B) = 0.0305$	1.0000

NOTAS Y COMENTARIOS

1. El teorema de Bayes se utiliza ampliamente en el análisis de decisiones. Las probabilidades previas suelen ser estimaciones subjetivas proporcionadas por quien toma decisiones. Se obtiene la información muestral y las probabilidades posteriores se calculan para usarlas en la elección de la mejor decisión.
2. Un evento y su complemento son mutuamente excluyentes, y su unión es todo el espacio muestral. Por tanto, el teorema de Bayes siempre se aplica al cálculo de las probabilidades posteriores de un evento y su complemento.

Ejercicios

Métodos

AUTOevaluación

39. Las probabilidades previas para los eventos A_1 y A_2 son $P(A_1) = 0.40$ y $P(A_2) = 0.60$. También se sabe que $P(A_1 \cap A_2) = 0$. Suponga que $P(B \mid A_1) = 0.20$ y $P(B \mid A_2) = 0.05$.
    a) ¿Los eventos A_1 y A_2 son mutuamente excluyentes? Explique su respuesta.
    b) Calcule $P(A_1 \cap B)$ y $P(A_2 \cap B)$.
    c) Calcule $P(B)$.
    d) Aplique el teorema de Bayes para calcular $P(A_1 \mid B)$ y $P(A_2 \mid B)$.

40. Las probabilidades previas de los eventos A_1, A_2 y A_3 son $P(A_1) = 0.20$; $P(A_2) = 0.50$, y $P(A_3) = 0.30$. Las probabilidades condicionales para el evento B, dados A_1, A_2 y A_3 son $P(B \mid A_1) = 0.50$; $P(B \mid A_2) = 0.40$, y $P(B \mid A_3) = 0.30$.
    a) Calcule $P(B \cap A_1)$, $P(B \cap A_2)$ y $P(B \cap A_3)$.
    b) Aplique el teorema de Bayes, la ecuación 4.19, para calcular la probabilidad posterior $P(A_2 \mid B)$.
    c) Utilice el método tabular para aplicar el teorema de Bayes al cálculo de $P(A_1 \mid B)$, $P(A_2 \mid B)$ y $P(A_3 \mid B)$.

Aplicaciones

41. Una firma de consultoría presentó una licitación para un proyecto de investigación grande. La gerencia de la firma pensó en un principio que tenía una probabilidad de 50-50 de ganar el proyecto. Sin embargo, la agencia donde presentó la licitación le solicitó después información sobre su propuesta. La experiencia previa indica que en 75% de las licitaciones exitosas y en 40% de las fallidas, la agencia solicitó información adicional.
    a) ¿Cuál es la probabilidad previa de que la licitación tenga éxito (es decir, previa a la solicitud de información adicional)?
    b) ¿Cuál es la probabilidad condicional de que se solicite información adicional si al final la licitación tiene éxito?
    c) Calcule la probabilidad posterior de que la licitación tenga éxito, dado que se solicita información adicional.

AUTOevaluación

42. Un banco local revisó su política de tarjetas de crédito con la intención de cancelar algunas tarjetas. En el pasado, aproximadamente 5% de los tarjetahabientes no cumplió con sus pagos, por lo que el banco no pudo cobrar los estados de cuenta pendientes. Por tanto, la gerencia estableció una probabilidad previa de 0.05 de que cualquier tarjetahabiente en particular no pague. El banco calculó también que la probabilidad de retraso en el pago mensual era de 0.20 para aquellos clientes que sí pagaban. Desde luego, la probabilidad de retrasarse en un pago mensual para aquellos que no pagaron es 1.
    a) Dado que un cliente no realizó uno o más pagos mensuales, calcule la probabilidad posterior de que no cumpla con el pago.
    b) Al banco le gustaría retirar su tarjeta si la probabilidad de que el cliente no cumpla con el pago es mayor que 0.20. ¿El banco debe retirar la tarjeta si el cliente no hace un pago mensual? ¿Por qué?

43. Los automóviles compactos obtienen un mejor millaje con respecto al consumo de gasolina, pero no son tan seguros como los grandes. Los compactos representaron 18% de los vehículos en la carretera, pero los accidentes que involucran unidades pequeñas causaron 11 898 muertes en un año reciente (*Reader's Digest,* mayo de 2000). Suponga que la probabilidad de que un automóvil compacto esté involucrado en un accidente es de 0.18. La probabilidad de un accidente fatal con el mismo tipo de vehículo es 0.128 y la probabilidad de un percance que no provoca una muerte con un automóvil compacto es 0.05. Suponga que se entera de un accidente fatal. ¿Cuál es la probabilidad de que esté involucrado un automóvil pequeño? Considere que la probabilidad de tener un percance es independiente del tamaño del vehículo.

44. El American Council of Education informó que 47% de los estudiantes de primer año universitario obtiene su título y se gradúa en cinco años (Associated Press, 6 de mayo de 2002). Suponga que los registros de graduación muestran que las mujeres constituyen 50% de los estudiantes que se graduaron en cinco años, pero sólo 45% de los que no se graduaron en este lapso. Quienes no se habían graduado en los cinco años abandonaron la escuela o siguieron estudiando su carrera.
    a) Sean A_1 = el estudiante graduado en cinco años
    A_2 = el estudiante que no se graduó en cinco años
    W = el estudiante es mujer
    Utilizando la información aportada, ¿cuáles son los valores para $P(A_1)$, $P(A_2)$, $P(W \mid A_1)$ y $P(W \mid A_2)$?
    b) ¿Cuál es la probabilidad de que una estudiante mujer se gradúe en cinco años?
    c) ¿Cuál es la probabilidad de que un hombre se gradúe en cinco años?
    d) Dados los resultados anteriores, ¿cuál es el porcentaje de mujeres y el de hombres que asisten a la clase de primer año?

45. En un artículo sobre las alternativas de inversión, la revista *Money* informó que las acciones de fármacos proporcionan un potencial de crecimiento a largo plazo, con más de 50% de la población estadounidense adulta que toma con regularidad medicamentos por prescripción médica. Para los adultos de 65 años y mayores, 82% toma fármacos con regularidad por prescripción. Para los adultos de 18 a 64 años de edad, 49% los ingiere con regularidad por prescripción. El grupo de edades de 18 a 64 años representa 83.5% de la población adulta (*Statistical Abstract of the United States*, 2008).
    a) ¿Cuál es la probabilidad de que un adulto seleccionado al azar tenga 65 años o más?
    b) Dado que un adulto toma medicamentos por prescripción de manera regular, ¿cuál es la probabilidad de que tenga 65 años o más?

Resumen

En este capítulo se presentaron los conceptos básicos de probabilidad y se ilustró cómo se utiliza el análisis de probabilidad para proporcionar información útil en la toma de decisiones. Se describió cómo se interpreta la probabilidad como una medida numérica de la posibilidad de que un evento ocurra. Además, se vio que las probabilidades de un evento se pueden calcular ya sea sumando las probabilidades de los resultados del experimento (puntos de la muestra) que comprenden el evento, o utilizando las relaciones establecidas por la suma, la probabilidad condicional y las leyes de multiplicación de la probabilidad. Para los casos en los que hay información adicional disponible, se mostró cómo se utiliza el teorema de Bayes para obtener las probabilidades revisadas o posteriores.

Glosario

Complemento de *A* Evento que consiste en todos los puntos de la muestra que no están en A.
Diagrama de árbol Representación gráfica que ayuda en la visualización de un experimento de pasos múltiples.
Diagrama de Venn Representación gráfica para ilustrar de manera simbólica el espacio muestral y las operaciones que involucran eventos en los cuales éste se representa por medio de un rectángulo y los eventos se dibujan como círculos dentro del espacio muestral.
Espacio muestral Conjunto de todos los resultados del experimento.

Evento Colección de puntos de la muestra.
Eventos independientes Dos eventos A y B donde $P(A \mid B) = P(A)$ o $P(B \mid A) = P(B)$; es decir, los eventos no se influyen entre sí.
Eventos mutuamente excluyentes Eventos que no tienen puntos de la muestra en común; es decir, $A \cap B$ es un conjunto vacío y $P(A \cap B) = 0$.
Experimento Proceso que genera resultados bien definidos.
Intersección de *A* y *B* Evento que contiene todos los puntos de la muestra que pertenecen tanto a A como a B. La intersección se denota por medio de $A \cap B$.
Ley de la adición Ley de la probabilidad empleada para calcular la probabilidad de la unión de dos eventos. Es $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Para eventos mutuamente excluyentes, $P(A \cap B) = 0$; en este caso la ley de la adición se reduce a $P(A \cup B) = P(A) + P(B)$.
Ley de la multiplicación Ley de la probabilidad utilizada para calcular la probabilidad de la intersección de dos eventos. Es $P(A \cap B) = P(B)P(A \mid B)$ o $P(A \cap B) = P(A)P(B \mid A)$. Para eventos independientes se reduce a $P(A \cap B) = P(A)P(B)$.
Método clásico Método de asignación de probabilidades que es apropiado cuando todos los resultados del experimento son igualmente probables.
Método de frecuencia relativa Método de asignación de probabilidades que es apropiado cuando los datos están disponibles para estimar la proporción del tiempo en que el resultado ocurrirá si el experimento se repite un gran número de veces.
Método subjetivo Método de asignación de probabilidades sobre la base del juicio.
Probabilidad Medida numérica de la probabilidad de que un evento ocurra.
Probabilidad condicional Probabilidad de un evento dado que otro evento ya ha ocurrido. La probabilidad condicional de A dado B es $P(A \mid B) = P(A \cap B)/P(B)$.
Probabilidad conjunta Probabilidad de que dos eventos ocurran; es decir, la probabilidad de la intersección de dos eventos.
Probabilidad marginal Valores en los márgenes de una tabla de probabilidad conjunta que proporciona las probabilidades de cada evento por separado.
Probabilidades posteriores Probabilidades revisadas de los eventos con base en información adicional.
Probabilidades previas Estimaciones iniciales de las probabilidades de eventos.
Punto de la muestra Elemento del espacio muestral. Un punto de la muestra representa un resultado experimental.
Requisitos básicos para la asignación de probabilidades Dos requisitos que restringen la manera de efectuar las asignaciones de probabilidad: 1) para el resultado experimental E_i se debe tener $0 \leq P(E_i) \leq 1$; (2) considerando todos los resultados del experimento, se debe tener $P(E_1) + P(E_2) + \cdots + P(E_n) = 1.0$.
Teorema de Bayes Método utilizado para calcular las probabilidades posteriores.
Unión de *A* y *B* Evento que contiene todos los puntos de la muestra que pertenecen a A o B, o a ambos. La unión se representa por $A \cup B$.

Fórmulas clave

Regla de conteo para combinaciones

$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!} \tag{4.1}$$

Regla de conteo para permutaciones

$$P_n^N = n!\binom{N}{n} = \frac{N!}{(N-n)!} \tag{4.2}$$

Cálculo de la probabilidad utilizando el complemento

$$P(A) = 1 - P(A^c) \tag{4.5}$$

Ley de la adición

$$P(A \cup B) = P(A) + P(B) + P(A \cap B) \tag{4.6}$$

Probabilidad condicional

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{4.7}$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{4.8}$$

Ley de la multiplicación

$$P(A \cap B) = P(B)P(A \mid B) \tag{4.11}$$

$$P(A \cap B) = P(A)P(B \mid A) \tag{4.12}$$

Ley de la multiplicación para eventos independientes

$$P(A \cap B) = P(A)P(B) \tag{4.13}$$

Teorema de Bayes

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(A_1)P(\mathrm{B} \mid A_1) + P(A_2)P(B \mid A_2) + \cdots + P(A_n)P(B \mid A_n)} \tag{4.19}$$

Ejercicios complementarios

46. La encuesta de *The Wall Street Journal*/Harris Personal Finance preguntó a 2 082 adultos si tenían casa propia (sitio web All Business, 23 de enero de 2008). Un total de 1 249 encuestados respondió Sí. De los 450 encuestados en el grupo de edades de 18 a 34 años, 117 respondieron Sí.
    a) ¿Cuál es la probabilidad de que un encuestado tenga casa propia?
    b) ¿Cuál es la probabilidad de que una persona del grupo de edades de 18 a 34 años tenga vivienda propia?
    c) ¿Cuál es la probabilidad de que un encuestado no tenga casa propia?
    d) ¿Cuál es la probabilidad de que una persona del grupo de edades de 18 a 34 años no tenga vivienda propia?

47. Un ejecutivo de finanzas hizo dos inversiones nuevas: una en la industria del petróleo y otra en bonos municipales. Después de un año, cada una de las inversiones se clasificará como exitosa o sin éxito. Considere como un experimento efectuar las dos inversiones.
    a) ¿Cuántos puntos de la muestra existen para este experimento?
    b) Muestre un diagrama de árbol y liste los puntos de la muestra.
    c) Sea O = al evento de que la inversión en la industria del petróleo es exitosa y M = el evento de que la inversión en bonos municipales es exitosa. Elabore una lista de los puntos de la muestra en O y M.
    d) Liste los puntos de la muestra en la unión de los eventos ($O \cup M$).
    e) Elabore una lista de los puntos de la muestra en la intersección de los eventos ($O \cap M$).
    f) ¿O y M son mutuamente excluyentes? Explique.

48. A principios de 2003, el presidente Bush propuso eliminar los impuestos de dividendos a los accionistas sobre la base de que era un doble gravamen. Las corporaciones pagan impuestos sobre las ganancias que después pagan en dividendos. En una encuesta a 671 estadounidenses, TechnoMetrica Market Intelligence encontró que 47% estuvo a favor de la propuesta, 44% se opuso y 9% no estaba seguro (*Investor's Business Daily,* 13 de enero de 2003). Al mirar

las respuestas en todas las políticas de partidos, la encuesta reveló que estaban a favor 29% de los demócratas, 64% de los republicanos y 48% de los independientes.

a) ¿Cuántos de los encuestados estaban a favor de la eliminación de los impuestos sobre los dividendos?

b) ¿Cuál es la probabilidad condicional a favor de la propuesta dado que la persona encuestada es un demócrata?

c) ¿La afiliación a un partido es independiente de si una persona está a favor de la propuesta?

d) Si asume que las respuestas de las personas concordaron con sus intereses personales, ¿cuál grupo cree que se beneficiará más de la aprobación de la propuesta?

49. Un estudio de 31 000 admisiones en los hospitales del estado de Nueva York reveló que 4% de los ingresos condujo a lesiones causadas por los tratamientos; un séptimo de estas lesiones desencadenadas por los tratamientos ocasionó la muerte, y un cuarto fue causado por negligencia. Las demandas por negligencia médica se presentan en uno de cada 7.5 casos y los pagos se efectúan en una de cada dos demandas.

a) ¿Cuál es la probabilidad de que una persona admitida en el hospital sufra una lesión causada por un tratamiento debido a negligencia?

b) ¿Cuál es la probabilidad de que una persona ingresada muera por una lesión causada por un tratamiento?

c) En el caso de una lesión ocasionada por un tratamiento, ¿cuál es la probabilidad de que una demanda por negligencia sea pagada?

50. Una encuesta por teléfono para determinar la respuesta de los espectadores a un nuevo programa de televisión arrojó los datos siguientes.

Calificación	Frecuencia
Mala	4
Debajo del promedio	8
Promedio	11
Arriba del promedio	14
Excelente	13

a) ¿Cuál es la probabilidad de que un espectador seleccionado al azar califique el nuevo programa como promedio o mejor?

b) ¿Cuál es la probabilidad de que un televidente seleccionado al azar califique el nuevo programa por debajo del promedio o peor?

51. La tabulación cruzada siguiente muestra los ingresos familiares por nivel educativo del jefe de familia (*Statistical Abstract of the United States*, 2008).

	Ingresos familiares (miles $)					
Nivel educativo	**Menos de 25**	**25.0–49.9**	**50.0–74.9**	**75.0–99.9**	**100 o más**	**Total**
Sin educación media	4 207	3 459	1 389	539	367	9 961
Educación media	4 917	6 850	5 027	2 637	2 668	22 099
Educación superior inconclusa	2 807	5 258	4 678	3 250	4 074	20 067
Educación superior	885	2 094	2 848	2 581	5 379	13 787
Maestría o doctorado	290	829	1 274	1 241	4 188	7 822
Total	13 106	18 490	15 216	10 248	16 676	73 736

a) Elabore una tabla de probabilidad conjunta.

b) ¿Cuál es la probabilidad de que un jefe de familia no cuente con educación media?

c) ¿Cuál es la probabilidad de que tenga educación superior o una maestría o doctorado?

d) ¿Cuál es la probabilidad de que una familia dirigida por una persona con educación superior gane $100 000 o más?

e) ¿Cuál es la probabilidad de que una familia tenga un ingreso inferior a $25 000?
f) ¿Cuál es la probabilidad de que una familia dirigida por alguien con educación superior gane menos de $25 000?
g) ¿El ingreso familiar es independiente del nivel educativo?

52. Una encuesta de los nuevos estudiantes inscritos en una maestría proporcionó los datos siguientes para 2 018 estudiantes.

Grupo de edades	Aplicado a más de una escuela: Sí	Aplicado a más de una escuela: No
23 y menores	207	201
24–26	299	379
27–30	185	268
31–35	66	193
36 y mayores	51	169

a) Para un estudiante de maestría seleccionado al azar, prepare una tabla de probabilidad conjunta del experimento que consiste en observar su edad y si presentó una solicitud de inscripción a una o más escuelas.
b) ¿Cuál es la probabilidad de que un aspirante seleccionado al azar tenga 23 años o menos?
c) ¿Cuál es la probabilidad de que sea mayor de 26?
d) ¿Cuál es la probabilidad de que haya presentado su solicitud en más de una escuela?

53. Vuelva a observar los datos del ejercicio 52 de la encuesta de estudiantes recién inscritos en la maestría.
a) Dado que una persona presentó su solicitud en más de una escuela, ¿cuál es la probabilidad de que tenga de 24 a 26 años de edad?
b) Dado que un estudiante está en el grupo de edades de 36 años y más, ¿cuál es la probabilidad de que haya presentado su solicitud en más de un colegio?
c) ¿Cuál es la probabilidad de que una persona tenga de 24 a 26 años de edad o haya presentado su solicitud en más de una escuela?
d) Suponga que se sabe que un aspirante ha presentado su solicitud de ingreso sólo en una escuela. ¿Cuál es la probabilidad de que éste tenga 31 años o más?
e) ¿El número de escuelas en las que se presenta una solicitud es independiente de la edad? Explique por qué.

54. Una encuesta de IBD/TIPP realizada para conocer las actitudes hacia la inversión y el retiro (*Investor's Business Daily,* 5 de mayo de 2000) preguntó a hombres y mujeres qué tan importante consideraban el nivel de riesgo en la elección de una inversión para el retiro. La tabla de probabilidad conjunta siguiente se construyó a partir de los datos facilitados. "Importante" significa que el encuestado dijo que el nivel de riesgo era importante o muy importante.

	Hombre	Mujer	Total
Importante	0.22	0.27	0.49
No importante	0.28	0.23	0.51
Total	0.50	0.50	1.00

a) ¿Cuál es la probabilidad de que un encuestado diga que el nivel de riesgo es importante?
b) ¿Cuál es la probabilidad de que un hombre consultado responda que el nivel de riesgo es importante?
c) ¿Cuál es la probabilidad de que una mujer encuestada responda que el nivel de riesgo es importante?
d) ¿El nivel de riesgo es independiente del sexo de la persona consultada? ¿Por qué?
e) ¿Difieren las actitudes de los hombres y de las mujeres hacia el riesgo?

55. Una empresa de bienes de consumo publicó un anuncio de televisión para uno de sus productos de jabón. Sobre la base de una encuesta que se realizó, se asignaron las probabilidades a los eventos siguientes.

$$B = \text{la persona compró el producto}$$
$$S = \text{la persona recuerda haber visto el anuncio}$$
$$B \cap S = \text{la persona adquirió el producto y recuerda haber visto el anuncio}$$

Las probabilidades asignadas fueron $P(B) = 0.20$; $P(S) = 0.40$, y $P(B \cap S) = 0.12$.

a) ¿Cuál es la probabilidad de que una persona adquiera el producto dado que recuerda haber visto el anuncio? ¿Ver el anuncio aumenta la probabilidad de que compre el producto? Como alguien que toma decisiones, ¿recomendaría seguir transmitiendo el anuncio (asumiendo que el costo es razonable)?

b) Suponga que las personas que no adquieren el producto de jabón de la empresa lo compran a sus competidores. ¿Cuál sería su estimación de la cuota de mercado de la empresa? ¿Esperaría usted que seguir transmitiendo el anuncio aumente su participación de mercado? ¿Por qué?

c) La empresa también probó otro anuncio y le asignó los valores de $P(S) = 0.30$ y $P(B \cap S) = 0.10$. ¿Cuál es la probabilidad conjunta $P(B \mid S)$ de este otro anuncio? ¿Cuál comercial parece haber tenido el efecto más grande sobre las compras de los clientes?

56. Cooper Realty es una pequeña compañía de bienes raíces ubicada en Albany, Nueva York, que se especializa principalmente en listados residenciales. Recientemente se interesó en determinar la probabilidad de que uno de sus listados se vendiera en cierto número de días. Un análisis de las ventas de la empresa de 800 casas en años anteriores arrojó los datos siguientes.

		Días en el listado hasta la venta			
		Menos de 30	**31–90**	**Más de 90**	**Total**
Precio de oferta inicial	**Menos de $150 000**	50	40	10	100
	$150 000–$199 999	20	150	80	250
	$200 000–$250 000	20	280	100	400
	Más de $250 000	10	30	10	50
	Total	100	500	200	800

a) Si A se define como el evento de que una casa aparezca en el listado por más de 90 días antes de ser vendida, calcule la probabilidad de A.

b) Si B se define como el evento de que el precio de oferta inicial sea menor de $150 000, calcule la probabilidad de B.

c) ¿Cuál es la probabilidad de $A \cap B$?

d) Suponiendo que un contrato se acaba de firmar para listar una casa con un precio inicial de menos de $150 000, ¿cuál es la probabilidad de que Cooper Realty tarde más de 90 días en venderla?

e) ¿Los eventos A y B son independientes?

57. Una empresa estudió el número de accidentes que generaron pérdida de tiempo en la planta de Brownsville, Texas. Los registros históricos muestran que 6% de los empleados tuvo accidentes que generaron una pérdida de tiempo el año pasado. La gerencia cree que un programa especial de seguridad reducirá los percances a 5% durante el año en curso. Además, estima que 15% de los empleados que sufrió este tipo de accidentes el año anterior sufrirá uno que generará pérdida de tiempo durante el año en curso.

a) ¿Qué porcentaje de los empleados tendrá accidentes que generen una pérdida de tiempo en los dos años?

b) ¿Qué porcentaje tendrá por lo menos un accidente que cause una pérdida de tiempo en el periodo de dos años?

58. Una encuesta reveló que 8% de los usuarios de Internet que tienen 18 años o más informan que mantienen un *blog*. Refiriéndose al grupo de edades de 18 a 29 años como adultos jóvenes, la encuesta reveló que, de los *bloggers*, 54% son adultos jóvenes y de los no *bloggers*, 24% son adultos jóvenes (Pew Internet & American Life Project, 19 de julio de 2006).
    a) Elabore una tabla de probabilidad conjunta para estos dos datos con dos filas (*bloggers* en comparación con no *bloggers*) y dos columnas (adultos jóvenes frente a adultos mayores).
    b) ¿Cuál es la probabilidad de que un usuario de Internet sea un adulto joven?
    c) ¿Cuál es la probabilidad de que un internauta mantenga un *blog* y sea un adulto joven?
    d) Suponga que en una encuesta telefónica de seguimiento se contactó a un adulto de 24 años de edad. ¿Cuál es la probabilidad de que esta persona mantenga un *blog*?

59. Una compañía petrolera compró un terreno en Alaska. Los estudios geológicos preliminares asignaron las probabilidades previas siguientes.

$$P(\text{petróleo de alta calidad}) = 0.50$$
$$P(\text{petróleo de calidad media}) = 0.20$$
$$P(\text{sin encontrar petróleo}) = 0.30$$

    a) ¿Cuál es la probabilidad de encontrar petróleo?
    b) Una prueba de suelo es tomada después de 200 pies de perforación del primer pozo. Las probabilidades de encontrar un tipo particular de suelo se identifica con una prueba de seguimiento.

$$P(\text{suelo} \mid \text{petróleo de alta calidad}) = 0.20$$
$$P(\text{suelo} \mid \text{petróleo de calidad media}) = 0.80$$
$$P(\text{suelo} \mid \text{sin encontrar petróleo}) = 0.20$$

    ¿Cómo debe interpretar la empresa la prueba de suelo? ¿Cuáles son las probabilidades revisadas y cuál es la nueva probabilidad de encontrar petróleo?

60. Las empresas que hacen negocios por Internet a menudo obtienen información acerca de los visitantes a la Web a partir de las páginas ya visitadas. El artículo "Internet Marketing" (*Interfaces,* marzo/abril de 2001) describe cómo se usan los datos sobre el flujo de *clics* en las páginas web consultadas, junto con un sistema bayesiano de actualización para determinar el género de un visitante. ParFore creó una página web para comercializar equipo y ropa de golf. A la gerencia le gustaría que cierta oferta apareciera para las mujeres que visitan la página y que se exhibiera una oferta distinta para los hombres. A partir una muestra de consultas anteriores a otras páginas web, la gerencia se enteró de que 60% de quienes consultan la página de ParFore son hombres y 40% mujeres.
    a) ¿Cuál es la probabilidad previa de que el próximo visitante de la página web sea una mujer?
    b) Suponga que sabe que el visitante actual al sitio web de ParFore antes de abrir su página visitó la página de Dillard's, y que las mujeres tienen tres veces más probabilidades de consultar la página de Dillard's que los hombres. ¿Cuál es la probabilidad revisada de que el visitante actual a la página web de ParFore sea del género femenino? ¿Debe usted mostrar la oferta que atrae a más visitantes mujeres o la que atrae a más visitantes varones?

Caso a resolver Jueces del condado de Hamilton

Los jueces (Judges) del condado de Hamilton procesan miles de casos al año. En la gran mayoría de los casos desechados, el veredicto permanece como se presentó. Sin embargo, algunos son apelados y de éstos algunos se revocan. Kristen DelGuzzi, del diario *Cincinnati Enquirer*, realizó un estudio de los casos manejados por los jueces del condado de Hamilton durante un periodo de tres años (*Cincinnati Enquirer*, 11 de enero de 1998). En la tabla 4.8 se muestran los resultados de 182 908 casos manejados (disposed) por 38 jueces del tribunal de primera instancia (Common Pleas Court), del tribunal de lo familiar (Domestic Relations Court) y del tribunal municipal (Municipal Court). Dos de los jueces (Dinkelacker y Hogan) no trabajaron en el mismo tribunal durante los tres años.

TABLA 4.8 Total de casos desechados, apelados y revocados en los tribunales del condado de hamilton

WEB archivo

Judge

Common Pleas Court

Judge	Total Cases Disposed	Appealed Cases	Reversed Cases
Fred Cartolano	3037	137	12
Thomas Crush	3372	119	10
Patrick Dinkelacker	1258	44	8
Timothy Hogan	1954	60	7
Robert Kraft	3138	127	7
William Mathews	2264	91	18
William Morrissey	3032	121	22
Norbert Nadel	2959	131	20
Arthur Ney, Jr.	3219	125	14
Richard Niehaus	3353	137	16
Thomas Nurre	3000	121	6
John O'Connor	2969	129	12
Robert Ruehlman	3205	145	18
J. Howard Sundermann	955	60	10
Ann Marie Tracey	3141	127	13
Ralph Winkler	3089	88	6
Total	43945	1762	199

Domestic Relations Court

Judge	Total Cases Disposed	Appealed Cases	Reversed Cases
Penelope Cunningham	2729	7	1
Patrick Dinkelacker	6001	19	4
Deborah Gaines	8799	48	9
Ronald Panioto	12970	32	3
Total	30499	106	17

Municipal Court

Judge	Total Cases Disposed	Appealed Cases	Reversed Cases
Mike Allen	6149	43	4
Nadine Allen	7812	34	6
Timothy Black	7954	41	6
David Davis	7736	43	5
Leslie Isaiah Gaines	5282	35	13
Karla Grady	5253	6	0
Deidra Hair	2532	5	0
Dennis Helmick	7900	29	5
Timothy Hogan	2308	13	2
James Patrick Kenney	2798	6	1
Joseph Luebbers	4698	25	8
William Mallory	8277	38	9
Melba Marsh	8219	34	7
Beth Mattingly	2971	13	1
Albert Mestemaker	4975	28	9
Mark Painter	2239	7	3
Jack Rosen	7790	41	13
Mark Schweikert	5403	33	6
David Stockdale	5371	22	4
John A. West	2797	4	2
Total	108464	500	104

El propósito del estudio del periódico es evaluar el desempeño de los jueces. Las apelaciones con frecuencia son el resultado de los errores cometidos por éstos, y el periódico quería saber cuáles de ellos hacían un buen trabajo y cuáles cometían demasiados errores. A usted le llaman para que ayude en el análisis de datos. Utilice sus conocimientos de probabilidad y probabilidad condicional para ayudar a calificar a los jueces. Tal vez pueda analizar la probabilidad de los casos manejados en los diferentes tribunales que fueron apelados y revocados.

Informe gerencial

Elabore un informe con sus calificaciones de los jueces. Incluya también un análisis de la probabilidad de apelación y la revocación de casos en los tres tribunales. Como mínimo, su informe debe incluir lo siguiente:

1. La probabilidad de casos apelados (Appealed Cases) y revocados (Reversed Cases) en los tres tribunales.
2. La probabilidad de que un caso sea apelado, por cada juez.
3. La probabilidad de que un caso sea revocado, por cada juez.
4. La probabilidad de una revocación, dada una apelación, por cada juez.
5. Una clasificación de los jueces dentro de cada tribunal. Establezca los criterios que manejó y las razones de su elección.

CAPÍTULO 5

Distribuciones de probabilidad discreta

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: CITIBANK

5.1 VARIABLES ALEATORIAS
Variables aleatorias discretas
Variables aleatorias continuas

5.2 DISTRIBUCIONES DE PROBABILIDAD DISCRETA

5.3 VALOR ESPERADO Y VARIANZA
Valor esperado
Varianza

5.4 DISTRIBUCIÓN DE PROBABILIDAD BINOMIAL
Un experimento binomial
El problema de Martin Clothing Store
Uso de tablas de probabilidades binomiales
Valor esperado y varianza de la distribución binomial

5.5 DISTRIBUCIÓN DE PROBABILIDAD DE POISSON
Un ejemplo con intervalos de tiempo
Un ejemplo con intervalos de longitud o de distancia

5.6 DISTRIBUCIÓN DE PROBABILIDAD HIPERGEOMÉTRICA

ESTADÍSTICA *en* LA PRÁCTICA

CITIBANK*

LONG ISLAND CITY, NUEVA YORK

Citibank, la división de banca minorista de Citigroup, presta una amplia gama de servicios financieros que incluyen cuentas corrientes y de ahorro, préstamos e hipotecas, seguros y servicios de inversión. Ofrece estos servicios por medio de un sistema único llamado *Citibanking*.

Citibank fue uno de los primeros bancos de Estados Unidos en introducir los cajeros automáticos (ATM). Estos dispositivos, ubicados en los centros bancarios Citicard (CBC), permiten a los clientes realizar todas sus operaciones bancarias en un solo lugar con el toque de un dedo, las 24 horas del día, los 7 días de la semana. Más de 150 funciones diferentes, que varían de depósitos a manejo de inversiones, pueden realizarse con facilidad. Los clientes de Citibank utilizan cajeros automáticos para 80% de sus transacciones.

Cada CBC opera como un sistema de fila de espera al que los clientes llegan en forma aleatoria a solicitar un servicio en uno de los cajeros automáticos. Si todos los cajeros están ocupados, los clientes que llegan esperan en fila. De manera periódica se realizan estudios de la capacidad del CBC para analizar los tiempos de espera de los usuarios y determinar si se requieren más cajeros automáticos.

Los datos recabados por Citibank mostraron que la llegada aleatoria de los clientes sigue una distribución de probabilidad conocida como *distribución de Poisson*. Mediante esta distribución, Citibank puede calcular las probabilidades del número de personas que llegan a un CBC durante cualquier periodo y tomar decisiones sobre el número de cajeros automáticos que se necesitan. Por ejemplo, x es el número de personas que llegan durante un periodo de un minuto. Suponiendo que un CBC decompletado tiene una tasa media de dos clientes por minuto, la tabla siguiente muestra las probabilidades del número de usuarios que podrían llegar durante un periodo de un minuto.



Un cajero automático vanguardista de Citibank.
© Jeff Greenberg/Photo Edit.

x	Probabilidad
0	0.1353
1	0.2707
2	0.2707
3	0.1804
4	0.0902
5 o más	0.0527

Las distribuciones de probabilidad discreta como la utilizada por Citibank son el tema de este capítulo. Además de la distribución de Poisson, usted aprenderá acerca de las distribuciones binomial e hipergeométrica y cómo se utilizan para proporcionar información útil de probabilidad.

* Los autores agradecen a Stacey Karter, de Citibank, por proporcionar este artículo para *Estadística en la práctica*.

Este capítulo continúa con el estudio de la probabilidad mediante la introducción de los conceptos variables aleatorias y distribuciones de probabilidad. El tema central son las distribuciones de probabilidad discreta. En particular se cubren tres distribuciones de este tipo: binomial, de Poisson e hipergeométrica.

5.1 Variables aleatorias

En el capítulo 4 se define el concepto de *experimento* y los resultados experimentales correspondientes. Una variable aleatoria proporciona un medio para describir estos resultados con valores numéricos. Las variables aleatorias deben asumir valores numéricos.

Las variables aleatorias deben asumir valores numéricos.

VARIABLE ALEATORIA

Una **variable aleatoria** es una descripción numérica de los resultados de un experimento.

En efecto, una variable aleatoria asocia un valor numérico con cada resultado experimental posible. El valor numérico particular de la variable aleatoria depende del resultado del experimento. Ésta se clasifica como *discreta* o *continua* en función de los valores numéricos que asume.

Variables aleatorias discretas

Una variable aleatoria que puede asumir cualquier número finito de valores o una sucesión infinita de valores como 0, 1, 2, . . . se conoce como **variable aleatoria discreta**. Por ejemplo, considere el experimento de un sujeto que presenta el examen de certificación de contador público, el cual consta de cuatro partes. Una variable aleatoria se define como x = el número de partes del examen aprobadas. Se trata de una variable aleatoria discreta, ya que puede asumir un número finito de valores 0, 1, 2, 3 o 4.

En otro ejemplo, considere el experimento de los automóviles que llegan a una caseta de cobro. La variable aleatoria de interés es x = el número de vehículos que llegan durante un periodo de un día. Los valores posibles para x provienen de la secuencia de números enteros 0, 1, 2, etc. Por consiguiente, x es una variable aleatoria discreta que asume uno de los valores de esta secuencia infinita.

Aunque los resultados de muchos experimentos se describen de manera natural por medio de valores numéricos, otros no pueden describirse así. Por ejemplo, en una encuesta se podría preguntar a una persona si recuerda el mensaje de un comercial de televisión reciente. Este experimento tendría dos resultados posibles: la persona no recuerda el mensaje y la persona recuerda el mensaje. También es posible describir numéricamente estos resultados experimentales mediante la definición de la variable aleatoria discreta x como sigue: sea $x = 0$ si la persona no recuerda el mensaje y $x = 1$ si la persona recuerda el mensaje. Los valores numéricos de esta variable son arbitrarios (se podría usar 5 y 10), pero son aceptables con base en la definición de una variable, es decir, x *es* una variable aleatoria, ya que proporciona una descripción numérica de los resultados del experimento.

La tabla 5.1 muestra algunos ejemplos de variables aleatorias discretas. Tenga en cuenta que en cada ejemplo la variable asume un número finito de valores o una secuencia infinita de valores como 0, 1, 2, . . . Estos tipos de variables se estudian con detalle en este capítulo.

TABLA 5.1 Ejemplos de variables aleatorias discretas

Experimento	Variable aleatoria (x)	Valores posibles de la variable aleatoria
Llamar a cinco clientes	Número de clientes que hacen un pedido	0, 1, 2, 3, 4, 5
Inspeccionar un embarque de 50 radios	Número de radios defectuosos	0, 1, 2, $\cdots$, 49, 50
Encargarse de un restaurante por un día	Número de clientes	0, 1, 2, 3, $\cdots$
Vender un automóvil	Género del cliente	0 si es hombre, 1 si es mujer

Variables aleatorias continuas

Una variable aleatoria que asume cualquier valor numérico en un intervalo o conjunto de intervalos se llama **variable aleatoria continua**. Los resultados experimentales basados en escalas de medición como el tiempo, el peso, la distancia y la temperatura se describen por medio de este tipo de variable. Por ejemplo, considere un experimento en el que se monitorean las llamadas telefónicas que llegan a la oficina de reclamaciones de una compañía de seguros importante. Suponga que la variable aleatoria de interés es x = tiempo entre las llamadas entrantes consecutivas en minutos. Esta variable puede asumir cualquier valor en el intervalo $x \geq 0$. En realidad, x puede asumir un número infinito de valores, incluidos algunos como 1.26 minutos, 2.751 minutos, 4.3333 minutos, etc. Otro ejemplo es un tramo de 90 millas de la carretera interestatal I-75 al norte de Atlanta, Georgia. Para un servicio de ambulancias de emergencia ubicado en Atlanta, la variable aleatoria podría definirse como x = número de millas al lugar del siguiente accidente de tránsito a lo largo del tramo de la carretera I-75. En este caso, x sería una variable aleatoria continua que asume cualquier valor en el intervalo $0 \leq x \leq 90$. La tabla 5.2 presenta otros ejemplos de variables aleatorias continuas. Observe que cada ejemplo describe una variable que asume cualquier valor en un intervalo de valores. Las variables aleatorias continuas y sus distribuciones de probabilidad serán el tema del capítulo 6.

TABLA 5.2 Ejemplos de variables aleatorias continuas

Experimento	Variable aleatoria (x)	Valores posibles de la variable aleatoria
Operar un banco	Tiempo entre las llegadas de los clientes, en minutos	$x \geq 0$
Llenar una lata de refresco (máx. = 12.1 onzas)	Cantidad de onzas	$0 \leq x \leq 12.1$
Construir una biblioteca	Porcentaje del proyecto completado después de seis meses	$0 \leq x \leq 100$
Probar un proceso químico nuevo	Temperatura a la que ocurre la reacción (mín. 150 °F; máx. 212 °F)	$150 \leq x \leq 212$

NOTAS Y COMENTARIOS

Una forma de determinar si una variable aleatoria es discreta o continua es pensar en sus valores como puntos en un segmento de recta. Elija dos puntos que representen valores de la variable aleatoria. Si todo el segmento de recta entre los dos puntos representa también los valores posibles de la variable aleatoria, entonces ésta es continua.

Ejercicios

Métodos

AUTOevaluación

1. Considere el experimento de lanzar una moneda dos veces.
   a) Elabore una lista de los resultados experimentales.
   b) Defina una variable aleatoria que represente el número de caras que caen en los dos lanzamientos.
   c) Muestre el valor que la variable aleatoria asumiría en cada uno de los resultados experimentales.
   d) ¿Esta variable aleatoria es discreta o continua?

2. Considere el experimento de un trabajador que ensambla un producto.
   a) Defina una variable aleatoria que represente el tiempo en minutos requerido para ensamblar el producto.
   b) ¿Qué valores puede asumir la variable aleatoria?
   c) ¿La variable es discreta o continua?

Aplicaciones

AUTOevaluación

3. Tres estudiantes programaron entrevistas para un empleo de verano en el Instituto Brookwood. En cada caso el resultado de la entrevista será una oferta de empleo o ninguna oferta. Los resultados experimentales se definen en función de los resultados de las tres entrevistas.
   a) Prepare una lista de los resultados experimentales.
   b) Defina una variable aleatoria que representa el número de ofertas de empleo formuladas. ¿La variable aleatoria es continua?
   c) Muestre el valor de la variable aleatoria para cada uno de los resultados experimentales.
4. En noviembre la tasa de desempleo estadounidense fue de 4.5% (*USA Today,* 4 de enero de 2007). La Oficina del Censo incluye nueve estados de la región noreste. Suponga que la variable aleatoria de interés es el número de estados que tuvieron una tasa de desempleo en noviembre menor de 4.5%. ¿Qué valores puede tomar esta variable aleatoria?
5. Para realizar cierto tipo de análisis de sangre, los técnicos deben llevar a cabo dos procedimientos. El primero requiere uno o dos pasos, y el segundo requiere ya sea uno, dos o tres pasos.
   a) Elabore una lista de los resultados experimentales asociados con el análisis de sangre.
   b) Si la variable aleatoria de interés es el número total de pasos requeridos para hacer el análisis completo (ambos procedimientos), determine qué valor asumirá la variable aleatoria en cada uno de los resultados experimentales.
6. Enseguida se proporciona una serie de experimentos y sus variables aleatorias asociadas. En cada caso, determine los valores que la variable aleatoria puede asumir y si es discreta o continua.

Experimento	Variable aleatoria (x)
a) Presentar un examen de 20 preguntas	Número de preguntas respondidas correctamente
b) Observar los automóviles que llegan a una caseta de cobro durante 1 hora	Número de automóviles que llegan a la caseta
c) Auditar 50 devoluciones de impuestos	Número de devoluciones que contienen errores
d) Observar el trabajo de un empleado	Número de horas improductivas en una jornada de 8 horas
e) Pesar un embarque de mercancías	Número de libras

5.2 Distribuciones de probabilidad discreta

La **distribución de probabilidad** de una variable aleatoria describe cómo se distribuyen las probabilidades entre los valores de la misma. Para una variable aleatoria discreta x, la distribución de probabilidad se define por medio de una **función de probabilidad**, denotada por $f(x)$. La función de probabilidad proporciona la probabilidad para cada valor que puede asumir la variable aleatoria.

Como ejemplo de una variable aleatoria discreta y su distribución de probabilidad, considere las ventas de automóviles en DiCarlo Motors, con sede en Saratoga, Nueva York. Durante los últimos 300 días de operación, los datos de ventas mostraron que en 54 días no se vendió ningún automóvil, en 117 días se vendió 1 automóvil, en 72 días se vendieron 2, en 42 días se vendieron 3, en 12 días se vendieron 4 y en 3 días se vendieron 5. Suponga que se considera el experimento de seleccionar un día de operación en DiCarlo Motors y se define la variable aleatoria de interés como x = número de automóviles vendidos en un día. A partir de los datos

históricos, sabemos que x es una variable aleatoria discreta que puede asumir los valores 0, 1, 2, 3, 4 o 5. En la notación de la función de probabilidad, $f(0)$ es la probabilidad de vender 0 unidades, $f(1)$ es la probabilidad de vender 1 automóvil, y así sucesivamente. Dado que los datos históricos muestran que en 54 de los 300 días se vendieron 0 unidades, se asigna el valor $54/300 = 0.18$ a $f(0)$, lo que indica que la probabilidad de que se vendan 0 automóviles en un día es de 0.18. Asimismo, como en 117 de los 300 días se vendió un vehículo, se asigna el valor $117/300 = 0.39$ a $f(1)$, indicando que la probabilidad de que se venda exactamente 1 automóvil en un día es de 0.39. Si se continúa de esta manera para los otros valores de la variable aleatoria, obtenemos los valores de $f(2)$, $f(3)$, $f(4)$ y $f(5)$ como muestra la tabla 5.3, que es la distribución de probabilidad para el número de vehículos vendidos durante un día en DiCarlo Motors.

Una de las principales ventajas de definir una variable aleatoria y su distribución de probabilidad es que, una vez que se conoce esta última, es relativamente fácil determinar la probabilidad de una variedad de eventos que pueden ser útiles para quien toma decisiones. Por ejemplo, utilizando la distribución de probabilidad para DiCarlo Motors que aparece en la tabla 5.3, vemos que el número de automóviles que es más probable vender en un día es 1, con una probabilidad de $f(1) = 0.39$. Además, hay una probabilidad de $f(3) + f(4) + f(5) = 0.14 + 0.04 + 0.01 = 0.19$ de vender 3 o más unidades durante un día. Estas probabilidades, además de otras que quien toma decisiones puede solicitar, proporcionan información que le ayudan a entender el proceso de la venta de automóviles en DiCarlo Motors.

Cuando se desarrolla una función de probabilidad para una variable aleatoria discreta, se deben satisfacer las dos condiciones siguientes.

Estas condiciones son análogas a los dos requerimientos básicos para asignar probabilidades a los resultados experimentales presentados en el capítulo 4.

CONDICIONES REQUERIDAS PARA UNA FUNCIÓN DE PROBABILIDAD DISCRETA

$$f(x) \geq 0 \qquad \textbf{(5.1)}$$

$$\Sigma f(x) = 1 \qquad \textbf{(5.2)}$$

La tabla 5.3 muestra que las probabilidades de la variable aleatoria x satisfacen la ecuación (5.1); $f(x)$ es mayor o igual que 0 para todos los valores de x. Además, como estas probabilidades suman 1, la ecuación (5.2) también se satisface. Por tanto, la función de probabilidad de DiCarlo Motors es una función de probabilidad discreta válida.

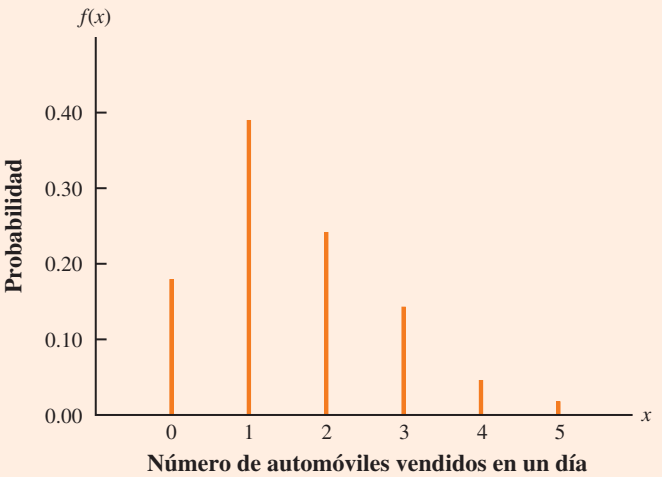
También se presentan las distribuciones de probabilidad de manera gráfica. En la figura 5.1 los valores de la variable aleatoria x para DiCarlo Motors aparecen en el eje horizontal y la probabilidad asociada con estos valores se muestra en el eje vertical.

Además de tablas y gráficas para describir las distribuciones de probabilidad, con frecuencia se utiliza una fórmula que proporciona la función de probabilidad, $f(x)$, para cada valor de

TABLA 5.3 Distribución de probabilidad para el número de automóviles vendidos durante un día en Dicarlo Motors

x	$f(x)$
0	0.18
1	0.39
2	0.24
3	0.14
4	0.04
5	0.01
Total	1.00

FIGURA 5.1 Representación gráfica de la distribución de probabilidad para el número de automóviles vendidos durante un día en Dicarlo Motors



x. El ejemplo más sencillo de una distribución de probabilidad discreta dada una fórmula, es la **distribución de probabilidad uniforme discreta**. Su función de probabilidad se define por medio de la ecuación (5.3).

FUNCIÓN DE PROBABILIDAD UNIFORME DISCRETA

$$f(x) = 1/n \tag{5.3}$$

Donde:

n = número de valores que la variable aleatoria puede asumir.

Por ejemplo, suponga que para el experimento de lanzar un dado la variable aleatoria x se define como el número de puntos en la cara que queda hacia arriba. Para este experimento, $n = 6$ valores son posibles para la variable aleatoria; $x = 1, 2, 3, 4, 5, 6$. Por tanto, la función de probabilidad para esta variable aleatoria uniforme discreta es

$$f(x) = 1/6 \qquad x = 1, 2, 3, 4, 5, 6$$

Los valores posibles de la variable aleatoria y las probabilidades asociadas se muestran en seguida.

x	$f(x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Como otro ejemplo, considere la variable aleatoria x con la distribución de probabilidad siguiente.

x	$f(x)$
1	1/10
2	2/10
3	3/10
4	4/10

Esta distribución de probabilidad se define por medio de la fórmula

$$f(x) = \frac{x}{10} \qquad \text{para } x = 1, 2, 3 \text{ o } 4$$

La evaluación de $f(x)$ para un valor dado de la variable aleatoria proporciona la probabilidad asociada. Por ejemplo, usando la función de probabilidad anterior, vemos que $f(2) = 2/10$ proporciona la probabilidad de que la variable aleatoria asuma el valor 2.

Las distribuciones de probabilidad discretas de uso más común por lo general se especifican por medio de fórmulas. Tres casos importantes son las distribuciones binomial, de Poisson e hipergeométrica, las cuales se estudian posteriormente en este capítulo.

Ejercicios

Métodos

AUTO evaluación

7. La distribución de probabilidad para la variable aleatoria x se presenta enseguida.

x	$f(x)$
20	0.20
25	0.15
30	0.25
35	0.40

a) ¿Es válida esta distribución de probabilidad? Explique por qué.
b) ¿Cuál es la probabilidad de que $x = 30$?
c) ¿Qué probabilidad existe de que x sea menor o igual que 25?
d) ¿Cuál es la probabilidad de que x sea mayor que 30?

Aplicaciones

AUTO evaluación

8. Los datos siguientes se obtuvieron por conteo del número de salas de operaciones en uso en el Hospital General Tampa durante un periodo de 20 días: en tres de estos días sólo se usó una sala de cirugía; en cinco de estos días se usaron dos; en ocho días se utilizaron tres, y en cuatro días se usaron las cuatro salas de operaciones del hospital.
   a) Use el método de frecuencia relativa a efecto de construir una distribución de probabilidad para el número de salas de operación en uso en cualquier día dado.
   b) Trace una gráfica de la distribución de probabilidad.
   c) Muestre que su distribución de probabilidad satisface las condiciones requeridas para una distribución de probabilidad discreta válida.

9. En Estados Unidos, 38% de los alumnos de cuarto grado de primaria no puede leer un libro apropiado para su edad. Los datos siguientes muestran el número de sujetos, por edad, que se identificaron como niños con problemas de aprendizaje que requieren educación especial. La mayoría tiene problemas de lectura que debieron identificarse y corregirse antes del tercer grado. La ley federal estadounidense actual prohíbe que la mayoría de los niños reciba ayuda adicional de programas de educación especial hasta que el retraso sea de aproximadamente dos años de aprendizaje, y por lo general eso significa hasta tercer grado o grados superiores (*USA Today,* 6 de septiembre, 2001).

Edad	Número de niños
6	37 369
7	87 436
8	160 840
9	239 719
10	286 719
11	306 533
12	310 787
13	302 604
14	289 168

Suponga que se desea seleccionar una muestra de menores con problemas de aprendizaje y que deben tomar educación especial a efecto de incluirlos en un programa diseñado para mejorar su capacidad de lectura. Sea x una variable aleatoria que indica la edad de un niño seleccionado al azar.

a) Use los datos para elaborar una distribución de probabilidad para x. Especifique los valores de la variable aleatoria y los valores correspondientes de la función de probabilidad $f(x)$.
b) Trace una gráfica de la distribución de probabilidad.
c) Muestre que la distribución de probabilidad satisface las ecuaciones (5.1) y (5.2).

10. A continuación se presentan las distribuciones de frecuencias porcentuales de la satisfacción laboral para una muestra de altos directivos y gerentes de rango medio en el área de sistemas de información (SI). Las puntaciones varían de baja, 1 (muy insatisfecho), a alta, 5 (muy satisfecho).

Puntuación de satisfacción laboral	Altos directivos de SI (%)	Gerentes de rango medio de SI (%)
1	5	4
2	9	10
3	3	12
4	42	46
5	41	28

a) Elabore una distribución de probabilidad para la puntuación de satisfacción laboral de un alto directivo.
b) Prepare una distribución de probabilidad para la puntuación de satisfacción laboral de un gerente de rango medio.
c) ¿Cuál es la probabilidad de que un alto directivo reporte una puntuación de satisfacción laboral de 4 o 5?
d) ¿Cuál es la probabilidad de que un gerente de rango medio esté muy satisfecho?
e) Compare la satisfacción laboral general de los altos directivos con la de los gerentes de rango medio.

11. Un técnico proporciona servicio a las máquinas de correo en algunas empresas del área de Phoenix. Dependiendo del tipo de falla, la visita de servicio puede durar 1, 2, 3 o 4 horas. Los distintos tipos de falla ocurren aproximadamente con la misma frecuencia.

a) Elabore una distribución de probabilidad para la duración de una visita de servicio.
b) Trace una gráfica de la distribución de probabilidad.
c) Muestre que su distribución de probabilidad satisface las condiciones requeridas para una función de probabilidad discreta.

d) ¿Cuál es la probabilidad de que una visita de servicio dure tres horas?
e) El técnico acaba de llegar a una visita de servicio, pero desconoce el tipo de falla. Son las 3:00 p.m. y los técnicos de servicio trabajan sólo hasta las 5:00 p.m. ¿Cuál es la probabilidad de que tenga que trabajar tiempo extra para reparar la máquina hoy?

12. Los dos proveedores de cable principales en Estados Unidos son Comcast Cable Communications, con 21.5 millones de suscriptores, y Time Warner Cable, con 11.0 millones de clientes (*The New York Times Almanac,* 2007). Suponga que la gerencia de Time Warner Cable evalúa de manera subjetiva una distribución de probabilidad del número de suscriptores nuevos el año siguiente en el estado de Nueva York como sigue.

x	$f(x)$
100 000	0.10
200 000	0.20
300 000	0.25
400 000	0.30
500 000	0.10
600 000	0.05

a) ¿Es válida esta distribución de probabilidad? Explique por qué.
b) ¿Cuál es la probabilidad de que Time Warner obtenga más de 400 000 suscriptores nuevos?
c) ¿Qué probabilidad existe de que Time Warner obtenga menos de 200 000 suscriptores nuevos?

13. Un psicólogo determinó que el número de sesiones requeridas para ganarse la confianza de un paciente nuevo es de 1, 2 o 3 sesiones. Sea x una variable aleatoria que indica el número de sesiones requeridas para ganarse la confianza de un paciente. Se ha propuesto la función de probabilidad siguiente.

$$f(x) = \frac{x}{6} \qquad \text{para } x = 1, 2 \text{ o } 3$$

a) ¿Esta función de probabilidad es válida? Explique por qué.
b) ¿Cuál es la probabilidad de que se requieran *exactamente* 2 sesiones para ganarse la confianza de un paciente?
c) ¿Cuál es la probabilidad de que sean necesarias *por lo menos* 2 sesiones para ganarse la confianza de un paciente?

14. La tabla siguiente es una distribución de probabilidad parcial para las utilidades proyectadas de MRA Company (x = utilidades en miles de dólares) para el primer año de operación (el valor negativo denota una pérdida).

x	$f(x)$
−100	0.10
0	0.20
50	0.30
100	0.25
150	0.10
200	

a) ¿Cuál es el valor apropiado para $f(200)$? ¿Cuál es su interpretación de este valor?
b) ¿Qué probabilidad existe de que MRA sea rentable?
c) ¿Cuál es la probabilidad de que obtenga por lo menos $100 000?

5.3 Valor esperado y varianza

Valor esperado

El **valor esperado**, o media, de una variable aleatoria es una medida de su posición central. La fórmula para el valor esperado de una variable aleatoria discreta x se indica enseguida.

El valor esperado es un promedio ponderado de los valores que asume la variable aleatoria cuando los pesos son las probabilidades.

VALOR ESPERADO DE UNA VARIABLE ALEATORIA DISCRETA

$$E(x) = \mu = \Sigma x f(x) \qquad (5.4)$$

Ambas notaciones, $E(x)$ y μ se usan para denotar el valor esperado de una variable aleatoria.

La ecuación (5.4) muestra que para calcular el valor esperado de una variable aleatoria discreta se debe multiplicar cada valor de la variable por su probabilidad correspondiente $f(x)$, y después se suman los productos que resultan. Utilizando el ejemplo de la venta de automóviles de DiCarlo Motors de la sección 5.2, en la tabla 5.4 se muestra el cálculo del valor esperado para el número de vehículos vendidos durante un día. La suma de las entradas de la columna $xf(x)$ muestra que el valor esperado es 1.50 unidades por día. Por consiguiente, aunque se sabe que en un día cualquiera las ventas pueden ser de 0, 1, 2, 3, 4 o 5 automóviles, DiCarlo anticipa que con el tiempo se venderá un promedio diario de 1.50. Suponiendo que un mes tiene 30 días de operación, se usa el valor esperado de 1.50 para pronosticar el promedio de ventas mensuales de $30(1.50) = 45$ vehículos.

El valor esperado no tiene que ser un valor que la variable aleatoria pueda asumir.

Varianza

Aun cuando el valor esperado proporciona el valor medio de la variable aleatoria, a menudo necesitamos una medida de variabilidad o dispersión. Así como la varianza se usó en el capítulo 3 para resumir la variabilidad en los datos, ahora la **varianza** se usa para resumir la variabilidad en los valores de una variable aleatoria. A continuación se presenta la fórmula para la varianza de una variable aleatoria discreta.

La varianza es un promedio ponderado de las desviaciones al cuadrado de una variable aleatoria de su media. Los pesos son las probabilidades.

VARIANZA DE UNA VARIABLE ALEATORIA DISCRETA

$$\text{Var}(x) = \sigma^2 = \Sigma(x - \mu)^2 f(x) \qquad (5.5)$$

Como muestra la ecuación (5.5), una parte esencial de la fórmula de la varianza es la desviación, $x - \mu$, la cual mide a qué distancia está el valor esperado, o la media, μ, de un valor particular de la variable aleatoria. Para calcular la varianza de una variable aleatoria, las desviaciones se elevan al cuadrado y luego se ponderan por el valor correspondiente de la función de probabilidad. La suma de estas desviaciones al cuadrado ponderadas para todos los valores de la variable aleatoria se conocen como la *varianza*. Las notaciones $Var(x)$ y σ^2 se usan para denotar la varianza de una variable aleatoria.

TABLA 5.4 Cálculo del valor esperado para el número de automóviles que se venden en un día en Dicarlo Motors

x	$f(x)$	$xf(x)$
0	0.18	0(0.18) = 0.00
1	0.39	1(0.39) = 0.39
2	0.24	2(0.24) = 0.48
3	0.14	3(0.14) = 0.42
4	0.04	4(0.04) = 0.16
5	0.01	5(0.01) = 0.05
		1.50

$E(x) = \mu = \Sigma x f(x)$

TABLA 5.5 Cálculo de la varianza para el número de automóviles que se venden en un día en Dicarlo Motors

x	$x - \mu$	$(x - \mu)^2$	$f(x)$	$(x - \mu)^2 f(x)$
0	$0 - 1.50 = -1.50$	2.25	0.18	$2.25(.18) = 0.4050$
1	$1 - 1.50 = -0.50$	0.25	0.39	$0.25(.39) = 0.0975$
2	$2 - 1.50 = 0.50$	0.25	0.24	$0.25(.24) = 0.0600$
3	$3 - 1.50 = 1.50$	2.25	0.14	$2.25(.14) = 0.3150$
4	$4 - 1.50 = 2.50$	6.25	0.04	$6.25(.04) = 0.2500$
5	$5 - 1.50 = 3.50$	12.25	0.01	$12.25(.01) = 0.1225$
				1.2500

$\sigma^2 = \Sigma(x - \mu)^2 f(x)$

El cálculo de la varianza para la distribución de probabilidad del número de automóviles vendidos durante un día en DiCarlo Motors se resume en la tabla 5.5. Vemos que la varianza es 1.25. La **desviación estándar**, σ, se define como la raíz cuadrada positiva de la varianza. Por tanto, la desviación estándar para el número de automóviles vendidos durante un día es

$$\sigma = \sqrt{1.25} = 1.118$$

La desviación estándar se mide en las mismas unidades que la variable aleatoria ($\sigma = 1.118$ automóviles) y por tanto a menudo se prefiere para describir la variabilidad de una variable aleatoria. La varianza σ^2 se mide en unidades cuadradas y, por tanto, es más difícil de interpretar.

Ejercicios

Métodos

15. La tabla siguiente proporciona una distribución de probabilidad para la variable aleatoria x.

x	$f(x)$
3	0.25
6	0.50
9	0.25

a) Calcule $E(x)$, el valor esperado de x.
b) Estime σ^2, la varianza de x.
c) Calcule σ, la desviación estándar de x.

AUTOevaluación

16. La tabla siguiente proporciona una distribución de probabilidad para la variable aleatoria y.

y	$f(y)$
2	0.20
4	0.30
7	0.40
8	0.10

a) Calcule $E(y)$.
b) Calcule Var(y) y σ.

Aplicaciones

17. El número de estudiantes que presentan la prueba de aptitudes escolares SAT ha aumentado a una cifra sin precedente de 1.5 millones (Consejo del Colegio, 26 de agosto de 2008). Se permite que los estudiantes repitan la prueba con la esperanza de que mejoren la calificación que se envía a las oficinas de admisión de los colegios y universidades. El número de veces que la SAT fue presentada y el número de estudiantes son los siguientes.

Número de veces	Número de estudiantes
1	721 769
2	601 325
3	166 736
4	22 299
5	6 730

a) Sea x una variable aleatoria que indica el número de veces que un estudiante presenta el SAT. Muestre la distribución de probabilidad para esta variable aleatoria.
b) ¿Cuál es la probabilidad de que un estudiante presente el SAT más de una vez?
c) ¿Cuál es la probabilidad de que un estudiante lo presente tres o más veces?
d) ¿Cuál es el valor esperado del número de veces que se presenta el SAT? ¿Cuál es su interpretación del valor esperado?
e) ¿Cuáles son la varianza y la desviación estándar para el número de veces que se presenta el SAT?

AUTOevaluación

18. El estudio American Housing Survey reportó los datos siguientes sobre el número de recámaras ocupadas en casas propias y rentadas en las ciudades centrales (sitio web de la Oficina del Censo de Estados Unidos, 31 de marzo de 2003).

	Número de casas (miles)	
Recámaras	Rentadas	Propias
0	547	23
1	5 012	541
2	6 100	3 832
3	2 644	8 690
4 o más	557	3 783

a) Defina una variable aleatoria x = número de recámaras en las casas rentadas y elabore una distribución de probabilidad para la variable aleatoria ($x = 4$ representa 4 o más recámaras.)
b) Calcule el valor esperado y la varianza del número de recámaras en las casas rentadas.
c) Defina una variable aleatoria y = número de recámaras en las casas propias, y elabore una distribución de probabilidad para la variable aleatoria ($y = 4$ representa 4 o más recámaras.)
d) Calcule el valor esperado y la varianza para el número de recámaras en las casas propias.
e) ¿Qué observaciones puede hacer de la comparación del número de recámaras en casas rentadas en comparación con las casas propias?

19. La NBA (National Basketball Association) lleva un registro de una variedad de estadísticas para cada equipo. Dos de éstas registran el porcentaje de tiros de campo y el porcentaje de tiros de tres puntos efectuados por equipo. Los registros de tiros de los 29 equipos de la NBA para una parte de la temporada 2004 mostraban que la probabilidad de anotar dos puntos en un tiro de

campo era de 0.44, y la probabilidad de anotar tres puntos al hacer un tiro de tres puntos era de 0.34 (sitio web de la NBA, 3 de enero de 2004).

a) ¿Cuál es el valor esperado de un tiro de dos puntos para estos equipos?
b) ¿Cuál es el valor esperado de un tiro de tres puntos para estos equipos?
c) Si la probabilidad de hacer un tiro de dos puntos es mayor que la de hacer un tiro de tres puntos, ¿por qué los entrenadores permiten que algunos jugadores lancen tiros de tres puntos si tienen la oportunidad? Use el valor esperado para explicar su respuesta.

20. La distribución de probabilidad de las reclamaciones por daños que pagó Newton Automobile Insurance Company por seguro contra choques es la siguiente.

Pago ($)	Probabilidad
0	0.85
500	0.04
1 000	0.04
3 000	0.03
5 000	0.02
8 000	0.01
10 000	0.01

a) Use el pago de choque esperado para determinar la prima del seguro contra colisiones que permitiría a la empresa no ganar ni perder.
b) La compañía de seguros cobra una tarifa anual de $520 por la cobertura de choques. ¿Cuál es el valor esperado del seguro contra choques para un asegurado? (*Pista:* son los pagos esperados de la empresa menos el costo de cobertura.) ¿Por qué el cliente compra un seguro contra colisiones con este valor esperado?

21. Las siguientes distribuciones de probabilidad de las puntuaciones de satisfacción laboral para una muestra de altos directivos y gerentes de rango medio del área de sistemas de información (SI) varía de un valor bajo de 1 (muy insatisfecho) a un valor alto de 5 (muy satisfecho).

	Probabilidad	
Puntuación de satisfacción laboral	**Altos directivos de SI**	**Gerentes de rango medio de SI**
1	0.05	0.04
2	0.09	0.10
3	0.03	0.12
4	0.42	0.46
5	0.41	0.28

a) ¿Cuál es el valor esperado de la puntuación de satisfacción laboral para los altos directivos?
b) ¿Cuál es el valor esperado de dicha puntuación para los gerentes de rango medio?
c) Calcule la varianza de las puntuaciones de satisfacción laboral para los directivos y los gerentes de rango medio.
d) Estime la desviación estándar de las calificaciones de satisfacción laboral en las dos distribuciones de probabilidad.
e) Compare la satisfacción laboral de los altos directivos con la de los gerentes de nivel medio.

22. La demanda de un producto de Carolina Industries varía mucho cada mes. La distribución de probabilidad en la tabla siguiente, con base en los datos de años pasados, muestra la demanda mensual de la empresa.

Demanda de unidades	Probabilidad
300	0.20
400	0.30
500	0.35
600	0.15

a) Si la empresa basa los pedidos de cada mes en el valor esperado de la demanda mensual, ¿cuál debe ser la cantidad de pedidos mensuales de Carolina para este producto?
b) Suponga que cada unidad demandada genera ingresos de $70 y que cada una cuesta $50. ¿Cuánto ganará o perderá la empresa en un mes si hace un pedido con base en su respuesta al inciso *a*) y la demanda real del artículo es 300 unidades?

23. La Encuesta de Viviendas y Unidades Desocupadas de la Ciudad de Nueva York mostró un total de 59 324 unidades de vivienda bajo control de rentas y 236 263 unidades bajo renta regulada construidas en 1947 o después. Las distribuciones de probabilidad del número de personas que viven en estas viviendas rentadas se proporcionan a continuación (sitio web de la Oficina del Censo de Estados Unidos, 12 de enero de 2004).

Número de personas	Control de rentas	Renta regulada
1	0.61	0.41
2	0.27	0.30
3	0.07	0.14
4	0.04	0.11
5	0.01	0.03
6	0.00	0.01

a) ¿Cuál es el valor esperado del número de personas que viven en cada tipo de unidad?
b) ¿Cuál es la varianza del número de personas que viven en cada tipo de unidad?
c) Haga algunas comparaciones entre el número de personas que viven en viviendas bajo rentas controladas y el número de personas que viven en unidades de renta regulada.

24. J. R. Ryland Computer Company considera la expansión de una planta para permitir a la empresa comenzar la fabricación de una computadora nueva. El presidente de la firma debe determinar si el proyecto de expansión se realiza a mediana o a gran escala. La demanda para la computadora nueva es incierta, y para propósitos de planeación puede ser baja, mediana o alta. Las probabilidades estimadas para la demanda son 0.20, 0.50 y 0.30, respectivamente; x y y indican las utilidades anuales en miles de dólares. Los encargados de la planeación en la empresa elaboraron los pronósticos de utilidades siguientes para los proyectos de expansión a mediana y gran escala.

		Utilidades de la expansión a mediana escala		Utilidades de la expansión a gran escala	
		x	$f(x)$	y	$f(y)$
Demanda	**Baja**	50	0.20	0	0.20
	Mediana	150	0.50	100	0.50
	Alta	200	0.30	300	0.30

a) Calcule el valor esperado para las utilidades asociadas con las dos alternativas de expansión. ¿Cuál decisión es preferible para el objetivo de maximizar las utilidades esperadas?
b) Calcule la varianza para la utilidad asociada con las dos alternativas de expansión. ¿Cuál decisión es preferible para el objetivo de minimizar el riesgo o la incertidumbre?

5.4 Distribución de probabilidad binomial

La distribución de probabilidad binomial es una distribución de probabilidad discreta que proporciona muchas aplicaciones. Se asocia con un experimento de múltiples pasos que se llama *experimento binomial*.

Un experimento binomial

Un **experimento binomial** tiene las cuatro propiedades siguientes.

PROPIEDADES DE UN EXPERIMENTO BINOMIAL

1. El experimento consiste de una secuencia de n ensayos idénticos.
2. En cada ensayo hay dos resultados posibles. A uno de ellos se le llama *éxito* y al otro, *fracaso*.
3. La probabilidad de éxito, denotada por p, no cambia de un ensayo a otro. Por consiguiente, la probabilidad de fracaso, denotada por $1 - p$, tampoco cambia de un ensayo a otro.
4. Los ensayos son independientes.

Jakob Bernoulli (1654-1705), el primero de una familia de matemáticos suizos, publicó un tratado sobre probabilidad que contenía la teoría de permutaciones y combinaciones, así como el teorema binomial.

Si están presentes las propiedades 2, 3 y 4, se dice que los ensayos son generados por un proceso de Bernoulli. Si, además, la propiedad 1 está presente, se dice que tenemos un experimento binomial. La figura 5.2 representa una secuencia posible de éxitos y fracasos para un experimento binomial que consta de ocho ensayos.

En un experimento binomial, lo que interesa es el *número de éxitos que ocurren en los n ensayos*. Si x denota el número de éxitos que ocurren en n ensayos, vemos que x puede asumir los valores 0, 1, 2, 3..., n. Debido a que el número de valores es finito, x es una variable aleatoria *discreta*. La distribución de probabilidad asociada con esta variable se llama **distribución de probabilidad binomial**. Por ejemplo, considere el experimento de lanzar una moneda cinco veces y en cada lanzamiento observe si la moneda cae con cara o cruz en el lado superior. Suponga que queremos contar el número de caras que aparecen durante los cinco lanzamientos. ¿Este ejemplo muestra las propiedades de un experimento binomial? ¿Cuál es la variable aleatoria de interés? Observe que:

1. El experimento consta de cinco ensayos idénticos; cada uno consiste en el lanzamiento de una moneda.
2. En cada ensayo hay dos resultados posibles: cara o cruz. Se puede designar cara como un éxito y cruz como un fracaso.
3. La probabilidad de obtener cara y la probabilidad de obtener cruz son iguales para cada ensayo, con $p = 0.5$ y $1 - p = 0.5$.
4. Los ensayos o lanzamientos son independientes debido a que el resultado de cualquier ensayo no se ve afectado por lo que ocurre con otros ensayos o lanzamientos.

FIGURA 5.2 Secuencia posible de éxitos y fracasos para un experimento binomial de ocho ensayos

Propiedad 1. El experimento consta de $n = 8$ ensayos idénticos.

Propiedad 2. Cada ensayo da como resultado un éxito (S) o un fracaso (F).

Ensayos →	1	2	3	4	5	6	7	8
Resultados →	*S*	*F*	*F*	*S*	*S*	*F*	*S*	*S*

Por tanto, las propiedades de un experimento binomial se satisfacen. La variable aleatoria que interesa es x = número de caras que ocurren en cinco ensayos. En este caso, x puede tomar los valores 0, 1, 2, 3, 4 o 5.

En otro ejemplo, considere a una vendedora de seguros que visita a 10 familias seleccionadas al azar. El resultado asociado con cada visita se clasifica como un éxito si la familia compra un seguro y un fracaso si no lo compra. A partir de su experiencia, la vendedora sabe que la probabilidad de que una familia seleccionada al azar compre un seguro es de 0.10. Al revisar las propiedades de un experimento binomial se observa que:

1. El experimento consta de 10 ensayos idénticos; cada uno consiste en visitar a una familia.
2. En cada ensayo hay dos resultados posibles: la familia compra el seguro (éxito) o no lo compra (fracaso).
3. Se asume que las probabilidades de que haya una compra o no la haya son iguales para cada visita, con $p = 0.10$ y $1 - p = 0.90$.
4. Los ensayos son independientes, porque las familias se eligen al azar.

Como estos cuatro supuestos se cumplen, este ejemplo es un experimento binomial. La variable aleatoria de interés es el número de ventas obtenidas al hacer contacto con las 10 familias. En este caso, x puede asumir los valores 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 y 10.

La propiedad 3 del experimento binomial se llama *supuesto de estacionariedad* y a veces se confunde con la propiedad 4, la independencia de los ensayos. Para ver cómo difieren, considere de nuevo el caso de la vendedora que visita a las familias para ofrecer seguros. Si, a medida que el día avanza, la empleada se cansa y pierde entusiasmo, la probabilidad de éxito (vender un seguro) para el décimo contacto podría disminuir a 0.05, por ejemplo. En este caso, la propiedad 3 (estacionariedad) no se cumpliría y el experimento no sería binomial. Incluso si la propiedad 4 se cumple, es decir, que las decisiones de compra de cada familia se realizaran en forma independiente, el experimento no sería binomial si la propiedad 3 no se satisface.

En las aplicaciones con experimentos binomiales se usa una fórmula matemática especial, llamada *función de probabilidad binomial,* para calcular la probabilidad de x éxitos en n ensayos. Enseguida se mostrará cómo se desarrolla la fórmula, en el contexto de un problema ilustrativo, usando los conceptos de probabilidad presentados en el capítulo 4.

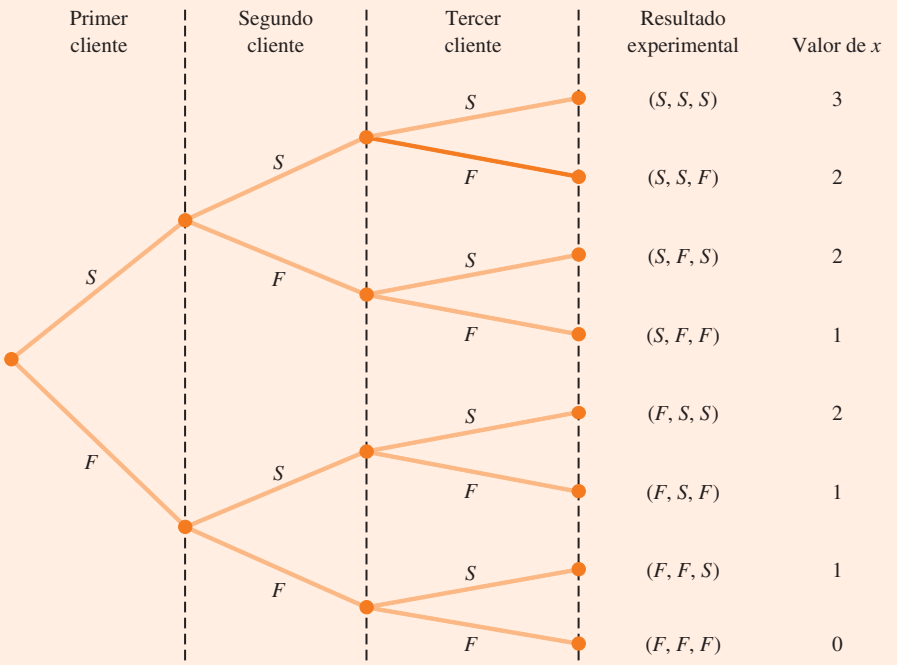
El problema de Martin Clothing Store

Considere las decisiones de compra de los tres clientes siguientes que entran en la tienda de ropa Martin Clothing Store. Con base en su experiencia, el gerente de la tienda estima que la probabilidad de que un cliente cualquiera haga una compra es de 0.30. ¿Cuál es la probabilidad de que dos de los tres clientes siguientes realicen una compra?

Un diagrama de árbol (figura 5.3) permite ver que en el experimento de observar a tres clientes que toman una decisión de compra, cada uno tiene ocho resultados posibles. Si S denota éxito (una compra) y F denota fracaso (no hay compra), se tiene interés en los resultados experimentales que consisten en dos éxitos en los tres ensayos (decisiones de compra). A continuación se verificará que el experimento con una secuencia de tres decisiones de compra puede verse como binomial. Al revisar los cuatro requerimientos para un experimento binomial, observamos que:

1. El experimento se describe como una secuencia de tres ensayos idénticos, uno para cada uno de los tres clientes que entran en la tienda.
2. Para cada ensayo hay dos resultados posibles: el cliente efectúa una compra (éxito) o el cliente no efectúa una compra (fracaso).
3. Se asume que la probabilidad de que el cliente realice una compra (0.30) o no la realice (0.70) es la misma para todos los clientes.
4. La decisión de compra de cada sujeto es independiente de las decisiones que tomen los otros clientes.

FIGURA 5.3 Diagrama de árbol para el problema de Martin Clothing Store



S = Hay compra
F = No hay compra
x = Número de clientes que efectúan una compra

Por consiguiente, están presentes las propiedades de un experimento binomial.

El número de resultados experimentales que producen exactamente x éxitos en n ensayos se calcula usando la fórmula siguiente.[1]

NÚMERO DE RESULTADOS EXPERIMENTALES QUE PROPORCIONAN EXACTAMENTE x ÉXITOS EN n ENSAYOS

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{5.6}$$

donde

$$n! = n(n-1)(n-2)\cdots(2)(1)$$

y por definición,

$$0! = 1$$

Ahora regresemos al experimento de Martin Clothing Store que consiste en las decisiones de compra de tres clientes. La ecuación (5.6) permite determinar el número de resultados que

[1] Esta fórmula, presentada en el capítulo 4, determina el número de combinaciones de n objetos seleccionados x a la vez. Para el experimento binomial, esta fórmula combinatoria proporciona el número de resultados experimentales (secuencias de n ensayos), lo que da como resultado x éxitos.

involucran dos compras; es decir, el número de maneras de obtener $x = 2$ éxitos en $n = 3$ ensayos. A partir de la ecuación (5.6) tenemos

$$\binom{n}{x} = \binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{(3)(2)(1)}{(2)(1)(1)} = \frac{6}{2} = 3$$

La ecuación (5.6) muestra que tres de los resultados experimentales produjeron dos éxitos. A partir de la figura 5.3, vemos que estos tres resultados se denotan por (*S, S, F*), (*S, F, S*) y (*F, S, S*).

Usando la ecuación (5.6) para determinar cuántos resultados experimentales tienen tres éxitos (compras) en los tres ensayos, obtenemos

$$\binom{n}{x} = \binom{3}{3} = \frac{3!}{3!(3-3)!} = \frac{3!}{3!0!} = \frac{(3)(2)(1)}{3(2)(1)(1)} = \frac{6}{6} = 1$$

A partir de la figura 5.3 observamos que el resultado experimental con tres éxitos se identifica por *(S, S, S).*

Se sabe que la ecuación (5.6) se utiliza para determinar el número de resultados experimentales que dan lugar a x éxitos. Si se determinará la probabilidad de x éxitos en n ensayos, no obstante, también debemos conocer la probabilidad asociada con cada uno de estos resultados. Como los ensayos de un experimento binomial son independientes, sencillamente es posible multiplicar las probabilidades asociadas con el resultado de cada ensayo para encontrar la probabilidad de una secuencia particular de éxitos y fracasos.

La probabilidad de que los dos primeros clientes compren y que el tercero no compre, denotada por *(S, S, F),* está dada por

$$pp(1-p)$$

Con una probabilidad de 0.30 de una compra en cualquier ensayo, la probabilidad de una compra en los primeros dos ensayos y ninguna compra en el tercero está dada por

$$(0.30)(0.30)(0.70) = (0.30)^2(0.70) = 0.063$$

Otros dos resultados experimentales también dan lugar a dos éxitos y un fracaso. Las probabilidades de tres resultados que tienen dos éxitos se presentan a continuación.

Resultados de los ensayos				
Primer cliente	**Segundo cliente**	**Tercer cliente**	**Resultado experimental**	**Probabilidad del resultado experimental**
Compra	Compra	No compra	(*S, S, F*)	$pp(1-p) = p^2(1-p) = (0.30)^2(0.70) = 0.063$
Compra	No compra	Compra	(*S, F, S*)	$p(1-p)p = p^2(1-p) = (0.30)^2(0.70) = 0.063$
No compra	Compra	Compra	(*F, S, S*)	$(1-p)pp = p^2(1-p) = (0.30)^2(0.70) = 0.063$

Observe que los tres resultados experimentales con dos éxitos tienen exactamente la misma probabilidad. Esta observación es válida en general. En cualquier experimento binomial, todas las secuencias de resultados de ensayos que producen x éxitos en n ensayos tienen la *misma probabilidad* de ocurrencia. La probabilidad de cada secuencia de ensayos que producen x éxitos en n ensayos se presenta a continuación.

$$\begin{pmatrix}\text{Probabilidad de una secuencia particular de resultados de con } x \text{ éxitos en } n \text{ ensayos}\end{pmatrix} = p^x(1-p)^{(n-x)} \quad \textbf{(5.7)}$$

En el caso de la tienda Martin Clothing Store, esta fórmula indica que cualquier resultado experimental con dos éxitos tiene una probabilidad de $p^2(1-p)^{(3-2)} = p^2(1-p)^1 = (0.30)^2(0.70)^1 = 0.063$.

Como la ecuación (5.6) muestra el número de resultados de un experimento binomial con x éxitos y la ecuación (5.7) proporciona la probabilidad de cada secuencia con x éxitos, las ecuaciones (5.6) y (5.7) se combinan para obtener la **función de probabilidad binomial** siguiente.

FUNCIÓN DE PROBABILIDAD BINOMIAL

$$f(x) = \binom{n}{x} p^x(1-p)^{(n-x)} \quad \textbf{(5.8)}$$

donde

x = número de éxitos
p = probabilidad de un éxito en un ensayo
n = número de ensayos
$f(x)$ = probabilidad de x éxitos en n ensayos
$\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$

Para la distribución de probabilidad binomial, x es una variable aleatoria discreta con la función de probabilidad $f(x)$ aplicable para los valores de $x = 0, 1, 2, \ldots, n$.

En el ejemplo de Martin Clothing Store, se usa la ecuación (5.8) para calcular la probabilidad de que ningún cliente realice una compra; exactamente un cliente haga una compra; exactamente dos clientes efectúen una compra, y los tres clientes compren. Los cálculos se resumen en la tabla 5.6, que proporciona la distribución de probabilidad del número de sujetos que realizan una compra. La figura 5.4 es una gráfica de esta distribución de probabilidad.

La función de probabilidad binomial se aplica a *cualquier* experimento binomial. Si una situación demuestra las propiedades de un experimento binomial y se conocen los valores de n y p, se puede usar la ecuación (5.8) para calcular la probabilidad de x éxitos en n ensayos.

TABLA 5.6 Distribución de probabilidad para el número de clientes que efectúan una compra

x	$f(x)$
0	$\frac{3!}{0!3!}(0.30)^0(0.70)^3 = 0.343$
1	$\frac{3!}{1!2!}(0.30)^1(0.70)^2 = 0.441$
2	$\frac{3!}{2!1!}(0.30)^2(0.70)^1 = 0.189$
3	$\frac{3!}{3!0!}(0.30)^3(0.70)^0 = \frac{0.027}{1.000}$

FIGURA 5.4 Representación gráfica de la distribución de probabilidad para el número de clientes que efectúan una compra



Si se consideran variaciones del experimento de Martin, por ejemplo que 10 clientes en vez de tres entren en la tienda, la función de probabilidad binomial dada la ecuación (5.8) sigue siendo válida. Suponga que se tiene un experimento binomial con $n = 10$, $x = 4$ y $p = 0.30$. La probabilidad de que exactamente cuatro de los 10 clientes que entran en la tienda realicen una compra es

$$f(4) = \frac{10!}{4!6!}(0.30)^4(0.70)^6 = 0.2001$$

Uso de tablas de probabilidades binomiales

Se han desarrollado tablas que proporcionan la probabilidad de x éxitos en n ensayos para un experimento binomial. Por lo general son fáciles de usar y más rápidas que la ecuación (5.8). La tabla 5 del apéndice B es una tabla de probabilidades binomiales de este tipo. Una parte de ella se reproduce en la tabla 5.7. Para usarla, se deben especificar los valores de n, p y x según el experimento binomial de que se trate. En el ejemplo que se presenta en la parte superior de la tabla 5.7, vemos que la probabilidad de que $x = 3$ éxitos en un experimento binomial con $n = 10$ y $p = 0.40$ es de 0.2150. Se puede recurrir a la ecuación (5.8) para verificar que se obtendría el mismo resultado si se usa directamente la función de probabilidad binomial.

Ahora se usará la tabla 5.7 para verificar la probabilidad de cuatro éxitos en 10 ensayos en el problema de Martin Clothing Store. Note que el valor de $f(4) = 0.2001$ se lee directamente de la tabla de probabilidades binomiales, según la cual $n = 10$, $x = 4$ y $p = 0.30$.

Con las calculadoras modernas, estas tablas son casi innecesarias. Es fácil evaluar directamente la ecuación (5.8).

Aun cuando las tablas de probabilidades binomiales son relativamente fáciles de usar, es imposible contar con tablas que muestren todos los valores posibles de n y p que podrían encontrarse en un experimento binomial. Sin embargo, con las calculadoras actuales, el uso de la ecuación (5.8) para calcular la probabilidad buscada no es difícil, en especial si el número de ensayos no es grande. En los ejercicios de esta sección tendrá la oportunidad de practicar con la ecuación (5.8) para calcular las probabilidades binomiales, a menos que el problema requiera que de manera específica se utilice la tabla de probabilidades binomiales.

TABLA 5.7 Valores seleccionados del ejemplo de la tabla de probabilidad binomial: $n = 10$; $x = 3$; $p = .040$; $f(3) = 0.2150$

		p									
n	***x***	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
9	0	0.6302	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020
	1	0.2985	0.3874	0.3679	0.3020	0.2253	0.1556	0.1004	0.0605	0.0339	0.0176
	2	0.0629	0.1722	0.2597	0.3020	0.3003	0.2668	0.2162	0.1612	0.1110	0.0703
	3	0.0077	0.0446	0.1069	0.1762	0.2336	0.2668	0.2716	0.2508	0.2119	0.1641
	4	0.0006	0.0074	0.0283	0.0661	0.1168	0.1715	0.2194	0.2508	0.2600	0.2461
	5	0.0000	0.0008	0.0050	0.0165	0.0389	0.0735	0.1181	0.1672	0.2128	0.2461
	6	0.0000	0.0001	0.0006	0.0028	0.0087	0.0210	0.0424	0.0743	0.1160	0.1641
	7	0.0000	0.0000	0.0000	0.0003	0.0012	0.0039	0.0098	0.0212	0.0407	0.0703
	8	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0013	0.0035	0.0083	0.0176
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0008	0.0020
10	0	0.5987	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010
	1	0.3151	0.3874	0.3474	0.2684	0.1877	0.1211	0.0725	0.0403	0.0207	0.0098
	2	0.0746	0.1937	0.2759	0.3020	0.2816	0.2335	0.1757	0.1209	0.0763	0.0439
	3	0.0105	0.0574	0.1298	0.2013	0.2503	0.2668	0.2522	**0.2150**	0.1665	0.1172
	4	0.0010	0.0112	0.0401	0.0881	0.1460	0.2001	0.2377	0.2508	0.2384	0.2051
	5	0.0001	0.0015	0.0085	0.0264	0.0584	0.1029	0.1536	0.2007	0.2340	0.2461
	6	0.0000	0.0001	0.0012	0.0055	0.0162	0.0368	0.0689	0.1115	0.1596	0.2051
	7	0.0000	0.0000	0.0001	0.0008	0.0031	0.0090	0.0212	0.0425	0.0746	0.1172
	8	0.0000	0.0000	0.0000	0.0001	0.0004	0.0014	0.0043	0.0106	0.0229	0.0439
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005	0.0016	0.0042	0.0098
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0010

El software para estadística, como Minitab, y los programas de hoja de cálculo, como Excel, también permiten calcular probabilidades binomiales. Considere el ejemplo de Martin Clothing Store con $n = 10$ y $p = 0.30$. La figura 5.5 muestra las probabilidades binomiales generadas por Minitab para todos los valores posibles de x. Note que estos valores son los mismos que aquellos encontrados en la columna $p = 0.30$ de la tabla 5.7. En el apéndice 5.1 se explica el procedimiento paso por paso para usar Minitab con la finalidad de generar el resultado que se exhibe en la figura 5.5. En el apéndice 5.2 se describe cómo usar Excel para calcular probabilidades binomiales.

Valor esperado y varianza de la distribución binomial

En la sección 5.3 se proporcionaron las fórmulas para calcular el valor esperado y la varianza de una variable aleatoria discreta. En el caso especial en que la variable tiene una distribución binomial con un número conocido de ensayos n y una probabilidad conocida de éxitos p, las fórmulas generales para el valor esperado y la varianza se simplifican. Los resultados se muestran a continuación.

VALOR ESPERADO Y VARIANZA DE LA DISTRIBUCIÓN BINOMIAL

$$E(x) = \mu = np \tag{5.9}$$

$$\text{Var}(x) = \sigma^2 = np(1 - p) \tag{5.10}$$

FIGURA 5.5 Resultado de Minitab que muestra las probabilidades binomiales para el problema de Martin Clothing Store

x	P(X = x)
0.00	0.0282
1.00	0.1211
2.00	0.23350
3.00	0.2668
4.00	0.2001
5.00	0.1029
6.00	0.0368
7.00	0.0090
8.00	0.0014
9.00	0.0001
10.00	0.0000

En el caso del problema de Martin Clothing Store con tres clientes, se usa la ecuación (5.9) para calcular el número esperado de clientes que realizarán una compra.

$$E(x) = np = 3(0.30) = 0.9$$

Suponga que para el mes siguiente Martin Clothing Store pronostica que 1 000 clientes entrarán en la tienda. ¿Cuál es el número esperado de personas que realizarán una compra? La respuesta es $\mu = np = (1\,000)(0.3) = 300$. Por tanto, para aumentar el número esperado de compras, la empresa debe lograr que más clientes entren en el establecimiento y/o aumentar de alguna manera la probabilidad de que un cliente realice una compra cuando esté adentro.

En este problema con tres clientes, vemos que la varianza y la desviación estándar del número de ellos que harán una compra es

$$\sigma^2 = np(1 - p) = 3(0.3)(0.7) = 0.63$$
$$\sigma = \sqrt{0.63} = 0.79$$

Para los próximos 1 000 clientes que entren en la tienda, la varianza y la desviación estándar del número de personas que harán una compra son

$$\sigma^2 = np(1 - p) = 1\,000(0.3)(0.7) = 210$$
$$\sigma = \sqrt{210} = 14.49$$

NOTAS Y COMENTARIOS

1. La tabla binomial del apéndice B muestra valores de p hasta $p = 0.95$, inclusive. Algunas fuentes de la tabla binomial sólo muestran valores de p hasta $p = 0.50$. Parecería que una tabla como ésta no puede usarse cuando la probabilidad de éxito rebasa $p = 0.50$. No obstante, puede utilizarse si se considera que la probabilidad de $n - x$ fracasos es también la probabilidad de x éxitos. Por tanto, cuando la probabilidad de éxito es mayor que $p = 0.50$, se calcula la probabilidad de $n - x$ fracasos en vez de la probabilidad de éxitos. La probabilidad de fracasos, $1 - p$, es menor que 0.50 cuando $p > 0.50$.

2. Algunas fuentes presentan las tablas binomiales en forma acumulada. Al usarlas para encontrar exactamente x éxitos en n ensayos, se deben restar las entradas de la tabla correspondiente. Por ejemplo, $f(2) = P(x \leq 2) - P(x \leq 1)$. La tabla binomial del apéndice B proporciona $f(2)$ directamente. Para calcular las probabilidades acumuladas usando las tablas binomiales del apéndice B, se suman las entradas de la tabla correspondiente. Por ejemplo, para determinar la probabilidad acumulada $P(x \leq 2)$, calcule la suma $f(0) + f(1) + f(2)$.

Ejercicios

Métodos

AUTO evaluación

25. Considere un experimento binomial con dos ensayos y $p = 0.4$.
    a) Trace un diagrama de árbol para este experimento (vea la figura 5.3).
    b) Calcule la probabilidad de un éxito, $f(1)$.
    c) Calcule $f(0)$.
    d) Estime $f(2)$.
    e) Calcule la probabilidad de por lo menos un éxito.
    f) Determine el valor esperado, la varianza y la desviación estándar.

26. Considere un experimento binomial con $n = 10$ y $p = 0.10$.
    a) Calcule $f(0)$.
    b) Estime $f(2)$.
    c) Calcule $P(x \leq 2)$.
    d) Determine $P(x \geq 1)$.
    e) Calcule $E(x)$.
    f) Estime $\text{Var}(x)$ y σ.

27. Considere un experimento binomial con $n = 20$ y $p = 0.70$.
    a) Calcule $f(12)$.
    b) Determine $f(16)$.
    c) Calcule $P(x \geq 16)$.
    d) Estime $P(x \leq 15)$.
    e) Calcule $E(x)$.
    f) Defina $\text{Var}(x)$ y σ.

Aplicaciones

28. Un estudio de Harris Interactive para Intercontinental Hotels & Resorts preguntó a los encuestados: "Cuando viaja por el mundo, ¿se aventura por cuenta propia para experimentar la cultura, o sigue con su grupo del *tour* y los itinerarios? El sondeo reveló que 23% de los encuestados se queda con su grupo de viaje (*USA Today,* 21 de enero de 2004).
    a) En una muestra de seis viajeros internacionales, ¿cuál es la probabilidad de que dos se queden con el grupo del *tour*?
    b) En una muestra de seis viajeros, ¿cuál es la probabilidad de que por lo menos dos permanezcan con su grupo de viaje?
    c) En una muestra de 10 viajeros, ¿cuál es la probabilidad de que ninguno se quede con el grupo del *tour*?

29. En San Francisco, 30% de los trabajadores toma diario el transporte público (*USA Today*, 21 de diciembre de 2005).
    a) En una muestra de 10 trabajadores, ¿cuál es la probabilidad de que exactamente tres tomen el transporte público todos los días?
    b) En una muestra de 10 trabajadores, ¿cuál es la probabilidad de que por lo menos tres aborden el transporte público todos los días?

AUTO evaluación

30. Cuando una máquina nueva funciona adecuadamente, sólo 3% de los artículos producidos resulta con defectos. Suponga que seleccionamos al azar dos partes producidas en la máquina y que nos interesa el número de partes defectuosas encontradas.
    a) Describa las condiciones bajo las cuales esta situación sería un experimento binomial.
    b) Trace un diagrama de árbol parecido al de la figura 5.3 que muestra este problema como un experimento de dos ensayos.
    c) ¿En cuántos resultados experimentales se encuentra exactamente un defecto?
    d) Calcule las probabilidades asociadas con no encontrar defecto, y hallar exactamente uno y dos defectos.

31. El 9% de los estudiantes universitarios en Estados Unidos tiene estados de cuenta de sus tarjetas de crédito mayores a $7 000 (*Reader's Digest,* julio de 2002). Suponga que 10 estudiantes fueron seleccionados al azar para entrevistarlos sobre el uso de tarjetas de crédito.
    a) ¿La selección de 10 estudiantes es un experimento binomial? Explique por qué.
    b) ¿Cuál es la probabilidad de que dos de los consultados tengan un estado de cuenta de su tarjeta de crédito mayor de $7 000?
    c) ¿Cuál es la probabilidad de que ninguno tenga un estado de cuenta mayor de $7 000?
    d) ¿Cuál es la probabilidad de que por lo menos tres tengan un estado de cuenta mayor de $7 000?

32. Los radares militares y sistemas de detección de misiles están diseñados para advertir a un país de un ataque enemigo. Una pregunta de fiabilidad de un sistema de este tipo permite determinar si éste es capaz de identificar un ataque y emitir una advertencia. Suponga que un sistema de detección particular tiene una probabilidad 0.90 de detectar un ataque con misiles. Use la distribución de probabilidad binomial para responder las preguntas siguientes.
    a) ¿Cuál es la probabilidad de que un solo sistema de detección capte un ataque?
    b) Si dos sistemas de detección se instalan en la misma zona y trabajan de forma independiente, ¿cuál es la probabilidad de que por lo menos uno detecte el ataque?
    c) Si se instalan tres sistemas, ¿cuál es la probabilidad de que por lo menos uno de ellos identifique el ataque?
    d) ¿Recomendaría el uso de sistemas de detección múltiple? Explique sus razones.

33. En 2001, el 50% de los estadounidenses creía que el país atravesaba por una recesión aunque técnicamente la economía no había mostrado dos trimestres consecutivos de crecimiento negativo (*Business Week*, 30 de julio de 2001). Para una muestra de 20 estadounidenses, realice los cálculos siguientes.
    a) Estime la probabilidad de que exactamente 12 personas creían que el país estaba en recesión.
    b) Calcule la probabilidad de que no más de cinco personas creían que el país pasaba por una recesión.
    c) ¿Cuántas personas esperaría que dijeran que el país atravesaba por una recesión?
    d) Calcule la varianza y la desviación estándar del número de personas que creían que el país estaba en recesión.

34. La Encuesta de Población actual de la Oficina del Censo muestra que 28% de los individuos, con edades de 25 y mayores, han completado cuatro años de universidad (*The New York Times Almanac,* 2006). Para una muestra de 15 individuos con edades de 25 y mayores, responda las preguntas siguientes.
    a) ¿Cuál es la probabilidad de que cuatro hayan completado cuatro años de universidad?
    b) ¿Cuál es la probabilidad de que tres o más hayan completado cuatro años de universidad?

35. Una universidad encontró que 20% de sus estudiantes se retiró sin completar el curso introductorio de estadística. Suponga que 20 alumnos se registraron para el curso.
    a) Calcule la probabilidad de que dos o menos estudiantes se retirarán.
    b) Determine la probabilidad de que exactamente cuatro abandonarán el curso.
    c) Calcule la probabilidad de que tres se retirarán.
    d) Estime el número esperado de retiros.

36. Una encuesta realizada por TD Ameritrade encontró que uno de cada cuatro inversionistas dispone de fondos cotizados en bolsa en sus portafolios (*USA Today*, 11 de enero de 2007). Considere una muestra de 20 inversionistas.
    a) Calcule la probabilidad de que exactamente cuatro inversionistas disponen de fondos cotizados en bolsa en sus portafolios.
    b) Calcule la probabilidad de que por lo menos dos tienen fondos cotizados en bolsa en sus portafolios.
    c) Si usted encuentra que exactamente 12 inversionistas disponen de fondos cotizados en bolsa en sus portafolios, ¿dudaría de la exactitud de los resultados de la encuesta?
    d) Calcule el número esperado de inversionistas que tienen fondos cotizados en bolsa en sus portafolios.

37. El 23% de los automóviles no cuenta con un seguro (CNN, 23 de febrero de 2006). En un fin de semana en particular, hubo 35 automóviles involucrados en accidentes de tráfico.
    a) ¿Cuál es el número esperado de estos vehículos que no cuenta con un seguro?
    b) ¿Cuáles son la varianza y la desviación estándar?

5.5 Distribución de probabilidad de Poisson

La distribución de probabilidad de Poisson a menudo se utiliza para modelar las llegadas aleatorias en situaciones de línea de espera.

En esta sección consideramos una variable aleatoria discreta que a menudo es útil para estimar el número de ocurrencias en un intervalo específico de tiempo o espacio. Por ejemplo, la variable aleatoria de interés podría ser el número de llegadas a un centro de lavado automotriz en una hora, el número de reparaciones necesarias en 10 millas de una autopista o el número de fugas en 100 millas de tubería. Si las dos propiedades siguientes se satisfacen, el número de ocurrencias es una variable aleatoria descrita por la **distribución de probabilidad de Poisson**.

PROPIEDADES DE UN EXPERIMENTO DE POISSON

1. La probabilidad de ocurrencia es la misma para cualesquiera dos intervalos de igual longitud.
2. La ocurrencia o no ocurrencia en cualquier intervalo es independiente de la ocurrencia o no ocurrencia en cualquier otro intervalo.

La **función de probabilidad de Poisson** se define por medio de la ecuación (5.11).

Simeón Poisson impartió matemáticas en la Ecole Polytechnique de París de 1802 a 1808. En 1837 publicó un trabajo titulado "Investigación sobre la probabilidad de los veredictos en materia penal y civil", que incluye un análisis de lo que más tarde se conoció como la distribución de Poisson.

FUNCIÓN DE PROBABILIDAD DE POISSON

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} \qquad \textbf{(5.11)}$$

donde

$f(x)$ = probabilidad de x ocurrencias en un intervalo
μ = valor esperado o número medio de ocurrencias en un intervalo
e = 2.71828

Para la distribución de probabilidad de Poisson, x es una variable aleatoria discreta que indica el número de ocurrencias en el intervalo. Como no hay un límite superior establecido para el número de ocurrencias, la función de probabilidad $f(x)$ es aplicable para los valores $x = 0, 1, 2, \ldots$ sin límite. En las aplicaciones prácticas, x a la larga se volverá lo suficientemente grande para que $f(x)$ sea aproximadamente cero y la probabilidad de cualquier valor mayor que x se vuelva insignificante.

Un ejemplo con intervalos de tiempo

Bell Labs usó la distribución de Poisson para modelar la entrada de llamadas telefónicas.

Suponga que le interesa conocer el número de llegadas al autocajero de un banco en las mañanas de lunes a viernes durante un periodo de 15 minutos. Si se asume que la probabilidad de un automóvil que llega es la misma para cualquiera de dos periodos de igual duración y que la llegada o no llegada de un vehículo en cualquier periodo es independiente del arribo o no en cualquier otro periodo, la función de probabilidad de Poisson es aplicable. Suponga que estos supuestos se cumplen y que un análisis de los datos históricos muestra que el número medio de automóviles que llega en un periodo de 15 minutos es 10; en este caso, se aplica la función de probabilidad siguiente.

$$f(x) = \frac{10^x e^{-10}}{x!}$$

La variable aleatoria aquí es x = número de automóviles que llega en un periodo de 15 minutos.

Si la gerencia quisiera conocer la probabilidad de exactamente cinco llegadas en 15 minutos, se establecería que $x = 5$ y por tanto obtendríamos

$$\text{Probabilidad de exactamente cinco llegadas en 15 minutos} = f(5) = \frac{10^5 e^{-10}}{5!} = 0.0378$$

Aunque esta probabilidad se determinó al evaluar la función de probabilidad con $\mu = 10$ y $x = 5$, a menudo es más fácil remitirse a una tabla para la distribución de Poisson, la cual proporciona probabilidades para valores específicos de x y μ. Se incluyó una similar a la tabla 7 del apéndice B. Por conveniencia, reproducimos una parte de ésta en la tabla 5.8. Observe que para usar la tabla de probabilidades de Poisson necesitamos conocer sólo los valores de x y μ. A partir de la tabla 5.8 vemos que la probabilidad de cinco llegadas en un periodo de 15 minutos se encuentra ubicando el valor en la fila de la tabla que corresponde a $x = 5$ y la columna que corresponde a $\mu = 10$. Por consiguiente, obtenemos $f(5) = 0.0378$.

Una propiedad de la distribución de Poisson consiste en que la media y la varianza son iguales.

En el ejemplo anterior, la media de la distribución de Poisson es $\mu = 10$ llegadas por un periodo de 15 minutos. Una propiedad de la distribución de Poisson consiste en que la media de la distribución y la varianza de la distribución son *iguales*. Por tanto, la varianza para el número de llegadas durante un periodo de 15 minutos es $\sigma^2 = 10$. La desviación estándar es $\sigma = \sqrt{10} = 3.16$.

El ejemplo involucra un periodo de 15 minutos, pero se pueden usar otros. Suponga que se quiere calcular la probabilidad de una llegada en un periodo de 3 minutos. Dado que 10 es el número esperado de llegadas en 15 minutos, vemos que $10/15 = 2/3$ es el número esperado de llegadas en 1 minuto y que $(2/3)(3 \text{ minutos}) = 2$ es el número esperado de arribos en 3 minutos. Por tanto, la probabilidad de x llegadas en un periodo de 3 minutos con $\mu = 2$ está dada por la función de probabilidad de Poisson siguiente.

$$f(x) = \frac{2^x e^{-2}}{x!}$$

TABLA 5.8 Valores seleccionados del ejemplo de las tablas de probabilidad de Poisson: $\mu = 10; x = 5; f(5) = 0.0378$

	μ									
x	**9.1**	**9.2**	**9.3**	**9.4**	**9.5**	**9.6**	**9.7**	**9.8**	**9.9**	**10**
0	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000
1	0.0010	0.0009	0.0009	0.0008	0.0007	0.0007	0.0006	0.0005	0.0005	0.0005
2	0.0046	0.0043	0.0040	0.0037	0.0034	0.0031	0.0029	0.0027	0.0025	0.0023
3	0.0140	0.0131	0.0123	0.0115	0.0107	0.0100	0.0093	0.0087	0.0081	0.0076
4	0.0319	0.0302	0.0285	0.0269	0.0254	0.0240	0.0226	0.0213	0.0201	0.0189
5	0.0581	0.0555	0.0530	0.0506	0.0483	0.0460	0.0439	0.0418	0.0398	**0.0378**
6	0.0881	0.0851	0.0822	0.0793	0.0764	0.0736	0.0709	0.0682	0.0656	0.0631
7	0.1145	0.1118	0.1091	0.1064	0.1037	0.1010	0.0982	0.0955	0.0928	0.0901
8	0.1302	0.1286	0.1269	0.1251	0.1232	0.1212	0.1191	0.1170	0.1148	0.1126
9	0.1317	0.1315	0.1311	0.1306	0.1300	0.1293	0.1284	0.1274	0.1263	0.1251
10	0.1198	0.1210	0.1219	0.1228	0.1235	0.1241	0.1245	0.1249	0.1250	0.1251
11	0.0991	0.1012	0.1031	0.1049	0.1067	0.1083	0.1098	0.1112	0.1125	0.1137
12	0.0752	0.0776	0.0799	0.0822	0.0844	0.0866	0.0888	0.0908	0.0928	0.0948
13	0.0526	0.0549	0.0572	0.0594	0.0617	0.0640	0.0662	0.0685	0.0707	0.0729
14	0.0342	0.0361	0.0380	0.0399	0.0419	0.0439	0.0459	0.0479	0.0500	0.0521
15	0.0208	0.0221	0.0235	0.0250	0.0265	0.0281	0.0297	0.0313	0.0330	0.0347
16	0.0118	0.0127	0.0137	0.0147	0.0157	0.0168	0.0180	0.0192	0.0204	0.0217
17	0.0063	0.0069	0.0075	0.0081	0.0088	0.0095	0.0103	0.0111	0.0119	0.0128
18	0.0032	0.0035	0.0039	0.0042	0.0046	0.0051	0.0055	0.0060	0.0065	0.0071
19	0.0015	0.0017	0.0019	0.0021	0.0023	0.0026	0.0028	0.0031	0.0034	0.0037
20	0.0007	0.0008	0.0009	0.0010	0.0011	0.0012	0.0014	0.0015	0.0017	0.0019
21	0.0003	0.0003	0.0004	0.0004	0.0005	0.0006	0.0006	0.0007	0.0008	0.0009
22	0.0001	0.0001	0.0002	0.0002	0.0002	0.0002	0.0003	0.0003	0.0004	0.0004
23	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002
24	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001

La probabilidad de una llegada en un periodo de 3 minutos se calcula como sigue:

$$\text{Probabilidad de exactamente 1 llegada en 3 minutos} = f(1) = \frac{2^1 e^{-2}}{1!} = 0.2707$$

Previamente se calculó la probabilidad de cinco llegadas en un periodo de 15 minutos; fue 0.0378. Observe que la probabilidad de un arribo en 3 minutos (0.2707) no es la misma. Cuando se estima una probabilidad de Poisson para un intervalo de tiempo distinto, primero se debe convertir la tasa media de llegadas al periodo de interés y luego calcular la probabilidad.

Un ejemplo con intervalos de longitud o de distancia

Se demostrará una aplicación que no tiene intervalos de tiempo en la que es útil la distribución de Poisson. Suponga que le interesa saber cuál es la ocurrencia de defectos importantes en una autopista un mes después de repavimentarla. Considere que la probabilidad de un defecto es la misma en cualquiera de dos intervalos de igual longitud de la autopista, y que la ocurrencia o no ocurrencia de defectos en cualquier intervalo es independiente de su ocurrencia o no en cualquier otro intervalo. Por ende, la distribución de Poisson puede aplicarse.

Suponga que se enteró de que los principales defectos después de un mes de repavimentar ocurren a una tasa media de 2 por milla. En seguida se determinará la probabilidad de que no hay defectos importantes en un tramo particular de 3 millas de la autopista. Como nos interesa un intervalo con esta longitud, $\mu = (2 \text{ defectos/milla})(3 \text{ millas}) = 6$ representa el número esperado de anomalías importantes en este tramo de la autopista. Mediante la ecuación (5.11), la probabilidad de que no haya alguna avería importante es $f(0) = 6^0 e^{-6}/0! = 0.0025$. Por tanto, es poco probable que ningún defecto importante se presente en la sección de las 3 millas. De hecho, este ejemplo indica que $1 - 0.0025 = 0.9975$ es la probabilidad de por lo menos un defecto importante en la sección de 3 millas de la autopista.

Ejercicios

Métodos

38. Considere una distribución de Poisson con $\mu = 3$.
    a) Escriba una función de probabilidad de Poisson apropiada.
    b) Calcule $f(2)$.
    c) Determine $f(1)$.
    d) Calcule $P(x \geq 2)$.

AUTOevaluación

39. Considere una distribución de Poisson con una media de dos ocurrencias por periodo.
    a) Escriba una función de probabilidad de Poisson apropiada.
    b) ¿Cuál es el número esperado de ocurrencias en tres periodos?
    c) Escriba una función de probabilidad de Poisson apropiada para determinar la probabilidad de ocurrencias en tres periodos.
    d) Calcule la probabilidad de dos ocurrencias en un periodo.
    e) Estime la probabilidad de seis ocurrencias en tres periodos.
    f) Calcule la probabilidad de cinco ocurrencias en dos periodos.

Aplicaciones

40. Las llamadas telefónicas entran a una razón de 48 por hora en la oficina de reservaciones de Regional Airways.
    a) Calcule la probabilidad de recibir tres llamadas en un intervalo de 5 minutos.
    b) Estime la probabilidad de recibir exactamente 10 llamadas en 15 minutos.
    c) Suponga que actualmente no hay llamada en espera. Si el agente tarda 5 minutos en terminar la llamada actual, ¿cuántas personas estimaría que estuvieran esperando en el teléfono para ese entonces? ¿Cuál es la probabilidad de que no haya llamada en espera?
    d) Si no se procesa actualmente alguna llamada, ¿cuál es la probabilidad de que el agente tarde 3 minutos en un asunto personal sin ser interrumpido por una llamada?

41. Durante el periodo en que una universidad local hace registros por teléfono, las llamadas entran a una razón de una cada 2 minutos.
    a) ¿Cuál es el número esperado de llamadas en una hora?
    b) ¿Cuál es la probabilidad de tres llamadas en 5 minutos?
    c) ¿Cuál es la probabilidad de que no haya llamadas en un periodo de 5 minutos?

AUTOevaluación

42. Cada año más de 50 millones de huéspedes se hospedan en hoteles que ofrecen alojamiento y desayuno. El sitio web para Bed and Breakfast Inns de Norteamérica, que recibe un promedio de siete visitantes por minuto, permite a muchos hoteles de este tipo atraer clientes (*Time,* septiembre de 2001).
    a) Calcule la probabilidad de que nadie visite el sitio web en un periodo de un minuto.
    b) Estime la probabilidad de dos o más visitantes al sitio web en un periodo de un minuto.
    c) Calcule la probabilidad de uno o más visitantes en un periodo de 30 segundos.
    d) Determine la probabilidad de cinco o más visitantes en un periodo de un minuto.

43. Los pasajeros de una línea aérea llegan al azar y de manera independiente a la instalación de revisión de pasajeros en un aeropuerto internacional. La razón media de llegadas es de 10 personas por minuto.
    a) Calcule la probabilidad de que no haya llegadas en un periodo de un minuto.
    b) Determine la probabilidad de que tres pasajeros o menos lleguen en un periodo de un minuto.
    c) Calcule la probabilidad de que no haya llegadas en un periodo de 15 segundos.
    d) Estime la probabilidad de cuando menos una llegada en un periodo de 15 segundos.

44. Cada año ocurre un promedio de 15 accidentes aéreos (*The World Almanac and Book of Facts,* 2004).
    a) Calcule el número medio de accidentes aéreos por mes.
    b) Determine la probabilidad de que no ocurran percances durante un mes.
    c) Calcule la probabilidad de exactamente un accidente al mes.
    d) Estime la probabilidad de que ocurra más de un accidente mensual.

45. El Consejo de Seguridad Nacional de Estados Unidos estima que los accidentes fuera del trabajo le cuestan a las empresas del país casi $200 000 millones al año en productividad perdida (Consejo de Seguridad Nacional, marzo de 2006). Con base en las estimaciones de la institución, se espera que las empresas con 50 empleados promedien tres accidentes fuera del trabajo por año. Responda las preguntas siguientes para las empresas con 50 empleados.
    a) ¿Cuál es la probabilidad de que no ocurran accidentes fuera del trabajo durante un periodo de un año?
    b) ¿Cuál es la probabilidad de que ocurran por lo menos dos percances fuera del trabajo durante un periodo de un año?
    c) ¿Cuál es el número esperado de accidentes fuera del trabajo durante seis meses?
    d) ¿Cuál es la probabilidad de accidentes fuera del trabajo durante los seis meses siguientes?

5.6 Distribución de probabilidad hipergeométrica

La **distribución de probabilidad hipergeométrica** mantiene una relación estrecha con la distribución binomial, pero difiere de ésta en dos puntos esenciales: sus ensayos no son independientes y su probabilidad de éxito cambia de un ensayo a otro.

En la notación usual para la distribución hipergeométrica, r denota el número de elementos en la población de tamaño N considerados como éxitos, y $N - r$ denota el número de elementos en la población considerados fracasos. La **función de probabilidad hipergeométrica** se usa para calcular la probabilidad de que en una muestra aleatoria de n elementos, seleccionados sin remplazo, se obtengan x elementos etiquetados como éxitos y $n - x$ elementos marcados como fracasos. Para que este resultado ocurra, se deben obtener x éxitos de los r éxitos que hay en la población y $n - x$ fracasos de los $N - r$ fracasos. La función de probabilidad hipergeométrica siguiente proporciona $f(x)$, la probabilidad de obtener x éxitos en n ensayos.

FUNCIÓN DE PROBABILIDAD HIPERGEOMÉTRICA

$$f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \qquad \textbf{(5.12)}$$

donde

x = número de éxitos
n = número de ensayos
$f(x)$ = probabilidad de x éxitos en n ensayos
N = número de elementos en la población
r = número de elementos en la población etiquetados como éxitos

Observe que $\binom{N}{n}$ representa el número de maneras en que n elementos pueden seleccionarse de una población de tamaño N; $\binom{r}{x}$ expresa el número de formas en que x éxitos pueden seleccionarse de un total de r éxitos en la población, y $\binom{N-r}{n-x}$ representa el número de maneras en que $n - x$ fracasos pueden elegirse de un total de $N - r$ fracasos en la población.

Para la distribución de probabilidad hipergeométrica, x es una variable aleatoria discreta, y la función de probabilidad $f(x)$ dada por la ecuación (5.12) por lo general se aplica a los valores de $x = 0, 1, 2, \ldots, n$. Sin embargo, sólo son válidos los valores de x donde el número de éxitos observados es *menor o igual* que el número de éxitos en la población ($x \leq r$) y donde el número de fracasos observados es *menor o igual* que el número de fracasos en la población ($n - x \leq N - r$). Si estas dos condiciones no son válidas para uno o más valores de x, la $f(x) = 0$ correspondiente indica que la probabilidad de este valor de x es cero.

Para ilustrar los cálculos que implica el uso de la ecuación (5.12), considere la siguiente aplicación de control de calidad. Los fusibles eléctricos producidos por Ontario Electric se empacan en cajas de 12 unidades cada una. Suponga que un inspector selecciona al azar tres de los 12 fusibles de una caja para probarlos. Si ésta contiene exactamente cinco fusibles averiados, ¿cuál es la probabilidad de que el inspector encuentre exactamente un fusible defectuoso en los tres que seleccionó? En esta aplicación $n = 3$ y $N = 12$. Con $r = 5$ fusibles defectuosos en la caja, la probabilidad de encontrar $x = 1$ fusible defectuoso es

$$f(1) = \frac{\binom{5}{1}\binom{7}{2}}{\binom{12}{3}} = \frac{\left(\frac{5!}{1!4!}\right)\left(\frac{7!}{2!5!}\right)}{\left(\frac{12!}{3!9!}\right)} = \frac{(5)(21)}{220} = 0.4773$$

Ahora suponga que quiere conocer la probabilidad de encontrar *por lo menos* 1 fusible defectuoso. La manera más fácil de responder esta pregunta consiste en calcular primero la probabilidad de que el inspector no encuentre un fusible en mal estado. La probabilidad de $x = 0$ es

$$f(0) = \frac{\binom{5}{0}\binom{7}{3}}{\binom{12}{3}} = \frac{\left(\frac{5!}{0!5!}\right)\left(\frac{7!}{3!4!}\right)}{\left(\frac{12!}{3!9!}\right)} = \frac{(1)(35)}{220} = 0.1591$$

Con una probabilidad de cero fusibles defectuosos $f(0) = 0.1591$, concluimos que la probabilidad de encontrar por lo menos uno debe ser $1 - 0.1591 = 0.8409$. Por tanto, hay una probabilidad razonablemente alta de que el inspector encuentra por lo menos 1 fusible defectuoso.

La media y la varianza de una distribución hipergeométrica son las siguientes.

$$E(x) = \mu = n\left(\frac{r}{N}\right) \tag{5.13}$$

$$\text{Var}(x) = \sigma^2 = n\left(\frac{r}{N}\right)\left(1 - \frac{r}{N}\right)\left(\frac{N-n}{N-1}\right) \tag{5.14}$$

En el ejemplo anterior, $n = 3$, $r = 5$ y $N = 12$. Por tanto, la media y la varianza para el número de fusibles defectuosos son

$$\mu = n\left(\frac{r}{N}\right) = 3\left(\frac{5}{12}\right) = 1.25$$

$$\sigma^2 = n\left(\frac{r}{N}\right)\left(1 - \frac{r}{N}\right)\left(\frac{N-n}{N-1}\right) = 3\left(\frac{5}{12}\right)\left(1 - \frac{5}{12}\right)\left(\frac{12-3}{12-1}\right) = 0.60$$

La desviación estándar es $\sigma = \sqrt{0.60} = 0.77$.

NOTAS Y COMENTARIOS

Considere una distribución hipergeométrica con n ensayos. Sea $p = (r/N)$ que denota la probabilidad de un éxito en el primero ensayo. Si el tamaño de la población es grande, el término $(N - n)/(N - 1)$ en la ecuación (5.14) se aproxima a 1. Como resultado, el valor esperado y la varianza se escriben $E(x) = np$ y $\text{Var}(x) = np(1 - p)$. Note que estas expresiones son las mismas que las usadas para calcular el valor esperado y la varianza de una distribución binomial, como en las ecuaciones (5.9) y (5.10).

Cuando el tamaño de la población es grande, una distribución hipergeométrica puede aproximarse por una distribución binomial con n ensayos y una probabilidad de éxito de $p = (r/N)$.

Ejercicios

Métodos

AUTOevaluación

46. Suponga que $N = 10$ y $r = 3$. Calcule las probabilidades hipergeométricas para los valores siguientes de n y x.
    a) $n = 4, x = 1$.
    b) $n = 2, x = 2$.
    c) $n = 2, x = 0$.
    d) $n = 4, x = 2$.
    e) $n = 4, x = 4$.

47. Suponga que $N = 15$ y $r = 4$. ¿Cuál es la probabilidad de $x = 3$ para $n = 10$?

Aplicaciones

48. En un estudio realizado por Gallup Organization se preguntó a los encuestados: "¿Cuál es su deporte favorito para ver?" El futbol americano y el basquetbol clasificaron como número uno y dos respectivamente en cuanto a preferencia (sitio web de Gallup, 3 de enero de 2004). Suponga que en un grupo de 10 individuos, siete prefieren el futbol americano y tres el basquetbol. Seleccionemos una muestra al azar de tres de estos individuos.
    a) ¿Cuál es la probabilidad de que exactamente dos prefieran el futbol americano?
    b) ¿Cuál es la probabilidad de que la mayoría (ya sea dos o tres) prefiera el futbol americano?

49. El blackjack o veintiuno, como se le llama con frecuencia, es un juego de apuestas popular en los casinos de Las Vegas. A un jugador se le reparten dos cartas. Las figuras (jotas, reinas y reyes) y los dieces tienen un valor de 10. Los ases tienen un valor de 1 u 11. Una baraja de 52 cartas contiene 16 con un valor de puntos de 10 (jotas, reinas, reyes y dieces) y cuatro ases.
   a) ¿Cuál es la probabilidad de que las dos cartas repartidas sean ases o cartas de 10 puntos?
   b) ¿Cuál es la probabilidad de que ambas sean ases?
   c) ¿Cuál es la probabilidad de que las dos tengan un valor de 10?
   d) Un blackjack es una carta de 10 puntos y un as que dan un valor de 21. Use las respuestas de los incisos *a*), *b*) y *c*) para determinar la probabilidad de que a un jugador le repartan un blackjack. [*Pista*. El inciso *d*) no es un problema hipergeométrico. Elabore una relación lógica propia de cómo las probabilidades hipergeométricas de los incisos *a*), *b*) y *c*) pueden combinarse para responder esta pregunta.]

AUTO evaluación

50. Axline Computers fabrica computadoras personales en dos plantas, una en Texas y la otra en Hawaii. La planta de Texas cuenta con 40 empleados y la de Hawaii con 20. A una muestra aleatoria de 10 empleados se le pedirá que llene un cuestionario de beneficios.
   a) ¿Cuál es la probabilidad de que ninguno de los empleados de la muestra trabaje en la planta de Hawaii?
   b) ¿Cuál es la probabilidad de que uno de estos empleados trabaje en la planta de Hawaii?
   c) ¿Cuál es la probabilidad de que dos o más sujetos de la muestra labore en la planta de Hawaii?
   d) ¿Cuál es la probabilidad de que nueve de los empleados trabajen en la planta de Texas?

51. La encuesta de restaurantes de ZAGAT proporciona las calificaciones de los platillos, la decoración y el servicio de algunos restaurantes de Estados Unidos. Para 15 establecimientos ubicados en Boston, el precio medio de una cena, incluyendo una bebida y la propina, es de $48.60. Usted está de viaje de negocios en Boston y cenará en tres de estos restaurantes. Su empresa le rembolsará un máximo de $50 por cena. Los socios de negocios familiarizados con estos establecimientos le han dicho que el costo de la cena en un tercio de los restaurantes de la encuesta rebasa los $50. Suponga que selecciona al azar tres de estos negocios para comer.
   a) ¿Cuál es la probabilidad de que ninguna de las cenas rebase el costo que cubre su empresa?
   b) ¿Cuál es la probabilidad de que una de las cenas supere el costo que cubre su empresa?
   c) ¿Cuál es la probabilidad de que dos de las cenas rebasen tal costo?
   d) ¿Cuál es la probabilidad de que tres de las cenas rebasen dicho costo?

52. El Troubled Asset Relief Program (TARP), aprobado por el Congreso de Estados Unidos en octubre de 2008, aportó $700 000 millones como apoyo financiero para que la economía del país saliera adelante. Más de $200 000 millones se destinaron a instituciones financieras con problemas con la esperanza de que hubiera un incremento en los créditos para ayudar a reactivar la economía. Pero tres meses después, una encuesta de la Reserva Federal reveló que dos tercios de los bancos que recibieron fondos del TARP habían restringido las condiciones de los créditos empresariales (*The Wall Street Journal*, 3 de febrero de 2009). De los 10 principales bancos receptores de fondos del TARP, sólo tres incrementaron realmente los créditos durante el periodo.

Incremento en los créditos	**Disminución en los créditos**
BB&T	Bank of America
Sun Trust Banks	Capital One
U.S. Bancorp	Citigroup
	Fifth Third Bancorp
	J.P. Morgan Chase
	Regions Financial
	U.S. Bancorp

En este ejercicio, suponga que se seleccionán al azar tres de estos 10 bancos para efectuar un estudio que permitirá seguir supervisando las prácticas crediticias de estas instituciones. Sea x una variable aleatoria que indica el número de bancos en el estudio que incrementaron sus créditos.

a) ¿Cuánto es $f(0)$? ¿Cuál es su interpretación de este valor?
b) ¿Cuánto es $f(3)$? ¿Cuál es su interpretación de este valor?
c) Calcule $f(1)$ y $f(2)$. Muestre la distribución de probabilidad para el número de bancos en el estudio que incrementaron sus créditos. ¿Qué valor de x tiene la mayor probabilidad?
d) ¿Cuál es la probabilidad de que el estudio muestre por lo menos un banco que incrementó sus créditos?
e) Calcule el valor esperado, la varianza y la desviación estándar para la variable aleatoria.

Resumen

Una variable aleatoria proporciona una descripción numérica del resultado de un experimento. La distribución de probabilidad de esta variable describe cómo se distribuyen las probabilidades entre los valores que la misma puede asumir. Para cualquier variable aleatoria discreta x, la distribución de probabilidad se define por medio de una función de probabilidad, denotada por $f(x)$, que proporciona la probabilidad asociada con cada valor de la variable aleatoria. Una vez que la función de probabilidad se define, puede calcularse el valor esperado, la varianza y desviación estándar de dicha variable.

La distribución binomial se utiliza para determinar la probabilidad de x éxitos en n ensayos siempre que el experimento tenga las propiedades siguientes:

1. El experimento consiste de una secuencia de n ensayos idénticos.
2. En cada ensayo dos resultados son posibles: uno llamado *éxito* y el otro, *fracaso*.
3. La probabilidad de un éxito p no cambia de un ensayo a otro. En consecuencia, la probabilidad de fracaso, $1 - p$, tampoco cambia de un ensayo a otro.
4. Los ensayos son independientes.

Cuando las cuatro propiedades se cumplen, la función de probabilidad binomial se utiliza para determinar la probabilidad de obtener x éxitos en n ensayos. En este capítulo también se presentaron las fórmulas para la media y la varianza de la distribución binomial.

La distribución de Poisson se usa cuando se desea determinar la probabilidad de obtener x ocurrencias en un intervalo de tiempo o espacio. Es necesario que se cumplan los supuestos siguientes para que la distribución de Poisson pueda aplicarse.

1. La probabilidad de una ocurrencia del evento es la misma para dos intervalos cualesquiera de igual longitud.
2. La ocurrencia o no ocurrencia del evento en cualquier intervalo es independiente de su ocurrencia o no ocurrencia en cualquier otro intervalo.

Una tercera distribución de probabilidad discreta, la hipergeométrica, se presentó en la sección 5.6. Al igual que la binomial, esta distribución se utiliza para calcular la probabilidad de x éxitos en n ensayos. Pero, a diferencia de la binomial, la probabilidad de éxito cambia de ensayo a ensayo.

Glosario

Desviación estándar Raíz cuadrada positiva de la varianza.
Distribución de probabilidad Descripción de cómo se distribuyen las probabilidades entre los valores de una variable aleatoria.
Distribución de probabilidad de Poisson Muestra la probabilidad de x ocurrencias de un evento a lo largo de un intervalo de tiempo o espacio específicos.

Distribución de probabilidad binomial Muestra la probabilidad de x éxitos en n ensayos de un experimento binomial.
Distribución de probabilidad hipergeométrica Describe la probabilidad de x éxitos en n ensayos de una población con r éxitos y $N - r$ fracasos.
Distribución de probabilidad uniforme discreta Distribución de probabilidad para la cual cada valor posible de la variable aleatoria tiene la misma probabilidad.
Experimento binomial Experimento que tiene las cuatro propiedades establecidas al principio de la sección 5.4.
Función de probabilidad Función, denotada por $f(x)$, que proporciona la probabilidad de que x asuma un valor particular para una variable aleatoria discreta.
Función de probabilidad binomial Se utiliza para calcular las probabilidades binomiales.
Función de probabilidad de Poisson Función usada para determinar las probabilidades de Poisson.
Función de probabilidad hipergeométrica Función utilizada para calcular las probabilidades hipergeométricas.
Valor esperado Medida de la ubicación central de una variable aleatoria.
Variable aleatoria Descripción numérica del resultado de un experimento.
Variable aleatoria continua Variable que puede asumir cualquier valor numérico en un intervalo o conjunto de intervalos.
Variable aleatoria discreta Variable que puede asumir cualquier número finito de valores o una secuencia infinita de valores.
Varianza Medida de la variabilidad, o dispersión, de una variable aleatoria.

Fórmulas clave

Función de probabilidad uniforme discreta

$$f(x) = 1/n \tag{5.3}$$

Valor esperado de una variable aleatoria discreta

$$E(x) = \mu = \Sigma x f(x) \tag{5.4}$$

Varianza de una variable aleatoria discreta

$$\text{Var}(x) = \sigma^2 = \Sigma(x - \mu)^2 f(x) \tag{5.5}$$

Número de resultados experimentales que proporcionan exactamente x éxitos en n ensayos

$$\binom{n}{x} = \frac{n!}{x!(n - x)!} \tag{5.6}$$

Función de probabilidad binomial

$$f(x) = \binom{n}{x} p^x (1 - p)^{(n-x)} \tag{5.8}$$

Valor esperado de una distribución binomial

$$E(x) = \mu = np \tag{5.9}$$

Varianza de una distribución binomial

$$\text{Var}(x) = \sigma^2 = np(1 - p) \tag{5.10}$$

Función de probabilidad de Poisson

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} \tag{5.11}$$

Función de probabilidad hipergeométrica

$$f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \tag{5.12}$$

Valor esperado de una distribución hipergeométrica

$$E(x) = \mu = n\left(\frac{r}{N}\right) \tag{5.13}$$

Varianza de una distribución hipergeométrica

$$\text{Var}(x) = \sigma^2 = n\left(\frac{r}{N}\right)\left(1 - \frac{r}{N}\right)\left(\frac{N-n}{N-1}\right) \tag{5.14}$$

Ejercicios complementarios

53. El estudio de Big Money de *Barron's* preguntó a 131 gerentes de inversiones de todo Estados Unidos su perspectiva sobre la inversión a corto plazo (*Barron's,* 28 de octubre de 2002). Sus respuestas mostraron los siguientes indicadores: 4% eran muy optimistas; 39% optimistas; 29% neutrales; 21% pesimistas, y 7% muy pesimistas. Sea x la variable aleatoria que refleja el nivel de optimismo con respecto al mercado y que asume los valores $x = 5$ para muy optimista hasta $x = 1$ para muy pesimista.
    a) Elabore una distribución de probabilidad para el nivel de optimismo de los gerentes de inversiones.
    b) Calcule el valor esperado para el nivel de optimismo.
    c) Calcule su varianza y desviación estándar.
    d) Comente qué indican sus resultados sobre el nivel de optimismo y su variabilidad.

54. La Asociación Estadounidense de Inversionistas Individuales publica una guía anual para los principales fondos de inversión (*The Individual Investor's Guide to the Top Mutual Funds,* 22a. ed., American Association of Individual Investors, 2003). La clasificación del riesgo total para 29 categorías de fondos de inversión se muestra a continuación.

Riesgo total	Número de categorías de fondos
Bajo	7
Por debajo del promedio	6
Promedio	3
Por encima del promedio	6
Alto	7

    a) Sea $x = 1$ para el riesgo bajo y hasta $x = 5$ para el riesgo alto; elabore una distribución de probabilidad para el nivel de riesgo.
    b) ¿Cuáles son el valor esperado y la varianza para el riesgo total?
    c) Resulta que 11 de las categorías eran fondos de bonos. Para estos últimos, siete categorías se clasificaron como bajas, y cuatro por debajo del promedio. Compare el riesgo total de los fondos de bonos con las 18 categorías de los fondos de acciones.

55. La preparación del presupuesto de una universidad de la región central de Estados Unidos generó los siguientes pronósticos de gastos para el año próximo (en millones de dólares): $9, $10, $11, $12 y $13. Como se conocen los gastos actuales, se asignaron las probabilidades respectivas siguientes: 0.3, 0.2, 0.25, 0.05 y 0.2.
    a) Muestre la distribución de probabilidad para el pronóstico de gastos.
    b) ¿Cuál es el valor esperado de este pronóstico para el año próximo?
    c) ¿Cuál es la varianza del pronóstico de gastos para el año próximo?
    d) Si las proyecciones de ingresos estimadas para el año son $12 millones, comente cuál es la posición financiera del colegio.

56. Un estudio reveló que en promedio una persona tarda alrededor de 26 minutos en trasladarse de su casa al trabajo o viceversa. Además, 5% de los encuestados informó que tarda más de una hora en ir o regresar del trabajo (sitio web de Bureau of Transportation Statistics, 11 de enero de 2004).
    a) Si 20 personas se encuestan un día en particular, ¿cuál es la probabilidad de que tres de ellas informen que tardan más de una hora en trasladarse?
    b) Si 20 personas se encuestan un día en particular, ¿cuál es la probabilidad de que ninguna informe que tarda más de una hora en trasladarse?
    c) Si una empresa tiene 2 000 empleados, ¿cuál es el número esperado de empleados que tardan más de una hora en trasladarse de su trabajo a su casa o viceversa?
    d) Si una empresa tiene 2 000 empleados, ¿cuáles son la varianza y la desviación estándar del número de ellos que tardan más de una hora en trasladarse?

57. Un grupo de acción política prevé entrevistar a los propietarios de casas para evaluar el impacto causado por una caída reciente de los precios de la vivienda. Según el estudio de finanzas personales de *The Wall Street Journal*/Harris Interactive, 26% de los individuos de 18–34 años, 50% del grupo de 35-44 años y 88% de los individuos mayores de 55 años son propietarios de una vivienda (sitio web de All Business, 23 de enero de 2008).
    a) ¿Cuántas personas del grupo de edades de entre 18 y 34 años deben incluirse en la muestra para encontrar un número esperado de al menos 20 propietarios de una casa?
    b) ¿Cuántas personas del grupo de 35-44 años de edad deben incluirse en la muestra para encontrar un número esperado de al menos 20 propietarios de una vivienda?
    c) ¿Cuántos sujetos de 55 años y más deben considerarse para encontrar un número esperado de al menos 20 propietarios de una vivienda?
    d) Si el número de 18-34 años de la muestra es igual al valor identificado en el inciso *a*), ¿cuál es la desviación estándar del número de personas que serán propietarias?
    e) Si el número de 35-44 años de la muestra es igual al valor indicado en el inciso *b*), ¿cuál es la desviación estándar del número de personas que serán propietarias de una vivienda?

58. Muchas empresas usan una técnica de control de calidad conocida como *muestreo de aceptación* para monitorear los envíos entrantes de partes, materias primas, etc. En la industria electrónica, los proveedores por lo general envían los componentes en lotes grandes. La inspección de una muestra de n componentes se considera como los n ensayos de un experimento binomial. El resultado de la prueba de cada componente (ensayo) es que éste se clasifique como bueno o defectuoso. Reynolds Electronics acepta un lote de cierto proveedor si los componentes defectuosos del lote no rebasan 1%. Suponga que se prueba una muestra aleatoria de cinco artículos de un embarque reciente.
    a) Asuma que 1% del embarque está defectuoso. Calcule la probabilidad de que ningún componente de la muestra está averiado.
    b) Suponga que 1% del embarque está defectuoso. Calcule la probabilidad de que exactamente uno de los componentes de la muestra tenga defectos.
    c) ¿Cuál es la probabilidad de observar una o más partes defectuosas en la muestra si 1% del embarque lo está?
    d) ¿Se sentiría cómodo al aceptar el embarque si se encontró que un componente estaba defectuoso? ¿Por qué?

59. La tasa de desempleo en el estado de Arizona es de 4.1% (sitio web de CNN Money, 2 de mayo de 2007). Suponga que 100 personas disponibles para un empleo en Arizona son seleccionadas al azar.
    a) ¿Cuál es el número esperado de personas desempleadas?
    b) ¿Cuáles son la varianza y la desviación estándar del número de personas sin empleo?

60. Un estudio realizado por Zogby International reveló que de aquellos estadounidenses para quienes la música desempeña un papel "muy importante" en su vida, 30% dijeron que sus estaciones de radio locales "siempre" transmiten el tipo de música que les gusta (sitio web de Zogby, 12 de enero de 2004). Suponga que se toma una muestra de 800 personas para quienes la música desempeña un papel importante en su vida.
    a) ¿Cuántas personas esperaría que dijeran que sus estaciones de radio locales siempre transmiten el tipo de música que les gusta?
    b) ¿Cuál es la desviación estándar del número de encuestados que piensa que sus estaciones de radio locales siempre transmiten el tipo de música que les agrada?
    c) ¿Cuál es la desviación estándar del número de encuestados que no piensa que sus estaciones de radio locales difunden la música de su preferencia?

61. En un lavado automotriz los automóviles llegan de manera aleatoria e independiente; la probabilidad de un arribo es la misma para cualesquier dos intervalos de tiempo de igual duración. La tasa de llegada media es 15 vehículos por hora. ¿Cuál es la probabilidad de que durante una hora cualquiera de operación lleguen 20 o más automóviles?

62. En un nuevo proceso de producción automatizada hay un promedio de 1.5 interrupciones por día. Debido a los costos asociados con una interrupción, la gerencia está preocupada por la posibilidad de que haya tres o más durante el día. Suponga que éstas ocurren aleatoriamente, que la probabilidad de interrupción es la misma para cualesquiera dos intervalos de tiempo de igual duración, y que las interrupciones en un lapso son independientes de las que ocurren en otro lapso. ¿Cuál es la probabilidad de que haya tres o más durante un día?

63. Un director regional responsable del desarrollo de negocios en el estado de Pennsylvania está preocupado por el número de quiebras de las empresas pequeñas. Si el número medio de estas quiebras por mes es 10, ¿cuál es la probabilidad de que exactamente cuatro empresas pequeñas incurran en esta situación durante un mes determinado? Suponga que la probabilidad de una quiebra es la misma para dos meses cualesquiera y que su ocurrencia o no ocurrencia en algún mes es independiente de las quiebras en cualquier otro mes.

64. Las llegadas de los clientes a un banco son aleatorias e independientes, y la probabilidad de un arribo en un periodo de un minuto es la misma que en cualquier otro periodo de un minuto. Responda las preguntas siguientes suponiendo una tasa media de llegadas de tres clientes por minuto.
    a) ¿Cuál es la probabilidad de exactamente tres llegadas en un periodo de un minuto?
    b) ¿Cuál es la probabilidad de por lo menos tres llegadas en un periodo de un minuto?

65. Una baraja contiene 52 cartas, cuatro de las cuales son ases. ¿Cuál es la probabilidad de que al repartir las cartas en una mano de cinco se obtengan los siguientes casos?
    a) Un par de ases.
    b) Exactamente un as.
    c) Ningún as.
    d) Por lo menos un as.

66. Durante la semana que terminó el 16 de septiembre de 2001, Tiger Woods fue el golfista que más dinero ganó en el PGA Tour. Sus ganancias sumaron un total de $5 517 777. De los 10 principales golfistas mejor remunerados, siete usaron pelotas de golf de la marca Titleist (sitio web de PGA Tour). Suponga que seleccionan al azar a dos de los 10 principales golfistas que ganan más dinero.
    a) ¿Cuál es la probabilidad de que exactamente uno use una pelota de golf Titleist?
    b) ¿Cuál es la probabilidad de que ambos usen pelotas Titleist?
    c) ¿Cuál es la probabilidad de que ninguno use esta marca de pelota?

Apéndice 5.1 Distribuciones de probabilidad discretas con Minitab

El software estadístico de Minitab ofrece un procedimiento relativamente fácil y eficiente para calcular probabilidades binomiales. En este apéndice se describe paso a paso el procedimiento para determinar las probabilidades binomiales para el problema de Martin Clothing Store de la sección 5.4. Recuerde que las probabilidades binomiales buscadas se basan en $n = 10$ y $p = 0.30$. Antes de comenzar con la rutina de Minitab, el usuario debe introducir los valores deseados de la variable aleatoria x en una columna de la hoja de trabajo. En el ejemplo de la figura 5.5 se introdujeron los valores 0, 1, 2, . . . , 10 en la columna 1 para generar la distribución de probabilidad binomial completa. Los pasos de Minitab para obtener las probabilidades deseadas se describen a continuación.

Paso 1. Seleccione el menú **Calc**.
Paso 2. Elija **Probability Distributions**.
Paso 3. Seleccione **Binomial**.
Paso 4. Cuando aparezca el cuadro de diálogo Binomial Distribution:
Seleccione **Probability**.
Introduzca 10 en el cuadro **Number of trials**.
Introduzca 0.3 en el cuadro **Event probability**.
Introduzca C1 en el cuadro **Input column**.
Haga haga clic en **OK**.

La salida de Minitab con las probabilidades binomiales aparecerá como se muestra en la figura 5.5.

Minitab proporciona probabilidades de Poisson e hipergeométricas de una manera parecida. Por ejemplo, para calcular las probabilidades de Poisson, las únicas diferencias se encuentran en el paso 3, donde se seleccionaría la opción **Poisson**, y en el paso 4, donde se introduciría **Mean** en vez del número de ensayos y la probabilidad de éxito.

Apéndice 5.2 Distribuciones de probabilidad discretas con Excel

Excel contiene funciones para calcular probabilidades de las distribuciones binomial, de Poisson e hipergeométrica presentadas en este capítulo. La función de Excel para calcular probabilidades binomiales es BINOMDIST. Tiene cuatro argumentos: x (núm_éxito), n (núm_ensayos), p (prob_éxito) y acumulado. FALSE se usa para el cuarto argumento (acumulado) si se busca la probabilidad de x éxitos, y TRUE se utiliza para el cuarto argumento si se quiere la probabilidad acumulada de x o menos éxitos. Aquí se muestra cómo calcular las probabilidades de 0 a 10 éxitos para el problema de la tienda Martin Clothing Store de la sección 5.4 (figura 5.5).

Cuando se describa el desarrollo de la hoja de trabajo, revise la figura 5.6; la hoja de trabajo de fórmulas se coloca en segundo plano, y la hoja de trabajo de valores aparece en primer plano. El número de ensayos (10) se introduce en la celda B1, la probabilidad de éxito en la celda B2 y los valores para la variable aleatoria en las celdas B5:B15. Los pasos siguientes generarán las probabilidades buscadas.

Paso 1. Use la función BINOMDIST para calcular la probabilidad de $x = 0$ al introducir la fórmula siguiente en la celda C5:

=BINOMDIST(B5,$B$1,$B$2,FALSE)

Paso 2. Copie la fórmula de la celda C5 en las celdas C6:C15.

FIGURA 5.6 Hoja de trabajo de Excel para calcular las probabilidades binomiales

	A	B	C	D
1	**Number of Trials (*n*)**	10		
2	**Probability of Success (*p*)**	0.3		
3				
4		***x***	***f*(*x*)**	
5		0	=BINOMDIST(B5,$B$1,$B$2,FALSE)	
6		1	=BINOMDIST(B6,$B$1,$B$2,FALSE)	
7		2	=BINOMDIST(B7,$B$1,$B$2,FALSE)	
8		3	=BINOMDIST(B8,$B$1,$B$2,FALSE)	
9		4	=BINOMDIST(B9,$B$1,$B$2,FALSE)	
10		5	=BINOMDIST(B10,$B$1,$B$2,FALSE)	
11		6	=BINOMDIST(B11,$B$1,$B$2,FALSE)	
12		7	=BINOMDIST(B12,$B$1,$B$2,FALSE)	
13		8	=BINOMDIST(B13,$B$1,$B$2,FALSE)	
14		9	=BINOMDIST(B14,$B$1,$B$2,FALSE)	
15		10	=BINOMDIST(B15,$B$1,$B$2,FALSE)	
16				

	A	B	C	D
1	**Number of Trials (*n*)**	10		
2	**Probability of Success (*p*)**	0.3		
3				
4		***x***	***f*(*x*)**	
5		0	0.0282	
6		1	0.1211	
7		2	0.2335	
8		3	0.2668	
9		4	0.2001	
10		5	0.1029	
11		6	0.0368	
12		7	0.0090	
13		8	0.0014	
14		9	0.0001	
15		10	0.0000	
16				

La hoja de trabajo de valores de la figura 5.6 muestra que las probabilidades obtenidas son las mismas que las de la figura 5.5. Las probabilidades de Poisson e hipoergeométricas se calculan de modo parecido. Se usan las funciones POISSON e HYPGEOMDIST. El cuadro de diálogo Insert Function (insertar función) de Excel ayuda al usuario a introducir los argumentos apropiados para estas funciones (vea el apéndice E).

CAPÍTULO 6

Distribuciones de probabilidad continua

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: PROCTER & GAMBLE

6.1 DISTRIBUCIÓN DE PROBABILIDAD UNIFORME
- El área como medida de la probabilidad

6.2 DISTRIBUCIÓN DE PROBABILIDAD NORMAL
- Curva normal
- Distribución de probabilidad normal estándar
- Cálculo de probabilidades para cualquier distribución de probabilidad normal
- El problema de Grear Tire Company

6.3 APROXIMACIÓN NORMAL DE LAS PROBABILIDADES BINOMIALES

6.4 DISTRIBUCIÓN DE PROBABILIDAD EXPONENCIAL
- Cálculo de probabilidades para la distribución exponencial
- Relación entre las distribuciones de Poisson y exponencial

ESTADÍSTICA *en* LA PRÁCTICA

PROCTER & GAMBLE*

CINCINNATI, OHIO

Procter & Gamble (P&G) produce y comercializa detergentes, pañales desechables, fármacos que no requieren receta médica, dentífricos, jabones de tocador, enjuagues bucales y toallas de papel, entre otros artículos. En todo el mundo, P&G tiene la marca líder en más categorías de productos de consumo que cualquier otra empresa. Desde su fusión con Gillette, también fabrica y comercializa rastrillos, navajas de afeitar y muchos otros artículos para el cuidado personal.

Como líder en la aplicación de métodos estadísticos en la toma de decisiones, P&G emplea a personas con diversas formaciones académicas: ingenieros, expertos en estadística, investigadores de operaciones y administradores de empresas. Las principales tecnologías cuantitativas en que estos especialistas aplican sus conocimientos son las decisiones probabilísticas y el análisis de riesgos, la simulación avanzada, la mejora de la calidad y los métodos cuantitativos (por ejemplo, programación lineal, análisis de regresión y análisis de probabilidad).

La División de Productos Químicos Industriales de P&G es un proveedor importante de alcoholes grasos derivados de sustancias naturales como el aceite de coco y el petróleo. La división quería conocer los riesgos económicos y las oportunidades de ampliar sus instalaciones de producción de alcoholes grasos, por lo que solicitó la ayuda de los expertos de P&G en decisiones probabilísticas y análisis de riesgos. Después de estructurar y modelar el problema, se determinó que la clave de la rentabilidad radicaba en la diferencia entre los costos de las materias primas derivadas del petróleo y del coco. No era posible determinar los costos futuros, pero los analistas pudieron aproximarlos utilizando las variables aleatorias continuas siguientes.

x = precio del aceite de coco por libra de alcoholes grasos

y

y = precio de la materia prima derivada del petróleo por libra de alcoholes grasos

Como la clave de la rentabilidad radicaba en la diferencia entre estas dos variables aleatorias, se empleó una tercera variable, $d = x - y$, en el análisis. Se entrevistó a varios expertos para determinar las distribuciones de probabilidad para x y y. A su vez, esta información se utilizó para elaborar una distribución de probabilidad de la diferencia en los precios d. Esta distribución de probabilidad continua mostró una probabilidad de 0.90 de que la diferencia en los precios fuera de \$0.0655 o menos y una probabilidad de 0.50 de que esta diferencia fuera de \$0.035 o menos. Además, sólo había una probabilidad de 0.10 de que tal diferencia fuera de \$0.0045 o menos.†



Algunos de los muchos productos conocidos de Procter & Gamble. © Robert Sullivan/AFP/Getty Images.

La División de Productos Químicos Industriales pensó que para llegar a un consenso era fundamental cuantificar el efecto de las diferencias en los precios de las materias primas. Las probabilidades obtenidas se usaron en un análisis de sensibilidad de tales diferencias. El análisis reveló información suficiente para fundamentar una recomendación a la gerencia.

El uso de las variables aleatorias continuas y sus distribuciones de probabilidad ayudó a P&G en el análisis de los riesgos económicos asociados con la producción de alcoholes grasos. Al leer este capítulo, usted comprenderá las variables aleatorias continuas y sus distribuciones de probabilidad, incluida una de las más importantes en la estadística: la distribución normal.

---

* Los autores agradecen a Joel Kahn, de Procter & Gamble, por proporcionar este artículo para *Estadística en la práctica.*

† Las diferencias en los precios establecidas aquí se modificaron para proteger los datos confidenciales.

En el capítulo anterior se estudiaron las variables aleatorias discretas y sus distribuciones de probabilidad. Este capítulo se dedica al estudio de las variables aleatorias continuas; en particular, se abordarán tres distribuciones de probabilidad continua: uniforme, normal y exponencial.

Una diferencia fundamental entre las variables aleatorias discretas y las continuas radica en la manera de calcular las probabilidades. Para las primeras, la función de probabilidad $f(x)$ proporciona la probabilidad de que la variable aleatoria asuma un valor particular. Con las segundas, el homólogo de la función de probabilidad es la **función de densidad de probabilidad**, que también se denota por medio de $f(x)$. La diferencia estriba en que la función de densidad de probabilidad no proporciona las probabilidades directamente. Sin embargo, el área bajo la gráfica $f(x)$ que corresponde a un intervalo dado representa la probabilidad de que la variable aleatoria continua x asuma un valor dentro de ese intervalo. De esta manera, cuando se calculan las probabilidades de las variables aleatorias continuas en realidad se está determinando la probabilidad de que la variable aleatoria asuma cualquier valor dentro de un intervalo.

Dado que el área bajo la gráfica $f(x)$ en cualquier punto en particular es cero, una de las implicaciones de la definición de probabilidad para las variables aleatorias continuas estriba en que la probabilidad de cualquier valor particular de la variable aleatoria sea cero. En la sección 6.1 se muestran estos conceptos para una variable aleatoria continua con una distribución uniforme.

Gran parte del capítulo se dedica a describir y mostrar las aplicaciones de la distribución normal. Ésta es de fundamental importancia debido a que tiene amplias aplicaciones y su uso está muy extendido en la inferencia estadística. El capítulo concluye con un análisis de la distribución exponencial, la cual es útil en las aplicaciones en que intervienen factores como los tiempos de espera y de servicio.

6.1 Distribución de probabilidad uniforme

Siempre que la probabilidad sea proporcional a la longitud del intervalo, la variable aleatoria está distribuida de manera uniforme.

Considere la variable aleatoria x que representa el tiempo de vuelo de un avión que viaja de Chicago a Nueva York. Suponga que este tiempo puede ser cualquier valor en el intervalo de 120 a 140 minutos. Dado que la variable aleatoria x puede asumir cualquier valor en ese intervalo, x es una variable aleatoria continua más que una variable aleatoria discreta. Suponga además que cuenta con suficientes datos reales sobre los vuelos para concluir que la probabilidad de que el tiempo de vuelo esté dentro de cualquier intervalo de 1 minuto es igual a la probabilidad de que esté dentro de cualquier otro intervalo de 1 minuto contenido dentro del intervalo mayor de 120 a 140 minutos. Como cada intervalo de 1 minuto es igualmente probable, se dice que la variable aleatoria x tiene una **probabilidad de distribución uniforme**. La función de densidad de probabilidad, que define la distribución uniforme para la variable aleatoria del tiempo de vuelo es

$$f(x) = \begin{cases} 1/20 & \text{para } 120 \le x \le 140 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

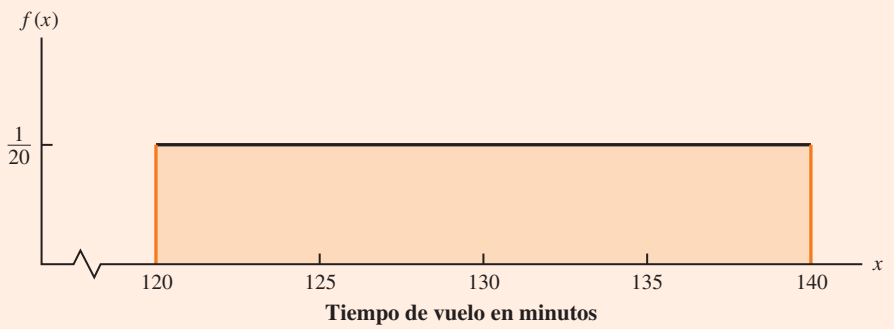
La figura 6.1 es una gráfica de esta función de densidad de probabilidad. En general, la función de densidad de probabilidad uniforme para una variable aleatoria x se define por medio de la fórmula siguiente.

FUNCIÓN DE DENSIDAD DE PROBABILIDAD UNIFORME

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{para } a \le x \le b \\ 0 & \text{en cualquier otro caso} \end{cases} \qquad \textbf{(6.1)}$$

Para la variable aleatoria del tiempo de vuelo, $a = 120$ y $b = 140$.

FIGURA 6.1 Distribución de probabilidad uniforme para el tiempo de vuelo

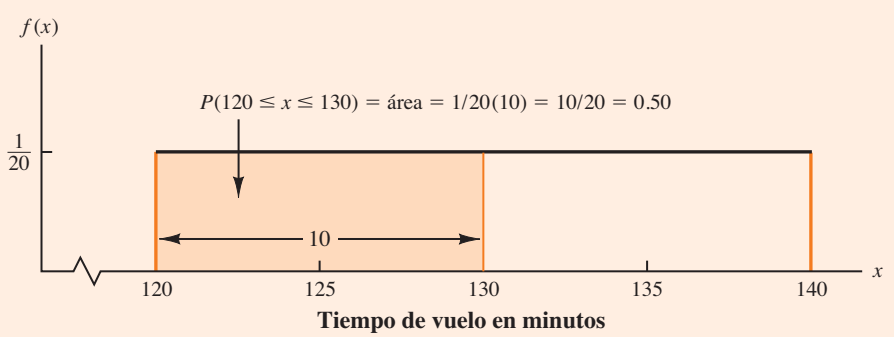


Como se observó en la introducción, en el caso de una variable aleatoria continua, la probabilidad sólo se considera en términos de la posibilidad de que la variable aleatoria tome un valor dentro de un intervalo determinado. En el ejemplo del tiempo de vuelo, una pregunta de probabilidad aceptable es: ¿cuál es la probabilidad de que el tiempo de vuelo se encuentre entre 120 y 130 minutos? Es decir, ¿cuánto es $P(120 \leq x \leq 130)$? Debido a que dicho tiempo debe estar entre 120 y 140 minutos y la probabilidad se describe como uniforme a lo largo de este intervalo, es factible decir que $P(120 \leq x \leq 130) = 0.50$. En la subsección siguiente se muestra que esta probabilidad se calcula como el área bajo la gráfica $f(x)$ de 120 a 130 (figura 6.2).

El área como medida de la probabilidad

Como una observación de la gráfica de la figura 6.2, considere que el área bajo la gráfica $f(x)$ en el intervalo de 120 a 130 es rectangular, y el área de un rectángulo es sencillamente el ancho multiplicado por la altura. Si se considera que el ancho del intervalo es igual a $130 - 120 = 10$, y la altura es igual al valor de la función de densidad de probabilidad $f(x) = 1/20$, se tiene el área = ancho × altura = $10(1/20) = 10/20 = 0.50$.

FIGURA 6.2 El área proporciona la probabilidad de que el tiempo de vuelo esté entre 120 y 130 minutos



¿Qué observaciones puede hacer sobre el área bajo la gráfica $f(x)$ y la probabilidad? ¡Son idénticas! De hecho, esta observación es válida para todas las variables aleatorias continuas. Una vez que la función de densidad de probabilidad $f(x)$ se identifica, la probabilidad de que x tome un valor entre uno inferior x_1 y uno superior x_2 se obtiene al calcular el área bajo la gráfica $f(x)$ en el intervalo de x_1 a x_2.

Dada la distribución uniforme para el tiempo de vuelo y usando la interpretación del área como una medida de probabilidad, es posible responder cualquier cantidad de preguntas de probabilidad sobre los tiempos de vuelo. Por ejemplo, ¿cuál es la probabilidad de un tiempo de vuelo entre 128 y 136 minutos? El ancho del intervalo es $136 - 128 = 8$. Con la altura uniforme de $f(x) = 1/20$, se ve que $P(128 \leq x \leq 136) = 8(1/20) = 0.40$.

Observe que $P(120 \leq x \leq 140) = 20(1/20) = 1$; es decir, el área total bajo la gráfica $f(x)$ es igual a 1. Esta propiedad es válida para todas las distribuciones de probabilidad continua y es el análogo de la condición que indica que la suma de las probabilidades debe ser igual a 1 para una función de probabilidad discreta. En el caso de una función de densidad de probabilidad continua, se requiere también que $f(x) \geq 0$ para todos los valores de x. Este requerimiento es el análogo del requisito de $f(x) \geq 0$ para las funciones de probabilidad discretas.

Hay dos diferencias importantes entre el tratamiento de la variable aleatoria continua y el tratamiento de sus homólogas discretas.

1. Ya no se alude a la probabilidad de que una variable aleatoria asuma un valor particular. En su lugar, se habla de la probabilidad de que asuma un valor dentro de cierto intervalo.
2. La probabilidad de que una variable aleatoria continua asuma un valor dentro de un intervalo dado de x_1 a x_2 se define como el área bajo la gráfica de la función de densidad de probabilidad entre x_1 y x_2. Como cada punto es un intervalo cuyo ancho es igual a cero, esto implica que la probabilidad de que una variable aleatoria continua asuma cualquier valor particular es exactamente cero; también significa que la probabilidad de que asuma un valor en cualquier intervalo es la misma, ya sea que se incluyan o no los puntos finales.

Para ver que la probabilidad de que cualquier punto individual sea 0, remítase a la figura 6.2 y calcule la probabilidad de un punto individual, es decir, $x = 125$. $P(x = 125) = P(125 \leq x \leq 125) = 0(1/20) = 0$.

El cálculo del valor esperado y de la varianza de una variable aleatoria continua es análogo al de la variable aleatoria discreta. Sin embargo, como el procedimiento para determinarlo requiere cálculo integral, la deducción de las fórmulas apropiadas se deja para libros más avanzados.

En el caso de la distribución de probabilidad continua uniforme presentada en esta sección, las fórmulas para el valor esperado y la varianza son

$$E(x) = \frac{a + b}{2}$$

$$\text{Var}(x) = \frac{(b - a)^2}{12}$$

En estas fórmulas, a es el valor menor y b es el valor mayor que la variable aleatoria puede asumir.

Al aplicar estas fórmulas a la distribución uniforme de los tiempos de vuelo de Chicago a Nueva York obtenemos

$$E(x) = \frac{(120 + 140)}{2} = 130$$

$$\text{Var}(x) = \frac{(140 - 120)^2}{12} = 33.33$$

La desviación estándar de los tiempos de vuelo se obtiene al calcular la raíz cuadrada de la varianza. Por tanto, $\sigma = 5.77$ minutos.

NOTAS Y COMENTARIOS

Para ver con mayor claridad por qué la altura de una función de densidad de probabilidad no es una probabilidad, considere la variable aleatoria con la distribución de probabilidad uniforme siguiente.

$$f(x) = \begin{cases} 2 & \text{para } 0 \leq x \leq 0.5 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

La altura de la función de densidad de probabilidad, $f(x)$, es 2 para valores de x entre 0 y 0.5. No obstante, se sabe que las probabilidades nunca pueden ser mayores que 1. Por tanto, se ve que $f(x)$ no se interpreta como la probabilidad de x.

Ejercicios

Métodos

AUTOevaluación

1. Se sabe que la variable aleatoria x está distribuida de manera uniforme entre 1.0 y 1.5.
   a) Trace la gráfica de la función de densidad de probabilidad.
   b) Calcule $P(x = 1.25)$.
   c) Determine $P(1.0 \leq x \leq 1.25)$.
   d) Calcule $P(1.20 < x < 1.5)$.

2. La variable aleatoria x está distribuida de manera uniforme entre 10 y 20.
   a) Trace la gráfica de la función de densidad de probabilidad.
   b) Calcule $P(x < 15)$.
   c) Estime $P(12 < x < 18)$.
   d) Calcule $E(x)$.
   e) Determine Var(x).

Aplicaciones

3. Delta Airlines ofrece un tiempo de 2 horas, 5 minutos para sus vuelos de Cincinnati a Tampa. Suponga que se piensa que los tiempos de vuelo reales están distribuidos uniformemente entre 2 horas y 2 horas, 20 minutos.
   a) Trace la gráfica de la función de densidad de probabilidad para el tiempo de vuelo.
   b) ¿Cuál es la probabilidad de que el vuelo no se retrase más de 5 minutos?
   c) ¿Cuál es la probabilidad de que se retrase más de 10 minutos?
   d) ¿Cuál es el tiempo esperado de vuelo?

AUTOevaluación

4. La mayoría de los lenguajes de cómputo incluye una función para generar números aleatorios. En Excel, la función RAND se utiliza para generar números aleatorios entre 0 y 1. Si x denota un número aleatorio generado por medio de RAND, entonces x es una variable aleatoria continua con la función de densidad de probabilidad siguiente.

$$f(x) = \begin{cases} 1 & \text{para } 0 \leq x \leq 1 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

   a) Trace la gráfica de la función de densidad de probabilidad.
   b) ¿Cuál es la probabilidad de generar un número aleatorio entre 0.25 y 0.75?
   c) ¿Cuál es la probabilidad de que el número aleatorio generado tenga un valor menor o igual que 0.30?
   d) ¿Cuál es la probabilidad de generar un número aleatorio con un valor mayor que 0.60?
   e) Genere 50 números aleatorios al introducir =RAND() en 50 celdas de una hoja de trabajo de Excel.
   f) Calcule la media y la desviación estándar de los números aleatorios en el inciso *e*).

5. La distancia de lanzamiento de los 100 mejores golfistas del tour PGA está entre 284.7 y 310.6 yardas (*Golfweek*, 29 de marzo de 2003). Suponga que la distancia de lanzamiento de estos deportistas está distribuida de manera uniforme a lo largo de este intervalo.
   a) Proporcione una expresión matemática para la función de densidad de probabilidad de la distancia de lanzamiento.
   b) ¿Cuál es la probabilidad de que la distancia de lanzamiento de uno de estos golfistas sea menor de 290 yardas?
   c) ¿Cuál es la probabilidad de que esta distancia de lanzamiento sea como mínimo de 300 yardas?
   d) ¿Cuál es la probabilidad de que la distancia de lanzamiento esté entre 290 y 305 yardas?
   e) ¿Cuántos de estos golfistas lanzan la pelota cuando menos 290 yardas?

6. En promedio, las comedias de 30 minutos que se transmiten por televisión tienen 22 minutos de programación (CNBC, 23 de febrero de 2006). Suponga que la distribución de probabilidad de los minutos de programación se aproxima por medio de una distribución uniforme de 18 a 26 minutos.
   a) ¿Cuál es la probabilidad de que una comedia tenga 25 o más minutos de programación?
   b) ¿Cuál es la probabilidad de que tenga entre 21 y 25 minutos de programación?
   c) ¿Cuál es la probabilidad de que incluya más de 10 minutos de comerciales o de otras interrupciones que no forman parte de la programación?

7. Suponga que le interesa adquirir un terreno y sabe que hay otros compradores interesados en él.[1] El vendedor anuncia que aceptará la oferta más alta mayor de $10 000. Considere que la oferta del competidor x es una variable aleatoria que está distribuida uniformemente entre $10 000 y $15 000.
   a) Suponga que usted propone $12 000. ¿Cuál es la probabilidad de que su oferta sea aceptada?
   b) Considere que ofrece $14 000. ¿Cuál es la probabilidad de que se acepte su postura?
   c) ¿Qué cantidad debe proponer para maximizar la probabilidad de comprar la propiedad?
   d) Suponga que conoce a alguien que está dispuesto a pagarle $16 000 por la propiedad. ¿Consideraría ofrecer menos de la cantidad del inciso *c*)? ¿Por qué?

6.2 Distribución de probabilidad normal

Abraham de Moivre, matemático francés que publicó La doctrina de las probabilidades *en 1733, dedujo la distribución normal.*

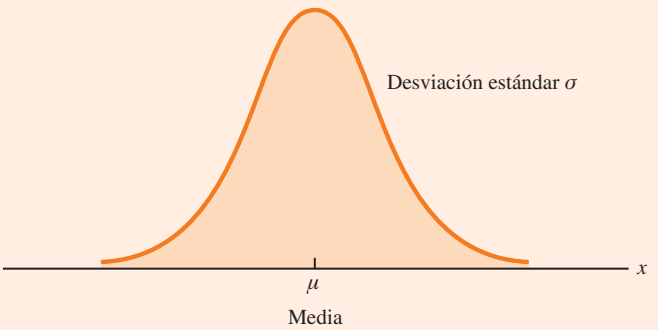
La distribución de probabilidad más importante para describir una variable aleatoria continua es la **distribución de probabilidad normal**. Ésta se ha utilizado en una amplia variedad de aplicaciones en las cuales las variables aleatorias son la altura y el peso de las personas, las calificaciones de los exámenes, las mediciones científicas, la precipitación pluvial y otros valores parecidos. También tiene un uso muy extendido en la inferencia estadística, la cual es el tema principal del resto de este libro. En estas aplicaciones, la distribución normal describe qué tan probables son los resultados obtenidos de un muestreo.

Curva normal

La forma de la distribución normal se ilustra por medio una curva con forma de campana que exhibe la figura 6.3. La función de densidad de probabilidad que define la curva de la distribución normal se muestra en seguida.

[1] Este ejercicio se basa en un problema sugerido por el profesor Roger Myerson, de la Northwestern University.

FIGURA 6.3 Curva con forma de campana de la distribución normal



FUNCIÓN DE DENSIDAD DE PROBABILIDAD NORMAL

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad \textbf{(6.2)}$$

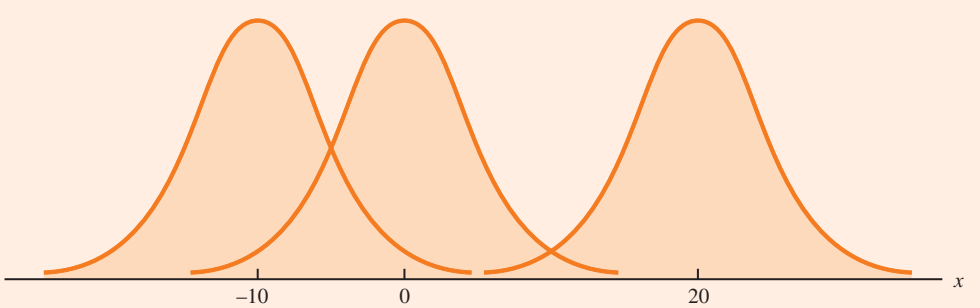
Donde:

μ = media
σ = desviación estándar
π = 3.14159
e = 2.71828

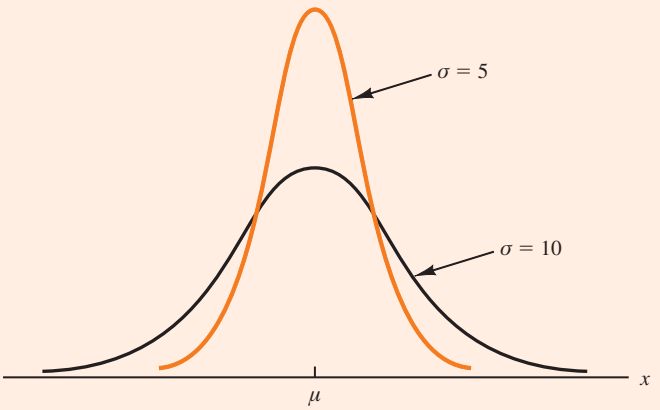
Se formulan varias observaciones acerca de las características de la distribución normal.

La curva normal tiene dos parámetros, μ y σ, que determinan la ubicación y la forma de la distribución normal.

1. La familia completa de distribuciones normales se diferencia por medio de dos parámetros: la media μ y la desviación estándar σ.
2. El punto más alto de una curva normal se encuentra sobre la media, el cual coincide con la mediana y la moda de la distribución.
3. La media de una distribución normal puede tener cualquier valor numérico: negativo, cero o positivo. A continuación se muestran tres distribuciones normales que tienen la misma desviación estándar pero tres medias diferentes (−10, 0 y 20).



4. La distribución normal es simétrica: la forma de la curva normal a la izquierda de la media es una imagen de espejo de la forma de la curva a la derecha de la media. Los extremos de la curva normal se extienden hacia el infinito en ambas direcciones y en teoría nunca tocan el eje horizontal. Como son simétricas, las distribuciones normales no están sesgadas; la medida de su sesgo es cero.
5. La desviación estándar determina qué tan plana y ancha es la curva normal. Los valores grandes de la desviación estándar dan como resultado curvas más anchas y planas, mostrando mayor variabilidad en los datos. En seguida se muestran dos distribuciones normales con la misma media, pero con desviaciones estándar diferentes.



6. Las probabilidades para la variable aleatoria normal están representadas por las áreas bajo la curva normal. El área total bajo la curva de una distribución normal es 1. Como la distribución es simétrica, el área bajo la curva a la izquierda de la media es 0.50 y el área a la derecha también es 0.50.
7. Los porcentajes de los valores en algunos intervalos de uso común son los siguientes.
   a) 68.3% de los valores de una variable aleatoria normal se sitúan más o menos a una desviación estándar de su media.
   b) 95.4% de los valores de una variable aleatoria normal se encuentran más o menos a dos desviaciones estándar de su media.
   c) 99.7% de los valores de una variable aleatoria normal están más o menos dentro de tres desviaciones estándar de su media.

Estos porcentajes son la base para la regla empírica que se presentó en la sección 3.3.

La figura 6.4 muestra una gráfica de las propiedades *a*), *b*) y *c*).

Distribución de probabilidad normal estándar

Se dice que una variable aleatoria que muestra una distribución normal con una media de cero y una desviación estándar de uno tiene una **distribución de probabilidad normal estándar**. La letra z se usa comúnmente para designar esta variable aleatoria normal. La figura 6.5 muestra la gráfica general de la distribución normal estándar, la cual tiene la misma apariencia que otras distribuciones normales, pero con las propiedades especiales de $\mu = 0$ y $\sigma = 1$.

FIGURA 6.4 Áreas bajo la curva de cualquier distribución normal

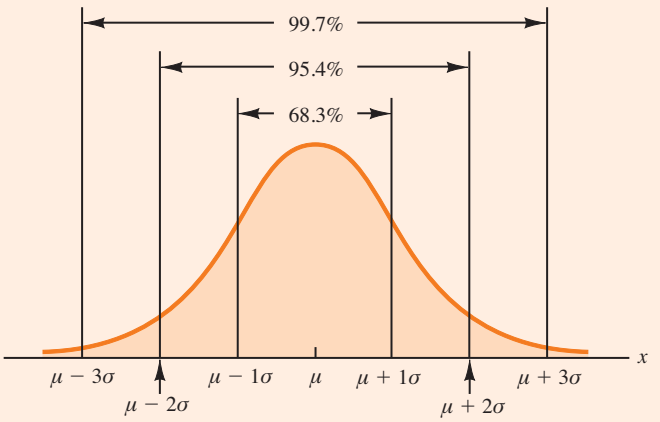
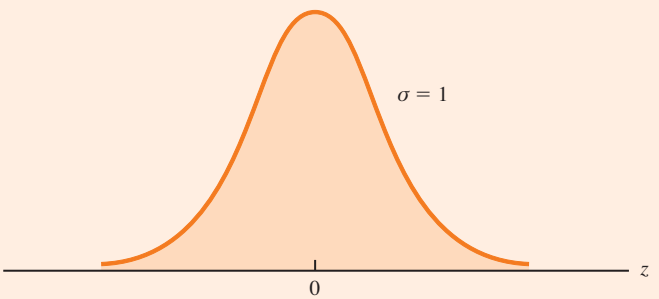


FIGURA 6.5 Distribución normal estándar



Como $\mu = 0$ y $\sigma = 1$, la fórmula para la función de densidad de probabilidad normal estándar es una versión más sencilla de la ecuación (6.2).

FUNCIÓN DE DENSIDAD NORMAL ESTÁNDAR

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

Para la función de densidad de probabilidad normal, la altura de la curva normal varía, y se requieren matemáticas más avanzadas para calcular las áreas que representan la probabilidad.

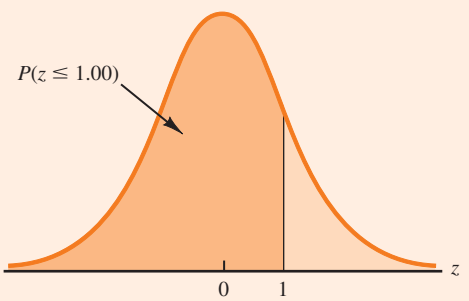
Como ocurre con otras variables aleatorias continuas, los cálculos de la probabilidad con cualquier distribución normal se efectúan al obtener las áreas bajo la gráfica de la función de densidad de probabilidad. Por tanto, para encontrar la probabilidad de que una variable aleatoria normal esté dentro de cualquier intervalo específico, debe calcularse el área bajo la curva normal en ese intervalo.

Para la distribución normal estándar, las áreas bajo la curva normal ya se han estimado y están disponibles en tablas que se utilizan para el cálculo de probabilidades. Una tabla como éstas aparece en las dos guardas de la cubierta anterior del libro. La de la página izquierda contiene las áreas o probabilidades acumuladas correspondientes a los valores de z menores o iguales a la media de cero. La tabla de la página derecha contiene las áreas o probabilidades acumuladas que corresponden a los valores de z superiores o iguales a la media de cero.

Los tres tipos de probabilidades que se necesita calcular incluyen: 1) la probabilidad de que la variable aleatoria normal estándar z sea menor o igual que un valor determinado; 2) la probabilidad de que z esté entre dos valores dados, y 3) la probabilidad de que z sea mayor o igual que un valor determinado. Para conocer cómo se usa la tabla de probabilidad acumulada de la distribución normal estándar con el propósito de calcular estos tres tipos de probabilidades, considere algunos ejemplos.

Como la variable aleatoria normal estándar es continua, $P(z \leq 1.00) = P(z < 1.00)$.

Primero se mostrará cómo calcular la probabilidad de que z sea menor o igual que 1.00, esto es, $P(z \leq 1.00)$. Esta probabilidad acumulada es el área bajo la curva normal a la izquierda de $z = 1.00$ en la gráfica siguiente.

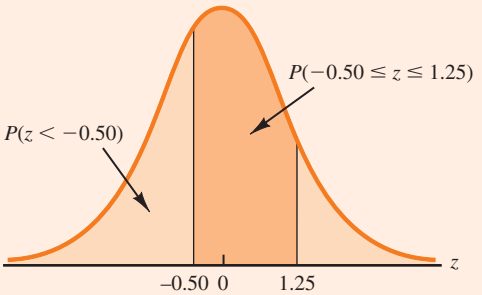


Revise la tabla de probabilidad normal estándar en la página derecha de las guardas de la cubierta anterior del libro. La probabilidad acumulada que corresponde a $z = 1.00$ es el valor ubicado en la intersección de la fila cuyo encabezado es 1.0 y la columna cuyo encabezado es 0.00. Primero se localiza 1.0 en la columna izquierda de la tabla y luego 0.00 en la fila superior. Al observar el cuerpo de la tabla, encontramos que la fila 1.0 y la columna 0.00 se intersecan en el valor 0.8413; por tanto, $P(z \leq 1.00) = 0.8413$. El extracto siguiente de la tabla de probabilidad muestra estos pasos.

z	**0.00**	**0.01**	**0.02**
.			
.			
.			
0.9	0.8159	0.8186	0.8212
1.0	**0.8413**	0.8438	0.8461
1.1	0.8643	0.8665	0.8686
1.2	0.8849	0.8869	0.8888
.			
.			
.			

$P(z \leq 1.00)$

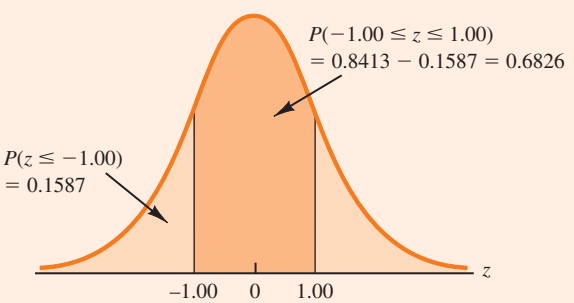
Para ilustrar el segundo tipo de cálculo de la probabilidad, suponga que se quiere determinar la probabilidad de que z esté en el intervalo entre -0.50 y 1.25; es decir, $P(-0.50 \leq z \leq 1.25)$. La gráfica siguiente muestra esta área, o probabilidad.



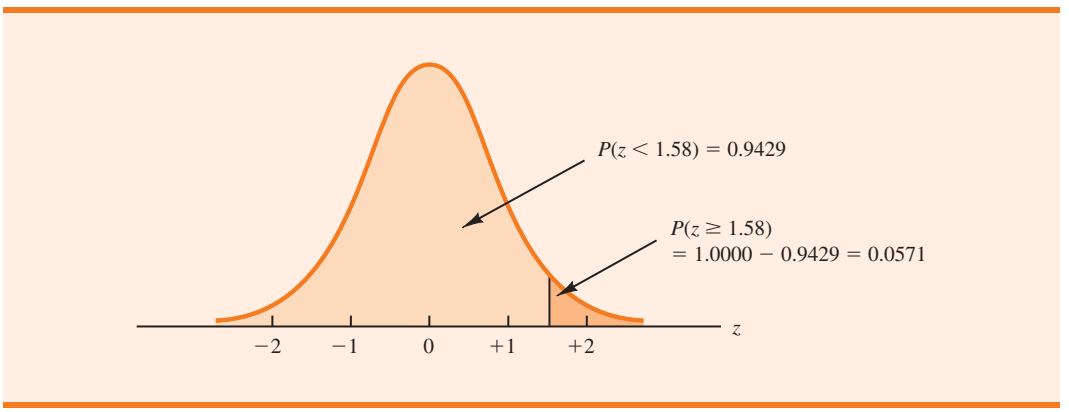
Se requieren tres pasos para calcular esta probabilidad. Primero se encuentra el área bajo la curva normal a la izquierda de $z = 1.25$. Segundo, se obtiene el área bajo la curva normal a la izquierda de $z = -0.50$. Y por último, se resta el área a la izquierda de $z = -0.50$, del área a la izquierda de $z = 1.25$ para obtener $P(-0.50 \le z \le 1.25)$.

Para calcular el área bajo la curva normal a la izquierda de $z = 1.25$, primero se localiza la fila 1.2 en la tabla de probabilidad normal estándar y luego se avanza hasta la columna 0.05. Como el valor que aparece en la fila 1.2 y en la columna 0.05 es 0.8944, $P(z \le 1.25) = 0.8944$. De manera similar, cuando se quiere determinar el área bajo la curva a la izquierda de $z = -0.50$, se usa la tabla de la página izquierda para localizar el valor de la fila -0.5 y la columna 0.00; como el valor es 0.3085, $P(z \le -0.50) = 0.3085$. Por tanto, $P(-0.50 \le z \le 1.25) = P(z \le 1.25) - P(z \le -0.50) = 0.8944 - 0.3085 = 0.5859$.

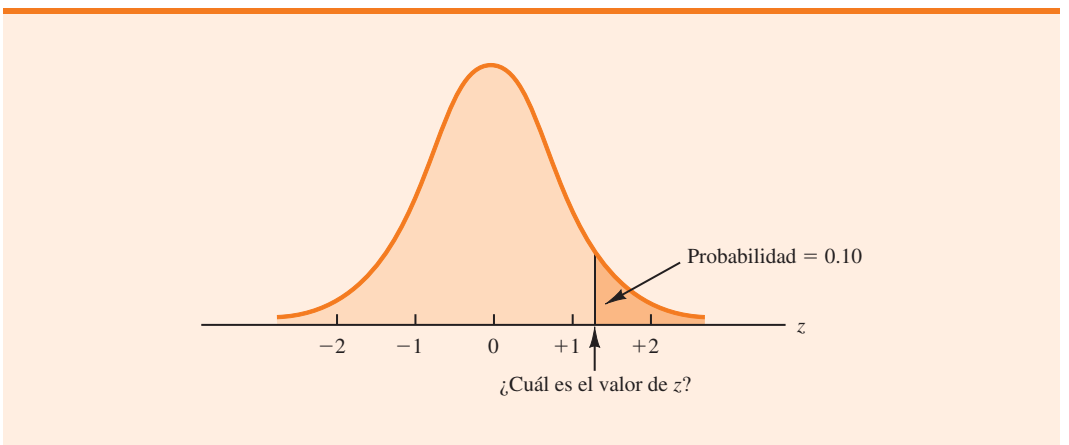
Considere otro ejemplo del cálculo de la probabilidad de que z esté en el intervalo entre dos valores dados. A menudo se quiere conocer la probabilidad de que una variable aleatoria normal asuma un valor dentro de cierto número de desviaciones estándar de la media. Suponga que queremos calcular la probabilidad de que la variable aleatoria normal estándar esté dentro de una desviación estándar de la media; es decir, $P(-1.00 \le z \le 1.00)$. Para ello, primero se obtiene el área bajo la curva entre -1.00 y 1.00. Antes se encontró que $P(z \le 1.00) = 0.8413$. Si observa de nuevo la tabla de las guardas de la cubierta anterior del libro, se ve que el área bajo la curva a la izquierda de $z = -1.00$ es 0.1587; por tanto $P(z \le -1.00) = 0.1587$. De ahí que $P(-1.00 \le z \le 1.00) = P(z \le 1.00) - P(z \le -1.00) = 0.8413 - 0.1587 = 0.6826$. Esta probabilidad se muestra gráficamente en la figura siguiente.



Para explicar cómo se efectúa el tercer tipo de cálculo de probabilidad, suponga que se quiere determinar la probabilidad de obtener un valor z por lo menos igual a 1.58; es decir, $P(z \geq 1.58)$. El valor en la fila $z = 1.5$ y la columna 0.08 de la tabla normal acumulada es 0.9429; por tanto, $P(z < 1.58) = 0.9429$. Sin embargo, como el área total bajo la curva normal es 1, $P(z \geq 1.58) = 1 - 0.9429 = 0.0571$. Esta probabilidad se muestra en la figura siguiente.



En los ejemplos anteriores se mostró cómo calcular las probabilidades cuando se proporcionan valores de z específicos. En algunas situaciones se da una probabilidad y se quiere trabajar a la inversa para encontrar el valor de z correspondiente. Suponga que quiere determinar un valor de z tal que la probabilidad de obtener un valor de z mayor sea 0.10. La figura siguiente muestra esta situación de manera gráfica.



Dada una probabilidad, se puede usar la tabla normal estándar en modo inverso para encontrar el valor de z correspondiente.

Este problema es el inverso de las situaciones presentadas en los ejemplos anteriores, en los cuales se especificó el valor de z y luego se calculó la probabilidad, o área, correspondiente. En este ejemplo se proporciona la probabilidad, o área, y luego se pide determinar el valor z respectivo. Para hacerlo, se usa la tabla de probabilidad normal estándar de una manera un poco distinta.

Recuerde que esta tabla proporciona el área bajo la curva a la izquierda de un valor de z determinado. Se tiene la información de que el área en el extremo superior de la curva es 0.10. Por consiguiente, el área bajo la curva a la izquierda del valor de z desconocido debe ser igual a 0.9000. Al revisar el cuerpo de la tabla, encontramos que 0.8997 es el valor de probabilidad acumulada más cercano a 0.9000. La sección de la tabla que muestra este resultado se reproduce a continuación.

z	0.06	0.07	0.08	0.09
.				
.				
.				
1.0	0.8554	0.8577	0.8599	0.8621
1.1	0.8770	0.8790	0.8810	0.8830
1.2	0.8962	0.8980	**0.8997**	0.9015
1.3	0.9131	0.9147	0.9162	0.9177
1.4	0.9279	0.9292	0.9306	0.9319
.				
.				
.				

Valor de probabilidad acumulada más cercano a 0.9000

Al leer el valor de z en la columna del extremo izquierdo y la fila superior de la tabla, encontramos que es 1.28. Por tanto, un área de aproximadamente 0.9000 (en realidad, 0.8997) estará a la izquierda de $z = 1.28$.[2] Respecto de la pregunta formulada originalmente, hay una probabilidad aproximada de 0.10 de que el valor de z sea mayor que 1.28.

Estos ejemplos ilustran que la tabla de probabilidades acumuladas para la distribución de probabilidad normal estándar es útil para encontrar las probabilidades asociadas con los valores de la variable aleatoria normal estándar z. Se pueden plantear dos tipos de preguntas. El primero especifica un valor, o valores, para z y pide usar la tabla para determinar las áreas o probabilidades correspondientes. El segundo proporciona un área, o probabilidad, y pide usar la tabla para determinar el valor de z correspondiente. Por tanto, se requiere flexibilidad en el uso de la tabla de probabilidad normal estándar para responder la pregunta de probabilidad deseada. En la mayoría de los casos el trazo de una gráfica de distribución de probabilidad normal estándar y el sombreado del área apropiada ayudan a visualizar la situación y a encontrar la respuesta correcta.

Cálculo de probabilidades para cualquier distribución de probabilidad normal

La razón para estudiar la distribución normal estándar de manera exhaustiva estriba en que ésta se utiliza para calcular las probabilidades de todas las distribuciones normales. Es decir, cuando se tiene una distribución normal con cualquier media μ y cualquier desviación estándar σ, las preguntas de probabilidad acerca de la distribución se responden convirtiendo primero a la distribución normal estándar. Luego se usa la tabla de probabilidad normal estándar y los valores de z apropiados para obtener las probabilidades buscadas. La fórmula para convertir cualquier variable aleatoria normal x con media μ y desviación estándar σ a la variable aleatoria normal estándar z se presenta a continuación.

La fórmula para la variable aleatoria normal estándar es similar a la fórmula para calcular los valores z de un conjunto de datos, presentada en el capítulo 3.

CONVERSIÓN A LA VARIABLE ALEATORIA NORMAL ESTÁNDAR

$$z = \frac{x - \mu}{\sigma} \qquad (6.3)$$

[2] Se podría haber hecho una interpolación en el cuerpo de la tabla para obtener una aproximación más exacta del valor de z que corresponde al área de 0.9000. Si se hace esto para obtener una posición decimal más precisa, produciría un valor de z de 1.282. No obstante, en la mayoría de las situaciones prácticas es suficiente con la precisión que se obtiene simplemente utilizando el valor de la tabla más cercano a la probabilidad buscada.

Un valor de x igual a su media μ da como resultado $z = (\mu - \mu)/\sigma = 0$. Por tanto, vemos que un valor de x igual a su media μ corresponde a $z = 0$. Ahora suponga que x está a una desviación estándar por encima de su media; es decir, $x = \mu + \sigma$. Al aplicar la ecuación (6.3), vemos que el valor de z correspondiente es $z = [(\mu + \sigma) - \mu]/\sigma = \sigma/\sigma = 1$. En consecuencia, un valor de x que está a una desviación estándar sobre su media corresponde a $z = 1$. En otras palabras, *z puede interpretarse como el número de desviaciones estándar de la media μ a las que está la variable aleatoria normal x.*

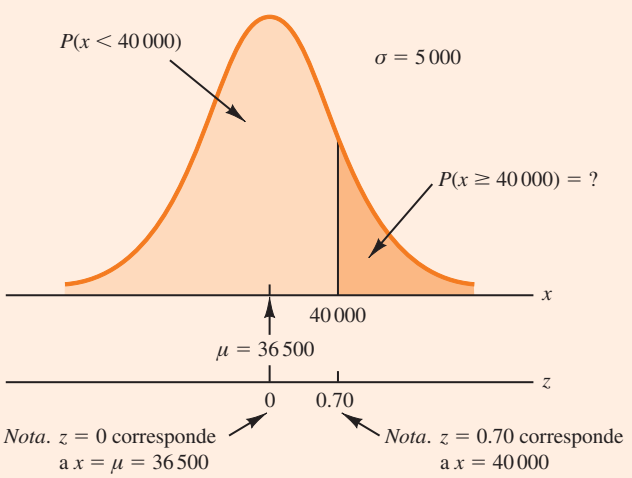
Para ver cómo esta conversión permite calcular las probabilidades de cualquier distribución normal, suponga que se tiene una distribución con $\mu = 10$ y $\sigma = 2$. ¿Cuál es la probabilidad de que la variable aleatoria x esté entre 10 y 14? Aplicando la ecuación (6.3) vemos que en $x = 10$, $z = (x - \mu)/\sigma = (10 - 10)/2 = 0$ y que en $x = 14$, $z = (14 - 10)/2 = 4/2 = 2$. Por tanto, la respuesta a nuestra pregunta sobre la probabilidad de que x esté entre 10 y 14 está dada por la probabilidad equivalente de que z esté entre 0 y 2 para la distribución normal estándar. En otras palabras, la probabilidad que se busca estriba en que la variable aleatoria x esté entre su media y a dos desviaciones estándar sobre la media. Al usar $z = 2.00$ y la tabla de probabilidad normal estándar de las guardas de la cubierta anterior del libro, $P(z \leq 2) = 0.9772$. Como $P(z \leq 0) = 0.5000$, podemos calcular $P(0.00 \leq z \leq 2.00) = P(z \leq 2) - P(z \leq 0) = 0.9772 - 0.5000 = 0.4772$. De ahí que la probabilidad de que x esté entre 10 y 14 sea 0.4772.

El problema de Grear Tire Company

Ahora veremos una aplicación de la distribución de probabilidad normal. Suponga que Grear Tire Company desarrolló un nuevo neumático radial con cinturón de acero que se vende a través de una cadena nacional de tiendas de descuento. Debido a que el neumático es un nuevo producto, los gerentes de Grear creen que la garantía de millaje ofrecida con la llanta será un factor importante para su aceptación. Antes de que la póliza de garantía de millaje de los neumáticos caduque, los gerentes de Grear quieren información de probabilidad sobre los x = número de millas que éstos durarán.

A partir de las pruebas de carretera reales con los neumáticos, el grupo de ingeniería estimó que su millaje es $\mu = 36\,500$ millas y que la desviación estándar es $\sigma = 5\,000$. Además, los datos recabados indican que una distribución normal es una suposición razonable. ¿Qué porcentaje de las llantas se espera que dure más de 40 000 millas? En otras palabras, ¿cuál es la probabilidad de que el millaje de los neumáticos, x, supere la cifra de 40 000? Esta pregunta puede responderse al calcular el área de la región sombreada de la figura 6.6.

FIGURA 6.6 Distribución de millaje de Grear Tire Company



En $x = 40\,000$ tenemos

$$z = \frac{x - \mu}{\sigma} = \frac{40\,000 - 36\,500}{5\,000} = \frac{3\,500}{5\,000} = 0.70$$

Remítase ahora a la parte inferior de la figura 6.6. Vemos que un valor de $x = 40\,000$ en la distribución normal de Grear Tire corresponde al valor de $z = 0.70$ en la distribución normal estándar. Consultando la tabla de probabilidad normal estándar, constatamos que el área bajo la curva normal estándar a la izquierda de $z = 0.70$ es 0.7580. Por tanto, $1.000 - 0.7580 = 0.2420$ es la probabilidad de que z exceda 0.70, y por consiguiente x excederá de 40 000. Podemos concluir que alrededor de 24.2% de los neumáticos superará las 40 000 millas.

Ahora suponga que Grear considera una garantía que proporcionará un descuento sobre los neumáticos de remplazo si los originales no proporcionan el millaje garantizado. ¿Cuál debe ser este millaje si Grear quiere que no más de 10% de los neumáticos sean aptos para la garantía de descuento? Esta pregunta se interpreta gráficamente en la figura 6.7.

Con base en la figura 6.7, el área bajo la curva a la izquierda del millaje de garantía desconocido debe ser 0.10. Así que primero se debe calcular el valor de z que recorta un área de 0.10 en el extremo izquierdo de una distribución normal estándar. Utilizando la tabla de probabilidad normal estándar vemos que $z = -1.28$ recorta un área de 0.10 en el extremo inferior. Por consiguiente, $z = -1.28$ es el valor de la variable aleatoria normal estándar que corresponde a la garantía de millaje buscada en la distribución normal de Grear Tire. Para encontrar el valor de x que corresponde a $z = -1.28$, tenemos

El millaje de garantía que se debe encontrar es 1.28 desviaciones estándar por debajo de la media. Por tanto, $x = \mu - 1.28\sigma$.

$$z = \frac{x - \mu}{\sigma} = -1.28$$

$$x - \mu = -1.28\sigma$$

$$x = \mu - 1.28\sigma$$

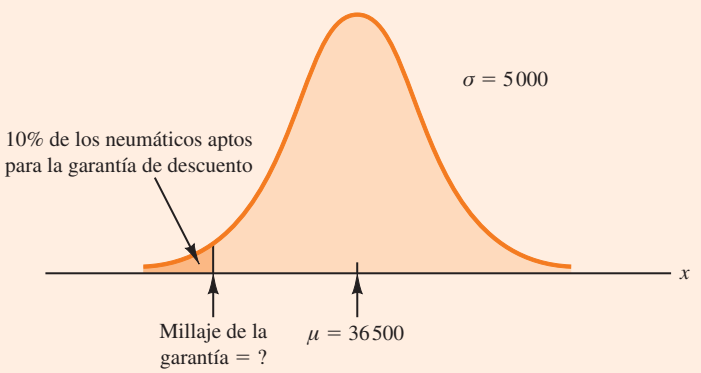
Con $\mu = 36\,500$ y $\sigma = 5\,000$,

$$x = 36\,500 - 1.28(5\,000) = 30\,100$$

Con la garantía establecida en 30 000 millas, el porcentaje real apto para la garantía será 9.68.

Por tanto, una garantía de 30 100 millas cumplirá con el requerimiento de que aproximadamente 10% de los neumáticos serán aptos para la promoción. Quizá con esta información la empresa establecerá su garantía de millaje en 30 000 millas.

FIGURA 6.7 Garantía de descuento de Grear



De nuevo, vemos el importante papel que las distribuciones de probabilidad desempeñan en proporcionar información para la toma de decisiones. En concreto, una vez que se establece una distribución de probabilidad para una aplicación en particular, se puede usar para obtener información de probabilidad sobre el problema. La probabilidad no hace directamente una recomendación de decisión, pero proporciona información que ayuda a quien la toma a comprender mejor los riesgos y las incertidumbres asociados con el problema. En definitiva, esta información ayuda a los ejecutivos a llegar a una buena decisión.

Ejercicios

Métodos

8. Utilizando la figura 6.4 como guía, trace una curva normal para la variable aleatoria x que tenga una media de $\mu = 100$ y una desviación estándar de $\sigma = 10$. Marque el eje horizontal con los valores 70, 80, 90, 100, 110, 120 y 130.

9. Una variable aleatoria está normalmente distribuida con una media de $\mu = 50$ y una desviación estándar de $\sigma = 5$.
   a) Trace una curva normal para la función de densidad de la probabilidad. Marque el eje horizontal con los valores 35, 40, 45, 50, 55, 60 y 65. La figura 6.4 muestra que la curva normal casi toca el eje horizontal en tres desviaciones estándar bajo la media y tres desviaciones estándar sobre la media (en este caso en 35 y 65).
   b) ¿Cuál es la probabilidad de que la variable aleatoria asuma un valor entre 45 y 55?
   c) ¿Cuál es la probabilidad de que asuma un valor entre 40 y 60?

10. Trace una gráfica para la distribución normal estándar. Rotule el eje horizontal con los valores -3, -2, -1, 0, 1, 2 y 3. Luego use la tabla de probabilidades para la distribución normal estándar incluida en el libro para calcular las probabilidades siguientes.
    a) $P(z \leq 1.5)$.
    b) $P(z \leq 1)$.
    c) $P(1 \leq z \leq 1.5)$.
    d) $P(0 < z < 2.5)$.

11. Dado que z es una variable aleatoria normal estándar, calcule las probabilidades siguientes.
    a) $P(z \leq -1.0)$.
    b) $P(z \geq -1)$.
    c) $P(z \geq -1.5)$.
    d) $P(-2.5 \leq z)$.
    e) $P(-3 < z \leq 0)$.

12. Puesto que z es una variable aleatoria normal estándar, calcule las probabilidades siguientes.
    a) $P(0 \leq z \leq 0.83)$.
    b) $P(-1.57 \leq z \leq 0)$.
    c) $P(z > 0.44)$.
    d) $P(z \geq -0.23)$.
    e) $P(z < 1.20)$.
    f) $P(z \leq -0.71)$.

AUTOevaluación

13. Dado que z es una variable aleatoria normal estándar, calcule las probabilidades siguientes.
    a) $P(-1.98 \leq z \leq 0.49)$.
    b) $P(0.52 \leq z \leq 1.22)$.
    c) $P(-1.75 \leq z \leq -1.04)$.

14. Considerando que z es una variable aleatoria normal estándar, calcule z para cada situación.
    a) El área a la izquierda de z es 0.9750.
    b) El área entre 0 y z es 0.4750.
    c) El área a la izquierda de z es 0.7291.
    d) El área a la derecha de z es 0.1314.
    e) El área a la izquierda de z es 0.6700.
    f) El área a la derecha de z es 0.3300.

AUTOevaluación

15. Dado que z es una variable aleatoria normal estándar, calcule z para cada situación.
    a) El área a la izquierda de z es 0.2119.
    b) El área entre $-z$ y z es 0.9030.
    c) El área entre $-z$ y z es 0.2052.
    d) El área a la izquierda de z es 0.9948.
    e) El área a la derecha de z es 0.6915.

16. Considerando que z es una variable aleatoria normal estándar, calcule z para cada situación.
    a) El área a la derecha de z es 0.01.
    b) El área a la derecha de z es 0.025.
    c) El área a la derecha de z es 0.05.
    d) El área a la derecha de z es 0.10.

Aplicaciones

17. Para los deudores con buenas calificaciones de crédito, la deuda media de las cuentas revolventes y a plazos es de $15 015 (*BusinessWeek*, 20 de marzo de 2006). Suponga que la desviación estándar es $3 540 y que los montos de la deuda se distribuyen de manera normal.
    a) ¿Cuál es la probabilidad de que la deuda para un deudor con un buen crédito sea mayor de $18 000?
    b) ¿Cuál es la probabilidad de que la deuda para dicho deudor sea menor de $10 000?
    c) ¿Cuál es la probabilidad de que esta deuda esté entre $12 000 y $18 000?
    d) ¿Cuál es la probabilidad de que la deuda no sea mayor de $14 000?

AUTOevaluación

18. El precio medio de las acciones de las empresas que forman el S&P 500 es $30, y la desviación estándar es $8.20 (*BusinessWeek*, publicación anual especial, primavera de 2003). Suponga que los precios de las acciones se distribuyen normalmente.
    a) ¿Cuál es la probabilidad de que las acciones de una empresa tengan un precio mínimo de $40?
    b) ¿Cuál es la probabilidad de que el precio de las acciones no supere $20?
    c) ¿Qué tan alto debe ser el precio de las acciones de una firma para situarla en el 10% de las principales empresas?

19. En un artículo sobre el costo de la asistencia médica, la revista *Money* informó que una visita a la sala de urgencias de un hospital por algo tan simple como un dolor de garganta tiene un costo medio de $328 (*Money*, enero de 2009). Suponga que el costo de este tipo de visitas se distribuye normalmente con una desviación estándar de $92. Responda las preguntas siguientes sobre el costo de una visita a la sala de urgencias de un hospital para este servicio médico.
    a) ¿Cuál es la probabilidad de que el costo sea mayor que $500?
    b) ¿Cuál es la probabilidad de que sea menor que $250?
    c) ¿Cuál es la probabilidad de que esté entre $300 y $400?
    d) Si el costo para un paciente está en el 8% más bajo de cargos para este servicio médico, ¿cuál fue el costo de la visita a la sala de urgencias?

20. En enero de 2003, el empleado estadounidense pasó un promedio de 77 horas conectado a Internet mientras trabajaba (CNBC, 15 de marzo de 2003). Suponga que la media poblacional es 77 horas, los tiempos están distribuidos normalmente y la desviación estándar es de 20 horas.
    a) ¿Cuál es la probabilidad de que en enero de 2003 un empleado seleccionado al azar pasara menos de 50 horas conectado a Internet?
    b) ¿Qué porcentaje de empleados pasó más de 100 horas conectado a Internet en dicha fecha?
    c) Una persona es clasificada como usuario intensivo si está en el 20% superior de uso. En el mes de referencia, ¿cuántas horas tuvo que conectarse un empleado para que se le considerara un usuario intensivo?

21. Una persona debe estar en el 2% más alto de la población en una prueba de IQ para aspirar a la membresía de Mensa, la sociedad internacional de IQ alto (*U.S. Airways Attaché*, septiembre de 2000). Si las calificaciones del IQ están normalmente distribuidas con una media de 100 y una desviación estándar de 15, ¿qué puntaje debe tener una persona que desea calificar para Mensa?

22. La tarifa media de pago por hora para los directores de finanzas en la región central del noreste de Estados Unidos es de $32.62, y la desviación estándar es $2.32 (Bureau of Labor Statistics, septiembre de 2005). Suponga que las tarifas de pago están distribuidas normalmente.
    a) ¿Cuál es la probabilidad de que un director de finanzas gane entre $30 y $35 por hora?
    b) ¿Qué tan alta debe ser la tarifa por hora para ubicar a un director de finanzas en el 10% superior con respecto al pago?
    c) Para un director de finanzas seleccionado al azar, ¿cuál es la probabilidad de que gane menos de $28 por hora?

23. El tiempo necesario para completar un examen final en un curso universitario particular está distribuido normalmente con una media de 80 minutos y una desviación estándar de 10 minutos. Responda las preguntas siguientes.
    a) ¿Cuál es la probabilidad de completar el examen en una hora o menos?
    b) ¿Cuál es la probabilidad de que un estudiante termine el examen en más de 60 minutos pero en menos de 75?
    c) Suponga que la clase tiene 60 estudiantes y el periodo de examen dura 90 minutos. ¿Cuántos estudiantes esperaría usted que lo completaran en el tiempo asignado?

24. El volumen negociado en la Bolsa de Valores de Nueva York es más intenso durante la primera media hora (temprano por la mañana) y en la última media hora (tarde en la tarde) del día de negociación. Los volúmenes negociados temprano en la mañana (millones de acciones) durante 13 días en enero y febrero se muestran enseguida (*Barron's*, 23 de enero de 2006; 13 y 17 de febrero de 2006).

WEB archivo
Volume

214	163	265	194	180
202	198	212	201	
174	171	211	211	

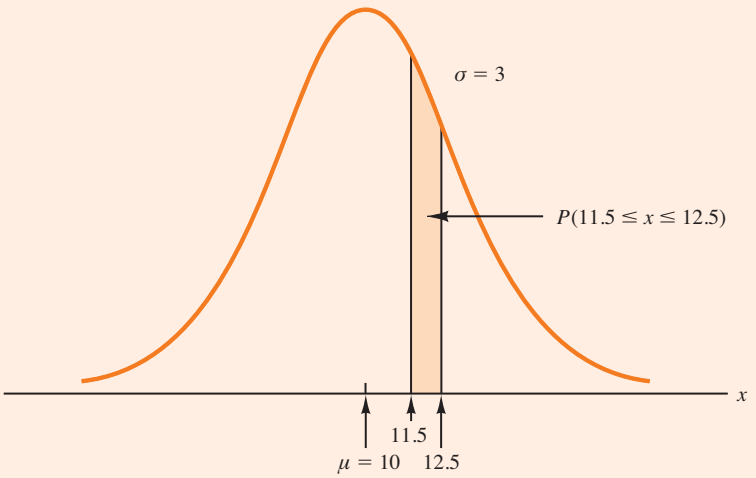
    La distribución de probabilidad del volumen negociado es aproximadamente normal.
    a) Calcule la media y la desviación estándar para usarla como estimaciones de la media poblacional y la desviación estándar.
    b) ¿Cuál es la probabilidad de que, en un día seleccionado al azar, el volumen negociado por la mañana sea menor de 180 millones de acciones?
    c) ¿Cuál es la probabilidad de que este volumen exceda los 230 millones de acciones?
    d) ¿Cuántas acciones deberán negociarse para que el volumen negociado por la mañana en un día determinado esté entre el 5% más ocupado de los días?

25. Según Sleep Foundation, el promedio de sueño nocturno es de 6.8 horas (*Fortune*, 20 de marzo de 2006). Suponga que la desviación estándar es 0.6 horas y que la distribución de probabilidad es normal.
    a) ¿Cuál es la probabilidad de que una persona seleccionada al azar duerma más de 8 horas?
    b) ¿Cuál es la probabilidad de que duerma 6 horas o menos?
    c) Los médicos sugieren dormir entre 7 y 9 horas cada noche. ¿Qué porcentaje de la población se toma este tiempo?

6.3 Aproximación normal de las probabilidades binomiales

En la sección 5.4 se presentó la distribución binomial discreta. Recuerde que un experimento binomial consiste en una secuencia de n ensayos independientes idénticos cada uno con dos resultados posibles: un éxito o un fracaso. La probabilidad de éxito es la misma para todos los ensayos y se denota como p. La variable aleatoria binomial es el número de éxitos en los n ensayos y las preguntas de probabilidad pertenecen a la probabilidad de x éxitos en los n ensayos.

FIGURA 6.8 Aproximación normal para una distribución de probabilidad normal con $n = 100$ y $p = 0.10$ que muestra la probabilidad de 12 errores



Cuando el número de ensayos es grande, es difícil evaluar la función de probabilidad binomial a mano o con una calculadora. En los casos en que $np \geq 5$ y $n(1 - p) \geq 5$, la distribución normal proporciona una aproximación fácil de usar de las probabilidades binomiales. Cuando se usa la aproximación normal a la binomial, se establece $\mu = np$ y $\sigma = \sqrt{np(1 - p)}$ en la definición de la curva normal.

La aproximación normal a la binomial se explicará mediante el ejemplo de una empresa particular que tiene una historia de cometer errores en 10% de sus facturas. Se tomó una muestra de 100 facturas y se quiere calcular la probabilidad de que 12 contengan errores. Es decir, se desea determinar la probabilidad binomial de 12 éxitos en 100 ensayos. Al aplicar la aproximación normal en este caso, se establece $\mu = np = (100)(0.1) = 10$ y $\sigma = \sqrt{np(1 - p)} = \sqrt{(100)(0.1)(0.9)} = 3$. Una distribución normal con $\mu = 10$ y $\sigma = 3$ se muestra en la figura 6.8.

Recuerde que, con una distribución de probabilidad continua, las probabilidades se calculan como las áreas bajo la función de densidad de probabilidad. Como resultado, la probabilidad de cualquier valor único para la variable aleatoria es cero. Por tanto, para aproximar la probabilidad binomial de 12 éxitos, se calcula el área bajo la curva normal correspondiente entre 11.5 y 12.5. El 0.5 que se suma y resta de 12 se llama **factor de corrección de continuidad**. Este concepto se introdujo porque se está utilizando una distribución continua para aproximar una distribución discreta. Por tanto, $P(x = 12)$ para la distribución binomial *discreta* se aproxima por $P(11.5 \leq x \leq 12.5)$ para la distribución normal *continua*.

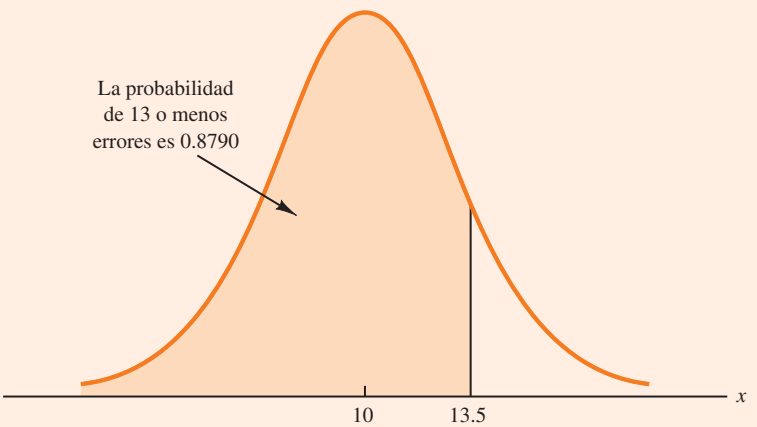
Al convertir a la distribución normal estándar para calcular $P(11.5 \leq x \leq 12.5)$, tenemos

$$z = \frac{x - \mu}{\sigma} = \frac{12.5 - 10.0}{3} = 0.83 \quad \text{en } x = 12.5$$

y

$$z = \frac{x - \mu}{\sigma} = \frac{11.5 - 10.0}{3} = 0.50 \quad \text{en } x = 11.5$$

FIGURA 6.9 Aproximación normal a una distribución de probabilidad binomial con $n = 100$ y $p = 0.10$ que muestra la probabilidad de 13 o menos errores



Al usar la tabla de probabilidad normal estándar, vemos que el área bajo la curva (figura 6.8) a la izquierda de 12.5 es 0.7967. Del mismo modo, el área bajo la curva a la izquierda de 11.5 es 0.6915. Por tanto, el área entre 11.5 y 12.5 es $0.7967 - 0.6915 = 0.1052$. La aproximación normal a la probabilidad de 12 éxitos en 100 ensayos es 0.1052.

En otro ejemplo, suponga que se desea calcular la probabilidad de 13 o menos errores en la muestra de 100 facturas. La figura 6.9 muestra el área bajo la curva normal que se aproxima a esta probabilidad. Tenga en cuenta que el uso del factor de corrección de continuidad da como resultado el valor de 13.5 utilizado para calcular la probabilidad deseada. El valor de z que corresponde a $x = 13.5$ es

$$z = \frac{13.5 - 10.0}{3.0} = 1.17$$

La tabla de probabilidad normal estándar muestra que el área bajo la curva normal estándar a la izquierda de $z = 1.17$ es 0.8790. El área bajo la curva normal que se aproxima a la probabilidad de 13 o menos errores está dada por la porción sombreada de la gráfica de la figura 6.9.

Ejercicios

Métodos

AUTOevaluación

26. Una distribución de probabilidad binomial tiene $p = 0.20$ y $n = 100$.
    a) ¿Cuál es la media y la desviación estándar?
    b) ¿Esta situación es una en la cual las probabilidades binomiales pueden aproximarse por medio de la distribución de probabilidad normal? Explique por qué.
    c) ¿Cuál es la probabilidad de exactamente 24 éxitos?
    d) ¿Cuál es la probabilidad de 18 a 22 éxitos?
    e) ¿Cuál es la probabilidad de 15 o menos éxitos?

27. Suponga que la distribución de probabilidad binomial tiene $p = 0.60$ y $n = 200$.
    a) ¿Cuáles son la media y la desviación estándar?
    b) ¿Esta situación es del tipo en que las probabilidades binomiales pueden aproximarse por medio de la distribución de probabilidad normal? Explique por qué.

c) ¿Cuál es la probabilidad de 100 a 110 éxitos?
d) ¿Y la probabilidad de 130 o más éxitos?
e) ¿Cuál es la ventaja de usar la distribución de probabilidad normal para aproximar las probabilidades binomiales? Use el inciso *d*) para explicarlo.

Aplicaciones

AUTOevaluación

28. Aunque continúan los estudios para mostrar que fumar ocasiona problemas de salud importantes, en Estados Unidos 20% de los adultos fuma. Considere un grupo de 250 sujetos.
    a) ¿Cuál es el número esperado de adultos que fuman?
    b) ¿Cuál es la probabilidad de que menos de 40 fumen?
    c) ¿Qué probabilidad hay de que de 55 a 60 adultos fumen?
    d) ¿Cuál es la probabilidad de que 70 o más fumen?

29. Un estudio del Consejo de Control de los Ingresos Internos encontró que 82% de los contribuyentes consideró que es muy importante para el Servicio de Administración Tributaria Estados Unidos (IRS) asegurar que los contribuyentes con ingresos altos no mientan sobre sus declaraciones de impuestos (*The Wall Street Journal*, 11 de febrero de 2009).
    a) Para una muestra de ocho contribuyentes, ¿cuál es la probabilidad de que por lo menos seis digan que es muy importante garantizar que los contribuyentes con ingresos altos no mientan sobre sus declaraciones de impuestos? Use la función de probabilidad de distribución binomial de la sección 5.4 para responder esta pregunta.
    b) Para una muestra de 80 contribuyentes, ¿cuál es la probabilidad de que por lo menos 60 digan que es muy importante garantizar que los contribuyentes con altos ingresos no mientan sobre sus declaraciones fiscales? Use la aproximación normal de la distribución binomial para responder esta pregunta.
    c) A medida que el número de ensayos en una aplicación de distribución binomial aumenta, ¿cuál es la ventaja de usar la aproximación normal de la distribución binomial para calcular las probabilidades?
    d) Cuando el número de ensayos para una aplicación de distribución binomial es grande, ¿los desarrolladores de *software* de estadística preferirían usar la función de probabilidad de distribución binomial mostrada en la sección 5.4 o la aproximación normal de la distribución binomial estudiada en la sección 6.3? Explique por qué.

30. Cuando usted firma un contrato para obtener una tarjeta de crédito, ¿lo lee detenidamente? En una encuesta de FindLaw.com se preguntó a las personas: "¿Cuánto cuidado pone al leer un contrato de tarjeta de crédito?" (*USA Today*, 16 de octubre de 2003). Los hallazgos arrojaron que 44% lee cada palabra, 33% lee lo suficiente para entender el contrato, 11% sólo le da un vistazo y 4% no lo lee.
    a) Para una muestra de 500 personas, ¿cuántas esperaría que dijeran que leen cada palabra de un contrato de tarjeta de crédito?
    b) Para una muestra de 500 personas, ¿cuál es la probabilidad de que 200 o menos digan que leen todas las palabras del contrato?
    c) Para una muestra de 500 personas, ¿cuál es la probabilidad de que por lo menos 15 digan que no leen los contratos?

31. Un hotel de un centro vacacional en Myrtle Beach tiene 120 habitaciones. En los meses de verano, la ocupación del hotel es de aproximadamente 75%.
    a) ¿Cuál es la probabilidad de que por lo menos la mitad de las habitaciones esté ocupada en un día determinado?
    b) ¿Cuál es la probabilidad de que 100 o más habitaciones estén ocupadas en un día determinado?
    c) ¿Cuál es la probabilidad de que 80 o menos estén ocupadas en tal día?

6.4 Distribución de probabilidad exponencial

La **distribución de probabilidad exponencial** puede usarse para variables aleatorias como el tiempo entre la llegada de un automóvil a un autolavado, el tiempo requerido para cargar un camión, la distancia entre los defectos importantes de una carretera, etc. La función de densidad de probabilidad exponencial se presenta a continuación.

FUNCIÓN DE DENSIDAD DE PROBABILIDAD EXPONENCIAL

$$f(x) = \frac{1}{\mu} e^{-x/\mu} \quad \text{para } x \geq 0 \tag{6.4}$$

donde μ = valor esperado o media

Como ejemplo de la distribución exponencial, suponga que x representa el tiempo de carga para un camión en el muelle Schips y sigue dicha distribución. Si la media, o promedio, del tiempo de carga es 15 minutos ($\mu = 15$), la función de densidad de probabilidad apropiada para x es

$$f(x) = \frac{1}{15} e^{-x/15}$$

La figura 6.10 es la gráfica de esta función de densidad de probabilidad.

Cálculo de probabilidades para la distribución exponencial

En las aplicaciones de línea en espera, la distribución exponencial a menudo se usa para el tiempo de servicio.

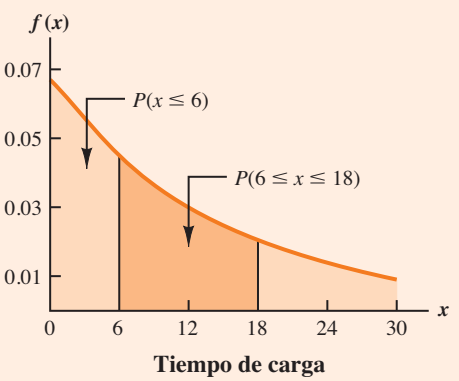
Al igual que con la distribución de probabilidad continua, el área bajo la curva correspondiente a un intervalo proporciona la probabilidad de que la variable aleatoria asuma un valor en ese intervalo. En el ejemplo del muelle Schips, la probabilidad de que cargar un camión tarde 6 minutos o menos $P(x \leq 6)$ se define como el área bajo la curva en la figura 6.10 de $x = 0$ a $x = 6$. De manera similar, la probabilidad de que dicho tiempo sea de 18 minutos o menos $P(x \leq 18)$ es el área bajo la curva de $x = 0$ a $x = 18$. Note también que la probabilidad de que el tiempo de carga esté entre 6 y 18 minutos $P(6 \leq x \leq 18)$ está dado por el área bajo la curva de $x = 6$ a $x = 18$.

Para calcular probabilidades exponenciales como las que se acaban de describir, se usa la fórmula siguiente, la cual proporciona la probabilidad acumulada de obtener un valor para la variable aleatoria exponencial menor o igual que un valor específico denotado por x_0.

DISTRIBUCIÓN EXPONENCIAL: PROBABILIDADES ACUMULADAS

$$P(x \leq x_0) = 1 - e^{-x_0/\mu} \tag{6.5}$$

FIGURA 6.10 Distribución exponencial para el ejemplo del muelle de carga Schips



Para el ejemplo del muelle Schips, x = tiempo de carga en minutos y $\mu = 15$ minutos. Usando la ecuación (6.5)

$$P(x \leq x_0) = 1 - e^{-x_0/15}$$

Por consiguiente, la probabilidad de que un camión tarde 6 minutos o menos es

$$P(x \leq 6) = 1 - e^{-6/15} = 0.3297$$

Utilizando la ecuación (6.5), calculamos la probabilidad de cargar un camión en 18 minutos o menos.

$$P(x \leq 18) = 1 - e^{-18/15} = 0.6988$$

Por tanto, la probabilidad de que la carga del camión tarde entre 6 y 18 minutos es igual a $0.6988 - 0.3297 = 0.3691$. Las probabilidades para cualquier otro intervalo pueden calcularse de manera similar.

Una propiedad de la distribución exponencial indica que la media y la desviación estándar son iguales.

En el ejemplo anterior, el tiempo medio que toma cargar un camión es $\mu = 15$ minutos. Una propiedad de la distribución exponencial indica que la media de la distribución y la desviación estándar de la distribución son *iguales*. Por tanto, la desviación estándar del tiempo que lleva cargar un camión es $\sigma = 15$ minutos. La varianza es $\sigma^2 = (15)^2 = 225$.

Relación entre las distribuciones de Poisson y exponencial

En la sección 5.5 se introdujo la distribución de Poisson como una distribución de probabilidad discreta que a menudo es útil cuando se examina el número de ocurrencias de un evento en un intervalo de tiempo o espacio específico. Recuerde que la función de probabilidad de Poisson es

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}$$

donde

$$\mu = \text{valor esperado o número medio de ocurrencias durante un intervalo especificado}$$

Si las llegadas siguen una distribución de Poisson, el tiempo entre las llegadas debe seguir una distribución exponencial.

La distribución de probabilidad exponencial continua está relacionada con la distribución de Poisson discreta. Si la distribución de Poisson proporciona una descripción apropiada del número de ocurrencias por intervalo, la distribución exponencial provee una descripción de la duración del intervalo entre ocurrencias.

Para ilustrar esta relación, suponga que el número de automóviles que llegan a un autolavado durante una hora se describe por medio de una distribución de probabilidad de Poisson con una media de 10 automóviles por hora. La función de probabilidad de Poisson que da la probabilidad de x llegadas por hora es

$$f(x) = \frac{10^x e^{-10}}{x!}$$

Como el número medio de arribos es 10 automóviles por hora, el tiempo promedio entre la llegada de los vehículos es

$$\frac{1 \text{ hora}}{10 \text{ automóviles}} = 0.1 \text{ hora/automóvil}$$

Por tanto, la distribución exponencial correspondiente que describe el tiempo entre las llegadas tiene una media de $\mu = 0.1$ hora por automóvil; como resultado, la función de densidad de probabilidad exponencial apropiada es

$$f(x) = \frac{1}{0.1} e^{-x/0.1} = 10e^{-10x}$$

NOTAS Y COMENTARIOS

Como se aprecia en la figura 6.10, la distribución exponencial está inclinada a la derecha. De hecho, la medición de la inclinación para este tipo de distribuciones es 2. La distribución exponencial nos da una buena idea de cómo se ve una distribución inclinada.

Ejercicios

Métodos

32. Considere la función de densidad de probabilidad exponencial siguiente.

$$f(x) = \frac{1}{8} e^{-x/8} \qquad \text{para } x \geq 0$$

a) Calcule $P(x \leq 6)$.
b) Encuentre $P(x \leq 4)$.
c) Calcule $P(x \geq 6)$.
d) Determine $P(4 \leq x \leq 6)$.

AUTOevaluación

33. Considere la función de densidad de probabilidad exponencial siguiente.

$$f(x) = \frac{1}{3} e^{-x/3} \qquad \text{para } x \geq 0$$

a) Escriba la fórmula para $P(x \leq x_0)$.
b) Calcule $P(x \leq 2)$.
c) Determine $P(x \geq 3)$.
d) Calcule $P(x \leq 5)$.
e) Encuentre $P(2 \leq x \leq 5)$.

Aplicaciones

34. El tiempo requerido para pasar la inspección de seguridad en el aeropuerto puede ser molesto para los viajeros. El tiempo de espera medio durante los periodos pico en el Aeropuerto Internacional de Cincinnati/norte de Kentucky es de 12.1 minutos (*The Cincinnati Enquirer*, 2 de febrero de 2006). Suponga que el tiempo para pasar la inspección de seguridad sigue una distribución exponencial.
a) ¿Cuál es la probabilidad de que tarde menos de 10 minutos pasar la inspección de seguridad durante un periodo pico?
b) ¿Cuál es la probabilidad de que pasar la inspección tarde más de 20 minutos?
c) ¿Cuál es la probabilidad de que la inspección tome entre 10 y 20 minutos?
d) Son las 8:00 a.m. (un periodo pico) y usted acaba de formarse en la fila de inspección. Para tomar su vuelo debe estar en la puerta en 30 minutos. Transcurren 12 minutos desde el momento en que pasa la inspección de seguridad hasta que llega a su puerta, ¿cuál es la probabilidad de que pierda el vuelo?

AUTOevaluación

35. El tiempo entre las llegadas de los vehículos en una intersección particular sigue una distribución de probabilidad exponencial con una media de 12 segundos.
a) Trace esta distribución de probabilidad exponencial.
b) ¿Cuál es la probabilidad de que el tiempo de llegada entre los vehículos sea de 12 o menos segundos?
c) ¿Cuál es la probabilidad de que este tiempo sea de 6 o menos segundos?
d) ¿Cuál es la probabilidad de 30 o más segundos entre las llegadas de vehículos?

36. Comcast Corporation es la compañía de televisión por cable más grande, el segundo proveedor de servicios de Internet más importante, y el cuarto proveedor de servicios telefónicos más grande de Estados Unidos. La empresa, generalmente conocida por su calidad y servicio confiable, experimenta periódicamente interrupciones de servicio inesperadas. El 14 de enero de 2009, una interrupción de este tipo ocurrió para los clientes que vivían en el suroeste de Florida. Cuando los clientes llamaron a la oficina de Comcast, un mensaje grabado les dijo que la empresa estaba al tanto del corte del servicio y que se anticipaba que éste sería restablecido en dos horas. Suponga que dos horas es el tiempo medio para efectuar la reparación y que el tiempo de reparación tiene una distribución de probabilidad exponencial.
    a) ¿Cuál es la probabilidad de que el servicio de cable sea reanudado en una hora o menos?
    b) ¿Cuál es la probabilidad de que la reparación tarde entre una y dos horas?
    c) Para un cliente que llama a la oficina de Comcast a la 1:00 p.m., ¿cuál es la probabilidad de que el servicio de cable no se haya reanudado todavía a las 5:00 p.m.?

37. Collina's Italian Café en Houston, Texas, anuncia que los pedidos tardan en llegar alrededor de 25 minutos (sitio web de Collina's, 27 de febrero de 2008). Suponga que el tiempo necesario para que un pedido esté listo a fin de que lo recoja el cliente tiene una distribución exponencial con una media de 25 minutos.
    a) ¿Cuál es la probabilidad de que un pedido para llevar esté listo en 20 minutos?
    b) Si un cliente llega 30 minutos después de hacer un pedido, ¿cuál es la probabilidad de que la orden no esté lista?
    c) Un cliente particular vive a 15 minutos del Collina's Italian Café. Si el cliente realiza un pedido telefónico a las 5:20 de la tarde, ¿cuál es la probabilidad de que el cliente pueda acudir en auto a la cafetería, recoger el pedido y regresar a casa antes de las 6:00 p.m.?

38. ¿Las interrupciones mientras usted trabaja reducen su productividad? Según un estudio de la Universidad de California, las personas de negocios son interrumpidas a una tasa de aproximadamente 5½ veces por hora (*Fortune*, 20 de marzo de 2006). Suponga que el número de interrupciones sigue una distribución de probabilidad de Poisson.
    a) Muestre la distribución de probabilidad para el tiempo entre interrupciones.
    b) ¿Cuál es la probabilidad de que una persona de negocios no tenga interrupciones durante un periodo de 15 minutos?
    c) ¿Cuál es la probabilidad de que la siguiente interrupción ocurra dentro de 10 minutos para una persona de negocios en particular?

Resumen

Este capítulo amplía el análisis de las distribuciones de probabilidad para el caso de las variables aleatorias continuas. La principal diferencia conceptual entre las distribuciones de probabilidad discretas y continuas involucra el método de calcular las probabilidades. Con las distribuciones discretas, la función de probabilidad $f(x)$ proporciona la probabilidad de que la variable aleatoria x asuma varios valores. Con las distribuciones continuas, la función de densidad de probabilidad $f(x)$ no proporciona los valores de probabilidad directamente. En su lugar, las probabilidades están dadas por las áreas bajo la curva o gráfica de la función de densidad de probabilidad $f(x)$. Debido a que el área bajo la curva encima de un solo punto es cero, se observa que la probabilidad de cualquier valor particular es cero para una variable aleatoria continua.

Tres distribuciones de probabilidad continua —uniforme, normal y exponencial— se tratan con detalle en este capítulo. La distribución normal se utiliza ampliamente en la inferencia estadística y su uso es abundante en el resto del libro.

Glosario

Distribución de probabilidad exponencial Distribución de probabilidad continua que se utiliza en el cálculo de probabilidades para el tiempo que toma completar una tarea.
Distribución de probabilidad normal Distribución de probabilidad continua. Su función de densidad de probabilidad tiene forma de campana y está determinada por su media μ y su desviación estándar σ.
Distribución de probabilidad normal estándar Distribución normal con una media de cero y una desviación estándar de uno.
Distribución de probabilidad uniforme Distribución de probabilidad continua para la cual la probabilidad de que la variable aleatoria asuma un valor en cualquier intervalo es la misma para cada intervalo de igual longitud.
Factor de corrección de continuidad Valor de 0.5 que se suma o resta de un valor de x cuando la distribución normal continua se usa para aproximar la distribución binomial discreta.
Función de densidad de probabilidad Función utilizada para calcular las probabilidades de una variable aleatoria continua. El área bajo la gráfica de una función de densidad de la probabilidad a lo largo de un intervalo representa la probabilidad.

Fórmulas clave

Función de densidad de probabilidad uniforme

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{para } a \le x \le b \\ 0 & \text{en cualquier otro caso} \end{cases} \tag{6.1}$$

Función de densidad de probabilidad normal

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{6.2}$$

Conversión a la variable aleatoria normal estándar

$$z = \frac{x-\mu}{\sigma} \tag{6.3}$$

Función de densidad de probabilidad exponencial

$$f(x) = \frac{1}{\mu} e^{-x/\mu} \qquad \text{para } x \ge 0 \tag{6.4}$$

Distribución exponencial: probabilidades acumuladas

$$P(x \le x_0) = 1 - e^{-x_0/\mu} \tag{6.5}$$

Ejercicios complementarios

39. Una ejecutiva de negocios, transferida de Chicago a Atlanta, necesita vender su casa en Chicago rápidamente. Un ejecutivo de la empresa ha ofrecido comprarla por $210 000, pero la oferta expira al final de la semana. La ejecutiva actualmente no tiene una mejor oferta, pero puede darse el lujo de dejar la casa en el mercado otro mes. De las conversaciones con su

agente de bienes raíces, el ejecutivo cree que el precio que obtendrá al dejar la casa en el mercado otro mes si está distribuido de manera uniforme estará entre $200 000 y $225 000.

a) Si deja la casa en el mercado otro mes, ¿cuál es la expresión matemática para la función de densidad de probabilidad del precio de venta?

b) Si la deja en el mercado otro mes, ¿cuál es la probabilidad de que obtenga por lo menos $215 000 por la casa?

c) Si la deja en el mercado otro mes, ¿cuál es la probabilidad de que obtenga menos de $210 000?

d) ¿La ejecutiva debe dejar la casa en el mercado otro mes? ¿Por qué?

40. La Oficina de Estadísticas Laborales de Estados Unidos informa que los gastos anuales promedio en alimentos y bebidas para todas las familias asciende a $5 700 (*Money*, diciembre de 2003). Suponga que dichos gastos anuales están distribuidos normalmente y que la desviación estándar es $1 500.

a) ¿Cuál es el rango de gastos de 10% de las familias con el gasto anual más bajo en alimentos y bebidas?

b) ¿Qué porcentaje de las familias erogó más de $7 000 al año ambos conceptos?

c) ¿Cuál es el rango de gastos para 5% de las familias con el gasto anual más alto en alimentos y bebidas?

41. Motorola utiliza la distribución normal para determinar la probabilidad de defectos y su número esperado en un proceso de producción. Suponga que este proceso genera artículos con un peso medio de 10 onzas. Calcule la probabilidad de un defecto y el número esperado de defectos en una corrida de producción de 1 000 unidades en las siguientes situaciones.

a) La desviación estándar del proceso es 0.15, y el control de procesos se establece en más o menos una desviación estándar. Las unidades con un peso inferior a 9.85 o superior a 10.15 onzas se clasificarán como defectos.

b) Por medio de mejoras en el diseño de procesos, la desviación estándar del proceso puede reducirse a 0.05. Suponga que el control de procesos sigue siendo el mismo, con pesos inferiores a 9.85 o superiores a 10.15 onzas que se clasificarán como defectos.

c) ¿Cuál es la ventaja de reducir la variación en el proceso, ocasionando así que los límites del control de procesos estén en un mayor número de desviaciones estándar de la media?

42. La cantidad media anual que las familias estadounidenses gastan en el transporte diario es $6 312 (*Money*, agosto de 2001). Considere que este monto está normalmente distribuido.

a) Suponga que se entera de que 5% de las familias estadounidenses gastó menos de $1000 en transporte diario. ¿Cuál es la desviación estándar de la cantidad erogada?

b) ¿Cuál es la probabilidad de que una familia gaste entre $4 000 y $6 000?

c) ¿Cuál es el rango de gasto para 3% de las familias con el costo de transporte diario más alto?

43. *Condé Nast Traveler* publica una lista de oro de los hoteles principales en todo el mundo. El hotel Broadmoor en Colorado Springs tiene 700 habitaciones y está en la lista de oro de 2004 (*Condé Nast Traveler*, enero de 2004). Suponga que el grupo de marketing de Broadmoor pronostica una demanda media de 670 habitaciones para el próximo fin de semana. Considere que esta demanda está distribuida normalmente con una desviación estándar de 30.

a) ¿Cuál es la probabilidad de que todas las habitaciones del hotel sean rentadas?

b) ¿Cuál es la probabilidad de que 50 o más habitaciones no sean rentadas?

c) ¿Recomendaría usted al hotel considerar la oferta de una promoción para aumentar la demanda? ¿Qué consideraciones serían importantes?

44. Ward Doering Auto Sales estudia ofrecer un contrato de servicio especial que cubra el costo total de cualquier trabajo de servicio requerido en los vehículos rentados. Por experiencia, el gerente de la empresa estima que los costos del servicio anuales están distribuidos normalmente de manera aproximada, con una media de $150 y una desviación estándar de $25.

a) Si la empresa ofrece el contrato de servicio a los clientes por un cargo anual de $200, ¿cuál es la probabilidad de que los costos del servicio al cliente rebasen el precio de contrato de $200?

b) ¿Cuáles son las utilidades esperadas de Ward's por contrato de servicio?

45. ¿La falta de sueño ocasiona muertes por tráfico? Un estudio realizado bajo los auspicios de la National Highway Traffic Safety Administration encontró que el número medio de accidentes fatales provocados por conductores somnolientos cada año es de 1 550 (*BusinessWeek*, 26 de enero de 2004). Suponga que el número anual de percances fatales se distribuye normalmente con una desviación estándar de 300.
    a) ¿Cuál es la probabilidad de menos de 1000 accidentes fatales en un año?
    b) ¿Cuál es la probabilidad de que el número de percances fatales esté entre 1 000 y 2 000 al año?
    c) Durante un año de estar en el 5% superior con respecto al número de accidentes fatales, ¿cuántos percances tendrían que ocurrir?

46. Asuma que los resultados del examen de admisión a la universidad tienen una distribución normal, con una media de 450 y una desviación estándar de 100.
    a) ¿Qué porcentaje de las personas que presentó la prueba obtuvo una calificación de puntuación entre 400 y 500?
    b) Suponga que alguien obtiene una calificación de 630. ¿Qué porcentaje de las personas que presentaron la prueba logró la mejor calificación? ¿Qué porcentaje obtuvo la peor?
    c) Si una universidad particular no admite alumnos con una calificación inferior a 480, ¿qué porcentaje de las personas que presentaron la prueba será aceptable para la universidad?

47. Según Salary Wizard, el sueldo base promedio de un gerente de marca en Houston, Texas, es de $88 592 y el de un gerente de marca en Los Ángeles, California, es de $97 417 (sitio web de Salary Wizard, 27 de febrero de 2008). Suponga que los sueldos están normalmente distribuidos y que la desviación estándar de los gerentes de marca en Houston es de $19 900 y en Los Ángeles es de $21 800.
    a) ¿Cuál es la probabilidad de que un gerente de marca en Houston tenga un sueldo base mayor de $100 000?
    b) ¿Cuál es la probabilidad de que su homólogo en Los Ángeles tenga un sueldo base que rebase los $100 000?
    c) ¿Cuál es la probabilidad de que un gerente de marca en Los Ángeles tenga un sueldo base inferior a $75 000?
    d) ¿Cuánto tendría que ganar un gerente en Los Ángeles para tener un sueldo superior a 99% de sus homólogos de Houston?

48. Una máquina llena envases de un producto en particular. Se sabe a partir de datos previos que la desviación estándar de los pesos de llenado es 0.6 oz. Si sólo a 2% de los envases contiene menos de 18 onzas, ¿cuál es la media del peso de llenado de la máquina? Es decir, ¿a qué debe ser igual μ? Suponga que los pesos de llenado tienen una distribución normal.

49. Considere un examen de opción múltiple con 50 preguntas. Cada interrogante tiene cuatro respuestas posibles. Suponga que un estudiante que hizo la tarea y asistió a conferencias tiene una probabilidad de 75% de responder correctamente cualquier pregunta.
    a) Un estudiante debe responder correctamente 43 o más preguntas para obtener una calificación de A. ¿Qué porcentaje de los que hicieron su tarea y asistieron a conferencias obtendrá una calificación de A en este examen de opción múltiple?
    b) Un alumno que responde correctamente de 35 a 39 preguntas recibirá una calificación de C. ¿Qué porcentaje de los que realizaron su tarea y asistieron a conferencias obtendrá una calificación de C en este examen?
    c) Un estudiante debe responder correctamente 30 o más preguntas para aprobar el examen. ¿Qué porcentaje de los que efectuaron su tarea y asistieron a las conferencias lo aprobará?
    d) Considere que un estudiante no asistió a clases y no hizo la tarea para el curso. Por otra parte, suponga que éste sencillamente adivina la respuesta a cada pregunta. ¿Cuál es la probabilidad de que conteste 30 o más preguntas correctamente y apruebe el examen?

50. Un jugador de blackjack en un casino de Las Vegas se enteró de que la casa proporcionará una habitación gratis si el juego dura cuatro horas con una apuesta media de $50. La estrategia

del jugador proporciona una probabilidad de 0.49 de ganar en cualquier mano, y sabe que hay 60 manos por hora. Suponga que el sujeto juega durante cuatro horas con una apuesta de $50 por mano.

a) ¿Cuál es el pago esperado del jugador?
b) ¿Cuál es la probabilidad de que pierda $1 000 o más?
c) ¿Cuál es la probabilidad de que gane?
d) Suponga que el jugador comienza con $1 500. ¿Cuál es la probabilidad de que se quede sin dinero para apostar?

51. El tiempo en minutos durante el cual un estudiante utiliza una terminal de computadora en el centro informático de una universidad importante sigue una distribución de probabilidad exponencial con una media de 36 minutos. Suponga que un estudiante llega a la terminal al mismo tiempo que otro empieza a trabajar en ella.
a) ¿Cuál es la probabilidad de que la espera para el segundo estudiante sea de 15 minutos o menos?
b) ¿Cuál es la probabilidad de que deba aguardar entre 15 y 45 minutos?
c) ¿Cuál es la probabilidad de que tenga que esperar una hora o más?

52. El sitio web de Bed and Breakfast Inns of North America recibe aproximadamente a siete visitantes por minuto (*Time*, septiembre de 2001). Suponga que el número de visitantes al sitio web por minuto sigue una distribución de probabilidad de Poisson.
a) ¿Cuál es el tiempo medio entre visitas al sitio web?
b) Muestre la función de densidad de probabilidad exponencial para el tiempo entre las consultas al sitio.
c) ¿Cuál es la probabilidad de que nadie entre al sitio web en un periodo de 1 minuto?
d) ¿Cuál es la probabilidad de que nadie entre en un periodo de 12 segundos?

53. The American Community Survey reveló que los residentes de la ciudad de Nueva York tienen los tiempos de viaje más largos para transportarse al trabajo en comparación con los residentes de otras ciudades de Estados Unidos (sitio web de la Oficina del Censo de Estados Unidos, agosto de 2008). Con base en las últimas estadísticas disponibles, el tiempo medio de viaje para transportarse al trabajo para los habitantes de la ciudad de Nueva York es de 38.3 minutos.
a) Suponga que la distribución de probabilidad exponencial es aplicable y muestra la función de densidad de probabilidad del tiempo de viaje para transportarse al trabajo para un residente de esta ciudad.
b) ¿Cuál es la probabilidad de que un neoyorquino tarde entre 20 y 40 minutos para transportarse al trabajo?
c) ¿Cuál es la probabilidad de que tarde más de una hora?

54. El tiempo (en minutos) entre llamadas telefónicas en una oficina de reclamación de seguros tiene la distribución de probabilidad exponencial siguiente.

$$f(x) = 0.50e^{-0.50x} \qquad \text{para } x \geq 0$$

a) ¿Cuál es el tiempo medio entre llamadas telefónicas?
b) ¿Cuál es la probabilidad de tener 30 segundos o menos entre llamadas telefónicas?
c) ¿Cuál es la probabilidad de que se tenga 1 minuto o menos?
d) ¿Cuál es la probabilidad de que pasen 5 o más minutos sin una llamada?

Caso a resolver Specialty Toys

Specialty Toys, Inc. vende una variedad de juguetes infantiles nuevos e innovadores. La gerencia se enteró de que la temporada prenavideña es el mejor momento para introducirlos, porque muchas familias utilizan este tiempo para buscar nuevas ideas para sus regalos de navidad en diciembre. Cuando Specialty Toys descubre un juguete nuevo con buen potencial de mercado, elige una fecha de entrada al mercado en octubre.

Con el fin de que los productos estén en los estantes de las tiendas en octubre, Specialty hace un solo pedido con sus fabricantes en junio o julio de cada año. La demanda de juguetes infantiles puede ser muy volátil. Si un juguete nuevo se vuelve popular, una sensación de

escasez en el mercado aumenta a menudo la demanda a niveles altos y se pueden obtener grandes utilidades. Sin embargo, los juguetes nuevos también pueden ser un fracaso, dejando a Specialty atorado con altos niveles de inventario que debe vender a precio bajo. Lo más importante que la empresa enfrenta es decidir cuántas unidades de un juguete nuevo debe comprar para satisfacer la demanda anticipada de ventas. Si se adquieren muy pocos, las ventas se perderán, si se compran muchos, las utilidades se reducirán debido a los precios bajos de las ventas de liquidación.

Para la próxima temporada, Specialty planea introducir un producto nuevo llamado Weather Teddy. Esta variante de un osito de peluche que habla se fabrica en una empresa de Taiwán. Cuando un niño presiona la mano del osito, éste empieza a hablar. Un barómetro integrado selecciona una de cinco respuestas que predicen las condiciones del clima. Las respuestas van desde "¡Parece ser buen día! Diviértete" a "Creo que va a llover hoy. No olvides tu paraguas." Las pruebas del producto demostraron que, aunque no es un predictor del clima perfecto, sus pronósticos son sorprendentemente atinados. Varios directivos de Specialty afirmaron que Teddy hizo predicciones climáticas tan buenas como muchos pronosticadores meteorológicos locales de televisión.

Al igual que con otros productos, Specialty enfrenta la decisión de cuántas unidades de Teddy ordenar para la próxima temporada de vacaciones. Los miembros del equipo gerencial sugirieron solicitar cifras de 15 000, 18 000, 24 000 o 28 000 unidades. La amplia variedad de cantidades sugerida para el pedido indica un desacuerdo considerable respecto del potencial de mercado. El equipo de administración del producto le solicita tanto un análisis de las probabilidades de que las existencias se agoten para pedidos de varias cantidades, como una estimación del potencial de utilidades, y una recomendación de la cantidad del pedido. Specialty espera vender el oso Weather Teddy en $24 sobre la base de un costo de $16 por unidad. Si queda inventario después de la temporada de vacaciones, la tienda venderá todo el excedente en $5 por unidad. Después de revisar el historial de ventas de productos similares, el encargado de pronósticos de ventas adjunto de Specialty predijo una demanda esperada de 20 000 unidades con una probabilidad 0.95 de que se ubicara entre 10 000 y 30 000 unidades.

Informe gerencial

Prepare un informe gerencial que aborde los temas siguientes y recomiende una cantidad de pedido para el producto Wealher Teddy.

1. Use la predicción del pronosticador de ventas para describir una distribución de probabilidad normal que permita aproximar la distribución de la demanda. Trace la distribución y muestre su media y desviación estándar.
2. Calcule la probabilidad de quedarse sin existencias para las cantidades de pedido sugeridas por los miembros del equipo gerencial.
3. Calcule las utilidades proyectadas para las cantidades de pedido sugeridas por el equipo de administración bajo tres escenarios: el peor caso en el cual las ventas = 10 000 unidades; el caso más probable de ventas = 20 000 unidades, y el mejor caso en el cual las ventas = 30 000 unidades.
4. Uno de los gerentes de Specialty consideró que el potencial de utilidades es tan grande que la cantidad de pedidos debe tener una posibilidad de 70% de satisfacer la demanda y sólo una probabilidad de 30% de que se agoten las existencias. ¿Qué cantidad debe solicitarse bajo esta política, y cuál es la utilidad proyectada bajo los tres escenarios de ventas?
5. Proporcione su propia sugerencia para una cantidad del pedido y observe las proyecciones de las utilidades asociadas. Comente en qué se basa para hacer su recomendación.

Apéndice 6.1 Distribuciones de probabilidad continua con Minitab

En este apéndice se demostrará el procedimiento de Minitab para calcular las probabilidades continuas en relación con el problema de Grear Tire Company, donde el millaje de los neumá-

ticos se describió por medio de una distribución normal con $\mu = 36\,500$ y $\sigma = 5\,000$. Una pregunta formulada al respecto fue: ¿cuál es la probabilidad de que el millaje de neumáticos rebasará las 40 000 millas?

Para las distribuciones de probabilidad continua, Minitab proporciona una probabilidad acumulada; es decir, la probabilidad de que la variable aleatoria asuma un valor menor o igual que una constante especificada. Para la pregunta del millaje de Grear Tire, se usa este programa con el fin de determinar la probabilidad acumulada de que el millaje de los neumáticos sea menor o igual que 40 000 millas. (La constante especificada en este caso es 40 000.) Después de obtener la probabilidad acumulada de Minitab, debemos restarla de 1 para determinar la probabilidad de que el millaje de los neumáticos rebase la cifra de 40 000.

Antes de usar Minitab para calcular una probabilidad, debemos introducir la constante especificada en una columna de la hoja de trabajo. Para la pregunta del millaje de los neumáticos de Grear se introdujo la constante especificada de 40 000 en la columna C1 de la hoja de trabajo. Los pasos para usar Minitab con el propósito de calcular la probabilidad acumulada de la variable aleatoria normal asumiendo un valor menor o igual que 40 000 se muestran a continuación.

Paso 1. Seleccione el menú **Calc.**
Paso 2. Elija **Probability Distributions.**
Paso 3. Elija **Normal.**
Paso 4. Cuando aparezca el cuadro de diálogo Normal Distribution:
Seleccione **Cumulative probability.**
Introduzca 36 500 en el cuadro **Mean.**
Ingrese 5 000 en el cuadro **Standard deviation.**
Introduzca C1 en el cuadro **Input column** (la columna que contiene 40 000).
Haga clic en **OK.**

Después de que el usuario hace clic en **OK**, Minitab imprime la probabilidad acumulada de que la variable aleatoria normal asume un valor menor o igual que 40 000, y muestra que esta probabilidad es de 0.7580. Puesto que se quiere conocer la probabilidad de que el millaje de los neumáticos sea mayor que 40 000, la probabilidad deseada es $1 - 0.7580 = 0.2420$.

Una segunda pregunta en el problema de Grear Tire Company fue: ¿qué garantía de millaje debe establecer Grear para asegurar que no más de 10% de los neumáticos califique para hacerla válida? En seguida se proporciona una probabilidad y se quiere encontrar el valor correspondiente para la variable aleatoria. Minitab usa una rutina de cálculo inversa para encontrar el valor de la variable aleatoria asociada con una probabilidad acumulada determinada. Primero, debemos introducir esta última en una columna de la hoja de trabajo de Minitab (por ejemplo, C1). En este caso, la probabilidad acumulada buscada es 0.10. Después seguimos los primeros tres pasos del procedimiento de Minitab como ya se explicó. En el paso 4 se selecciona **Inverse cumulative probability** en vez de **Cumulative probability** y se completan las partes restantes del procedimiento. Minitab exhibe luego la garantía de 30 092 millas.

El programa es capaz de calcular las probabilidades para otras distribuciones de probabilidad continua, incluida la distribución de probabilidad exponencial. Para calcular las probabilidades exponenciales, siga el procedimiento explicado antes para la distribución de probabilidad normal y elija la opción **Exponential** en el paso 3. El paso 4 es como se muestra, con la excepción de que no es necesario introducir la desviación estándar. El resultado para las probabilidades acumuladas y las probabilidades acumuladas inversas es idéntico al descrito para la distribución de probabilidad normal.

Apéndice 6.2 Distribuciones de probabilidad continua con Excel

Excel permite calcular probabilidades para varias distribuciones de probabilidad continua, incluidas las distribuciones de probabilidad normal y exponencial. En este apéndice se describe

cómo se usa Excel para calcular probabilidades de cualquier distribución normal. Los procedimientos para las distribuciones exponencial y otras continuas son similares a los que se describen para la distribución normal.

Retomemos el problema de Grear Tire Company, donde el millaje de los neumáticos se describió por medio de una distribución normal con $\mu = 36\,500$ y $\sigma = 5\,000$. Suponga que le interesa la probabilidad de que el millaje de los neumáticos exceda las 40 000 millas.

La función NORMDIST de Excel proporciona las probabilidades acumuladas de una distribución normal. La forma general de la función es NORMDIST (x,μ,σ,cumulative). Para el cuarto argumento, se especifica TRUE si se desea obtener una probabilidad acumulada. Por tanto, para calcular la probabilidad acumulada de que el millaje de los neumáticos sea menor o igual que 40 000 se introduciría la fórmula siguiente en cualquier celda de una hoja de trabajo de Excel:

$$= \text{NORMDIST}(40000,36500,5000,\text{TRUE})$$

En este punto, aparecerá 0.7580 en la celda donde fue introducida la fórmula, lo que indica que la probabilidad de que el millaje de los neumáticos sea menor o igual que 40 000 es 0.7580. Por tanto, la probabilidad de que el millaje de los neumáticos sea superior a 40 000 es $1 - 0.7580 = 0.2420$.

La función NORMINV de Excel utiliza un cálculo inverso para encontrar el valor de x correspondiente a una determinada probabilidad acumulada. Por ejemplo, suponga que se quiere determinar el millaje garantizado que Grear debe ofrecer para que no más de 10% de los neumáticos sea apto para la garantía. Se introduce la fórmula siguiente en cualquier celda de una hoja de trabajo de Excel:

$$= \text{NORMINV}(.1,36500,5000)$$

En este punto, aparece 30092 en la celda donde se introdujo la fórmula, lo que indica que la probabilidad de que un neumático dura 30 092 millas o menos es 0.10.

La función de Excel para el cálculo de probabilidades exponenciales es EXPONDIST. Su uso es sencillo. Pero si usted necesita especificar los valores adecuados para los argumentos, el cuadro de diálogo Insert Function de Excel puede ayudarle (vea el apéndice E del libro).

CAPÍTULO 7

Muestreo y distribuciones de muestreo

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: MEADWESTVACO CORPORATION

7.1 EL PROBLEMA DE MUESTREO DE ELECTRONICS ASSOCIATES

7.2 SELECCIÓN DE UNA MUESTRA
- Muestreo de una población finita
- Muestreo de una población infinita

7.3 ESTIMACIÓN PUNTUAL
- Consejo práctico

7.4 INTRODUCCIÓN A LAS DISTRIBUCIONES MUESTRALES O DE MUESTREO

7.5 DISTRIBUCIÓN DE MUESTREO DE $\bar{x}$
- Valor esperado de $\bar{x}$
- Desviación estándar de $\bar{x}$
- Forma de la distribución de muestreo de $\bar{x}$
- Distribución de muestreo de $\bar{x}$ en el problema EAI
- Valor práctico de la distribución de muestreo de $\bar{x}$
- Relación entre el tamaño de la muestra y la distribución de muestreo de $\bar{x}$

7.6 DISTRIBUCIÓN DE MUESTREO DE $\bar{p}$
- Valor esperado de $\bar{p}$
- Desviación estándar de $\bar{p}$
- Forma de la distribución de muestreo de $\bar{p}$
- Valor práctico de la distribución de muestreo de $\bar{p}$

7.7 PROPIEDADES DE LOS ESTIMADORES PUNTUALES
- Insesgadez
- Eficiencia
- Consistencia

7.8 OTROS MÉTODOS DE MUESTREO
- Muestreo aleatorio estratificado
- Muestreo por conglomerados
- Muestreo sistemático
- Muestreo de conveniencia
- Muestreo subjetivo

ESTADÍSTICA *en* LA PRÁCTICA

MEADWESTVACO CORPORATION*

STAMFORD, CONNECTICUT

MeadWestvaco Corporation, líder mundial en la producción de embalajes y papeles especiales, bienes de consumo y de oficina y sustancias químicas especiales, emplea a más de 30 000 personas. Opera a nivel mundial en 29 países y atiende a clientes localizados en 100 países. La empresa tiene una posición líder en la producción de papel, con una capacidad de 1.8 millones de toneladas anuales. Entre los productos que comercializa se encuentran papel para libros y revistas, sistemas de embalaje para bebidas y productos de oficina. Los consultores internos de MeadWestvaco usan el muestreo para obtener información diversa que permite a la empresa ganar productividad y seguir siendo competitiva.

Por ejemplo, la firma posee bosques que le proporcionan los árboles, o la materia prima, para muchos de sus productos. Los directivos requieren información confiable y precisa acerca de sus bienes maderables para evaluar las posibilidades de satisfacción de las futuras necesidades de materia prima. ¿Cuál es el volumen actual de los bosques? ¿Cuál ha sido su crecimiento? ¿Cuál es su crecimiento proyectado? Las respuestas a estas preguntas permiten a los directivos elaborar los planes para el futuro, incluyendo proyecciones a largo plazo y calendarios para la tala de árboles.

¿Cómo recolecta MeadWestvaco la información acerca de los amplios bosques que requiere? Los datos que obtiene de puntos muestrales en las áreas forestales son la base para contar con información acerca de la población de árboles propiedad de la empresa. Para localizar estos puntos muestrales, primero se dividen los bosques en tres secciones con base en la localización y el tipo de especímenes. Mediante mapas y números aleatorios, los analistas identifican puntos muestrales aleatorios de 1/5 a 1/7 de acres en cada sección forestal. Los ingnieros de MeadWestvaco recogen los datos de estos puntos muestrales para obtener información acerca de la población forestal.



El muestreo aleatorio de los bosques de MeadWestvaco le permite a satisfacer necesidades futuras de materia prima. © Walter Hodges/CORBIS.

En el proceso de acopio de datos de campo también participan guardabosques a través de toda la organización. De manera periódica, equipos de dos personas recolectan la información de cada árbol en todos los puntos muestrales. Los datos se ingresan en el sistema computacional de inventario forestal continuo (IFC) de la empresa. Los reportes obtenidos del sistema IFC contienen información de distribuciones de frecuencia con estadísticos sobre los tipos de árboles, volumen actual de los bosques, tasas de crecimiento anteriores y crecimiento y volumen proyectados. El muestreo y los correspondientes resúmenes estadísticos de los datos muestrales proporcionan la información esencial para la adecuada administración de los bosques y selvas de MeadWestvaco.

En este capítulo se estudia el muestreo aleatorio simple y el proceso de selección de muestras. Se verá también el uso de estadísticos como la media muestral y la proporción muestral para estimar la media y la proporción de la población. También se presenta el importante concepto de *distribución de muestreo*.

* Los autores agradecen al Dr. Edward P. Winkofsky por proporcionar este artículo para *Estadística en la práctica*.

En el capítulo 1 se presentaron las siguientes definiciones de los términos *elemento*, *población* y *muestra*.

- Un *elemento* es la entrada en la que se recolectan los datos.
- Una *población* es el conjunto de todos los elementos de interés.
- Una *muestra* es un subconjunto de la población.

La razón por la que se selecciona una muestra estriba en recabar datos para realizar una inferencia y responder una pregunta de investigación acerca de una población.

Para empezar, se presentan dos ejemplos en los que se utiliza el muestreo para responder una pregunta de investigación acerca de una población.

1. Los miembros de un partido político en Texas consideraban postular a un determinado candidato para el Senado, y los dirigentes del partido querían estimar la proporción de votantes registrados en el estado que podían apoyarlo. Por tanto, se seleccionó una muestra de 400 votantes registrados en Texas, y 160 de este total indicaron estar a favor del candidato. Así, una estimación de la proporción de la población de votantes registrados a favor del candidato es 160/400 = 0.40.
2. Un fabricante de llantas está considerando producir un nuevo modelo que ofrezca mayor duración que los actuales neumáticos de línea de la empresa. Para estimar la duración media, en millas, el fabricante selecciona una muestra de 120 neumáticos nuevos para probarlos. De los resultados de esta prueba se obtiene una media muestral de 36 500 millas. Por tanto, una estimación de la vida útil media de la población de nuevas llantas es 36 500 millas.

Una media muestral proporciona una estimación de la media poblacional, y una proporción muestral suministra una estimación de la proporción poblacional. En ambos casos puede esperarse un cierto error de estimación. Este capítulo enseña las bases para determinar cuán grande puede ser ese error.

Es importante observar que los resultados muestrales sólo proporcionan una *estimación* de los valores de las características de la población. No se espera que exactamente 0.40, o 40%, de la población de los votantes registrados esté a favor del candidato, ni que la media muestral de 36 500 millas sea exactamente igual al millaje medio de la población de todos los nuevos neumáticos. La razón es simple: la muestra sólo contiene una parte de la población. Es de esperarse algún error de muestro. Con métodos adecuados, los resultados muestrales proporcionarán "buenas" estimaciones de los parámetros poblacionales. Pero ¿cuán buenos puede esperarse que sean estos resultados? Por fortuna, existen procedimientos estadísticos para responder esta pregunta.

Definamos algunos términos que se utilizan en el muestreo. La **población muestreada** es aquella de la cual se extrae la muestra, y un **marco** es la lista de los elementos de donde se seleccionará la muestra. En el primer ejemplo, la población muestreada son todos los votantes registrados en Texas, y el marco es una lista de todos los votantes registrados. Debido a que éstos constituyen un número finito, el primer ejemplo ilustra qué es un muestreo de una población finita. En la sección 7.2 se analiza cómo seleccionar una muestra aleatoria simple cuando se muestrea una población finita.

Definir la población muestreada del ejemplo del millaje de los neumáticos es más difícil, porque la muestra de 120 llantas se obtuvo de un proceso productivo en un punto particular en el tiempo. Podemos pensar la población muestreada como la población conceptual de todos los neumáticos que pueden ser fabricados en el proceso de producción en un punto particular en el tiempo. En este sentido, la población muestreada se considera infinita, siendo imposible construir un marco del cual trazar la muestra. En la sección 7.2 se analiza cómo seleccionar una muestra aleatoria simple en una situación como ésta.

En este capítulo mostramos cómo emplear el muestreo aleatorio simple para seleccionar una muestra de una población finita y cómo puede tomarse una muestra aleatoria de una población infinita generada por un proceso en marcha. Después se analiza cómo usar una muestra aleatoria simple para calcular estimaciones de una media poblacional, una desviación estándar poblacional y una proporción poblacional. También se introduce el importante concepto de *distribución de muestreo o distribución muestral*. Como se verá, el conocimiento de la distribución de muestreo adecuada permite establecer qué tan cerca se encuentran las estimaciones muestrales de los correspondientes parámetros poblacionales. En la última sección se estudian alternativas al muestreo aleatorio simple, empleadas con frecuencia en la práctica.

7.1 El problema de muestreo de Electronics Associates

Al director de personal de Electronics Associates, Inc. (EAI) se le ha encargado elaborar un perfil de los 2 500 gerentes de la empresa. Las características a determinar son su sueldo medio anual y la proporción de ellos que ha completado el programa de capacitación de la empresa.

WEB archivo
EAI

Utilizando los 2 500 gerentes de la empresa como población para este estudio, es posible determinar el sueldo anual y la situación respecto del programa de capacitación de cada sujeto al consultar los archivos del personal. El conjunto de datos que contiene esta información para cada uno de los 2 500 gerentes que forman la población se encuentra en el archivo denominado EAI.

Con los datos de EAI y las fórmulas presentadas en el capítulo 3, se calcula la media poblacional y la desviación estándar poblacional de los sueldos anuales.

$$\text{Media poblacional} \quad \mu = \$51\,800$$
$$\text{Desviación estándar poblacional} \quad \sigma = \$4\,000$$

Con frecuencia los costos de recolectar información de una muestra son significativamente menores que si se acopian de una población, en especial cuando se deben realizar entrevistas personales para recabar la información.

Los datos sobre la situación de la capacitación indican que 1 500 de los 2 500 gerentes han completado el programa respectivo.

A las características numéricas de una población, como la media y la desviación estándar, se les llama **parámetros**. Si p denota la proporción de la población que ha completado el programa de capacitación, se tiene que $p = 1\,500/2\,500 = 0.60$. La media poblacional de los sueldos anuales ($\mu = \$51\,800$), la desviación estándar poblacional de los sueldos anuales ($\sigma = \$4\,000$) y la proporción poblacional de quienes han completado el programa de capacitación ($p = 0.60$) son parámetros de la población de gerentes de EAI.

Ahora suponga que la información necesaria acerca de todos los gerentes de EAI no esté disponible en la base de datos de la empresa. La pregunta que se considera ahora es: ¿cómo el director de personal de la empresa puede obtener estimaciones de los parámetros poblacionales utilizando una muestra de los gerentes, en lugar de estudiar a los 2 500 sujetos de la población? Asuma que se empleará una muestra de 30 gerentes. Es obvio que el tiempo y el costo de la elaboración de un perfil será mucho menor usando 30 sujetos que la población entera. Si el director de personal tuviera la certeza de que una muestra de 30 gerentes proporciona la información adecuada acerca de la población de 2 500, preferiría trabajar con una muestra que hacerlo con toda la población. Para explorar la posibilidad de usar una muestra en el estudio de EAI, primero se considerará cómo determinar la de 30 gerentes.

7.2 Selección de una muestra

En esta sección se describe cómo seleccionar una muestra. Primero se estudiará cómo seleccionarla de una población finita y luego de una población infinita.

Muestreo de una población finita

Los profesionales de la estadística recomiendan seleccionar una muestra de probabilidad cuando se muestree de una población finita, debido a que permite hacer inferencias estadísticas válidas acerca de la población. El tipo de muestra de probabilidad más simple es uno en el cual cada muestra de tamaño n tiene la misma probabilidad de ser seleccionada. Esto se llama *muestreo aleatorio simple*. Un muestreo aleatorio simple de tamaño n de una población finita de tamaño N se define como sigue.

En la sección 7.8 se describen otros métodos de muestreo de probabilidad.

MUESTREO ALEATORIO SIMPLE (POBLACION FINITA)

Una **muestra aleatoria simple** de tamaño n de una población finita de tamaño N es una muestra seleccionada de manera que cada posible muestra de tamaño n tenga la misma probabilidad de ser seleccionada.

Los números aleatorios generados por computadora también sirven para realizar el proceso de selección de una muestra aleatoria. Excel proporciona una función para generar números aleatorios en sus hojas de cálculo.

Un procedimiento para seleccionar una muestra aleatoria simple de una población finita es elegir los elementos para la muestra de uno en uno, de manera que, en cada paso, cada uno de los elementos que quedan en la población tenga la misma probabilidad de ser seleccionado. Al elegir n elementos de esta manera, será satisfecha la definición de muestra aleatoria simple seleccionada de una población finita.

Para elegir una muestra aleatoria simple de la población finita de gerentes de EAI, primero se le asigna un número a cada sujeto; por ejemplo, los números del 1 al 2 500 en el orden en

TABLA 7.1 Números aleatorios

63 271	59 986	71 744	51 102	15 141	80 714	58 683	93 108	13 554	79 945
88 547	09 896	95 436	79 115	08 303	01 041	20 030	63 754	08 459	28 364
55 957	57 243	83 865	09 911	19 761	66 535	40 102	26 646	60 147	15 702
46 276	87 453	44 790	67 122	45 573	84 358	21 625	16 999	13 385	22 782
55 363	07 449	34 835	15 290	76 616	67 191	12 777	21 861	68 689	03 263
69 393	92 785	49 902	58 447	42 048	30 378	87 618	26 933	40 640	16 281
13 186	29 431	88 190	04 588	38 733	81 290	89 541	70 290	40 113	08 243
17 726	28 652	56 836	78 351	47 327	18 518	92 222	55 201	27 340	10 493
36 520	64 465	05 550	30 157	82 242	29 520	69 753	72 602	23 756	54 935
81 628	36 100	39 254	56 835	37 636	02 421	98 063	89 641	64 953	99 337
84 649	48 968	75 215	75 498	49 539	74 240	03 466	49 292	36 401	45 525
63 291	11 618	12 613	75 055	43 915	26 488	41 116	64 531	56 827	30 825
70 502	53 225	03 655	05 915	37 140	57 051	48 393	91 322	25 653	06 543
06 426	24 771	59 935	49 801	11 082	66 762	94 477	02 494	88 215	27 191
20 711	55 609	29 430	70 165	45 406	78 484	31 639	52 009	18 873	96 927
41 990	70 538	77 191	25 860	55 204	73 417	83 920	69 468	74 972	38 712
72 452	36 618	76 298	26 678	89 334	33 938	95 567	29 380	75 906	91 807
37 042	40 318	57 099	10 528	09 925	89 773	41 335	96 244	29 002	46 453
53 766	52 875	15 987	46 962	67 342	77 592	57 651	95 508	80 033	69 828
90 585	58 955	53 122	16 025	84 299	53 310	67 380	84 249	25 348	04 332
32 001	96 293	37 203	64 516	51 530	37 069	40 261	61 374	05 815	06 714
62 606	64 324	46 354	72 157	67 248	20 135	49 804	09 226	64 419	29 457
10 078	28 073	85 389	50 324	14 500	15 562	64 165	06 125	71 353	77 669
91 561	46 145	24 177	15 294	10 061	98 124	75 732	00 815	83 452	97 355
13 091	98 112	53 959	79 607	52 244	63 303	10 413	63 839	74 762	50 289

que aparecen sus nombres en el archivo de personal. A continuación se revisa la tabla de dígitos aleatorios que figuran en la tabla 7.1. Al consultar la primera fila, se advierte que cada dígito, 6, 3, 2, . . . , es un número aleatorio con la misma oportunidad de aparecer que cualquier otro. Como el número mayor en la lista de la población de gerentes de EAI, 2 500, tiene cuatro dígitos, se seleccionarán números de la tabla en conjuntos o grupos de cuatro dígitos. Aun cuando para la selección de números aleatorios se puede empezar en cualquier lugar de la tabla y avanzar sistemáticamente en una de las cuatro direcciones, aquí se utilizará la primera fila y se avanzará de izquierda a derecha. Los primeros siete números aleatorios de cuatro dígitos son

Los números aleatorios en la tabla aparecen en grupos de cinco para facilitar su lectura.

6 327 1 599 8 671 7 445 1 102 1 514 1 807

Como los números de la tabla son aleatorios, estas cifras de cuatro dígitos son todas igualmente posibles.

Ahora se pueden usar estos números aleatorios de cuatro dígitos para darle a cada uno de los gerentes que constituyen la población la misma oportunidad de ser incluido en la muestra aleatoria. El primer número, 6 327, es mayor que 2 500. No corresponde a ninguno de los gerentes numerados que forman la población y, por tanto, se descarta. El segundo número, 1 599, está entre 1 y 2 500. Por tanto, el primer gerente seleccionado para la muestra aleatoria es el que tiene el número 1 599 en la lista de EAI. Siguiendo este proceso, se ignoran los números 8 671 y 7 445 antes de identificar a los gerentes con los números 1 102, 1 514 y 1 807 e incluirlos en la muestra. Este proceso continúa hasta que se tiene la muestra aleatoria de 30 gerentes de EAI.

Al realizar este proceso para la selección de una muestra aleatoria simple, es posible que un número que ya haya sido usado se encuentre de nuevo en la tabla antes de completar la muestra de los 30 gerentes. Como no se quiere seleccionar a un sujeto más de una vez, cualquier número aleatorio que ya ha sido usado se ignora, porque el gerente correspondiente ya se ha incluido en la muestra. A este tipo de selección se le conoce como **muestreo sin remplazo**.

Cuando se selecciona una muestra en la que se aceptan números aleatorios ya usados y los gerentes correspondientes son incluidos dos o más veces, se realiza un **muestreo con remplazo**. Muestrear con remplazo es una forma válida de identificar una muestra aleatoria simple; sin embargo, como es el procedimiento de muestreo más usado, cuando se hable de muestreo aleatorio simple se asumirá que éste es sin reemplazo.

Muestreo de una población infinita

Algunas veces se quiere seleccionar una muestra de una población, pero ésta es infinitamente grande o sus elementos están siendo generados por un proceso en marcha, por lo cual no hay límite para el número de elementos que pueden ser generados. Por tanto, no es posible hacer una lista de todos los elementos de la población. Esto se considera el caso de una población infinita, con la cual no se puede seleccionar una muestra aleatoria simple debido a que no es factible construir un marco constituido por todos los elementos. En el caso de una población infinita, los profesionales de la estadística recomiendan seleccionar lo que se llama una *muestra aleatoria*.

MUESTRA ALEATORIA (POBLACIÓN INFINITA)

Una **muestra aleatoria** de tamaño n de una población infinita es seleccionada de manera tal que se satisfagan las condiciones siguientes.

1. Cada elemento elegido proviene de la misma población.
2. Cada elemento es seleccionado de manera independiente.

La implementación del proceso de selección de una muestra aleatoria en una población infinita se debe efectuar con cuidado y criterio. Cada caso puede requerir un procedimiento de selección diferente. Considere dos ejemplos para ver qué significan las condiciones 1) cada elemento seleccionado proviene de la misma población, y 2) cada elemento se elige de manera independiente.

Una aplicación de control de calidad común involucra un proceso de producción donde no hay un límite en el número de elementos generados. La población conceptual que se muestrea son todos los elementos que se pueden producir (no sólo los que se producen) por el proceso de manufactura. Debido a que no es posible hacer una lista de todos ellos, se considera que la población es infinita. Para ser más precisos, considere una línea de producción diseñada para llenar cajas de un cereal para desayunar con un peso medio de 24 onzas por caja. De manera periódica, un inspector de control de calidad selecciona muestras de 12 cajas llenas con este proceso para determinar si éste funciona de manera apropiada o si, tal vez, un mal funcionamiento mecánico ha ocasionado que el proceso llene de forma insuficiente o excesiva los contenedores.

Con una operación productiva como ésta, la mayor preocupación en seleccionar una muestra aleatoria es asegurar que se satisfaga la condición 1 (los elementos de la muestra son seleccionados de la misma población). Para asegurar que se satisfaga esa condición, se deben elegir las cajas aproximadamente en el mismo punto en el tiempo. De esta manera el inspector evita la posibilidad de tomar algunas cajas cuando el proceso está funcionando de forma apropiada y otras cuando no funciona adecuadamente y las está llenando de manera insuficiente o excesiva. Con un proceso de producción como éste se satisface la segunda condición (cada elemento se selecciona en forma independiente), al haber diseñado un proceso en el que cada caja se llena individualmente. Con este supuesto, el inspector de control de calidad sólo necesita preocuparse por satisfacer la condición de que sean de la misma población.

En otro ejemplo de selección de una muestra aleatoria de una población infinita, piense en la población de clientes que llegan a un restaurante de comida rápida. Suponga que se le pide a un empleado que seleccione y entreviste una muestra para elaborar un perfil de los consumidores que visitan el restaurante. El proceso de arribo de los clientes está en marcha y no hay forma de obtener una lista de todos los consumidores de la población. Para fines prácticos, la población de este proceso en marcha se considera infinita. Se obtendrá una muestra aleatoria,

en la medida en que se diseñe un procedimiento de muestreo en el que todos los elementos de la muestra son clientes del restaurante y son seleccionados de manera independiente. En este caso, el empleado que obtiene la muestra necesita seleccionarla de las personas que llegan al establecimiento y realizan un consumo para asegurar que se satisfaga la condición de que sean de la misma población. Si, por ejemplo, elige alguna persona que llegó al restaurante sólo para entrar al sanitario, podría no ser un consumidor y se violaría la condición de que sean de la misma población. Así, en la medida en que el encuestador extrae la muestra de entre las personas que realizan un consumo en el restaurante, se satisface la condición 1. Asegurarse de que los consumidores son seleccionados en forma independiente puede ser más difícil.

El propósito de la segunda condición del procedimiento de selección de una muestra aleatoria (cada elemento se elige de manera independiente) consiste en prevenir el sesgo en la selección, que en este caso podría ocurrir si el encuestador fuera libre de tomar clientes para la muestra de manera arbitraria. Podría ser que éste se sintiera más cómodo seleccionando clientes de un grupo de edad particular y evitaría los de otros grupos de edad. El sesgo podría ocurrir también si eligiera un grupo de cinco comensales que entraron juntos al restaurante y le pidiera a todos participar en la muestra. Un grupo de clientes podría muy bien mostrar características similares, lo que generaría información engañosa acerca de la población. Se puede evitar un sesgo en una selección como ésta asegurando que la elección de un cliente específico no influya en la de cualquier otro. Esto es, los elementos (clientes) son escogidos de manera independiente.

McDonald's, el restaurante líder en comida rápida, realizó un muestreo aleatorio simple precisamente en una situación así. El procedimiento de muestreo se basó en el hecho de que algunos clientes presentaban cupones de descuento. Cada vez que una persona presentaba un cupón, a la siguiente que se atendía se le pedía que llenara un cuestionario sobre el perfil del cliente. Como quienes llegaban al restaurante presentaban cupones de descuento aleatoria e independientemente, este plan de muestreo garantizaba que los clientes fueran seleccionados de manera independiente. Por consiguiente, los dos requerimientos para un muestreo aleatorio simple de una población infinita fueron satisfechos.

Las poblaciones infinitas suelen asociarse con un proceso que opera continuamente a lo largo del tiempo. Algunos ejemplos son partes fabricadas en una línea de producción, repetidas pruebas experimentales en un laboratorio, transacciones en un banco, llamadas que llegan a un centro de asesoría técnica y clientes que entran en una tienda minorista. En cada caso, la situación puede verse como un proceso que genera elementos provenientes de una población infinita. En la medida en que los elementos de la muestra sean seleccionados de la misma población y en forma independiente, se considera que se trata de una muestra aleatoria de una población infinita.

NOTAS Y COMENTARIOS

1. En esta sección se ha tenido sumo cuidado en definir dos tipos de muestras: la muestra aleatoria simple de una población finita y la muestra aleatoria de una población infinita. En el resto de la obra se hará referencia a ellas como *muestra aleatoria* o sólo *muestra*. No se hará distinción de que sea una muestra aleatoria "simple" a menos que sea necesario para el ejercicio o el análisis.
2. Los profesionales de la estadística especializados en encuestas por muestreo de poblaciones finitas utilizan métodos que proporcionan muestras de probabilidad, con las cuales cada posible muestra tiene una probabilidad conocida de selección y se utiliza un proceso aleatorio para elegir sus elementos. El muestreo aleatorio simple es uno de esos métodos. En la sección 7.8 se describen algunos otros métodos de muestreo probabilístico: muestreo aleatorio estratificado, muestreo por conglomerados y muestreo sistemático. Se utiliza el término "simple" en el muestreo aleatorio simple para aclarar que es el método que asegura que cada muestra de tamaño n tiene la misma probabilidad de ser seleccionada.
3. El número de muestras aleatorias simples distintas de tamaño n que pueden seleccionarse de una población finita de tamaño N es

$$\frac{N!}{n!(N-n)!}$$

En esta expresión, $N!$ y $n!$ son las fórmulas factoriales estudiadas en el capítulo 4. Al utilizar esta expresión con los datos del problema de EAI,

en el que $N = 2\,500$ y $n = 30$, se ve que se pueden obtener aproximadamente 2.75×10^{69} muestras aleatorias simples distintas de 30 gerentes de EAI.

4. Para tomar una muestra aleatoria puede emplearse software. En los apéndices del capítulo se explica cómo usar Minitab y Excel para seleccionar una muestra aleatoria simple de una población finita.

Ejercicios

Métodos

AUTOevaluación

1. Tome una población finita con cinco elementos A, B, C, D y E. Se pueden seleccionar 10 muestras aleatorias simples de tamaño 2.
   a) Liste las 10 muestras empezando con AB, AC y así en lo sucesivo.
   b) Utilizando el muestreo aleatorio simple, ¿cuál es la probabilidad para cada muestra de tamaño 2 de ser seleccionada?
   c) Asuma que el número aleatorio 1 corresponde a A, el número 2 corresponde a B y así en lo sucesivo. Liste la muestra aleatoria de tamaño 2 que será seleccionada al usar los números aleatorios 8 0 5 7 5 3 2.
2. Suponga que una población finita tiene 350 elementos. A partir de los últimos tres dígitos de cada uno de los siguientes números aleatorios de cinco dígitos (por ejemplo: 601, 022, 448, . . .), determine los primeros cuatro elementos que se seleccionarán para una muestra aleatoria simple.

   98 601 73 022 83 448 02 147 34 229 27 553 84 147 93 289 14 209

Aplicaciones

AUTOevaluación

3. *Fortune* publica datos sobre ventas, valor del activo, valor de mercado y utilidades por acción de las 500 corporaciones industriales más grandes de Estados Unidos (*Fortune* 500, 2006). Suponga que usted desea seleccionar una muestra aleatoria simple de 10 corporaciones de la lista *Fortune* 500. Use los tres últimos dígitos de la novena columna de la tabla 7.1, empezando con 554. Leyendo hacia abajo por esa columna, identifique los números de las 10 corporaciones que se tomarán para la muestra.
4. A continuación se presentan las 10 acciones más activas en la Bolsa de Nueva York del 6 de marzo de 2006 (*The Wall Street Journal* 7 de marzo de 2006).

AT&T	Lucent	Nortel	Qwest	BellSouth
Pfizer	Texas Instruments	General Electric	iShrMSJpn	LSI Logic

   Las autoridades bursátiles decidieron investigar las prácticas de negociación utilizando una muestra de tres de estas acciones.
   a) Comenzando con el primer dígito aleatorio de la sexta columna de la tabla 7.1, lea los números descendiendo por esa columna para seleccionar una muestra aleatoria simple de tres acciones para las autoridades.
   b) Con la información aportada en la nota y comentario 3, determine cuántas muestras aleatorias simples diferentes de tamaño 3 pueden seleccionarse de una lista de 10 acciones.
5. Una organización estudiantil está interesada en estimar la proporción de estudiantes que está a favor de cierta disposición de la escuela. Se cuenta con una lista de los nombres y direcciones de los 645 estudiantes inscritos en el presente trimestre. Tomando números aleatorios de tres dígitos de la décima fila de la tabla 7.1 y avanzando por esa fila de izquierda a derecha, determine los 10 primeros estudiantes que serán seleccionados utilizando un muestreo aleatorio simple. Los números aleatorios de tres dígitos empiezan con 816, 283 y 610.
6. El *County and City Data Book* de la Oficina del Censo de Estados Unidos cuenta con información de los 3 139 condados de Estados Unidos. Suponga que para un estudio nacional se recogerán datos de 30 condados seleccionados de forma aleatoria. De la última columna de la tabla 7.1 extraiga números aleatorios de cuatro dígitos para determinar las cifras correspondientes a los primeros cinco condados seleccionados para la muestra. Ignore los primeros dígitos y empiece con los números aleatorios de cuatro dígitos 9 945, 8 364, 5 702 y así sucesivamente.

7. Suponga que se toma una muestra aleatoria simple de 12 de los 372 médicos de una determinada ciudad. Los nombres de los miembros de una organización médica local están disponibles. De la tabla 7.1 use la octava columna de números aleatorios de cinco dígitos para determinar cuáles serán los 12 médicos para la muestra. Ignore los primeros dos dígitos de cada grupo de cinco. Este proceso empieza con el número 108 y continúa descendiendo por la columna de números aleatorios.

8. Las siguientes acciones conforman el promedio industrial Dow Jones (*Barron's*, 23 de marzo de 2009).

1. 3M	11. Disney	21. McDonald's
2. AT&T	12. DuPont	22. Merck
3. Alcoa	13. ExxonMobil	23. Microsoft
4. American Express	14. General Electric	24. J. P. Morgan
5. Bank of America	15. Hewlett-Packard	25. Pfizer
6. Boeing	16. Home Depot	26. Procter & Gamble
7. Caterpillar	17. IBM	27. Travelers
8. Chevron	18. Intel	28. United Technologies
9. Cisco Systems	19. Johnson & Johnson	29. Verizon
10. Coca-Cola	20. Kraft Foods	30. Wal-Mart

   Suponga que se quiere seleccionar una muestra de seis de esas empresas para realizar un estudio a profundidad de prácticas administrativas. Utilice los primeros dos dígitos de cada fila de la novena columna de la tabla 7.1 para seleccionar una muestra aleatoria de seis empresas.

9. *The Wall Street Journal* proporciona el valor del activo neto, el rendimiento porcentual en lo que va del año y el rendimiento porcentual en tres años de 555 fondos de inversión (*The Wall Street Journal,* 25 de abril de 2003). Suponga que se usará una muestra aleatoria simple de 12 de estos 555 fondos para un estudio acerca de su tamaño y desempeño. Utilice la cuarta columna de números aleatorios de la tabla 7.1 comenzando con 51102, para seleccionar la muestra aleatoria simple de 12 fondos de inversión. Empiece con el fondo 102 y use los *últimos* tres dígitos de cada fila de la cuarta columna para el proceso de selección. ¿Cuáles son los números de los 12 fondos de inversión en esta muestra aleatoria simple?

10. Indique cuáles de las siguientes situaciones involucran muestreo de una población finita y cuáles muestreo de una población infinita. En los casos en que la población muestreada sea finita, describa cómo construiría un marco.
    a) Obtener una muestra de los conductores con licencia en el estado de Nueva York.
    b) Determinar una muestra de las cajas de cereal producidas por Breakfast Choice Company.
    c) Extraer una muestra de automóviles cruzando el puente Golden Gate en un fin de semana normal.
    d) Definir una muestra de estudiantes en un curso de estadística en la Universidad de Indiana.
    e) Obtener una muestra de las órdenes que son procesadas por una empresa de pedidos por correo.

7.3 Estimación puntual

Una vez descrito cómo seleccionar una muestra aleatoria simple, se vuelve al problema de EAI. En la tabla 7.2 se presenta una muestra aleatoria simple de 30 gerentes con sus respectivos datos de sueldo anual y participación en el programa de capacitación. La notación x_1, x_2, etc., se usa para denotar el sueldo anual del primer gerente de la muestra, del segundo, y así sucesivamente. La participación en el programa de capacitación se indica por un Sí en la columna "programa de capacitación".

Para estimar el valor de un parámetro poblacional se calcula la característica correspondiente de la muestra, a lo que se le conoce como **estadístico muestral**. Por ejemplo, para estimar la media poblacional μ y la desviación estándar poblacional σ de los sueldos anuales de los gerentes de EAI, se emplean los datos de la tabla 7.2 y se calculan los estadísticos mues-

TABLA 7.2 Sueldo anual y situación respecto del programa de capacitación para una muestra aleatoria simple de 30 gerentes de EAI

Sueldo anual ($)	Programa de capacitación	Sueldo anual ($)	Programa de capacitación
$x_1 = 49\,094.30$	Sí	$x_{16} = 51\,766.00$	Sí
$x_2 = 53\,263.90$	Sí	$x_{17} = 52\,541.30$	No
$x_3 = 49\,643.50$	Sí	$x_{18} = 44\,980.00$	Sí
$x_4 = 49\,894.90$	Sí	$x_{19} = 51\,932.60$	Sí
$x_5 = 47\,621.60$	No	$x_{20} = 52\,973.00$	Sí
$x_6 = 55\,924.00$	Sí	$x_{21} = 45\,120.90$	Sí
$x_7 = 49\,092.30$	Sí	$x_{22} = 51\,753.00$	Sí
$x_8 = 51\,404.40$	Sí	$x_{23} = 54\,391.80$	No
$x_9 = 50\,957.70$	Sí	$x_{24} = 50\,164.20$	No
$x_{10} = 55\,109.70$	Sí	$x_{25} = 52\,973.60$	No
$x_{11} = 45\,922.60$	Sí	$x_{26} = 50\,241.30$	No
$x_{12} = 57\,268.40$	No	$x_{27} = 52\,793.90$	No
$x_{13} = 55\,688.80$	Sí	$x_{28} = 50\,979.40$	Sí
$x_{14} = 51\,564.70$	No	$x_{29} = 55\,860.90$	Sí
$x_{15} = 56\,188.20$	No	$x_{30} = 57\,309.10$	No

trales correspondientes: media muestral y desviación estándar muestral s. Con las fórmulas para ambas categorías, presentadas en el capítulo 3, se obtiene que la media muestral es

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{1\,554\,420}{30} = \$51\,814$$

y la desviación estándar muestral es

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{325\,009\,260}{29}} = \$3\,348$$

Para estimar p, la proporción de gerentes en la población que completaron el programa de capacitación, se usa la proporción muestral correspondiente $\bar{p}$. Sea x que denota el número de gerentes en la muestra que completaron el programa de capacitación. Según la tabla 7.2, $x = 19$. Por tanto, como el tamaño de la muestra es $n = 30$, la proporción muestral es

$$\bar{p} = \frac{x}{n} = \frac{19}{30} = 0.63$$

Al efectuar los cálculos anteriores, se lleva a cabo el proceso estadístico conocido como *estimación puntual.* A la media muestral $\bar{x}$ se le identifica como **estimador puntual** de la media poblacional μ, a la desviación estándar muestral s como el estimador puntual de la desviación estándar poblacional σ y a la proporción muestral $\bar{p}$ como el estimador puntual de la proporción poblacional p. Al valor numérico obtenido de $\bar{x}$, s o $\bar{p}$ se le conoce como **estimación puntual**. Así, en la muestra aleatoria simple de 30 gerentes de EAI que se presenta en la tabla 7.2, $51 814 es la estimación puntual de μ, $3 348 es la estimación puntual de σ y 0.63 es la estimación puntual de p. En la tabla 7.3 se resumen los resultados muestrales y se comparan las estimaciones puntuales con los valores de los parámetros poblacionales.

Como se observa en la tabla 7.3, las estimaciones puntuales difieren un poco de los correspondientes parámetros poblacionales. Estas diferencias son de esperarse, ya que para elaborar las estimaciones muestrales se usa una muestra, y no un censo de toda la población. En el capítulo siguiente se verá cómo elaborar un intervalo de estimación para tener información respecto de qué tan cerca está la estimación muestral del parámetro poblacional.

TABLA 7.3 Resumen de las estimaciones puntuales obtenidas de una muestra aleatoria simple de 30 gerentes de EAI

Parámetro poblacional	Valor del parámetro	Estimador puntual	Estimación puntual
μ = Media poblacional de los sueldos anuales	\$51 800	$\bar{x}$ = Media muestral de los sueldos anuales	\$51 814
σ = Desviación estándar poblacional de los sueldos anuales	\$4 000	s = Desviación estándar muestral de los sueldos anuales	\$3 348
p = Proporción poblacional que ha completado el programa de capacitación	0.60	$\bar{p}$ = Proporción muestral que ha completado el programa de capacitación	0.63

Consejo práctico

El tema de la mayor parte del resto de este libro se relaciona con la inferencia estadística. La estimación puntual es una de sus formas. Se utiliza un estadístico de muestra para hacer una inferencia acerca de un parámetro poblacional. Al realizar inferencias acerca de una población basada en una muestra, es importante tener una correspondencia cerrada entre la población muestreada y la población objetivo. La **población objetivo** es aquella de la cual buscamos hacer inferencias, en tanto que la población muestreada es aquella de la cual se toma realmente la muestra. En esta sección se describe el proceso de tomar una muestra aleatoria simple de la población de gerentes en EAI y establecer puntos estimados de características de la misma población. Así, la población muestreada y la población objetivo son idénticas, que es la situación deseada. En otros casos, sin embargo, no es fácil obtener una correspondencia cerrada entre ambos tipos de poblaciones.

Piense en el caso de un parque temático seleccionando una muestra de sus clientes para conocer algunas de sus características, como la edad y el tiempo que pasan en el parque. Suponga que todos los elementos de la muestra se seleccionan en un día en que la entrada al parque está restringida a los empleados de una gran empresa. Entonces la población muestreada estaría compuesta de los empleados de dicha empresa y los miembros de su familia. Si la población objetivo se busca para realizar inferencias acerca de los clientes usuales durante un verano común, se podría encontrar una diferencia significativa entre la población muestreada y la población objetivo. En tal caso, se podría cuestionar la validez de los puntos de estimación que se están realizando. La gerencia del parque estaría en mejor posición para saber si una muestra tomada en un día específico parecería ser representativa de la población objetivo.

En resumen, cada vez que se utiliza una muestra para hacer inferencias acerca de una población, debemos estar seguros de que el estudio está diseñado para que la población muestreada y la población objetivo estén en un acuerdo cerrado. El buen juicio es un ingrediente necesario en una práctica estadística sólida.

Ejercicios

Métodos

AUTOevaluación

11. Los datos siguientes provienen de una muestra aleatoria simple.

    5 8 10 7 10 14

    a) ¿Cuál es la estimación puntual de la media poblacional?
    b) ¿Cuál es la estimación puntual de la desviación estándar poblacional?

12. Como respuestas a una pregunta de encuesta a una muestra de 150 individuos se obtuvieron 75 Sí, 55 No y 20 sujetos que no dieron su opinión.
    a) ¿Cuál es la estimación puntual de la proporción en la población que responde Sí?
    b) ¿Cuál es la estimación puntual de la proporción en la población que responde No?

Aplicaciones

AUTO evaluación

13. La siguiente información son datos obtenidos en una muestra aleatoria de las ventas de cinco meses:

Mes	1	2	3	4	5
Unidades vendidas	94	100	85	94	92

a) Calcule una estimación puntual de la media poblacional del número medio de unidades vendidas por mes.
b) Calcule una estimación puntual de la desviación estándar poblacional.

WEB archivo
MutualFund

14. *BusinessWeek* publicó información sobre 283 fondos de inversión (*BusinessWeek*, 26 de enero de 2004). En el conjunto de datos MutualFund se encuentra una muestra de 40 de estos fondos. Use este conjunto de datos para efectuar lo que se solicita en los incisos siguientes.
a) Calcule una estimación puntual de la proporción de fondos de inversión de *BusinessWeek* que son fondos de cargo.
b) Desarrolle una estimación puntual de la proporción de fondos clasificados como de alto riesgo.
c) Calcule una estimación puntual de la proporción de fondos con una puntuación abajo del promedio para el riesgo.

15. Muchos medicamentos empleados en la cura del cáncer son costosos. *Business Week* dio a conocer los costos de los tratamientos con Herceptin, un medicamento suministrado para el cáncer de mama (*BusinessWeek*, 30 de enero de 2006). Los siguientes son los costos (en doláres) de tratamientos comunes con Herceptin en una muestra aleatoria simple de 10 pacientes.

4 376	5 578	2 717	4 920	4 495
4 798	6 446	4 119	4 237	3 814

a) Calcule una estimación puntual del costo medio de un tratamiento con Herceptin.
b) Desarrolle una estimación puntual de la desviación estándar para los costos de los tratamientos con Herceptin.

16. En una muestra de 50 empresas de la lista *Fortune* 500 (*Fortune*, 14 de abril de 2003), cinco se encontraban en Nueva York, seis en California, dos en Minnesota y una en Wisconsin.
a) Calcule una estimación de la proporción de empresas de *Fortune* 500 con sede en Nueva York.
b) Desarrolle una estimación del número de empresas de *Fortune* 500 ubicadas en Minnesota.
c) Calcule una estimación de la proporción de empresas de *Fortune* 500 que no se encuentran en ninguno de estos estados.

17. La American Association of Individuals Investors (AAII) realiza sondeos semanales entre sus suscriptores para determinar cuántos se muestran optimistas, pesimistas o indiferentes respecto del mercado de acciones a corto plazo. Sus hallazgos en la semana que terminó el 2 de marzo de 2006 son consistentes con los resultados muestrales siguientes (sitio web de AAII, 7 de marzo de 2006).

Optimistas 409 Indiferentes 299 Pesimistas 291

Proporcione una estimación puntual de los parámetros poblacionales siguientes.
a) Proporción de suscriptores de AAII que son optimistas respecto del mercado de acciones.
b) Proporción de suscriptores que son indiferentes al mercado de acciones.
c) Proporción de suscriptores que son pesimistas acerca del mercado accionario.

7.4 Introducción a las distribuciones muestrales o de muestreo

En la sección anterior se dijo que la media muestral $\bar{x}$ es el estimador puntual de la media poblacional μ, y que la proporción muestral $\bar{p}$ es el estimador puntual de la proporción poblacional p. En la muestra aleatoria simple de los 30 gerentes de EAI que se presenta en la tabla 7.2, la estimación puntual de μ es $\bar{x} = \$51\,814$ y la estimación puntual de p es $\bar{p} = 0.63$. Suponga que se selecciona otra muestra aleatoria simple de 30 gerentes de EAI y se obtienen las estimaciones puntuales siguientes:

Media muestral: $\bar{x} = \$52\,670$
Proporción muestral: $\bar{p} = 0.70$

TABLA 7.4 Valores de $\bar{x}$ y de $\bar{p}$ obtenidos en 500 muestras aleatorias simples de 30 gerentes de EAI

Muestra número	Media muestral ($\bar{x}$)	Proporción muestral ($\bar{p}$)
1	51 814	0.63
2	52 670	0.70
3	51 780	0.67
4	51 588	0.53
.	.	.
.	.	.
.	.	.
500	51 752	0.50

Observe que se obtuvieron valores diferentes de $\bar{x}$ y de $\bar{p}$. En efecto, una segunda muestra aleatoria simple de 30 gerentes de EAI no se puede esperar que proporcione las mismas estimaciones puntuales que la primera.

Ahora suponga que el proceso de seleccionar una muestra aleatoria simple de 30 gerentes de EAI se repite una y otra vez, y que en cada ocasión se calculan los valores de $\bar{x}$ y de $\bar{p}$. La tabla 7.4 presenta una parte de los resultados obtenidos en 500 muestras aleatorias simples y la tabla 7.5 registra las distribuciones de frecuencia y de frecuencia relativa de los valores $\bar{x}$ de las 500. En la figura 7.1 se muestra el histograma de las frecuencias de los valores de $\bar{x}$.

La habilidad para entender el material de los capítulos siguientes depende en gran medida de comprender y usar las distribuciones muestrales que se presentan en este capítulo.

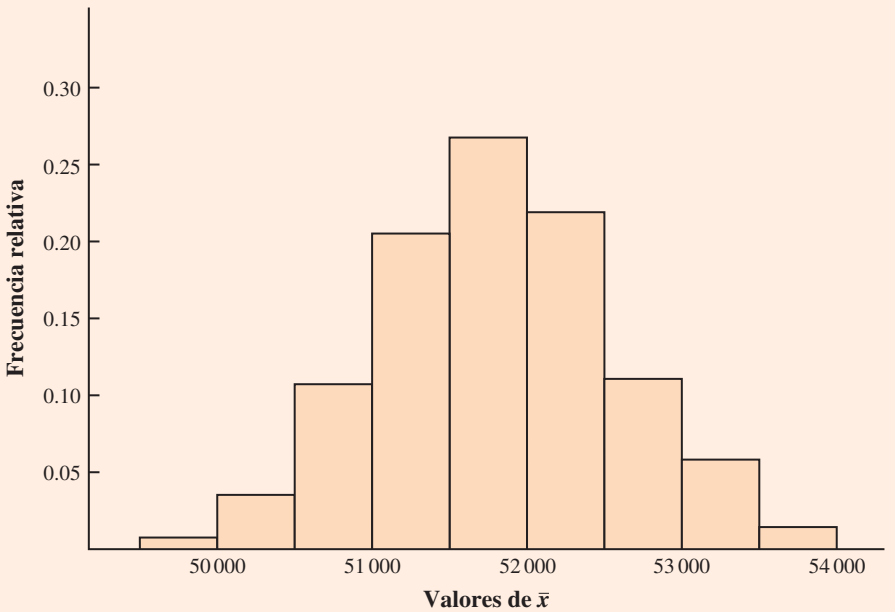
En el capítulo 5 se define una variable aleatoria como una descripción numérica del resultado de un experimento. Si el proceso de seleccionar una muestra aleatoria simple se considera un experimento, la media muestral $\bar{x}$ es la descripción numérica del resultado de ese experimento. Por tanto, la media muestral $\bar{x}$ es una variable aleatoria. Entonces, como ocurre con otras variables aleatorias, $\bar{x}$ tiene una media o valor esperado, una desviación estándar y una distribución de probabilidad. Como los distintos valores que toma $\bar{x}$ son resultado de distintas muestras aleatorias simples, a la distribución de probabilidad de $\bar{x}$ se le conoce como **distribución de muestreo** de $\bar{x}$. Conocer esta distribución y sus propiedades permitirá hacer declaraciones de probabilidad acerca de qué tan cerca está la media muestral $\bar{x}$ de la media poblacional μ.

Remítase a la figura 7.1. Se necesitaría enumerar todas las muestras posibles de 30 gerentes y calcular cada una de las medias muestrales para determinar totalmente la distribución de muestreo de $\bar{x}$. Sin embargo, el histograma de 500 valores de $\bar{x}$ provee una aproximación a esta distribución de muestreo. En esta aproximación se observa la apariencia de una curva de campana de esta distribución. Note además que la mayor concentración de valores de $\bar{x}$ y la

TABLA 7.5 Distribuciones de frecuencia y de frecuencia relativa de $\bar{x}$ en 500 muestras aleatorias simples de 30 gerentes de EAI

Sueldo anual medio ($)	Frecuencia	Frecuencia relativa
49 500.00–49 999.99	2	0.004
50 000.00–50 499.99	16	0.032
50 500.00–50 999.99	52	0.104
51 000.00–51 499.99	101	0.202
51 500.00–51 999.99	133	0.266
52 000.00–52 499.99	110	0.220
52 500.00–52 999.99	54	0.108
53 000.00–53 499.99	26	0.052
53 500.00–53 999.99	6	0.012
Totals	500	1.000

FIGURA 7.1 Histograma de la frecuencia relativa de los valores de $\bar{x}$ obtenidos en 500 muestras aleatorias simples de tamaño 30 cada una



media de los 500 valores de $\bar{x}$ se encuentran cerca de la media poblacional $\mu = \$51\,800$. En la sección siguiente se describirán más detalladamente las propiedades de la distribución de muestreo de $\bar{x}$.

Los 500 valores de la proporción muestral de $\bar{p}$ se resumen en el histograma de frecuencia relativa de la figura 7.2. Como ocurre con $\bar{x}$, $\bar{p}$ es una variable aleatoria. Si se tomara cada muestra posible de tamaño 30 y para cada una se calculara el valor de $\bar{p}$, la distribución de probabilidad que se obtuviera sería la distribución de muestreo de $\bar{p}$. En la figura 7.2, el histograma de frecuencia relativa de los 500 valores muestrales proporciona una idea general de la apariencia de la distribución de muestreo de $\bar{p}$.

En la práctica sólo se selecciona una muestra aleatoria simple de la población. En esta sección el proceso de muestreo se repitió 500 veces para ilustrar que es posible tomar muchas muestras diferentes y que distintas muestras darán valores diversos de los estadísticos muestrales $\bar{x}$ y $\bar{p}$. A la distribución de muestreo de cualquier estadístico determinado se le llama *distribución de muestreo del estadístico*. En la sección 7.5 se presentan las características de la distribución de muestreo de $\bar{x}$. En la sección 7.6 se describen las características de la distribución de muestreo de $\bar{p}$.

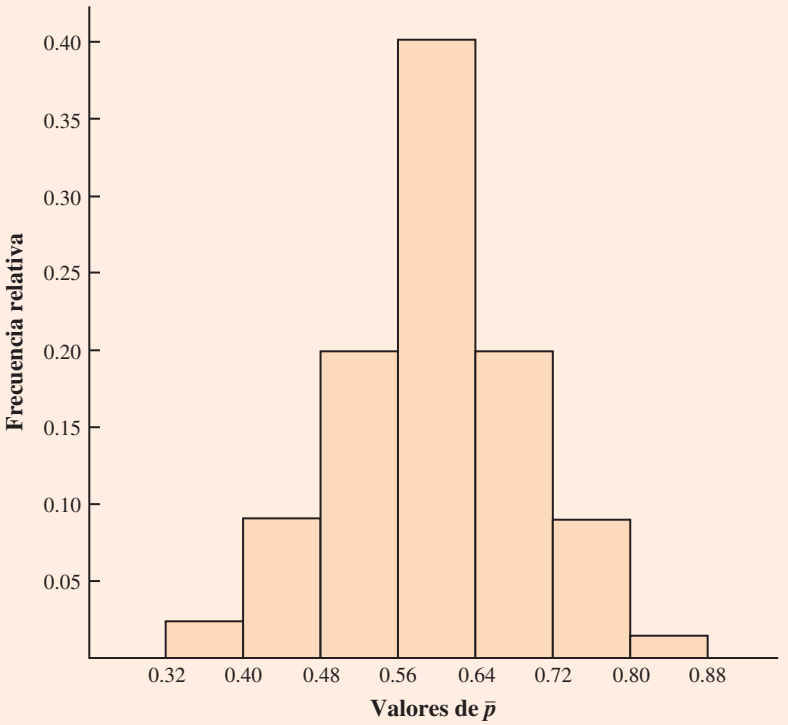
7.5 Distribución de muestreo de $\bar{x}$

En la sección anterior se dijo que la media muestral $\bar{x}$ es una variable aleatoria y que a su distribución de probabilidad se le llama *distribución de muestreo de* $\bar{x}$.

DISTRIBUCIÓN DE MUESTREO DE $\bar{x}$

La distribución muestral de $\bar{x}$ es la distribución de probabilidad de todos los posibles valores de la media muestral $\bar{x}$.

FIGURA 7.2 Histograma de la frecuencia relativa de los valores de $\bar{p}$ obtenidos en 500 muestras aleatorias simples de tamaño 30 cada una



En esta sección se describen las propiedades de la distribución de muestreo de $\bar{x}$. Como ocurre con otras distribuciones de probabilidad estudiadas, la distribución de muestreo de $\bar{x}$ tiene un valor esperado o media, una desviación estándar y una forma característica. Para empezar, se considerará la media de todos los valores posibles de $\bar{x}$, a la que se conoce como *valor esperado de* $\bar{x}$.

Valor esperado de $\bar{x}$

En el problema de muestreo de EAI se vio que en distintas muestras aleatorias simples se obtienen valores diferentes para la media muestral $\bar{x}$. Como la variable aleatoria $\bar{x}$ puede tener muchos valores diversos, suele ser de interés conocer la media de todos los valores de $\bar{x}$ que se obtienen con diferentes muestras aleatorias simples. La media de la variable aleatoria $\bar{x}$ es el valor esperado de $\bar{x}$; sea éste $E(\bar{x})$ y μ la media de la población de la que se selecciona una muestra aleatoria simple. Se puede demostrar que cuando se emplea el muestreo aleatorio simple, $E(\bar{x})$ y μ son iguales.

El valor esperado de $\bar{x}$ es igual a la media de la población de la cual se seleccionó la muestra.

VALOR ESPERADO DE $\bar{x}$

$$E(\bar{x}) = \mu \tag{7.1}$$

donde:

$E(\bar{x})$ = valor esperado de $\bar{x}$
μ = media poblacional

Este resultado enseña que utilizando el muestreo aleatorio simple, el valor esperado o media de la distribución de muestreo de $\bar{x}$ es igual a la media de la población. En la sección 7.1 se vio que el sueldo anual medio de los gerentes de EAI es $\mu = \$51\,800$. Por tanto, con base en la ecuación (7.1), la media de todas las medias muestrales posibles en el estudio de EAI es también $51 800.

Cuando el valor esperado de un estimador puntual es igual al parámetro poblacional, se dice que el estimador puntual es **insesgado**. Por tanto, la ecuación (7.1) indica que $\bar{x}$ es un estimador insesgado de la media poblacional μ.

Desviación estándar de $\bar{x}$

Ahora se definirá la desviación estándar de la distribución de muestreo de $\bar{x}$. Se empleará la notación siguiente.

$\sigma_{\bar{x}}$ = desviación estándar de $\bar{x}$
σ = desviación estándar de la población
n = tamaño de la muestra
N = tamaño de la población

Es posible demostrar que la fórmula de la desviación estándar de $\bar{x}$ depende de que la población sea finita o infinita. Las dos fórmulas para la desviación estándar de $\bar{x}$ son las siguientes.

DESVIACIÓN ESTÁNDAR DE $\bar{x}$

Población finita

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Población infinita

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \textbf{(7.2)}$$

Al comparar las dos fórmulas en (7.2) se ve que el factor $\sqrt{(N-n)/(N-1)}$ se requiere cuando la población es finita, pero no cuando es infinita. A este factor se le conoce como **factor de corrección para una población finita**. En muchas situaciones prácticas de muestreo se encuentra que, aunque la población sea finita, es "grande", mientras que el tamaño de la muestra es "pequeño". En estos casos el factor de corrección para una población finita $\sqrt{(N-n)/(N-1)}$ es cercano a 1. Por tanto, la diferencia entre el valor de la desviación estándar de $\bar{x}$ para las poblaciones finitas e infinitas se vuelve despreciable. Entonces $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ se convierte en una buena aproximación a la desviación estándar de $\bar{x}$ aun cuando la población sea finita. Esta observación lleva al siguiente lineamiento, o regla general, para calcular la desviación estándar de $\bar{x}$.

USAR LA EXPRESIÓN SIGUIENTE PARA CALCULAR LA DESVIACIÓN ESTÁNDAR DE $\bar{x}$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \textbf{(7.3)}$$

siempre que

1. La población sea infinita; o
2. La población sea finita y el tamaño de la muestra sea menor o igual a 5% del tamaño de la población; es decir, $n/N \leq 0.05$.

El problema 21 muestra que cuando $n/N \leq 0.05$, el factor de corrección para una población finita tiene poco efecto en el valor de $\sigma_{\bar{x}}$.

El término error estándar *se utiliza en la inferencia estadística para referirse a la desviación estándar de un estimador puntual.*

En los casos en que $n/N > 0.05$, para calcular $\sigma_{\bar{x}}$ debe usarse la versión para poblaciones finitas de la fórmula (7.2). En este libro, a menos que se indique otra cosa, se supondrá que el tamaño de la población es "grande", $n/N \leq 0.05$, y se utilizará la expresión (7.3) para calcular $\sigma_{\bar{x}}$.

Para calcular $\sigma_{\bar{x}}$ se necesita conocer σ, la desviación estándar de la población. Para subrayar, aún más, la diferencia entre $\sigma_{\bar{x}}$ y σ, a la desviación estándar de $\bar{x}$, $\sigma_{\bar{x}}$, se le llama **error estándar** de la media. En general, el término *error estándar* se refiere a la desviación estándar de un estimador puntual. Más adelante se verá que el valor del error estándar de la media ayuda a determinar qué tan lejos puede estar la media muestral de la media poblacional. Ahora, de nuevo con el ejemplo de EAI, se calcula el error estándar de la media correspondiente a las muestras aleatorias simples de 30 gerentes de EAI.

En la sección 7.1 vimos que la desviación estándar de los sueldos anuales en la población de los 2 500 gerentes de EAI era $\sigma = 4\,000$. En este caso la población es finita, $N = 2\,500$. Sin embargo, como el tamaño de la muestra es 30, se tiene $n/N = 30/2\,500 = 0.012$. Dado que el tamaño de la muestra es menor que 5% del tamaño de la población, se puede ignorar el factor de corrección para una población finita y usar la ecuación (7.3) para calcular el error estándar.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4\,000}{\sqrt{30}} = 730.3$$

Forma de la distribución de muestreo de $\bar{x}$

Los resultados anteriores respecto del valor esperado y la desviación estándar en la distribución de muestreo de $\bar{x}$ son aplicables a cualquier población. El paso final para identificar las características de la distribución de muestreo de $\bar{x}$ consiste en determinar la forma de la distribución de muestreo. Se considerarán dos casos: 1) La población tiene distribución normal, y 2) La población no tiene distribución normal.

La población tiene distribución normal. En muchas situaciones es razonable suponer que la población de la que se selecciona la muestra aleatoria simple tiene distribución normal o casi normal. Cuando esto ocurre, la distribución de muestreo de $\bar{x}$ está distribuida normalmente cualquiera que sea el tamaño de la muestra.

La población no tiene distribución normal. Cuando la población de la que se tomó la muestra aleatoria simple no tiene distribución normal, el **teorema del límite central** ayuda a determinar la forma de la distribución de muestreo de $\bar{x}$. El enunciado de este teorema aplicado a la distribución de muestreo de $\bar{x}$ dice lo siguiente.

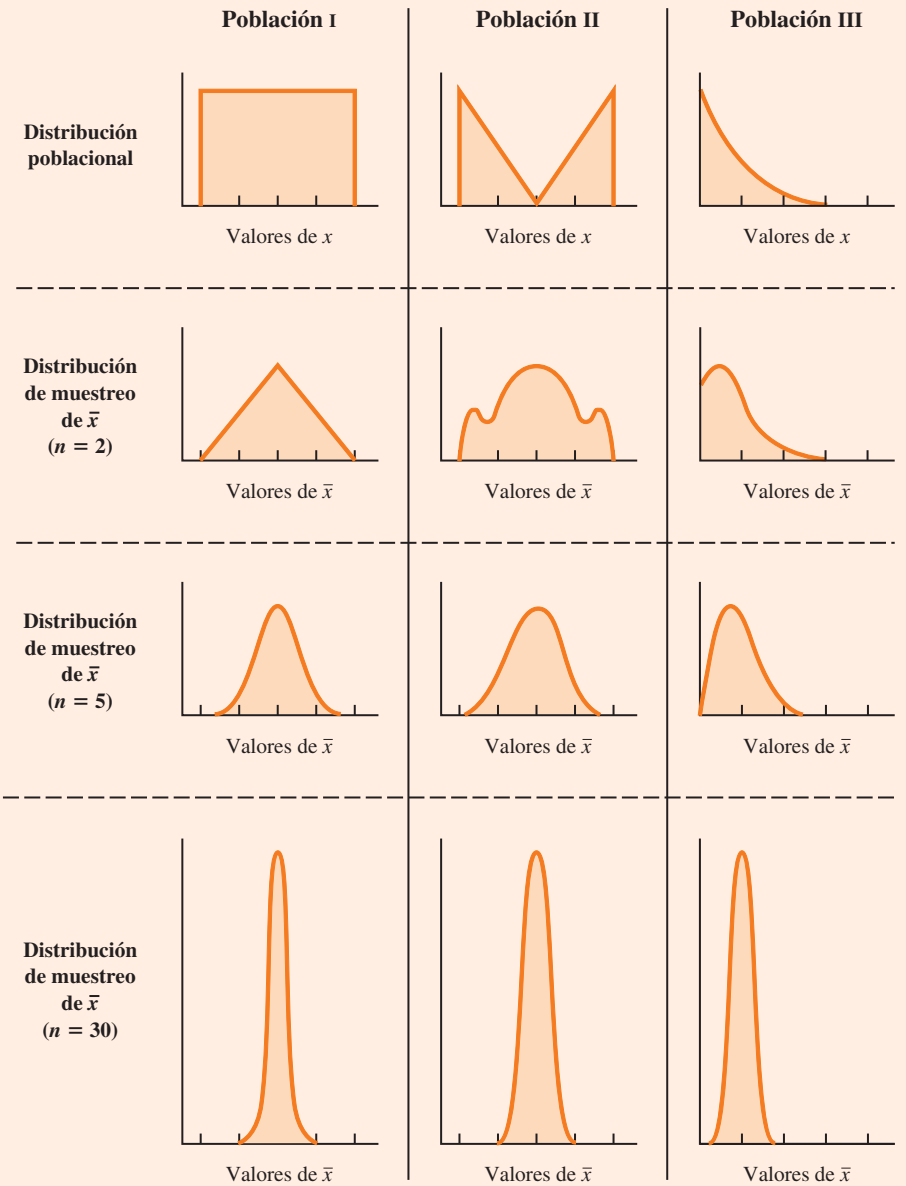
TEOREMA DEL LÍMITE CENTRAL

Cuando se seleccionan muestras aleatorias simples de tamaño n de una población, la distribución de muestreo de la media muestral $\bar{x}$ puede aproximarse mediante una *distribución normal* a medida que el tamaño de la muestra se hace grande.

En la figura 7.3 se ilustra cómo funciona el teorema del límite central en tres poblaciones diferentes; cada columna se refiere a una de ellas. En el panel superior de la figura se aprecia que ninguna de las tres poblaciones está distribuida normalmente. La población I tiene una distribución uniforme, y a la II se le conoce como *distribución de orejas de conejo*. Esta distribución es simétrica, pero los valores más probables se encuentran en las colas de la distribución. La forma de la población III se parece a una distribución exponencial y es sesgada a la derecha.

En los tres paneles superiores de la figura 7.3 se presentan las formas de las distribuciones de muestreo de tamaños $n = 2$, $n = 5$ y $n = 30$. Cuando el tamaño es 2, se observa que cada distribución de muestreo tiene una forma diferente a la distribución poblacional correspondiente.

FIGURA 7.3 Ilustración del teorema central del límite con tres poblaciones



Con el tamaño 5 vemos que las formas de las distribuciones de muestreo en los casos de las poblaciones I y II empiezan a parecerse a la forma de una distribución normal. En el caso de la población III, aun cuando la forma de la distribución de muestreo comienza a semejarse a una distribución normal, se observa todavía cierto sesgo a la derecha. Por último, para el tamaño 30, la forma de cada una de las tres distribuciones de muestreo es aproximadamente normal.

Desde un punto de vista práctico, con frecuencia se querrá saber qué tan grande debe ser el tamaño de la muestra antes de aplicar el teorema del límite central y suponer que la forma de la distribución de muestreo es aproximadamente normal. En las investigaciones estadísticas se ha estudiado este problema en distribuciones de muestreo de $\bar{x}$ de diversas poblaciones y tamaños de muestra. En la práctica estadística general se asume que, en la mayoría de las aplicaciones, la distribución de muestreo de $\bar{x}$ se puede aproximar mediante una distribución normal siempre que la muestra sea de tamaño 30 o mayor. En los casos en que la población es muy sesgada o existen

observaciones atípicas, pueden necesitarse muestras de tamaño 50. Por último, si la población es discreta, el tamaño de muestra necesario para la aproximación normal suele depender de la proporción poblacional. Se profundizará más en este tema cuando se estudie la distribución de muestreo de $\bar{p}$ en la sección 7.6.

Distribución de muestreo de $\bar{x}$ en el problema de EAI

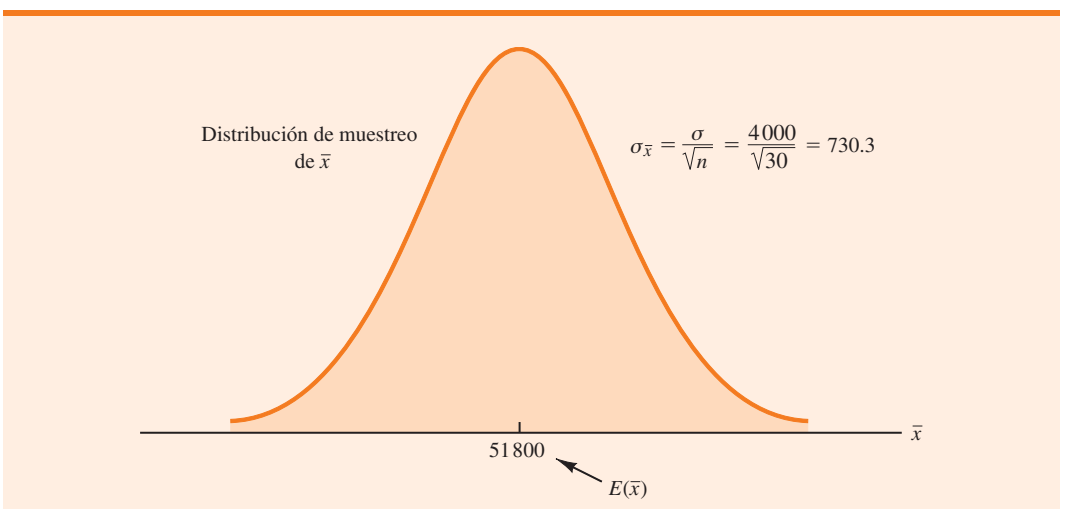
En el problema de EAI, para el que ya previamente se mostró que $E(\bar{x}) = \$51\,800$ y $\sigma_{\bar{x}} = 730.3$, no se cuenta con ninguna información acerca de la distribución de la población, que puede estar o no distribuida normalmente. Si se da el segundo caso, la distribución muestral de $\bar{x}$ estará distribuida normalmente. Si la población no tiene una distribución normal, la muestra aleatoria simple de 30 gerentes y el teorema del límite central permiten concluir que la distribución de muestreo de $\bar{x}$ puede aproximarse mediante una distribución normal. En cualquiera de los casos, se concluye que la distribución de muestreo de $\bar{x}$ se describe mediante una distribución normal como la que se muestra en la figura 7.4.

Valor práctico de la distribución de muestreo de $\bar{x}$

Siempre que se seleccione una muestra aleatoria simple y se use el valor de la media muestral para estimar el valor de la media poblacional μ, no se podrá esperar que la media muestral sea exactamente igual a la media poblacional. La razón práctica por la que interesa la distribución de muestreo de $\bar{x}$ estriba en que se puede usar para proporcionar información probabilística acerca de la diferencia entre la media muestral y la media poblacional. Para demostrar este uso, se retomará el problema de EAI.

Suponga que el director de personal cree que la media muestral será una estimación aceptable de la media poblacional si la primera está en un margen de $500 de la segunda. Sin embargo, no es posible garantizar que la media muestral esté en un margen de $500 de la media poblacional. En efecto, en la tabla 7.5 y en la figura 7.1 se observa que algunas de las 500 medias muestrales difieren en más de $2 000 de la media poblacional. Entonces hay que pensar en el requerimiento del director de personal en términos de probabilidad. Es decir, a éste le interesa la interrogante siguiente: ¿cuál es la probabilidad de que la media muestral obtenida usando una muestra aleatoria simple de 30 gerentes de EAI se encuentre en un margen de $500 de la media poblacional?

FIGURA 7.4 Distribución de muestreo de $\bar{x}$ para el sueldo medio anual de una muestra aleatoria simple de 30 gerentes de EAI



Como ya se identificaron las propiedades de la distribución de muestreo de $\bar{x}$ (figura 7.4), se utilizará esta distribución para contestar dicha interrogante probabilística. Observe la distribución de muestreo de $\bar{x}$ que se presenta nuevamente en la figura 7.5. Como la media poblacional es \$51 800, el director de personal desea saber cuál es la probabilidad de que $\bar{x}$ esté entre \$51 300 y \$52 300. Esta probabilidad corresponde al área sombreada de la distribución de muestreo de la figura 7.5. Como la distribución de muestreo está distribuida normalmente, su media es \$51 800 y el error estándar de la media es 730.3, se usa la tabla de probabilidad normal estándar para determinar el área o probabilidad.

Primero se calcula el valor de z en el extremo superior de este intervalo (52 300) y se usa la tabla para hallar el área bajo la curva a la izquierda de ese punto (hacia la cola izquierda). Después se determina el valor de z en el extremo inferior de este intervalo (51 300) y se usa la tabla para hallar el área bajo la curva a la izquierda de este punto (otra área hacia la cola izquierda). Al restar la segunda área de la primera, se obtiene la probabilidad buscada.

En $\bar{x} = 52\,300$ tenemos

$$z = \frac{52\,300 - 51\,800}{730.30} = 0.68$$

En la tabla de probabilidad normal estándar se encuentra que la probabilidad acumulada (área a la izquierda de $z = 0.68$) es 0.7517.

En $\bar{x} = 51\,300$ tenemos

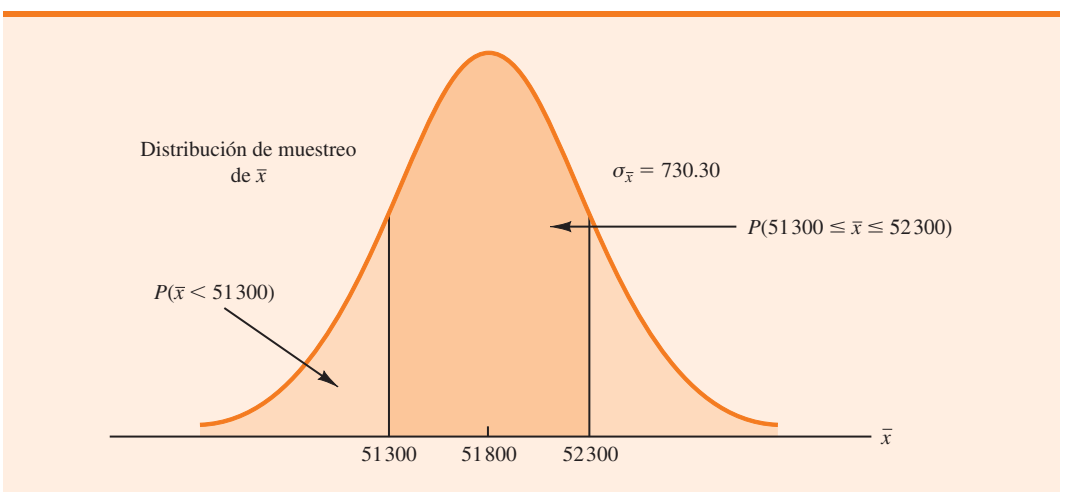
$$z = \frac{51\,300 - 51\,800}{730.30} = -0.68$$

La distribución de muestreo de $\bar{x}$ se utiliza para obtener información probabilística en torno a qué tan cerca se encuentra la media muestral $\bar{x}$ de la media poblacional μ.

El área bajo la curva a la izquierda de $z = -0.68$ es 0.2483. Por tanto, $P(51\,300 \leq \bar{x} \leq 52\,300) = P(z \leq 0.68) - P(z < -0.68) = 0.7517 - 0.2483 = 0.5034$.

Estos cálculos indican que hay una probabilidad de 0.5034 de que con una muestra aleatoria simple de 30 gerentes de EAI se obtenga una media muestral $\bar{x}$ que esté en un margen de \$500 de la media poblacional. Por tanto, la probabilidad de que la diferencia entre $\bar{x}$ y $\mu = \$51\,800$ sea superior a \$500 es $1 - 0.5034 = 0.4966$. En otras palabras, una muestra aleatoria simple de 30 gerentes de EAI tiene aproximadamente 50/50 oportunidades de tener una media muestral que no difiera de la media poblacional en más de los aceptables \$500. Quizá deba pensarse en

FIGURA 7.5 Probabilidad de que una media muestral se encuentre en un margen de \$500 de la media poblacional en una muestra aleatoria simple de 30 gerentes de EAI



una muestra de tamaño mayor. Se explorará esta posibilidad considerando la relación entre el tamaño de la muestra y la distribución de muestreo de $\bar{x}$.

Relación entre el tamaño de la muestra y la distribución de muestreo de $\bar{x}$

Suponga que en el problema de muestreo de EAI se toma una muestra aleatoria simple de 100 gerentes en lugar de los 30 considerados. La intuición indica que teniendo más datos proporcionados por una muestra mayor, la media muestral basada en $n = 100$ proporcionará una mejor estimación de la media poblacional que la basada en $n = 30$. Para ver cuán mejor es, se considerará la relación entre el tamaño de la muestra y la distribución de muestreo de $\bar{x}$.

Primero observe que $E(\bar{x}) = \mu$ independientemente del tamaño de la muestra. Entonces, la media de todos los valores posibles de $\bar{x}$ es igual a la media poblacional μ independientemente del tamaño n de la muestra. No obstante, el error estándar de la media, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, está relacionado con la raíz cuadrada del tamaño de la muestra. Siempre que este tamaño aumente, el error estándar de la media $\sigma_{\bar{x}}$ disminuirá. Con $n = 30$, el error estándar de la media en el problema de EAI es 730.3. Sin embargo, aumentando el tamaño de la muestra a $n = 100$, el error estándar de la media disminuye a

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4\,000}{\sqrt{100}} = 400$$

En la figura 7.6 se ilustran las distribuciones de muestreo de $\bar{x}$ correspondientes a $n = 30$ y a $n = 100$. Como la distribución muestral con $n = 100$ tiene un error estándar más pequeño, habrá menos variación entre los valores de $\bar{x}$ y éstos tenderán a estar más cerca de la media poblacional que los valores de $\bar{x}$ con $n = 30$.

La distribución de muestreo de $\bar{x}$, en el caso de $n = 100$, puede emplearse para calcular la probabilidad de que una muestra aleatoria simple de 100 gerentes de EAI dé una media muestral que no difiera de los $500 de la media poblacional. Como la distribución de muestreo es normal y su media es $51 800 y el error estándar de la media es 400, se emplea la tabla de probabilidad normal estándar para determinar el área o la probabilidad.

Para $\bar{x} = 52\,300$ (figura 7.7) tenemos

$$z = \frac{52\,300 - 51\,800}{400} = 1.25$$

FIGURA 7.6 Comparación entre las distribuciones de muestreo de $\bar{x}$ con muestras aleatorias simples de tamaño $n = 30$ y $n = 100$ gerentes de EAI

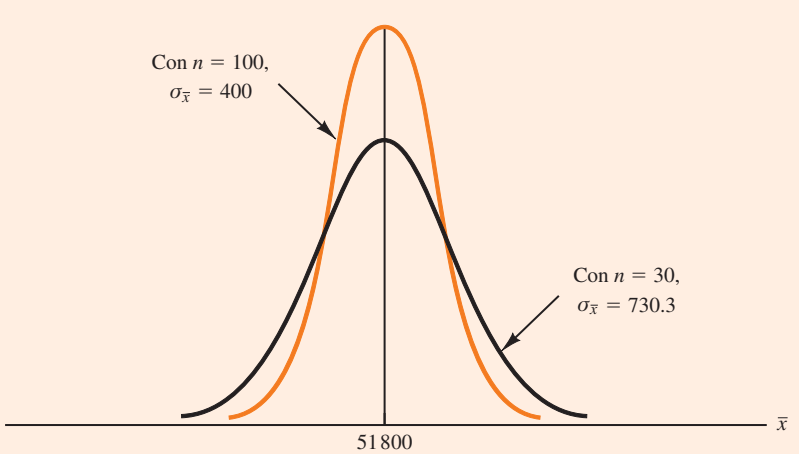
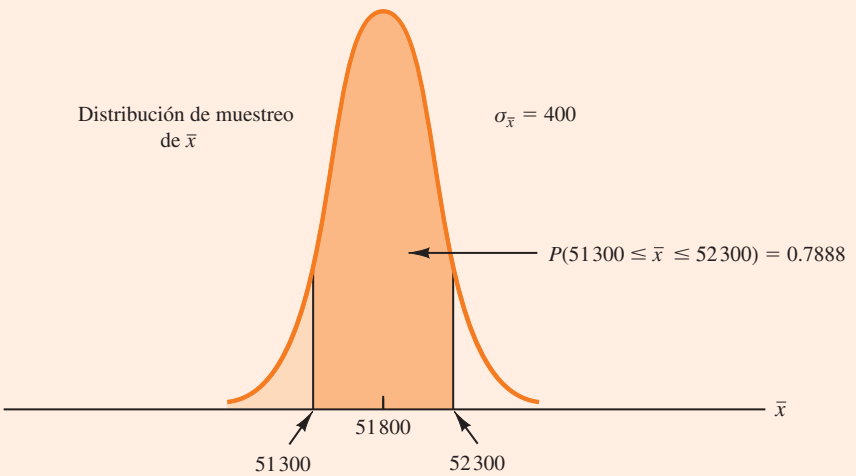


FIGURA 7.7 Probabilidad de que la media muestral esté en un margen de $500 de la media poblacional usando una muestra aleatoria simple de 100 gerentes de EAI



En la tabla de probabilidad normal estándar se encuentra que la probabilidad acumulada correspondiente a $z = 1.25$ es 0.8944.

Para $\bar{x} = 51\,300$ tenemos

$$z = \frac{51\,300 - 51\,800}{400} = -1.25$$

La probabilidad acumulada correspondiente a $z = -1.25$ es 0.1056. Por tanto, $P(51\,300 \leq \bar{x} \leq 52\,300) = P(z \leq 1.25) - P(z \leq -1.25) = 0.8944 - 0.1056 = 0.7888$. Entonces, al aumentar el tamaño de la muestra de 30 a 100 gerentes de EAI, la probabilidad de obtener una muestra aleatoria simple que esté entre los $500 de la media poblacional aumenta de 0.5034 a 0.7888.

El punto importante estriba en que cuando el tamaño de la muestra aumenta, el error estándar de la media disminuye. Como resultado, una muestra de mayor tamaño proporciona mayor probabilidad de que la media muestral esté dentro de una distancia determinada de la media poblacional.

NOTAS Y COMENTARIOS

1. Al presentar la distribución de muestreo de $\bar{x}$ para el problema de EAI, se aprovechó la ventaja de que se conocían la media poblacional $\mu = 51\,800$ y la desviación estándar poblacional $\sigma = 4\,000$. Sin embargo, lo usual es que los valores de la media y la desviación estándar poblacionales que se necesitan para determinar la distribución de muestreo de $\bar{x}$ no se conozcan. En el capítulo 8 se verá cómo se usan la media muestral $\bar{x}$ y la desviación estándar muestral s cuando no se conocen μ y σ.
2. La demostración del teorema del límite central requiere observaciones independientes en la muestra. Esta condición se satisface cuando se trata de poblaciones infinitas y poblaciones finitas si el muestreo se hace con remplazo. Aunque el teorema del límite central no se refiere directamente a muestreos sin remplazo de poblaciones finitas, se aplican sus hallazgos cuando la población es de tamaño grande.

Ejercicios

Métodos

18. La media de una población es 200 y su desviación estándar es 50. Se tomará una muestra aleatoria simple de tamaño 100 y se utilizará la media muestral $\bar{x}$ para estimar la media poblacional.
    a) ¿Cuál es el valor esperado de $\bar{x}$?
    b) ¿Cuál es la desviación estándar de $\bar{x}$?
    c) Ilustre la distribución de muestreo de $\bar{x}$.
    d) ¿Qué expresa la distribución de muestreo de $\bar{x}$?

AUTOevaluación

19. La media de una población es 200 y su desviación estándar es 50. Suponga que se selecciona una muestra aleatoria simple de tamaño 100 y que se usa $\bar{x}$ para estimar μ.
    a) ¿Cuál es la probabilidad de que la diferencia entre la media muestral y la media poblacional no sea mayor que ± 5?
    b) ¿Y de que la diferencia entre la media muestral y la media poblacional no sea mayor que ± 10?

20. Suponga que la desviación estándar poblacional es $\sigma = 25$. Calcule el error estándar de la media, $\sigma_{\bar{x}}$, con muestras de tamaño 50, 100, 150 y 200. ¿Qué puede decir acerca del tamaño del error estándar de la media conforme el tamaño de la muestra aumenta?

21. Suponga que se toma una muestra aleatoria simple de tamaño 50 a partir de una población en la que $\sigma = 10$. Determine el valor del error estándar de la media en cada uno de los casos siguientes (si es necesario, use el factor de corrección para una población finita).
    a) El tamaño de la población es infinito.
    b) El tamaño de la población es $N = 50\,000$.
    c) El tamaño de la población es $N = 5\,000$.
    d) El tamaño de la población *es* $N = 500$.

Aplicaciones

22. Regrese al problema de los gerentes de EAI. Suponga que se utiliza una muestra aleatoria simple de 60 gerentes.
    a) Dibuje la distribución de muestreo de $\bar{x}$ si se emplean muestras aleatorias simples de tamaño 60.
    b) ¿Qué sucede con la distribución de muestreo de $\bar{x}$ si se usan muestras aleatorias simples de tamaño 120?
    c) ¿Qué puede decir acerca de qué le sucede a la distribución de muestreo de $\bar{x}$ conforme el tamaño de la muestra aumenta? ¿Parece lógica esta generalización? Explique.

AUTOevaluación

23. En el problema de muestreo de EAI (figura 7.5), se indicó que con $n = 30$, la probabilidad de que la media muestral no difiriera más de $\pm \$500$ de la media poblacional era 0.5034.
    a) ¿Cuál es la probabilidad de que la media muestral no difiera más de \$500 de la media poblacional si se usa una muestra de tamaño 60?
    b) Responda el inciso *a*) si el tamaño de la muestra es 120.

24. *Barron's* reportó que el número promedio de semanas que un individuo está desempleado es de 17.5 (*Barron's*, 18 de febrero de 2008). Suponga que el tamaño de la media poblacional es 17.5 semanas para la población de todos los individuos desempleados, y que la desviación estándar poblacional es de cuatro semanas. Asuma que quiere seleccionar una muestra aleatoria de 50 individuos sin empleo para un estudio de seguimiento.
    a) Presente la distribución de muestreo de $\bar{x}$, la media muestral promedio de una muestra de 50 individuos desempleados.
    b) ¿Cuál es la probabilidad de que la muestra aleatoria simple de los 50 sujetos proporcione una media muestral que no difiera de la media poblacional en más de una semana?
    c) ¿Cuál es la probabilidad de que la muestra aleatoria simple referida proporcione una media muestral que no difiera de la media poblacional en más de ½ semana?

25. El College Board informó que se obtuvieron las siguientes puntuaciones medias en las tres partes del examen de admisión a las universidades (*The World Almanac*, 2009).

Lectura de comprensión	502
Matemáticas	515
Redacción	494

Suponga que la desviación estándar poblacional en cada parte es $\sigma = 100$.

a) ¿Cuál es la probabilidad de que en una muestra aleatoria de 90 sustentantes la media muestral de las puntuaciones no difiera más de 10 puntos de la media poblacional de 502 en la parte de lectura de comprensión en el examen?

b) ¿Cuál es la probabilidad de que en una muestra aleatoria de 90 estudiantes la media muestral de las puntuaciones no difiera más de 10 puntos de la media poblacional de 515 en la sección de matemáticas? Compare esta probabilidad con el valor calculado en el inciso *a*).

c) ¿Cuál es la probabilidad de que en una muestra aleatoria de 90 sustentantes la media muestral de las puntuaciones no difiera más de 10 puntos de la media poblacional de 494 en la parte de redacción en el examen? Compare esta probabilidad con el valor calculado en los incisos *a*) y *b*).

26. El costo medio anual de un seguro para automóvil es de $939 (*CNBC*, 23 de febrero de 2006). Suponga que la desviación estándar es $\sigma = \$245$.

a) ¿Cuál es la probabilidad de que en una muestra aleatoria simple de pólizas de seguros de automóvil la media muestral no difiera más de $25 de la media poblacional si el tamaño de la muestra es 30, 50, 100 y 400?

b) ¿Qué ventaja tiene una muestra más grande cuando se quiere estimar la media poblacional?

27. *BusinessWeek* realizó una encuesta entre los estudiantes que terminaban sus estudios en los 30 programas de una maestría (*BusinessWeek,* 22 de septiembre de 2003). Con base en esta encuesta el sueldo medio anual de un hombre y de una mujer 10 años después de terminar sus estudios asciende a $168 000 y $117 000, respectivamente. Suponga que la desviación estándar entre los sueldos de los hombres con grado es $40 000 y entre las mujeres con grado es $25 000.

a) ¿Cuál es la probabilidad de que en una muestra aleatoria simple de 40 hombres con grado la media muestral no difiera más de $10 000 de la media poblacional de $168 000?

b) ¿Cuál es la probabilidad de que en una muestra aleatoria simple de 40 mujeres graduadas la media muestral no difiera más de $10 000 de la media poblacional de $117 000?

c) ¿En cuál de los dos casos, inciso *a*) o inciso *b*), hay más probabilidad de obtener una media muestral que no difiera en más de $10 000 de la media poblacional? ¿Por qué?

d) ¿Cuál es la probabilidad de que en una muestra aleatoria simple de 100 hombres con grado, la media muestral no difiera en más de $4 000 de la media poblacional?

28. La puntuación promedio de golfistas hombres es de 95 y para las golfistas mujeres es de 106 (*Golf Digest,* abril de 2006). Considere estos valores como medias poblacionales de los hombres y las mujeres y suponga que la desviación estándar poblacional es $\sigma = 14$ golpes en ambos casos. Se tomará una muestra aleatoria simple de 30 golfistas hombres y otra muestra aleatoria simple de 45 golfistas mujeres.

a) Proporcione la distribución de muestreo de $\bar{x}$ correspondiente a los golfistas.

b) ¿Cuál es la probabilidad de que la media muestral no difiera en más de 3 golpes de la media poblacional en la muestra de hombres?

c) ¿Cuál es la probabilidad de que la media muestral no difiera en más de 3 golpes de la media poblacional en la muestra de golfistas mujeres?

d) ¿En cuál de los casos, inciso *a*) o inciso *b*), es mayor la probabilidad de que la media muestral no difiera en más de 3 golpes de la media poblacional? ¿Por qué?

29. El precio promedio de un galón de gasolina sin plomo era de $2.34 en el norte de Kentucky (*The Cincinnati Enquirer,* 21 de enero de 2006). Use este precio como media poblacional y suponga que la desviación estándar poblacional es $0.20.

a) ¿Cuál es la probabilidad de que el precio medio en una muestra de 30 gasolineras no difiera en más de $0.03 de la media poblacional?
b) ¿Cuál es la probabilidad de que el precio medio en una muestra de 50 gasolineras no difiera en más de $0.03 de la media poblacional?
c) ¿Cuál es la probabilidad de que el precio medio en una muestra de 100 gasolineras no difiera en más de $0.03 de la media poblacional?
d) ¿Recomendaría usted alguno de los tamaños muestrales de los incisos *a*), *b*) o *c*) para tener al menos 0.95 de probabilidad de que la media muestral esté dentro de $0.03 de la media poblacional?

30. Para estimar la edad media de una población de 4 000 empleados se selecciona una muestra aleatoria simple de 40 sujetos.
a) ¿Usaría el factor de corrección para una población finita en el cálculo del error estándar de la media? Explique.
b) Si la desviación estándar poblacional es $\sigma = 8.2$ años, calcule el error estándar con y sin el factor de corrección para una población finita. ¿Cuál es la base para ignorar el factor de corrección para la población finita si $n/N \leq 0.05$?
c) ¿Cuál es la probabilidad de que la media muestral de las edades de los empleados no difiera en más de ± 2 años de la media poblacional de las edades?

7.6 Distribución de muestreo de $\bar{p}$

La proporción muestral $\bar{p}$ es el estimador puntual de la proporción poblacional p. La fórmula para calcular la proporción muestral es

$$\bar{p} = \frac{x}{n}$$

donde

x = número de elementos de la muestra que poseen la característica de interés
n = tamaño de la muestra

Como se indica en la sección 7.4, la proporción muestral $\bar{p}$ es una variable aleatoria y su distribución de probabilidad se conoce como *distribución de muestreo de* $\bar{p}$.

DISTRIBUCIÓN DE MUESTREO DE $\bar{p}$

La distribución de muestreo de $\bar{p}$ es la distribución de probabilidad de todos los posibles valores de la proporción muestral $\bar{p}$.

Para determinar qué tan cerca está la proporción muestral $\bar{p}$ de la proporción poblacional p, es necesario entender las propiedades de la distribución de muestreo de $\bar{p}$: el valor esperado de $\bar{p}$, la desviación estándar de $\bar{p}$ y la forma de la distribución de muestreo de $\bar{p}$.

Valor esperado de $\bar{p}$

El valor esperado de $\bar{p}$, la media de todos los posibles valores de $\bar{p}$, es igual a la proporción poblacional p.

VALOR ESPERADO DE $\bar{p}$

$$E(\bar{p}) = p \tag{7.4}$$

donde

$E(\bar{p})$ = valor esperado de $\bar{p}$
p = proporción poblacional

Como $E(\bar{p}) = p$, $\bar{p}$ es un estimador insesgado de p. Recuerde que en la sección 7.1 se encontró que en la población de EAI, $p = 0.60$, siendo p la proporción de la población de gerentes que han participado en el programa de capacitación de la empresa. Por tanto, el valor esperado de $\bar{p}$ en el problema de muestreo de EAI es 0.60.

Desviación estándar de $\bar{p}$

Como en el caso de la desviación estándar de $\bar{x}$, la desviación estándar de $\bar{p}$ depende de si la población es finita o infinita. Las dos fórmulas para calcularla se presentan a continuación.

DESVIACIÓN ESTÁNDAR DE $\bar{p}$

Población finita

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}} \tag{7.5}$$

Población infinita

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{7.5}$$

Al comparar las dos fórmulas en (7.5) se aprecia que la única diferencia es el uso del factor de corrección para una población finita $\sqrt{(N-n)/(N-1)}$.

Como en el caso de la media muestral $\bar{x}$, la diferencia entre las expresiones para una población finita y una infinita es despreciable si el tamaño de la población finita es grande en comparación con el tamaño de la muestra. Se seguirá la misma regla recomendada para la media muestral. Es decir, si la población es finita y $n/N \leq 0.05$ se usará $\sigma_{\bar{p}} = \sqrt{p(1-p)/n}$. Pero si la población es finita y $n/N > 0.05$, entonces deberá utilizarse el factor de corrección para una población finita. También, a menos que se especifique otra cosa, en este libro se supondrá que el tamaño de la población es grande en comparación con el tamaño de la muestra y, por tanto, el factor de corrección para una población finita no será necesario.

En la sección 7.5 se utilizó el término *error estándar de la media* para referirse a la desviación estándar de $\bar{x}$. Se dijo que en general la expresión error estándar se refiere a la desviación estándar de un estimador puntual. Así, en el caso de proporciones, se usa *el error estándar de la proporción* para referirse a la desviación estándar de $\bar{p}$. Ahora se vuelve al ejemplo de EAI para calcular el error estándar de la proporción asociada con la muestra aleatoria simple de los 30 gerentes de EAI.

En el estudio de EAI se sabe que la proporción poblacional de gerentes que han participado en el programa de capacitación es $p = 0.60$. Como $n/N = 30/2\,500 = 0.012$, se puede ignorar el factor de corrección para una población finita al calcular el error estándar de la proporción. En la muestra aleatoria simple de 30 gerentes, $\sigma_{\bar{p}}$ es

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.60(1-0.60)}{30}} = 0.0894$$

Forma de la distribución de muestreo de $\bar{p}$

Ahora que se conoce la media y la desviación estándar de la distribución de muestreo de $\bar{p}$, el úl-timo paso es determinar la forma de esta distribución. La proporción muestral es $\bar{p} = x/n$. En una muestra aleatoria simple de una población grande, el valor de x es una variable aleatoria binomial que indica el número de los elementos de la muestra que tienen la característica de interés. Como n es una constante, la probabilidad de x/n es la misma que la probabilidad binomial de x, lo cual significa que la distribución de muestreo de $\bar{p}$ también es una distribución de probabilidad discreta y la probabilidad de cada x/n es la misma que la de x.

En el capítulo 6 se estableció que una distribución binomial se aproxima mediante una distribución normal, siempre que el tamaño de la muestra sea lo suficientemente grande para satisfacer las dos condiciones siguientes.

$$np \geq 5 \quad \text{y} \quad n(1 - p) \geq 5$$

Suponiendo que se satisfagan estas dos condiciones, la distribución de probabilidad de x en la proporción muestral, $\bar{p} = x/n$, puede aproximarse por medio de una distribución normal. Y como n es una constante, la distribución de muestreo de $\bar{p}$ también se aproxima mediante una distribución normal. Esta aproximación se formula como se indica enseguida:

La distribución de muestreo de $\bar{p}$ se aproxima mediante una distribución normal, siempre que $np \geq 5$ y $n(1 - p) \geq 5$.

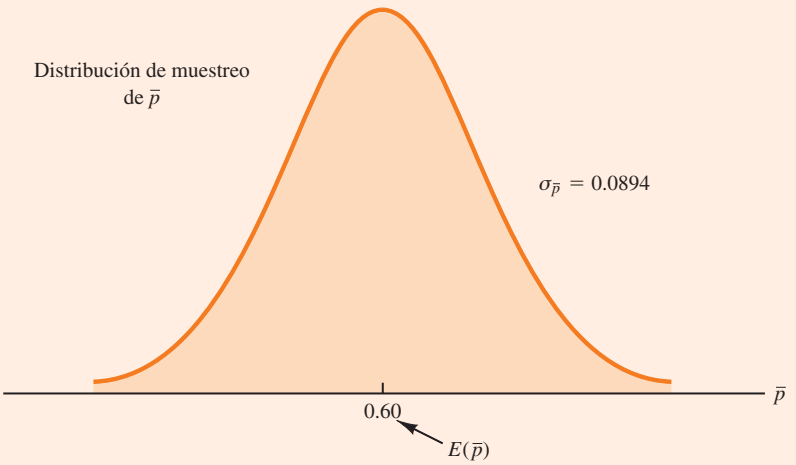
En las aplicaciones prácticas, cuando se requiere una estimación de la proporción poblacional, casi siempre se encuentra que el tamaño de la muestra es suficientemente grande para permitir usar la aproximación normal para la distribución de muestreo de $\bar{p}$.

Recuerde que en el problema de muestreo de EAI la proporción poblacional de gerentes que han participado en el programa de capacitación es $p = 0.60$. Con una muestra aleatoria simple de tamaño 30, se tiene $np = 30(0.60) = 18$ y $n(1 - p) = 30\ (0.40) = 12$. Por tanto, la distribución de muestreo de $\bar{p}$ se calcula mediante la distribución normal que se presenta en la figura 7.8.

Valor práctico de la distribución de muestreo de $\bar{p}$

El valor práctico de la distribución de muestreo de $\bar{p}$ radica en que permite obtener información probabilística acerca de la diferencia entre la proporción muestral y la proporción poblacional. Por ejemplo, en el problema de EAI, el director de personal desea saber cuál es la probabilidad de obtener un valor de $\bar{p}$ que no difiera en más de 0.05 de la proporción poblacional de los gerentes de EAI que han participado en el programa de capacitación. Es decir, ¿cuál es la probabilidad de tener una muestra en la que la proporción muestral $\bar{p}$ esté entre 0.55 y 0.65? El área sombreada de la figura 7.9 corresponde a esta probabilidad. A partir de que la distribución de muestreo de $\bar{p}$ se aproxima mediante una distribución normal con media 0.60 y un error estándar de la proporción $\sigma_{\bar{p}} = 0.0894$, se encuentra que la variable aleatoria normal estándar correspondiente a $\bar{p} = 0.65$ tiene el valor $z = (0.65 - 0.60)/0.0894 = 0.56$. En la tabla de probabilidad normal estándar aparece que la probabilidad acumulada que corresponde a $z = 0.56$ es 0.7123. De manera similar para $\bar{p} = 0.55$, se encuentra que $z = (0.55 - 0.60)/0.0894 = -0.56$. En la misma tabla se aprecia que la probabilidad acumulada correspondiente a $z = -0.56$ es 0.2877. De esta manera, la probabilidad de seleccionar una muestra en la cual el valor de $\bar{p}$ no difiera más de 0.05 de la proporción poblacional p está dada por $0.7123 - 0.2877 = 0.4246$.

FIGURA 7.8 Distribución de muestreo de $\bar{p}$ para la proporción de gerentes que ha participado en el programa de capacitación de EAI

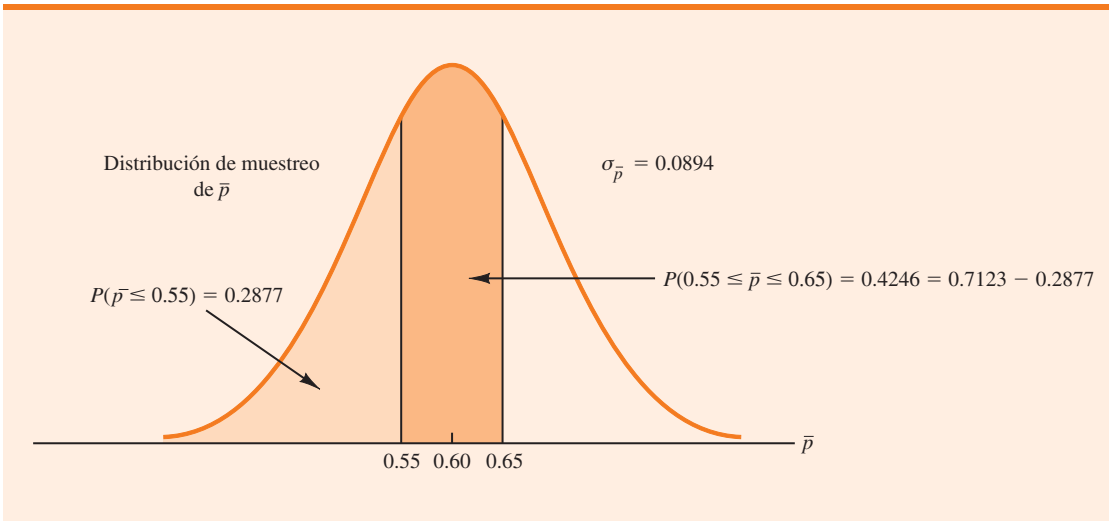


Si se aumenta el tamaño de la muestra a $n = 100$, el error estándar de la proporción se convierte en

$$\sigma_{\bar{p}} = \sqrt{\frac{0.60(1 - 0.60)}{100}} = 0.049$$

Con una muestra de 100 gerentes de EAI, se calcula ahora la probabilidad de que la proporción muestral tenga un valor que no difiera en más de 0.05 de la proporción poblacional. Como la distribución de muestreo es aproximadamente normal, con media 0.60 y desviación estándar 0.049, se puede usar la tabla de probabilidad normal estándar para determinar el área o probabilidad. Para $\bar{p} = 0.65$, se tiene $z = (0.65 - 0.60)/0.049 = 1.02$. La tabla de probabilidad normal estándar indica que la probabilidad acumulada correspondiente a $z = 1.02$ es 0.8461. De

FIGURA 7.9 Probabilidad de que $\bar{p}$ esté entre 0.55 y 0.65



manera similar, para $\bar{p} = 0.55$, se tiene que $z = (0.55 - 0.60)/0.049 = -1.02$. Se encuentra que la probabilidad acumulada correspondiente a $z = -1.02$ es 0.1539. Por tanto, si el tamaño de la muestra aumenta de 30 a 100, la probabilidad de que la proporción muestral $\bar{p}$ no difiera en más de 0.05 de la proporción poblacional p aumenta a $0.8461 - 0.1539 = 0.6922$.

Ejercicios

Métodos

31. Una muestra aleatoria de tamaño 100 es seleccionada de una población en la que $p = 0.40$.
    a) ¿Cuál es el valor esperado de $\bar{p}$?
    b) ¿Cuál es el error estándar de $\bar{p}$?
    c) Exprese la distribución de muestreo de $\bar{p}$.
    d) ¿Qué indica esta distribución?

AUTOevaluación

32. Una proporción poblacional es 0.40. Se toma una muestra aleatoria simple de tamaño 200 y la proporción muestral $\bar{p}$ se usa para estimar la proporción poblacional.
    a) ¿Cuál es la probabilidad de que la proporción muestral esté entre ± 0.03 de la proporción poblacional?
    b) ¿Cuál es la probabilidad de que la proporción muestral se encuentre entre ± 0.05 de la proporción poblacional?

33. Suponga que la proporción poblacional es 0.55. Calcule el error estándar de la proporción, $\sigma_{\bar{p}}$, para los tamaños de muestra 100, 200, 500 y 1 000. ¿Qué puede decir acerca del tamaño del error estándar a medida que el tamaño de la muestra aumenta?

34. La proporción poblacional es 0.30. ¿Cuál es la probabilidad de que las proporciones muestral y poblacional estén entre ± 0.04 con los tamaños de muestra siguientes?
    a) $n = 100$
    b) $n = 200$
    c) $n = 500$
    d) $n = 1\,000$
    e) ¿Qué ventaja tiene un tamaño grande de muestra?

Aplicaciones

AUTOevaluación

35. El director de Doerman Distributors, Inc. piensa que 30% de los pedidos proviene de nuevos clientes. Para ver la proporción de clientes nuevos se usará una muestra aleatoria simple de 100 pedidos.
    a) Suponga que el director está en lo cierto y que $p = 0.30$. ¿Cuál es la distribución de muestreo de $\bar{p}$ en este estudio?
    b) ¿Cuál es la probabilidad de que la proporción muestral de $\bar{p}$ esté entre 0.20 y 0.40?
    c) ¿Cuál es la probabilidad de que esté entre 0.25 y 0.35?

36. *The Cincinnati Enquirer* informa que en Estados Unidos 66% de los adultos y 87% de los jóvenes entre 12 y 17 años usan Internet (*The Cincinnati Enquirer*, 7 de febrero de 2006). Considere estos datos como proporciones poblacionales y suponga que se usará una muestra de 300 adultos y 300 jóvenes para obtener información respecto de su opinión acerca de la seguridad en Internet.
    a) Exponga la distribución de muestreo de $\bar{p}$, siendo $\bar{p}$ la proporción muestral de adultos que usan Internet.
    b) ¿Cuál es la probabilidad de que la diferencia entre la proporción muestral y la proporción poblacional de adultos que usan Internet no sea mayor que ± 0.04?
    c) ¿Cuál es la probabilidad de que la diferencia entre la proporción muestral y la proporción poblacional de jóvenes que usan Internet no sea mayor que ± 0.04?

d) ¿Son diferentes las probabilidades del inciso *b*) y del inciso *c*)? Si es así, ¿por qué?
e) Responda al inciso *b*) en el caso de que el tamaño de la muestra sea 600. ¿Es menor la probabilidad? ¿Por qué?

37. Las personas terminan por desechar 12% de lo que compran en el supermercado (*Reader's Digest*, marzo de 2009). Asuma que ésta es la verdadera proporción poblacional y que planea realizar una encuesta por muestreo de 450 compradores para investigar más acerca de su comportamiento.
a) Presente la distribución de muestreo de $\overline{p}$, la proporción de mercancía que desechan los encuestados de la muestra.
b) ¿Cuál es la probabilidad de que la encuesta genere una proporción muestral de ± 0.03 de la proporción poblacional?
c) ¿Cuál es la probabilidad de que la encuesta genere una proporción muestral de ± 0.015 de la proporción poblacional?

38. Roper ASW realizó una encuesta para obtener información acerca de la opinión de los estadounidenses respecto del dinero y la felicidad (*Money,* octubre de 2003). De los entrevistados, 56% dijo revisar el estado de su chequera por lo menos una vez al mes.
a) Suponga que se toma una muestra de 400 estadounidenses adultos. Indique la distribución de muestreo de la proporción de éstos que revisa el estado de su chequera por lo menos una vez al mes.
b) ¿Cuál es la probabilidad de que la diferencia entre las proporciones muestral y poblacional no sea mayor que ± 0.02?
c) ¿Cuál es la probabilidad de que dicha diferencia no sea mayor que ± 0.04?

39. En 2008, el Better Business Bureau resolvió 75% de las quejas que recibió (*USA Today*, 2 de marzo de 2009). Suponga que ha sido contratado por esta oficina para investigar los reclamos que recibió este año y que involucran a nuevos concesionarios automotrices. Usted planea seleccionar una muestra de las quejas de estos últimos para estimar la proporción que el Better Business Bureau está en posibilidad de resolver. Asuma que la proporción poblacional de quejas resueltas de nuevos concesionarios automotrices es 0.75, la misma que la proporción general de reclamos resueltos en 2008.
a) Suponga que selecciona una muestra de 450 quejas que involucran a nuevos concesionarios automotrices. Presente la distribución muestral de $\overline{p}$.
b) Con base en la muestra de 450 quejas, ¿cuál es la probabilidad de que la diferencia entre las proporciones muestral y poblacional no sea mayor que 0.04?
c) Suponga que selecciona una muestra de 200 quejas que involucran a nuevos concesionarios automotrices. Presente la distribución de muestreo de $\overline{p}$.
d) Con base en la muestra más pequeña de sólo 200 quejas, ¿cuál es la probabilidad de que la diferencia entre las proporciones muestral y poblacional no sea mayor que 0.04?
e) Con base en lo determinado por el incremento en la probabilidad, ¿qué tanto se ganaría en precisión si se tomara la muestra más grande en el inciso *b*)?

40. The Grocery Manufacturers of America informa que 76% de los consumidores lee los ingredientes que se mencionan en la etiqueta de un producto. Suponga que la proporción poblacional es $p = 0.76$ y que de la población de consumidores se selecciona una muestra de 400.
a) Exprese la distribución de muestreo de la proporción muestral $\overline{p}$, si $\overline{p}$ es la proporción de consumidores de la muestra que lee los ingredientes que se mencionan en la etiqueta.
b) ¿Cuál es la probabilidad de que la diferencia entre las proporciones muestral y poblacional no sea mayor que ± 0.03?
c) Conteste el inciso *b*) si el tamaño de la muestra es 750 consumidores.

41. El Food Marketing Institute informa que 17% de los hogares gasta más de $100 en productos de abarrotes. Suponga que la proporción poblacional es $p = 0.17$ y que de la población se toma una muestra aleatoria simple de 800 hogares.
a) Exprese la distribución de muestreo de $\overline{p}$, la proporción muestral de hogares que gastan más de $100 semanales en abarrotes.
b) ¿Cuál es la probabilidad de que la proporción poblacional no difiera en más de 0.02 de la proporción poblacional?
c) Conteste el inciso *b*) en caso de que el tamaño de la muestra sea de 1 600 hogares.

7.7 Propiedades de los estimadores puntuales

En este capítulo se ha explicado que los estadísticos muestrales, como la media muestral $\bar{x}$, la desviación estándar muestral s y la proporción muestral $\bar{p}$ sirven como estimadores puntuales de sus correspondientes parámetros poblacionales, μ, σ y p. Resulta interesante advertir que cada uno de estos estadísticos muestrales sean los estimadores puntuales de sus correspondientes parámetros poblacionales. Sin embargo, antes de usar un estadístico muestral como estimador puntual, se verifica si éste tiene ciertas propiedades que corresponden a un buen estimador puntual. En esta sección se estudian las propiedades que deben tener los buenos estimadores puntuales: insesgadez, eficiencia y consistencia.

Como hay distintos estadísticos muestrales que se utilizan como estimadores puntuales de sus diferentes parámetros poblacionales, en esta sección se usará la notación general siguiente.

$$\theta = \text{parámetro poblacional de interés}$$
$$\hat{\theta} = \text{estadístico muestral o estimador puntual de } \theta$$

En esta notación, θ es la letra griega *theta* y la notación $\hat{\theta}$ se lee "theta sombrero". En general, θ representa cualquier parámetro poblacional como, por ejemplo, la media poblacional, la desviación estándar poblacional, la proporción poblacional, etc., y $\hat{\theta}$ representa el correspondiente estadístico muestral, por ejemplo, la media muestral, la desviación estándar muestral y la proporción muestral.

Insesgadez

Si el valor esperado del estadístico muestral es igual al parámetro poblacional que se estima, se dice que el estadístico muestral es un *estimador insesgado* del parámetro poblacional.

INSESGADEZ

El estadístico muestral $\hat{\theta}$ es un estimador insesgado del parámetro poblacional θ si

$$E(\hat{\theta}) = \theta$$

donde

$$E(\hat{\theta}) = \text{valor esperado del estadístico muestral } \hat{\theta}$$

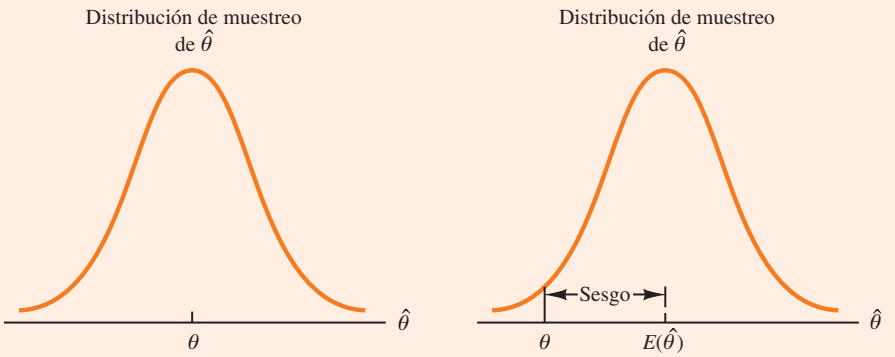
Por tanto, el valor esperado, o media, de todos los posibles valores de un estadístico muestral insesgado es igual al parámetro poblacional que se está estimando.

En la figura 7.10 se exponen los casos de los estimadores puntuales sesgado e insesgado. En la gráfica que ilustra el estimador insesgado, la media de la distribución de muestreo es igual al valor del parámetro poblacional. En este caso los errores de estimación se equilibran, ya que algunas veces el valor del estimador puntual $\hat{\theta}$ puede ser menor que θ y otras veces es mayor que θ. En el estimador sesgado, la media de la distribución de muestreo es menor o mayor que el valor del parámetro poblacional. En la gráfica *B* de la figura 7.10, $E(\hat{\theta})$ es mayor que θ; así, la probabilidad de que los estadísticos muestrales sobreestimen el valor del parámetro poblacional es grande. En la figura se muestra la amplitud de este sesgo.

Al estudiar las distribuciones de muestreo de la media muestral y de la proporción muestral, se vio que $E(\bar{x}) = \mu$ y que $E(\bar{p}) = p$. Por tanto, $\bar{x}$ y $\bar{p}$ son estimadores insesgados de sus correspondientes parámetros poblacionales μ y p.

En cuanto a la desviación estándar muestral s y la varianza muestral s^2, se puede demostrar que $E(s^2) = \sigma^2$. Por consiguiente, se concluye que la varianza muestral s^2 es un estimador insesgado de la varianza poblacional σ^2. En efecto, en el capítulo 3, cuando se presentaron las

FIGURA 7.10 Ejemplos de estimadores puntuales insesgados y sesgados



El parámetro θ se localiza en la media de la distribución de muestreo; $E(\hat{\theta}) = \theta$

Gráfica *A*. Estimador insesgado

El parámetro θ no se localiza en la media de la distribución de muestreo; $E(\hat{\theta}) \neq \theta$

Gráfica *B*. Estimador sesgado

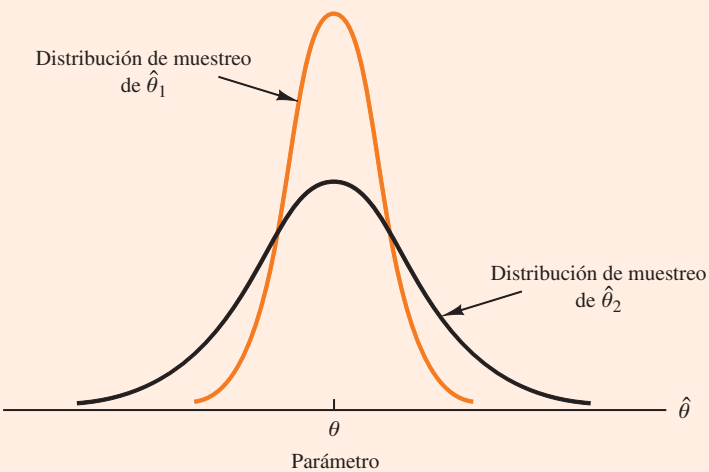
fórmulas para la varianza muestral y la desviación estándar muestral, en el denominador se usó $n - 1$ en lugar de n para que la varianza muestral fuera un estimador insesgado de la varianza poblacional.

Eficiencia

Suponga que se usa una muestra aleatoria simple de n elementos para obtener dos estimadores puntuales insesgados de un mismo parámetro poblacional. En estas circunstancias, se preferirá usar el estimador puntual con el menor error estándar, ya que tenderá a dar estimaciones más cercanas al parámetro poblacional. Se dice que el estimador puntual con menor error estándar tiene mayor **eficiencia relativa** que los otros.

En la figura 7.11 se presentan las distribuciones de muestreo de dos estimadores puntuales insesgados, $\hat{\theta}_1$ y $\hat{\theta}_2$. Observe que el error estándar de $\hat{\theta}_1$ es menor que el error estándar de $\hat{\theta}_2$; por

FIGURA 7.11 Distribuciones de muestreo de dos estimadores puntuales insesgados



Cuando se muestrean poblaciones normales, el error estándar de la media muestral es menor que el error estándar de la mediana muestral. Por tanto, la media muestral es más eficiente que la mediana muestral.

tanto, los valores de $\hat{\theta}_1$ tienen más posibilidades de estar cerca del parámetro $\hat{\theta}$ que los valores de $\hat{\theta}_2$. Como el error estándar del estimador puntual $\hat{\theta}_1$ es menor que el del estimador puntual $\hat{\theta}_2$, $\hat{\theta}_1$ es relativamente más eficiente que $\hat{\theta}_2$ y se prefiere como estimador puntual.

Consistencia

La tercera propiedad relacionada con un buen estimador puntual es la **consistencia**. Dicho de manera sencilla, un estimador puntual es consistente si su valor tiende a estar más cerca del parámetro poblacional a medida que el tamaño de la muestra aumenta. En otras palabras, una muestra grande tiende a proporcionar mejor estimación puntual que una pequeña. Observe que en el caso de la media muestral $\bar{x}$, el error estándar de $\bar{x}$ está dado por $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. Puesto que $\sigma_{\bar{x}}$ está vinculado con el tamaño de la muestra, de manera que muestras mayores dan valores menores de $\sigma_{\bar{x}}$, entonces las de tamaño grande tienden a proporcionar estimadores puntuales más cercanos a la media de la población μ. Mediante un razonamiento similar, también se puede concluir que la proporción muestral $\bar{p}$ es un estimador consistente de la proporción poblacional p.

NOTAS Y COMENTARIOS

En el capítulo 3 se dijo que la media y la mediana son dos medidas de localización central. En este capítulo sólo se estudió la media debido a que cuando se muestrea una población normal, en la cual la media y la mediana poblacionales son idénticas, el error estándar de la mediana es aproximadamente 25% mayor que el error estándar de la media. Recuerde que en el problema de EAI, con $n = 30$, el error estándar de la media fue $\sigma_{\bar{x}} = 730.3$, mientras que el de la mediana en este problema sería $1.25 \times (730.3) = 913$. Por tanto, la media muestral es más eficiente y tendrá más probabilidad de estar dentro de una determinada distancia de la media poblacional.

7.8 Otros métodos de muestreo

Esta sección proporciona una breve introducción a otros métodos de muestreo distintos del muestreo aleatorio simple.

Se describió el muestreo aleatorio simple como un procedimiento de muestreo de una población finita y se estudiaron las propiedades de las distribuciones de muestreo de $\bar{x}$ y de $\bar{p}$ cuando se usó el muestreo aleatorio simple. Sin embargo, no es el único método de muestreo que existe. Hay otros, como el muestreo aleatorio estratificado, el muestreo por conglomerados y el muestreo sistemático que, en ciertas situaciones, tienen ventajas sobre el aleatorio simple. En esta sección se presentan brevemente estos tres métodos. En el capítulo 22, que se encuentra en el sitio web del libro, se estudian con más detalle.

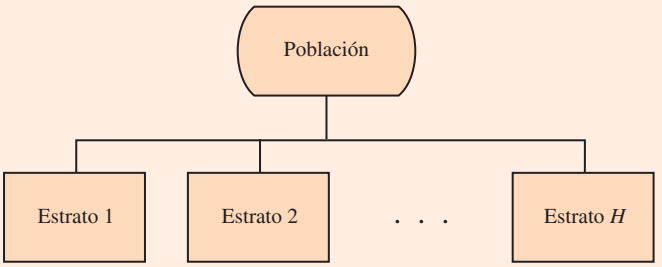
Muestreo aleatorio estratificado

El muestreo aleatorio estratificado funciona mejor cuando la varianza entre los elementos de cada estrato es relativamente pequeña.

En el **muestreo aleatorio estratificado** los elementos de la población primero se dividen en grupos, a los que se les llama *estratos,* de manera que cada elemento pertenezca a uno y sólo un estrato. La base para la formación de los estratos, que pueden ser departamento, edad, tipo de industria, etc., está a discreción de la persona que diseña la muestra. Sin embargo, se obtienen mejores resultados cuando los elementos que los forman son lo más parecidos posible. La figura 7.12 es el diagrama de una población dividida en H estratos.

Una vez formados los estratos, se toma una muestra aleatoria simple de cada uno. Existen fórmulas para combinar los resultados de las muestras de varios estratos individuales en una estimación del parámetro poblacional de interés. El valor del muestreo aleatorio estratificado depende de qué tan homogéneos sean los elementos dentro de cada grupo. Si los elementos de

FIGURA 7.12 Diagrama de un muestreo aleatorio estratificado



un estrato son parecidos, éste tendrá una varianza pequeña. Por tanto, con muestras relativamente pequeñas de los estratos se obtienen buenas estimaciones de sus características. Si éstos son homogéneos, el muestreo aleatorio estratificado proporciona resultados tan precisos como los de un muestreo aleatorio simple, pero con una muestra de tamaño total menor.

Muestreo por conglomerados

El muestreo por conglomerados o clusters funciona mejor cuando cada conglomerado proporciona una representación a menor escala de la población.

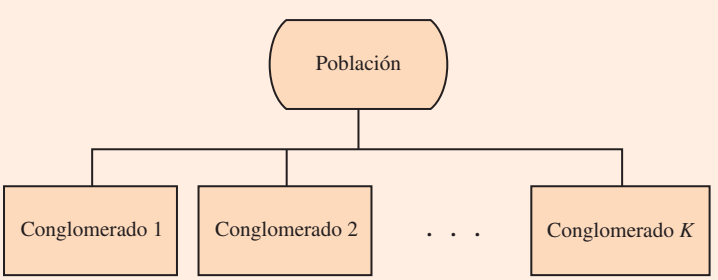
En el **muestreo por conglomerados** (o clusters) los elementos de la población primero se dividen en grupos separados, llamados *conglomerados* o *clusters*. Cada elemento pertenece a uno y sólo un conglomerado (vea la figura 7.13). Se toma una muestra aleatoria simple de los conglomerados. Todos los elementos en cada conglomerado muestreado forman la muestra. Este muestreo tiende a proporcionar mejores resultados cuando los elementos dentro de los conglomerados no son semejantes. Lo ideal es que cada conglomerado sea una representación, a pequeña escala, de la población completa. Si todos son semejantes en este aspecto, tomando en la muestra un número pequeño de conglomerados, se obtendrá una buena estimación de los parámetros poblacionales.

Una de las principales aplicaciones del muestro por conglomerados es el muestreo de áreas, en el que los conglomerados son las manzanas de una ciudad u otras zonas bien definidas. Por lo general, precisa tamaños de muestra mayores que los requeridos en el muestreo aleatorio simple o en el muestreo aleatorio estratificado. Sin embargo, permite reducir costos debido a que cuando se envía a un entrevistador a uno de los conglomerados de la muestra (por ejemplo, a una manzana de una ciudad), puede obtener muchas observaciones en poco tiempo. Por tanto, provee una muestra de tamaño grande a un costo significantemente menor.

Muestreo sistemático

Para ciertos muestreos, en especial en aquellos con poblaciones grandes, se necesita mucho tiempo para tomar una muestra aleatoria simple, pues se requiere determinar primero los nú-

FIGURA 7.13 Diagrama del muestreo por conglomerados



meros aleatorios y después contar y recorrer toda una lista de la población hasta encontrar los elementos correspondientes. Una alternativa al muestreo aleatorio simple es el **muestreo sistemático**. Por ejemplo, si se quiere una muestra de tamaño 50 de una población que tiene 5 000 elementos, se muestrea uno de cada 5 000/50 = 100 elementos de la población. En este caso, un muestreo sistemático consiste en seleccionar en forma aleatoria uno de los primeros 100 elementos de la lista de la población. Los otros se identifican empezando con el primer elemento muestreado y seleccionando cada 100o. elemento que siga en la lista. En efecto, los elementos de la muestra de 50 se identifican moviéndose sistemáticamente entre la población e identificando cada 100o. elemento después del primero seleccionado aleatoriamente. Por lo general, de esta manera es más fácil identificar la muestra de 50 que si se utilizara el muestreo aleatorio simple. Como el primer elemento que se selecciona es elegido al azar, se supone que una muestra sistemática tiene las propiedades de una muestra aleatoria simple. Este supuesto es aplicable, en especial, cuando la lista de los elementos de la población constituye un orden aleatorio de los elementos.

Muestreo de conveniencia

Los métodos de muestreo hasta ahora analizados se conocen como técnicas *probabilísticas de muestreo.* Los elementos seleccionados de una población tienen una probabilidad conocida de ser incluidos en la muestra. La ventaja del muestreo probabilístico estriba en que, por lo general, se identifica la distribución de muestreo del estadístico muestral correspondiente. Para determinar las propiedades de la distribución de muestreo se usan las fórmulas para el muestreo aleatorio simple presentadas en este capítulo. La distribución de muestreo permite plantear afirmaciones probabilísticas acerca del error asociado con el uso de los resultados muestrales al hacer inferencias de la población.

El **muestreo de conveniencia** es una técnica de *muestreo no probabilística.* Como el nombre lo indica, la muestra se determina principalmente por conveniencia. Los elementos se incluyen sin que haya una probabilidad previamente especificada o conocida de que sean incorporados en la muestra. Por ejemplo, un profesor que realiza una investigación en una universidad puede usar estudiantes voluntarios para que constituyan una muestra simplemente porque los tiene al alcance y participarán como sujetos a un costo bajo o sin costo. De manera similar, un inspector puede muestrear un cargamento de naranjas seleccionándolas al azar de varias cajas. Marcar cada naranja y usar un método probabilístico de muestreo puede no resultar práctico. Muestras como capturas en la vida salvaje y paneles de voluntarios en investigaciones del consumidor son también de conveniencia.

Esta técnica tiene la ventaja de que es relativamente fácil seleccionar la muestra y recabar los datos; sin embargo, es imposible evaluar su "bondad" en términos de representatividad de la población. Una muestra de conveniencia puede dar buenos resultados o no; ningún procedimiento justificado estadísticamente permite un análisis e inferencia probabilísticos acerca de la calidad de los resultados muestrales. Algunas veces los investigadores aplican los métodos estadísticos propios de muestras probabilísticas a las muestras de conveniencia con el argumento de que ésta se trata como si fuera una muestra probabilística. Sin embargo, estos argumentos no tienen fundamento y se debe tener cuidado al interpretar los resultados de muestreos de conveniencia que han sido utilizados para hacer inferencias acerca de las poblaciones.

Muestreo subjetivo

Otra técnica de muestreo no probabilística es el **muestreo subjetivo**. En este método la persona que más sabe sobre un asunto selecciona elementos de la población a los que considera los más representativos. Este método suele representar una manera relativamente fácil de seleccionar una muestra. Por ejemplo, un reportero puede elegir a dos o tres senadores considerando que éstos reflejan la opinión general de todos los senadores. Sin embargo, la calidad de los resultados muestrales depende de la persona que selecciona la muestra. Aquí también hay que tener mucho cuidado al hacer inferencias acerca de las poblaciones a partir de muestreos subjetivos.

NOTAS Y COMENTARIOS

Cuando se realizan muestreos de poblaciones finitas, se recomienda usar métodos de muestreo probabilístico: muestreo aleatorio simple, muestreo aleatorio estratificado, muestreo por conglomerados o muestreo sistemático. Existen fórmulas para evaluar la "bondad" de los resultados muestrales basadas en el uso de estos métodos en términos de la cercanía de los resultados a los parámetros poblacionales que se estiman. Con los muestreos de conveniencia o con los subjetivos no se puede estimar la bondad de los resultados. Por tanto, debe tenerse mucho cuidado al interpretar resultados basados en métodos de muestreo no probabilístico.

Resumen

En este capítulo se presentaron los conceptos de muestreo aleatorio simple y distribución de muestreo. Se describió cómo seleccionar una muestra aleatoria simple de una población finita y una muestra aleatoria de una población infinita. Los datos recolectados de tales muestras se pueden utilizar para obtener estimadores puntuales de los parámetros poblacionales. Ya que distintas muestras proporcionan valores diferentes de los estimadores puntuales, los estimadores puntuales como $\bar{x}$ y $\bar{p}$ son variables aleatorias. A la distribución de probabilidad de una variable aleatoria de este tipo se le conoce como *distribución de muestreo*. En particular, se describieron las distribuciones de muestreo de la media muestral $\bar{x}$ y de la proporción muestral $\bar{p}$.

Al estudiar las características de las distribuciones de muestreo de $\bar{x}$ y de $\bar{p}$, se estableció que $E(\bar{x}) = \mu$ y que $E(\bar{p}) = p$. Después de proporcionar las fórmulas para la desviación estándar o error estándar de dichos estimadores, se describieron las condiciones necesarias para que las distribuciones de muestreo de $\bar{x}$ y de $\bar{p}$ sigan una distribución normal. Otros métodos de muestreo que también se abordaron son el muestreo aleatorio estratificado, por conglomerados o clusters, sistemático, por conveniencia y subjetivo.

Glosario

Consistencia Propiedad de un estimador puntual que se hace presente siempre que muestras más grandes tienden a proporcionar estimaciones puntuales más cercanas al parámetro poblacional.

Distribución de muestreo o muestral Distribución de probabilidad que consta de todos los posibles valores de un estadístico muestral.

Eficiencia relativa Dados dos estimadores puntuales insesgados de un mismo parámetro poblacional, el estimador puntual con menor error estándar será más eficiente.

Error estándar Desviación estándar de un estimador puntual.

Estadístico muestral Característica muestral, por ejemplo, la media muestral $\bar{x}$, la desviación estándar muestral s, la proporción muestral $\bar{p}$, etc. El valor del estadístico muestral se utiliza para estimar el valor del parámetro poblacional correspondiente.

Estimación puntual Valor de un estimador que se utiliza en una situación particular como estimación del parámetro poblacional.

Estimador puntual Un estadístico muestral como $\bar{x}$, s o $\bar{p}$ que proporciona una estimación puntual del parámetro poblacional correspondiente.

Factor de corrección para una población finita Es el término $\sqrt{(N-n)/(N-1)}$ utilizado en las fórmulas de $\sigma_{\bar{x}}$ y $\sigma_{\bar{p}}$ siempre que se muestrea de una población finita y no de una población infinita. Sin embargo, hay una regla generalmente aceptada: ignorar el factor de corrección en una población finita siempre que $n/N \leq 0.05$.

Insesgadez Propiedad de un estimador puntual que se hace presente cuando el valor esperado del estimador es igual al parámetro poblacional que se estima.

Marco Lista de los elementos de donde se selecciona la muestra.

Muestreo aleatorio Muestra aleatoria de una población infinita seleccionada de manera tal que se satisfagan las condiciones siguientes: 1) cada elemento escogido proviene de la misma población y, 2) cada elemento se selecciona de manera independiente.

Muestreo aleatorio estratificado Método probabilístico en el que primero se divide la población en estratos y después se toma una muestra aleatoria simple de cada estrato.

Muestreo aleatorio simple Muestra aleatoria simple de tamaño n de una población finita de tamaño N seleccionada de manera que cada posible muestra de tamaño n tenga la misma probabilidad de ser seleccionada.

Muestreo con remplazo Una vez que un elemento se ha incluido en la muestra, se regresa a la población. Un elemento ya seleccionado puede nuevamente ser elegido y aparecer más de una vez en la muestra.

Muestreo de conveniencia Método no probabilístico en el que la selección de los elementos para la muestra es acorde con la conveniencia.

Muestreo por conglomerados o clusters Método probabilístico en el que primero se divide la población en conglomerados y después se toma una muestra aleatoria de éstos.

Muestreo sin remplazo Una vez que un elemento ha sido incluido en la muestra, se retira de la población y ya no se selecciona más.

Muestreo sistemático Método probabilístico en el que primero se selecciona uno de los primeros k elementos de una población y después cada k-ésimo elemento.

Muestreo subjetivo Método no probabilístico en el que la selección de los elementos para la muestra se realiza de acuerdo con la opinión de la persona que efectúa el estudio.

Parámetro Característica numérica de una población, por ejemplo, media poblacional μ, desviación estándar poblacional σ, proporción poblacional p, etcétera.

Población muestreada Población de la cual se extrae la muestra.

Población objetivo Es aquella de la cual se hacen inferencias estadísticas como estimaciones puntuales. Es importante que la población objetivo corresponda tan cercanamente como sea posible a la población muestreada.

Teorema del límite central Permite usar la distribución de probabilidad normal para aproximar la distribución de muestreo de $\bar{x}$ siempre que la muestra sea grande.

Fórmulas clave

Valor esperado de $\bar{x}$

$$E(\bar{x}) = \mu \tag{7.1}$$

Desviación estándar de $\bar{x}$ (error estándar)

Población finita

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Población infinita

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{7.2}$$

Valor esperado de $\bar{p}$

$$E(\bar{p}) = p \tag{7.4}$$

Desviación estándar de $\bar{p}$ (error estándar)

Población finita

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}}$$

Población infinita

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \qquad \textbf{(7.5)}$$

Ejercicios complementarios

42. *U. S. News & World Report* publica información extensa acerca de las mejores universidades de Estados Unidos (*America's Best Colleges*, ed. 2009). Entre otras cosas, proporciona una lista de las 133 mejores universidades a nivel nacional. Se desea tomar una muestra de tales instituciones para realizar un estudio de seguimiento de sus alumnos. Inicie en la parte inferior de la tercera columna de dígitos aleatorios de la tabla 7.1. Ignore los dos primeros dígitos de cada conjunto de cinco números usando números aleatorios de tres cifras. Empiece con 959, lea *hacia arriba* de la columna para identificar el número (de 1 a 133) de las siete primeras universidades a incluir en una muestra aleatoria simple. Continúe iniciando en la parte inferior de las columnas cuarta y quinta, y lea hacia arriba si es necesario.

43. Los estadounidenses están cada vez más preocupados por el aumento en los costos de Medicare. En 1990 el promedio de gastos anuales de un derechohabiente de Medicare ascendía a $3 267; en 2003 este promedio había aumentado a $6 883 (*Money*, otoño de 2003). Suponga que usted contrata a una firma de consultoría para tomar una muestra de 50 de los derechohabientes de Medicare en 2003 con objeto de investigar los gastos. Asuma que la desviación estándar poblacional en 2003 fue $2 000.
    a) Presente la distribución de muestreo de la cantidad media de los gastos de Medicare para una muestra de 50 derechohabientes en 2003.
    b) ¿Cuál es la probabilidad de que la media muestral no se aleje más de ±$300 de la media poblacional?
    c) ¿Cuál es la probabilidad de que la media muestral sea mayor que $7 500? Si la empresa que contrató le dice que la media muestral para los derechohabientes que entrevistó es $7 500, ¿dudaría de que la empresa contratada hubiera hecho un procedimiento de muestreo aleatorio simple adecuado? ¿Por qué?

44. *BusinessWeek* encuesta a exalumnos de administración 10 años después de terminados sus estudios *(BusinessWeek,* 22 de septiembre de 2003). Uno de sus hallazgos indica que gastan en promedio $115.50 semanales en comidas sociales. A usted se le pide que realice un estudio con una muestra de 40 de estos exalumnos. Asuma que la desviación estándar poblacional es $35.
    a) Presente la distribución de muestreo de $\bar{x}$, la media muestral de los gastos semanales de los 40 exalumnos de administración.
    b) ¿Cuál es la probabilidad de que la media muestral no se aleje en más o menos $10 de la media poblacional?
    c) Suponga que encuentra una media muestral de $100. ¿Cuál es la probabilidad de hallar una media muestral de $100 o menos? ¿Consideraría que los exalumnos de esta muestra son un grupo con un gasto inusualmente bajo? ¿Por qué?

45. El tiempo promedio que un estadounidense destina a ver televisión es de 15 horas por semana *(Money,* noviembre de 2003). Suponga que se toma una muestra de 60 estadounidenses para investigar con más detalle sus hábitos a este respecto. Asuma que la desviación estándar poblacional en las horas de televisión semanales es $\sigma = 4$ horas.
    a) ¿Cuál es la probabilidad de que la media muestral no se aleje más o menos de 1 hora de la media poblacional?
    b) ¿Cuál es la probabilidad de que la media muestral no se aleje más o menos de 45 minutos de la media poblacional?

46. Después de deducir los gastos necesarios, el costo promedio por asistir a la Universidad del Sur de California (USC) es de $27 175 (*U. S. News & World Report, America's Best Colleges*, ed. 2009). Suponga que la desviación estándar poblacional es $7 400. Asuma que se selecciona una muestra aleatoria de 60 estudiantes de la USC de esta población.
    a) ¿Cuál es el valor del error estándar de la media?
    b) ¿Cuál es la probabilidad de que la media muestral sea mayor que $27 175?

c) ¿Cuál es la probabilidad de que la media muestral no se aleje más o menos de $1 000 de la media poblacional?

d) ¿Qué tanto variaría la probabilidad del inciso *c*) si el tamaño de la muestra se aumentara a 100?

47. Tres empresas transportan inventarios de distintos tamaños. El inventario de la empresa *A* contiene 2 000 artículos, el de la empresa *B*, 5 000 artículos y el de la empresa *C*, 10 000 artículos. La desviación estándar poblacional de los costos de los artículos en los inventarios de estas empresas es $\sigma = 144$. Un consultor de estadística recomienda que cada compañía tome una muestra de 50 artículos de su inventario para obtener una estimación estadística válida del costo promedio por unidad. Los gerentes de la firma más pequeña opinan que, como su población es menor, se podrá hacer la estimación con una muestra mucho menor de la que se requiere para la empresa más grande. Sin embargo, el consultor opina que para tener el mismo error estándar y, por tanto, la misma precisión en los resultados muestrales, todas las compañías deberán emplear el mismo tamaño de muestra, sin importar el tamaño de la población.

a) Utilizando el factor de corrección para una población finita, calcule el error estándar de cada una de las tres empresas para un tamaño de muestra de 50.

b) ¿Cuál es la probabilidad para cada firma de que la media muestral $\bar{x}$ esté a no más de ± 25 de la media poblacional μ?

48. Un investigador reporta sus resultados diciendo que el error estándar de la media es 20 y la desviación estándar poblacional es 500.

a) ¿De qué tamaño fue la muestra utilizada en esta investigación?

b) ¿Cuál es la probabilidad de que la estimación puntual esté a no más de ± 25 de la media poblacional?

49. Un inspector de control de calidad vigila periódicamente un proceso de producción. El inspector selecciona muestras aleatorias simples de 30 artículos ya terminados y calcula la media muestral del peso del producto $\bar{x}$. Si en un periodo largo se encuentra que 5% de los valores de $\bar{x}$ son mayores que 2.1 libras y 5% son menores que 1.9 libras, ¿cuáles son la media y la desviación estándar de la población de los productos elaborados en este proceso?

50. Cerca de 28% de las empresas privadas tiene como propietario a una mujer (*The Cincinnati Enquirer,* 26 de enero de 2006). Responda estas preguntas con base en una muestra de 240 empresas privadas.

a) Desarrolle la distribución de muestreo de $\bar{p}$, la proporción muestral de las empresas propiedad de una mujer.

b) ¿Cuál es la probabilidad de que la proporción muestral esté a no más de ± 0.04 de la proporción poblacional?

c) ¿Cuál es la probabilidad de que la proporción muestral esté a no más de ± 0.02 de la proporción poblacional?

51. Una firma de investigación de mercados realiza encuestas telefónicas con una tasa histórica de respuesta de 40%. ¿Cuál es la probabilidad de que en una nueva muestra de 400 números telefónicos, por lo menos 150 personas cooperen y respondan las preguntas? En otras palabras, ¿cuál es la probabilidad de que la proporción muestral sea por lo menos $150/400 = 0.375$?

52. Los publicistas contratan a proveedores de servicios de Internet y motores de búsqueda para colocar sus anuncios en los sitios web. Pagan una cuota con base en el número de clientes potenciales que hacen clic en su publicidad. Desafortunadamente, el fraude por clic (la práctica de hacer clic en una publicidad con el solo objeto de aumentar las ganancias) se ha convertido en un problema. El 40% de los anunciantes se queja de haber sido víctima de fraude por clic (*BusinessWeek,* 13 de marzo de 2006). Suponga que se toma una muestra aleatoria de 380 publicistas con objeto de aprender más acerca de cómo son afectados por esta práctica.

a) ¿Cuál es la probabilidad de que la proporción muestral esté a no más de ± 0.04 de la proporción poblacional que ha experimentado fraude por clic?

b) ¿Cuál es la probabilidad de que la proporción muestral sea mayor que 0.45?

53. La proporción de personas aseguradas por All-Driver Automobile Insurance Company que contraen una multa de tráfico en el periodo de cinco años es 0.15.

a) Indique la distribución de muestreo de $\bar{p}$ si se emplea una muestra aleatoria de 150 asegurados para determinar la proporción de quienes han contraído por lo menos una multa.

b) ¿Cuál es la probabilidad de que la proporción muestral esté a no más de ± 0.03 de la proporción poblacional?

54. Lori Jeffrey es una exitosa representante de ventas de libros universitarios. Históricamente, ella consigue una adopción de libros de texto en 25% de sus llamadas de ventas. Considere sus telefonemas de ventas de un mes como muestra de todas sus posibles llamadas; suponga que en el análisis estadístico de los datos se encuentra que el error estándar de la proporción es 0.0625.
   a) ¿De qué tamaño fue la muestra que se utilizó en el análisis? Es decir, ¿cuántas llamadas hizo Lori Jeffrey en ese mes?
   b) Sea $\bar{p}$ la proporción muestral de adopciones de libros de texto en el mes. Presente la distribución de muestreo de $\bar{p}$.
   c) Mediante la distribución de muestreo de $\bar{p}$, calcule la probabilidad de que Lori logrará adopciones de libros de texto en 30% o más de sus llamadas de ventas en el lapso de un mes.

Apéndice 7.1 Valor esperado y desviación estándar de $\bar{x}$

En este apéndice se presentan las bases matemáticas de las expresiones $E(\bar{x})$, valor esperado de $\bar{x}$ dado en la ecuación (7.1), y $\sigma_{\bar{x}}$, la desviación estándar de $\bar{x}$ dada por la ecuación (7.2).

Valor esperado de $\bar{x}$

Se tiene una población con media μ y varianza σ^2. Se selecciona una muestra aleatoria simple de tamaño n cuyas observaciones individuales se denotan $x_1, x_2, \ldots, x_n$. La media muestral $\bar{x}$ se calcula como sigue.

$$\bar{x} = \frac{\Sigma x_i}{n}$$

Si se repiten los muestreos aleatorios simples de tamaño n, $\bar{x}$ será una variable aleatoria que tomará diferentes valores dependiendo de los n elementos que formen la muestra. El valor esperado de la variable aleatoria $\bar{x}$ es la media de todos los posibles valores de $\bar{x}$.

$$\begin{aligned}\text{Media de } \bar{x} = E(\bar{x}) &= E\left(\frac{\Sigma x_i}{n}\right) \\ &= \frac{1}{n}[E(x_1 + x_2 + \cdots + x_n)] \\ &= \frac{1}{n}[E(x_1) + E(x_2) + \cdots + E(x_n)]\end{aligned}$$

Para cada x_i se tiene $E(x_i) = \mu$; por tanto, escribimos

$$\begin{aligned}E(\bar{x}) &= \frac{1}{n}(\mu + \mu + \cdots + \mu) \\ &= \frac{1}{n}(n\mu) = \mu\end{aligned}$$

Este resultado indica que la media de todos los posibles valores de $\bar{x}$ es igual a la media poblacional μ. Es decir, $E(\bar{x}) = \mu$.

Desviación estándar de $\bar{x}$

Se tiene, de nuevo, una población con media μ y varianza σ^2, y una media muestral dada por

$$\bar{x} = \frac{\Sigma x_i}{n}$$

Se sabe que $\bar{x}$ es una variable aleatoria que toma distintos valores numéricos, con repetidas muestras aleatorias simples de tamaño *n,* dependiendo de los *n* elementos que integran la muestra. Lo que sigue es una derivación de la fórmula para la desviación estándar de los valores de $\bar{x}$, $\sigma_{\bar{x}}$, en el caso de que la población sea infinita. La deducción de la fórmula para $\sigma_{\bar{x}}$ cuando la población es finita y el muestreo se realiza sin remplazo es más complicada, y queda fuera de los alcances de este libro.

De vuelta al caso de una población infinita, recuerde que una muestra aleatoria simple de una población infinita consta de observaciones $x_1, x_2, \ldots, x_n$ que son independientes. Las dos expresiones siguientes son fórmulas generales para la varianza de variables aleatorias.

$$\text{Var}(ax) = a^2 \text{Var}(x)$$

donde a es una constante y x es una variable aleatoria, y

$$\text{Var}(x + y) = \text{Var}(x) + \text{Var}(y)$$

donde x y y son variables aleatorias *independientes*. Utilizando las dos ecuaciones anteriores, se puede deducir la fórmula para la varianza de la variable aleatoria $\bar{x}$ como sigue.

$$\text{Var}(\bar{x}) = \text{Var}\left(\frac{\Sigma x_i}{n}\right) = \text{Var}\left(\frac{1}{n}\Sigma x_i\right)$$

Entonces, como $1/n$ es una constante, tenemos

$$\begin{aligned}\text{Var}(\bar{x}) &= \left(\frac{1}{n}\right)^2 \text{Var}(\Sigma x_i) \\ &= \left(\frac{1}{n}\right)^2 \text{Var}(x_1 + x_2 + \cdots + x_n)\end{aligned}$$

En el caso de una población infinita, las variables aleatorias $x_1, x_2, \ldots, x_n$ son independientes, lo que permite escribir

$$\text{Var}(\bar{x}) = \left(\frac{1}{n}\right)^2 [\text{Var}(x_1) + \text{Var}(x_2) + \cdots + \text{Var}(x_n)]$$

Para toda x_i se tiene $\text{Var}(x_i) = \sigma^2$; por tanto, obtenemos

$$\text{Var}(\bar{x}) = \left(\frac{1}{n}\right)^2 (\sigma^2 + \sigma^2 + \cdots + \sigma^2]$$

Como en esta expresión hay n valores σ^2, tenemos

$$\text{Var}(\bar{x}) = \left(\frac{1}{n}\right)^2 (n\sigma^2) = \frac{\sigma^2}{n}$$

Calculando ahora la raíz cuadrada, se obtiene la fórmula de la desviación estándar de $\bar{x}$.

$$\sigma_{\bar{x}} = \sqrt{\text{Var}(\bar{x})} = \frac{\sigma}{\sqrt{n}}$$

Apéndice 7.2 Muestreo aleatorio con Minitab

Si en un archivo de Minitab se encuentra una lista con los elementos de una población, se puede usar dicho software para seleccionar una muestra aleatoria simple. Por ejemplo, en la columna 1 del conjunto de datos MetAreas se proporciona una lista de las 100 principales áreas metropolitanas de Estados Unidos y Canadá (*Places Rated Almanac–The Millenium Edition 2000*). La columna 2 contiene la puntuación general asignada a cada área. En la tabla 7.6 se presentan las primeras 10 áreas metropolitanas con sus puntuaciones correspondientes.

Suponga que pretende seleccionar una muestra aleatoria simple de 30 áreas metropolitanas con objeto de hacer un estudio sobre el costo de la vida en Estados Unidos y Canadá. Para seleccionar la muestra aleatoria se siguen los pasos que se indican a continuación.

Paso 1. Seleccione el menú desplegable **Calc**.
Paso 2. Elija **Random Data**.
Paso 3. Seleccione **Sample From Columns**.
Paso 4. Cuando el cuadro de diálogo Sample From Columns aparezca:
Ingrese 30 en el cuadro **Number of rows to sample**.
Introduzca Cl C2 en el cuadro **From columns** que se encuentra debajo.
Ingrese C3 C4 en el cuadro **Store samples in**.
Paso 5. Haga clic en **OK**.

La muestra aleatoria con las 30 áreas metropolitanas aparece en las columnas C3 y C4.

Apéndice 7.3 Muestreo aleatorio con Excel

Si en un archivo de Excel se encuentra una lista con los elementos de una población, se podrá usar dicho software para seleccionar una muestra aleatoria simple. Por ejemplo, en la columna A del conjunto de datos MetAreas se proporciona una lista de las 100 principales áreas metropolitanas de Estados Unidos y Canadá (*Places Rated Almanac–The Millenium Edition 2000*). La columna B contiene el rating general asignado a cada área. En la tabla 7.6 se presentan las primeras 10 áreas metropolitanas con sus puntuaciones correspondientes. Suponga que quiere seleccionar una muestra aleatoria simple de 30 áreas metropolitanas con objeto de hacer un estudio de profundidad sobre el costo de la vida en Estados Unidos y Canadá.

TABLA 7.6 Puntuación general para las primeras 10 áreas metropolitanas en el conjunto de datos MetAreas

Metropolitan Area	Rating
Albany, NY	64.18
Albuquerque, NM	66.16
Appleton, WI	60.56
Atlanta, GA	69.97
Austin, TX	71.48
Baltimore, MD	69.75
Birmingham, AL	69.59
Boise City, ID	68.36
Boston, MA	68.99
Buffalo, NY	66.10

WEB archivo
MetAreas

Las filas de cualquier conjunto de datos en Excel se pueden colocar en orden aleatorio agregando una columna al conjunto de datos y llenando la columna con números aleatorios mediante la función =RAND(). Después, con la herramienta de Excel para ordenar en forma ascendente aplicada a la columna de números aleatorios, las filas del conjunto de datos se reordenan de forma aleatoria. La muestra aleatoria de tamaño *n* aparecerá en las *n* primeras filas del conjunto de datos reordenado.

En el conjunto de datos MetAreas, los encabezados aparecen en la fila 1 y las 100 áreas metropolitanas se encuentran en las filas 2 a 101. Para seleccionar una muestra aleatoria de 30 áreas metropolitanas aplique los pasos siguientes.

Paso 1. Ingrese =RAND() en la celda C2.
Paso 2. Copie la celda C2 a las celdas C3:C101.
Paso 3. Seleccione cualquier celda de la columna C.
Paso 4. Haga clic en la ficha **Home** sobre la cinta.
Paso 5. En el grupo **Editing**, dé clic en **Sort & Filter**.
Paso 6. Haga clic en **Sort Smallest to Largest**.

La muestra aleatoria con 30 áreas metropolitanas aparecerá en las filas 2 a 31 del conjunto de datos reordenado. Los números aleatorios de la columna C ya no son necesarios y pueden borrarse si se desea.

Apéndice 7.4 Muestreo aleatorio con StatTools

WEB archivo
MetAreas

Si en un archivo de Excel se encuentra una lista con los elementos de una población, se podrá usar StatTools Random Sample Utility para seleccionar una muestra aleatoria simple. Por ejemplo, en la columna A del conjunto de datos MetAreas se proporciona una lista de las 100 principales áreas metropolitanas de Estados Unidos y Canadá (*Places Rated Almanac–The Millenium Edition 2000*). La columna B contiene la puntuación general asignada a cada área. Suponga que quiere seleccionar una muestra aleatoria simple de 30 áreas metropolitanas con objeto de hacer un estudio de profundidad sobre el costo de la vida en Estados Unidos y Canadá.

Se inicia con Data Set Manager a efecto de crear un conjunto de datos de StatTools utilizando el procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes se utilizan para generar una muestra aleatoria simple de 30 áreas metropolitanas.

Paso 1. Dé clic en la ficha **StatTools** sobre la cinta.
Paso 2. En el grupo **Data Group**, haga clic en **Data Utilities**.
Paso 3. Seleccione la opción **Random Sample**.
Paso 4. Cuando el cuadro de diálogo StatTools–Random Sample aparezca:
En la sección **Variables**:
Seleccione **Metropolitan Area**.
Elija **Rating**.
En la sección **Options**:
Ingrese 1 en el cuadro **Number of Samples**.
Ingrese 30 en el cuadro **Sample Size**.
Haga clic en **OK**.

La muestra aleatoria de 30 áreas metropolitanas aparecerá en las columnas A y B de la hoja de trabajo titulada *Random Sample*.

CAPÍTULO 8

Estimación por intervalo

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: FOOD LION

8.1 MEDIA POBLACIONAL: σ CONOCIDA
- Margen de error y estimación por intervalo
- Consejo práctico

8.2 MEDIA POBLACIONAL: σ DESCONOCIDA
- Margen de error y estimación por intervalo
- Consejo práctico
- Uso de una muestra pequeña
- Resumen de los procedimientos de estimación por intervalo

8.3 DETERMINACIÓN DEL TAMAÑO DE LA MUESTRA

8.4 PROPORCIÓN POBLACIONAL
- Determinación del tamaño de la muestra

ESTADÍSTICA *en* LA PRÁCTICA

FOOD LION*

SALISBURY, CAROLINA DEL NORTE

Food Lion, fundada en 1957 como Food Town, es una de las más grandes cadenas de supermercados de Estados Unidos, con 1 300 tiendas en 11 estados del sudeste y el Atlántico medio. La empresa vende más de 24 000 productos diferentes y ofrece mercancías de marcas publicitadas a nivel nacional y regional, así como una cantidad cada vez mayor de productos de gran calidad de marca propia especialmente fabricados por Food Lion. La cadena mantiene su liderazgo en precios bajos y asegura la calidad a partir de eficientes controles, entre ellos, formatos estándar de tienda, diseño innovador de los almacenes, instalaciones con uso eficiente de energía y sincronización de datos con los proveedores. Food Lion mira hacia un futuro de innovación continua, crecimiento, liderazgo en precios y servicios a sus clientes.

Siendo un negocio intensivo en inventarios, Food Lion decidió adoptar como forma de valuación de inventarios el método UEPS (último en entrar, primero en salir), el cual compara los costos y los ingresos actuales, minimizando los efectos de los cambios radicales de precios sobre los resultados de utilidad y pérdida. Además, el método UEPS reduce la utilidad neta, disminuyendo con ello los impuestos al ingreso o sobre la renta durante los periodos de inflación.

Food Lion establece un índice UEPS para cada uno de los siete grupos de inventario: abarrotes, papel/artículos para el hogar, artículos para mascotas, bienes para la salud y la belleza, lácteos, cigarros/tabaco y cervezas/vinos. Por ejemplo, un índice UEPS de 1.008 para el grupo de abarrotes indica que el valor de este inventario, a los costos actuales, refleja un aumento de 0.8% debido a la inflación en el último periodo de un año.

Un índice UEPS para cada grupo requiere que el inventario de final de año de cada producto sea valuado tanto al costo actual de final de año como al del año anterior. Para ahorrar tiempo y gastos excesivos por el conteo del inventario en las 1 200 tiendas, Food Lion selecciona una muestra aleatoria simple de 50 establecimientos. El inventario físico de final de año se realiza en cada una de las tiendas de la muestra. Para obtener el índice UEPS de cada uno de los grupos de inventario se utilizan los costos del año actual y del año anterior.



Pan fresco llegando al almacén de Food Lion.
© Jeff Greenberg/PhotoEdit.

En uno de los últimos años, la estimación muestral del índice UEPS para el inventario del grupo de productos de salud y belleza fue de 1.015. Con un nivel de confianza de 95%, Food Lion calculó un margen de error de 0.006 para la estimación muestral. Por tanto, el intervalo de 1.009 a 1.021 proporciona una estimación por intervalo de confianza de 95% del índice UEPS poblacional. Este nivel de precisión se consideró muy bueno.

En ese capítulo aprenderá cómo calcular el margen de error asociado con una estimación puntual. También verá cómo usar esta información para construir e interpretar estimaciones por intervalo para una media poblacional y una proporción poblacional.

* Los autores agradecen a Keith Cunningham, director de Impuestos, y a Bobby Harkey, del equipo de Contadores fiscales, por proporcionar este artículo para *Estadística en la práctica*.

En el capítulo 7 se dijo que un estimador puntual es un estadístico muestral que se usa para estimar un parámetro poblacional. Por ejemplo, la media muestral $\bar{x}$ es un estimador puntual de la media poblacional μ, y la proporción muestral $\bar{p}$ es un estimador puntual de la proporción poblacional p. Como no se puede esperar que dicho estadístico muestral suministre el valor exacto del parámetro poblacional, se suele calcular una **estimación por intervalo** al sumar y restar a la estimación puntual un cantidad llamada **margen de error**. La forma general de una estimación por intervalo es:

$$\text{Estimación puntual} \pm \text{margen de error}$$

El objetivo de la estimación por intervalo es aportar información sobre qué tan cerca se encuentra la estimación puntual obtenida de la muestra, del valor del parámetro poblacional.

En este capítulo se explica cómo obtener una estimación por intervalo para la media poblacional μ y para la proporción poblacional p. La fórmula general para obtener una estimación por intervalo de una media poblacional es la siguiente.

$$\bar{x} \pm \text{margen de error}$$

De manera similar, la fórmula general para obtener una estimación por intervalo de una proporción poblacional es la que se indica enseguida.

$$\bar{p} \pm \text{margen de error}$$

Las distribuciones muestrales o de muestreo de $\bar{x}$ y de $\bar{p}$ son clave para calcular estas estimaciones por intervalo.

8.1 Media poblacional: σ conocida

Para obtener una estimación por intervalo para la media poblacional se necesita la desviación estándar poblacional σ o la desviación estándar muestral s a efecto de calcular el margen de error. En la mayoría de los casos no se conoce σ, y para calcular el margen de error se emplea s. Sin embargo, en algunas aplicaciones se cuenta con una gran cantidad de datos anteriores (históricos) que se pueden usar para calcular la desviación estándar poblacional antes de tomar la muestra. También en aplicaciones sobre control de calidad, en las que se supone que el proceso se desarrolla correctamente o "en control", se considera que se conoce la desviación estándar. A tales situaciones se les denomina casos de ***σ* conocida**. En esta sección se presenta un ejemplo en el que es razonable considerar que se conoce σ y se muestra cómo construir una estimación por intervalo.

Cada semana, Lloyd's Department Store selecciona una muestra aleatoria simple de 100 clientes con objeto de conocer información acerca de la cantidad que gastan en cada visita a la tienda. Si x representa la cantidad gastada en cada visita a la tienda, la media muestral $\bar{x}$ es una estimación puntual de μ, la cantidad media gastada en cada visita a la tienda por la población integrada por los clientes de Lloyd's Department Store. La tienda ha realizado estos estudios semanales durante varios años. Con base en sus datos anteriores, supone que el valor conocido de la desviación estándar poblacional es $\sigma = \$20$. Los datos anteriores (históricos) indican también que la población tiene una distribución normal.

WEB archivo
Lloyd's

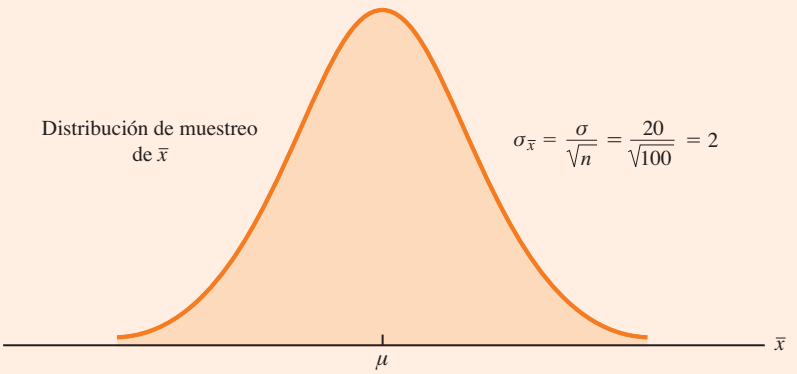
En la semana más reciente, en su estudio de 100 clientes ($n = 100$), Lloyd's obtuvo como media muestral $\bar{x} = \$82$. La media muestral de la cantidad gastada permite una estimación puntual de la media poblacional de la cantidad gastada en cada visita, μ. A continuación se explica cómo calcular un margen de error para esta estimación y cómo desarrollar una estimación por intervalo para la media poblacional.

Margen de error y estimación por intervalo

En el capítulo 7 se menciona que la distribución de muestreo de $\bar{x}$ sirve para calcular la probabilidad de que $\bar{x}$ esté dentro de una distancia dada de μ. En el ejemplo de Lloyd's, los datos históricos indican que la población constituida por las cantidades gastadas está distribuida normalmente y que su desviación estándar es $\sigma = 20$. De esta manera, utilizando lo aprendido en el capítulo 7, se puede concluir que la distribución de muestreo de $\bar{x}$ sigue una distribución normal con un error estándar de $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 20/\sqrt{100} = 2$. En la figura 8.1 se presenta esta distribución de muestreo.[1] Puesto que indica cómo están distribuidos los valores de $\bar{x}$ en torno a

[1] Se aprovecha que las cantidades gastadas tienen una distribución normal para concluir que la distribución de muestreo de $\bar{x}$ tiene una distribución normal. Si la población no la tuviera, se podría invocar el teorema del límite central, y el hecho de que el tamaño de la muestra es $n = 100$, para concluir que la distribución de muestreo de $\bar{x}$ es aproximadamente normal. De cualquier manera, esta distribución es como se observa en la figura 8.1.

FIGURA 8.1 Distribución de muestreo de la media muestral de las cantidades gastadas para muestras aleatorias simples de 100 clientes

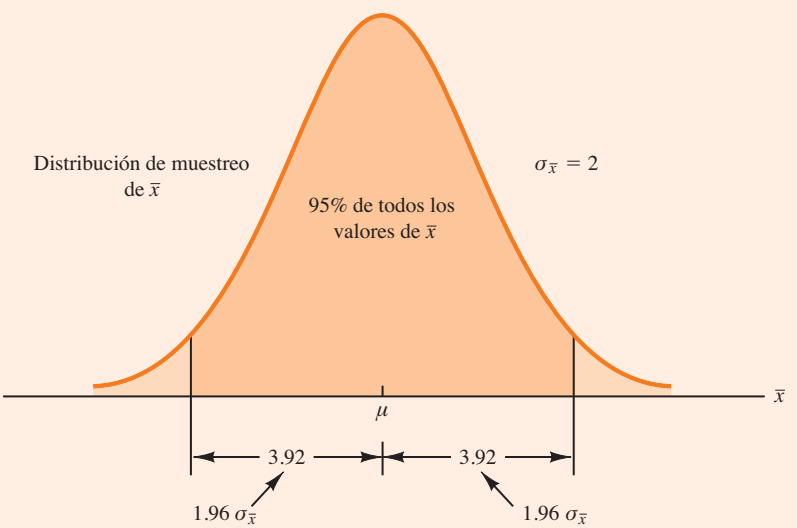


la media poblacional μ, la distribución de muestreo de $\bar{x}$ proporciona información acerca de la posible diferencia entre $\bar{x}$ y μ.

En la tabla de probabilidad normal estándar se encuentra que 95% de los valores de cualquier variable aleatoria distribuida normalmente aparecen dentro de ± 1.96 desviaciones estándar de la media. Por tanto, si la distribución de muestreo de $\bar{x}$ está distribuida normalmente, 95% de los valores de $\bar{x}$ deben estar dentro de $\pm 1.96\,\sigma_{\bar{x}}$ de la media μ. En el ejemplo de Lloyd's, se sabe que la distribución de muestreo de $\bar{x}$ está distribuida normalmente con un error estándar de $\sigma_{\bar{x}} = 2$. Como $\pm 1.96\,\sigma_{\bar{x}} = 1.96(2) = 3.92$, se puede concluir que 95% de los valores de $\bar{x}$ obtenidos usando muestras de $n = 100$ estarán dentro de ± 3.92 de la media poblacional μ. Vea la figura 8.2.

En la introducción a este capítulo se dijo que la fórmula general para estimar un intervalo

FIGURA 8.2 Distribución de muestreo de $\bar{x}$ que ilustra la ubicación de la media muestral que está dentro de 3,92 de μ

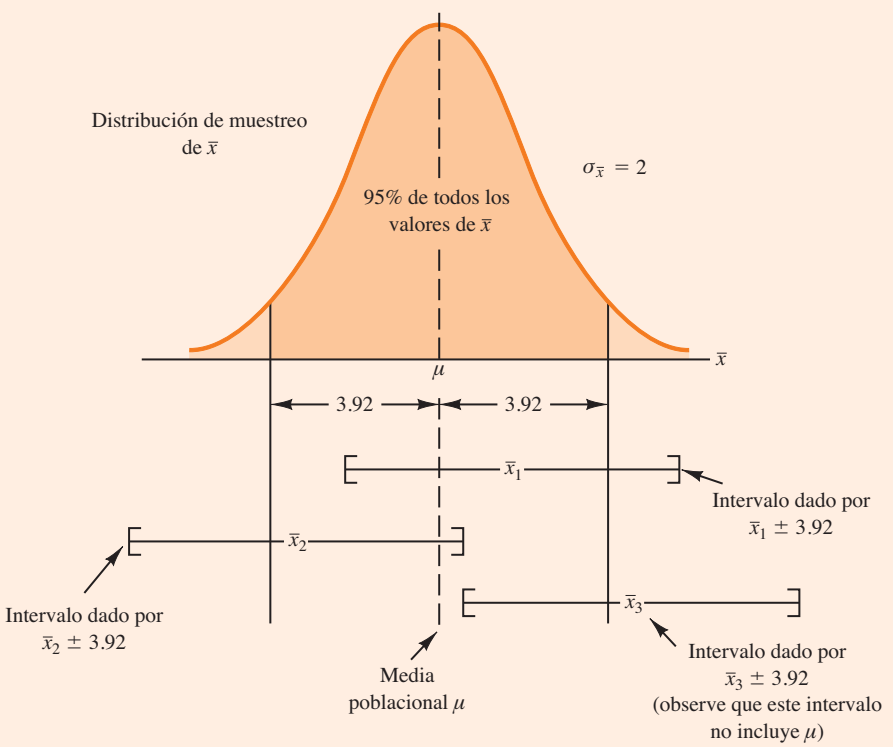


de la media poblacional μ es $\bar{x} \pm$ margen de error. En el ejemplo de Lloyd's, suponga que se establece 3.92 como margen de error y se calcula una estimación por intervalo para μ usando $\bar{x} \pm 3.92$. Para ver cómo se interpreta dicha estimación por intervalo, considere los valores de $\bar{x}$ que podrían obtenerse si se tomaran tres muestras aleatorias simples *diferentes*, cada una de 100 clientes de Lloyd's. La primera media muestral puede que dé el valor $\bar{x}_1$ de la figura 8.3. En este caso, como se ve en la figura, el intervalo que se obtiene al restar 3.92 de $\bar{x}_1$ y sumar 3.92 a $\bar{x}_1$ abarca la media poblacional μ. Ahora razone qué pasa si la segunda media muestral resulta tener el valor $\bar{x}_2$ que se observa en la figura 8.3. Aunque esta media muestral difiere de la primera, el intervalo obtenido al restar 3.92 de $\bar{x}_2$ y sumar 3.92 a $\bar{x}_2$ también comprende la media poblacional μ. Pero considere qué sucede si la tercera media muestral resulta tener el valor $\bar{x}_3$ que se indica en la figura 8.3. En este caso el intervalo obtenido al restar 3.92 de $\bar{x}_3$ y sumar 3.92 a $\bar{x}_3$ no abarca la media poblacional μ. Como $\bar{x}_3$ cae en la cola superior de la distribución de muestreo y dista más de 3.92 de μ, restando y sumando 3.92 a $\bar{x}_3$ se obtiene un intervalo que no incluye μ.

Con cualquier media muestral $\bar{x}$ que se encuentre dentro de la región sombreada en la figura 8.3 se obtendrá un intervalo que contenga la media poblacional μ. Como 95% de todas las posibles medias muestrales se ubican en la región sombreada más oscura, 95% de todos los intervalos que se obtengan al restar 3.92 de $\bar{x}$ y sumar 3.92 a $\bar{x}$ abarcarán la media poblacional μ.

Recuerde que en la última semana el equipo encargado de asegurar la calidad de Lloyd's encuestó a 100 clientes y obtuvo una media muestral de la cantidad gastada $\bar{x} = 82$. Utilizando $\bar{x} \pm 3.92$ para construir la estimación por intervalo, se obtiene 82 ± 3.92. Por tanto, la estima-

FIGURA 8.3 Intervalos obtenidos a partir de algunas medias muestrales localizadas en $\bar{x}_1$, $\bar{x}_2$ y $\bar{x}_3$



Este análisis ofrece una visión de porqué se le llama intervalo de confianza *de 95%.*

ción por intervalo de μ que se basa en los datos de la última semana va de $82 - 3.92 = 78.08$ a $82 + 3.92 = 85.92$. Como 95% todos los intervalos construidos usando $\bar{x} \pm 3.92$ contendrán la media poblacional, se tiene 95% de confianza de que el intervalo 78.08 a 85.92 contenga μ. Entonces dicho intervalo tiene un **nivel de confianza** de 95%. Al valor 0.95 se le conoce como **coeficiente de confianza**, y al intervalo 78.08 a 85.92 como **intervalo de confianza** de 95%.

Como el margen de error está dado por $z_{\alpha/2}(\sigma/\sqrt{n})$, la fórmula general de una estimación por intervalo de la media poblacional con σ conocida es la siguiente.

En el ejemplo de Lloyd's, mediante la expresión (8.1) se construye un intervalo de con-

ESTIMACIÓN POR INTERVALO DE LA MEDIA POBLACIONAL: σ CONOCIDA

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \textbf{(8.1)}$$

donde $(1 - \alpha)$ es el coeficiente de confianza y $z_{\alpha/2}$ es el valor de z que proporciona un área $\alpha/2$ en la cola superior de la distribución de probabilidad normal estándar.

fianza de 95% con un coeficiente de confianza $(1 - \alpha) = 0.95$ y, por tanto, $\alpha = 0.05$. En la tabla de distribución normal estándar se ve que un área de $\alpha/2 = 0.05/2 = 0.025$ en la cola superior corresponde a $z_{0.025} = 1.96$. Como en el ejemplo de Lloyd's, la media muestral es $\bar{x} = 82$, $\sigma = 20$ y el tamaño de la muestra es $n = 100$, se obtiene

$$82 \pm 1.96\frac{20}{\sqrt{100}}$$

$$82 \pm 3.92$$

Por tanto, al emplear la expresión (8.1), el margen de error es 3.92 y el intervalo de confianza de 95% va de $82 - 3.92 = 78.08$ a $82 + 3.92 = 85.92$.

Aunque a menudo se usa un nivel de confianza de 95%, también suelen utilizarse otros niveles, como 90 y 99%. En la tabla 8.1 se muestran los valores de $z_{\alpha/2}$ correspondientes a los niveles de confianza más utilizados. A partir de estos valores y de la expresión (8.1), el intervalo de confianza de 90% en el ejemplo de Lloyd's es

$$82 \pm 1.645\frac{20}{\sqrt{100}}$$

$$82 \pm 3.29$$

TABLA 8.1 Valores de $z_{\alpha/2}$ para los niveles de confianza más utilizados

Nivel de confianza	α	$\alpha/2$	$z_{\alpha/2}$
90%	0.10	0.05	1.645
95%	0.05	0.025	1.960
99%	0.01	0.005	2.576

Por tanto, para 90% de confianza, el margen de error es 3.29 y el intervalo de confianza es $82 - 3.29 = 78.71$ a $82 + 3.29 = 85.29$. De manera similar, el intervalo de 99% es

$$82 \pm 2.576\frac{20}{\sqrt{100}}$$

$$82 \pm 5.15$$

Entonces, para 99% de confianza el margen de error es 5.15 y el intervalo de confianza es $82 - 5.15 = 76.85$ a $82 + 5.15 = 87.15$.

Al comparar los resultados para los niveles de 90, 95 y 99%, es claro que para tener mayor grado de confianza, el margen de error, y con esto la amplitud del intervalo de confianza, debe ser mayor.

Consejo práctico

Si la población tiene una distribución normal, el intervalo de confianza que se obtiene con la expresión (8.1) es exacto. En otras palabras, si esta expresión se usa repetidas veces para generar intervalos de confianza de 95%, exactamente 95% de los intervalos generados contendrán la media poblacional. Si la población no tiene una distribución normal, el intervalo de confianza obtenido con la expresión (8.1) será aproximado. En tal caso, la calidad de la aproximación depende tanto de la distribución de la población como del tamaño de la muestra.

En la mayoría de las aplicaciones, cuando se utiliza la expresión (8.1), un tamaño de muestra $n \geq 30$ es adecuado para obtener una estimación por intervalo de la media poblacional. Si la población no está distribuida normalmente, pero es más o menos simétrica, puede esperarse que tamaños de muestra hasta de 15 proporcionen una buena aproximación del intervalo de confianza. Con tamaños menores, la expresión (8.1) sólo se debe usar si el analista cree, o está dispuesto a suponer, que la distribución de la población es al menos aproximadamente normal.

NOTAS Y COMENTARIOS

1. El procedimiento de estimación por intervalo estudiado en esta sección se basa en el supuesto de que la desviación estándar poblacional σ es conocida. Decir que σ es conocida significa que se cuenta con datos históricos o con otra información que permita obtener una buena estimación de la desviación estándar poblacional antes de tomar la muestra que se usará para obtener la estimación de la media poblacional. De manera que, técnicamente, esto no significa que σ se conozca con seguridad. Sólo significa que se obtuvo una buena estimación de la desviación estándar antes de tomar la muestra, y que de esta manera no se usará la misma muestra para estimar tanto la media como la desviación estándar poblacionales.

2. El tamaño de la muestra n aparece en el denominador de la expresión (8.1) para la estimación por intervalo. En consecuencia, si un determinado tamaño de muestra proporciona un intervalo demasiado amplio como para que tenga utilidad práctica, se debe considerar aumentar el tamaño de la muestra. Si n está en el denominador, con un tamaño de muestra mayor se obtendrá un margen de error menor, un intervalo más estrecho y mayor precisión. El procedimiento para determinar el tamaño de la muestra aleatoria simple que se necesita para obtener una determinada precisión se aborda en la sección 8.3.

Ejercicios

Métodos

1. En una muestra aleatoria simple de 40 artículos la media muestral obtenida es 25. La desviación estándar poblacional es $\sigma = 5$.
   a) ¿Cuál es el error estándar de la media, $\sigma_{\bar{x}}$?
   b) Con 95% de confianza, ¿cuál es el margen de error?

AUTOevaluación

2. En una muestra aleatoria simple de 50 artículos de una población en la que $\sigma = 6$, la media muestral resultante es 32.
   a) Proporcione un intervalo de confianza de 90% para la media poblacional.
   b) Calcule un intervalo de confianza de 95% para la media poblacional.
   c) Proporcione un intervalo de confianza de 99% para el mismo indicador.
3. En una muestra aleatoria simple de 60 artículos, la media muestral es 80. La desviación estándar poblacional es $\sigma = 15$.
   a) Calcule el intervalo de confianza de 95% para la media poblacional.
   b) Suponga que la misma media muestral se obtuvo de una muestra de 120 artículos. Proporcione el intervalo de confianza de 95% para la media poblacional.
   c) ¿Cuál es el efecto de una muestra de tamaño grande sobre la estimación por intervalo?
4. Para la media poblacional, el intervalo de confianza de 95% resultó de 152 a 160. Si $\sigma = 15$, ¿cuál es el tamaño de la muestra utilizada en este estudio?

Aplicaciones

AUTOevaluación

5. Con objeto de estimar la cantidad media que gasta un cliente en una comida en un importante restaurante de Atlanta, se recabaron los datos de una muestra de 49 comensales. Suponga que la desviación estándar de la población es $5.
   a) ¿Cuál es el margen de error para 95% de confianza?
   b) Si la media poblacional es $24.80, ¿cuál es el intervalo de confianza de 95% para la media poblacional?

WEBarchivo
Nielsen

6. Nielsen Media Research llevó a cabo un estudio para conocer cuánto tiempo se veía televisión en los hogares en el horario de 8:00 a 11:00 de la noche. Los datos que se encuentran en el archivo Nielsen son consistentes con los hallazgos reportados (*The World Almanac*, 2003). Con base en estudios anteriores, la desviación estándar poblacional se considera conocida y es $\sigma = 3.5$ horas. Proporcione una estimación mediante un intervalo de confianza de 95% para la media del tiempo que se ve televisión a la semana en el horario de referencia.
7. *The Wall Street Journal* informó que en 2008 los accidentes automovilísticos le costaron $162 mil millones a Estados Unidos (*The Wall Street Journal*, 5 de marzo de 2008). El costo promedio por persona de los accidentes automovilísticos en el área de Tampa, Florida, fue considerado de $1 599. Suponga que este costo promedio se basó en una muestra de 50 personas que estuvieron involucradas en dichos percances y que la desviación estándar poblacional es $\sigma = \$600$. ¿Cuál es el margen de error para un intervalo de 95% de confianza? ¿Qué recomendaría si el estudio requiriera un margen de error de $150 o menos?
8. The National Quality Research Center, de la Universidad de Michigan, proporciona medidas trimestrales de las opiniones de los consumidores acerca de ciertos bienes y servicios (*The Wall Street Journal*, 18 de febrero de 2003). En una encuesta sobre 10 restaurantes de comida rápida y pizza, la media muestral del índice de satisfacción del cliente fue 71. Datos anteriores indican que la desviación estándar poblacional ha sido relativamente estable, con $\sigma = 5$.
   a) ¿Qué debe estar dispuesto a asumir el investigador para considerar si un margen de error es deseable?
   b) Con 95% de confianza, ¿cuál es el margen de error?
   c) ¿Cuál es el margen de error si se desea 99% de confianza?

WEBarchivo
TaxReturn

9. La AARP dio a conocer un estudio para saber cuánto tardan las personas físicas en preparar su decla-ración federal de impuestos sobre la renta (*AARP Bulletin*, abril de 2008). Los datos contenidos en el archivo TaxReturn son congruentes con los resultados del estudio, y proporcionan el tiempo en horas requerido por 40 personas para completar su declaración federal de impuestos sobre la renta. Con base en datos de años anteriores, se asume que la desviación estándar poblacional es $\sigma = 9$ horas. ¿Cuál es la estimación mediante un intervalo de confianza de 95% para la media del tiempo que demoran las personas en completar su declaración fiscal?
10. La revista *Playbill* reportó que el ingreso familiar anual medio de sus suscriptores es $119 155 (*Playbill,* enero de 2006). Suponga que la estimación del ingreso familiar anual medio está basada en una muestra de 80 familias y que por datos de estudios anteriores la desviación estándar poblacional es conocida y es $\sigma = \$30\,000$.

a) Proporcione un intervalo de estimación de 90% de confianza para la media poblacional.
b) Calcule un intervalo de 95%.
c) Proporcione ahora un intervalo de estimación de 99%.
d) ¿Qué le sucede a la amplitud del intervalo de confianza a medida que el nivel de confianza aumenta? ¿Parece esto razonable? Explique.

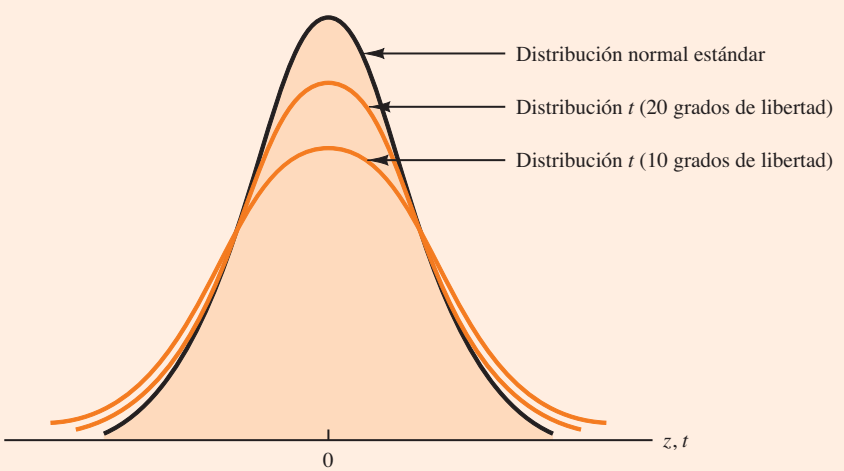
8.2 Media poblacional: σ desconocida

Cuando se calcula una estimación por intervalo para la media poblacional, suele no contarse con una buena estimación de la desviación estándar poblacional. En tales casos se usa la misma muestra para calcular μ y σ. Esta situación se conoce como **σ desconocida**. Cuando se utiliza s para estimar σ, el margen de error y la estimación por intervalo de la media poblacional se basan en una distribución de probabilidad conocida como **distribución *t***. Aunque el desarrollo matemático de esta última parte del supuesto de que la población muestreada tiene una distribución normal, las investigaciones han demostrado que la distribución t se aplica en muchas situaciones en que la población se desvía significantemente de la normal. Más adelante, en esta misma sección se proporcionan lineamientos para usar la distribución t cuando la población no está distribuida normalmente.

William Sealy Gosset, quien publicaba bajo el seudónimo "Student" es el creador de la distribución t. Gosset, que había estudiado matemáticas en Oxford, trabajaba para Guinness Brewery en Dublín, Irlanda. Desarrolló la distribución t cuando trabajaba sobre materiales a pequeña escala y con experimentos de temperatura.

La distribución t es una familia de distribuciones de probabilidad similar, y cada una depende de un parámetro conocido como **grados de libertad**. La distribución t para un grado de libertad es única, como lo es para dos grados o tres grados de libertad, etc. A medida que este número aumenta, la diferencia entre la distribución t y la distribución normal estándar se reduce. En la figura 8.4 se muestran las distribuciones t para 10 y 20 grados de libertad y su relación con la distribución de probabilidad normal estándar. Observe que una distribución t con más

FIGURA 8.4 Comparación de la distribución normal estándar con las distribuciones t para 10 y 20 grados de libertad



grados de libertad exhibe menos variabilidad y un mayor parecido con la distribución normal estándar. Note también que la media de toda distribución t es cero.

Para denotar el área en la cola superior de la distribución t, a la t se le coloca un subíndice. Por ejemplo, así como se usó $z_{0.025}$ para indicar el valor de z que deja en la cola superior de la distribución normal estándar un área de 0.025, también se usará $t_{0.025}$ para indicar el valor de t que deja en la cola superior de la distribución t un área de 0.025. En general, se manejará la notación $t_{\alpha/2}$ para representar el valor de t que deja un área de $\alpha/2$ en la cola superior de la distribución t (figura 8.5).

A medida que los grados de libertad aumentan, la distribución t se aproxima más a la distribución normal estándar.

La tabla 2 del apéndice B contiene una distribución t. En la tabla 8.2 se muestra una parte. Cada fila corresponde a una distribución t distinta con los grados de libertad que se indican. Por ejemplo, en la distribución t con 9 grados de libertad, $t_{0.025} = 2.262$. De manera similar, en la distribución t con 60 grados de libertad, $t_{0.025} = 2.000$. A medida que estos grados aumentan, $t_{0.025}$ se aproxima a $z_{0.025} = 1.96$. En efecto, el valor z de la distribución normal estándar se encuentra en la fila correspondiente a infinitos grados de libertad (etiquetado como ∞) de la tabla de distribuciones t. Si los grados de libertad son más de 100, se puede usar la fila correspondiente a infinitos grados para aproximar el verdadero valor de t; en otras palabras, para más de 100 grados de libertad, el valor z normal estándar proporciona una buena aproximación del valor t.

Margen de error y estimación por intervalo

En la sección 8.1 se mostró que la estimación por intervalo de la media poblacional cuando σ es conocida es

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Para calcular una estimación por intervalo de μ cuando no se conoce σ, se usa la desviación estándar muestral s para estimar σ, y $z_{\alpha/2}$ se sustituye por el valor $t_{\alpha/2}$ de la distribución t. El

FIGURA 8.5 Distribución t con un área o probabilidad $\alpha/2$ en la cola superior

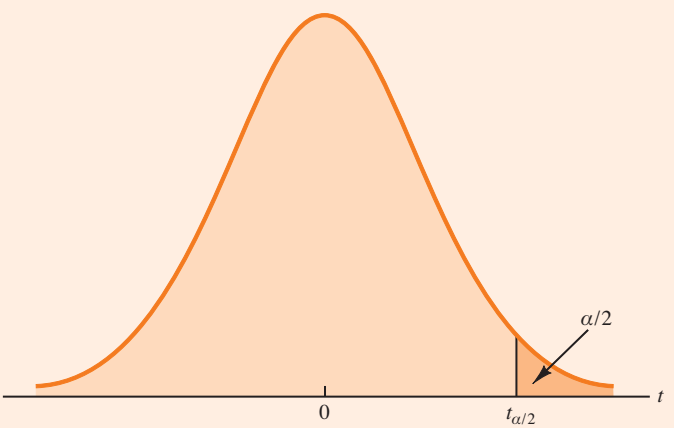
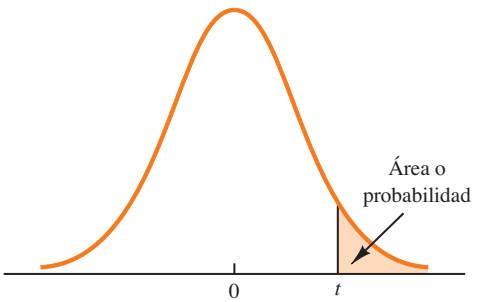


TABLA 8.2 Valores seleccionados de la tabla de distribución *t**



Grados de libertad	Área en la cola superior					
	0.20	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
1	1.376	3.078	6.314	12.706	31.821	63.656
2	1.061	1.886	2.920	4.303	6.965	9.925
3	0.978	1.638	2.353	3.182	4.541	5.841
4	0.941	1.533	2.132	2.776	3.747	4.604
5	0.920	1.476	2.015	2.571	3.365	4.032
6	0.906	1.440	1.943	2.447	3.143	3.707
7	0.896	1.415	1.895	2.365	2.998	3.499
8	0.889	1.397	1.860	2.306	2.896	3.355
9	0.883	1.383	1.833	2.262	2.821	3.250
⋮	⋮	⋮	⋮	⋮	⋮	⋮
60	0.848	1.296	1.671	2.000	2.390	2.660
61	0.848	1.296	1.670	2.000	2.389	2.659
62	0.847	1.295	1.670	1.999	2.388	2.657
63	0.847	1.295	1.669	1.998	2.387	2.656
64	0.847	1.295	1.669	1.998	2.386	2.655
65	0.847	1.295	1.669	1.997	2.385	2.654
66	0.847	1.295	1.668	1.997	2.384	2.652
67	0.847	1.294	1.668	1.996	2.383	2.651
68	0.847	1.294	1.668	1.995	2.382	2.650
69	0.847	1.294	1.667	1.995	2.382	2.649
⋮	⋮	⋮	⋮	⋮	⋮	⋮
90	0.846	1.291	1.662	1.987	2.368	2.632
91	0.846	1.291	1.662	1.986	2.368	2.631
92	0.846	1.291	1.662	1.986	2.368	2.630
93	0.846	1.291	1.661	1.986	2.367	2.630
94	0.845	1.291	1.661	1.986	2.367	2.629
95	0.845	1.291	1.661	1.985	2.366	2.629
96	0.845	1.290	1.661	1.985	2.366	2.628
97	0.845	1.290	1.661	1.985	2.365	2.627
98	0.845	1.290	1.661	1.984	2.365	2.627
99	0.845	1.290	1.660	1.984	2.364	2.626
100	0.845	1.290	1.660	1.984	2.364	2.626
∞	0.842	1.282	1.645	1.960	2.326	2.576

* *Nota.* Una versión más extensa es la tabla 2 del apéndice B.

margen de error está dado, entonces, por $t_{\alpha/2}s/\sqrt{n}$. Con este margen, la expresión general para una estimación por intervalo de la media poblacional cuando σ no se conoce es la siguiente.

ESTIMACIÓN POR INTERVALO DE LA MEDIA POBLACIONAL: σ DESCONOCIDA

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \qquad \textbf{(8.2)}$$

donde s es la desviación estándar muestral, $(1 - \alpha)$ es el coeficiente de confianza y $t_{\alpha/2}$ es el valor de t que proporciona un área $\alpha/2$ en la cola superior de la distribución t con $n - 1$ grados de libertad.

La razón por la que el número de grados de libertad para el valor de t en la expresión (8.2) sea $n - 1$ se debe al uso de s como estimación de la desviación estándar poblacional σ. La expresión para calcular la desviación estándar muestral es

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}}$$

Los grados de libertad se refieren al número de valores independientes en el cálculo de $\Sigma(x_i - \bar{x})^2$. Los n valores en este cálculo son los siguientes: $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$. En la sección 3.2 se indicó que en cualquier conjunto de datos $\Sigma(x_i - \bar{x}) = 0$. Por tanto, únicamente $n - 1$ de los valores $x_i - \bar{x}$ son independientes; es decir, si se conocen $n - 1$ de estos valores, el valor restante puede determinarse exactamente usando la condición de que $x_i - \bar{x}$ debe sumar 0. Entonces, $n - 1$ es el número de grados de libertad en la suma $\Sigma(x_i - \bar{x})^2$ y de ahí el número de grados de libertad para la distribución t en la expresión (8.2).

Para ilustrar la estimación por intervalo en el caso de σ desconocida, se considerará un estudio realizado para estimar la media del adeudo en las tarjetas de crédito en la población de familias de Estados Unidos. En la tabla 8.3 se presentan los saldos en las tarjetas de crédito de una muestra de $n = 70$ familias. En esta ocasión no se cuenta con una estimación previa de la desviación estándar poblacional σ. Por tanto, deberán utilizarse los datos muestrales para estimar tanto la media como la desviación estándar poblacionales. Con los datos de la tabla 8.3 calculamos la media muestral $\bar{x} = \$9312$ y la desviación estándar muestral $s = \$4007$. Con 95% de confianza y $n - 1 = 69$ grados de libertad podemos usar la tabla 8.2 para obtener

TABLA 8.3 Saldos en las tarjetas de crédito de una muestra de 70 familias

9430	14661	7159	9071	9691	11032
7535	12195	8137	3603	11448	6525
4078	10544	9467	16804	8279	5239
5604	13659	12595	13479	5649	6195
5179	7061	7917	14044	11298	12584
4416	6245	11346	6817	4353	15415
10676	13021	12806	6845	3467	15917
1627	9719	4972	10493	6191	12591
10112	2200	11356	615	12851	9743
6567	10746	7117	13627	5337	10324
13627	12744	9465	12557	8372	
18719	5742	19263	6232	7445	

WEB archivo
NewBalance

el valor apropiado de $t_{0.025}$. El valor de t que se necesita está en la fila que indica 69 grados de libertad y en la columna correspondiente a 0.025 en la cola superior. El valor que se encuentra en $t_{0.025} = 1.995$.

Con la expresión (8.2) para calcular la estimación por intervalo de la media poblacional de los saldos en las tarjetas de crédito tenemos:

$$9\,312 \pm 1.995\frac{4\,007}{\sqrt{70}}$$

$$9\,312 \pm 955$$

La estimación puntual de la media poblacional es $9 312, el margen de error es $955 y el intervalo de confianza de 95% va de 9 312 − 955 = $8 357 a 9 312 + 955 = $10 267. En consecuencia, se tiene 95% de confianza de que la media de los saldos en las tarjetas de crédito de la población de todas las familias está entre $8 357 y $10 267.

En los apéndices 8.1, 8.2 y 8.3 se describen los procedimientos para obtener un intervalo de confianza para la media poblacional usando Minitab, Excel y StatTools. En la figura 8.6 se presentan los resultados para el estudio de los saldos en las tarjetas de crédito que da el procedimiento de Minitab para la estimación por intervalo. Con la muestra de 70 familias se obtiene una media muestral de $9 312 para los saldos en las tarjetas de crédito, una desviación estándar muestral de $4 007, un error estándar de la media de $479 (valor redondeado) y un intervalo de confianza de 95%, que va de $8 357 a $10 267.

Consejo práctico

Si la población tiene una distribución normal, el intervalo de confianza suministrado en la expresión (8.2) es exacto y se puede usar con cualquier tamaño de muestra. Si la población no sigue una distribución normal, el intervalo de confianza en la expresión (8.2) será aproximado. En este caso la calidad de la aproximación depende tanto de la distribución de la población como del tamaño de la muestra.

Cuando la distribución de la población es altamente sesgada o hay observaciones atípicas, se requieren muestras grandes.

En la mayoría de las aplicaciones, un tamaño de muestra $n \geq 30$ es suficiente al usar la expresión (8.2) para obtener una estimación por intervalo de la media poblacional. Sin embargo, si la distribución de la población es muy sesgada o si hay observaciones atípicas, la mayoría de los especialistas en estadística recomienda un tamaño de muestra de 50 o más. Si la población no tiene una distribución normal pero es más o menos simétrica, con un tamaño de muestra de 15 puede esperarse una buena aproximación al intervalo de confianza. Con muestras más pequeñas la expresión (8.2) sólo debe usarse si el analista cree, o está dispuesto a suponer, que la distribución de la población es por lo menos aproximadamente normal.

Uso de una muestra pequeña

En el ejemplo siguiente se desarrolla una estimación por intervalo para una media poblacional manejando una muestra pequeña. Como ya se indicó, conocer la distribución de la población es importante para decidir si mediante una estimación por intervalo se obtendrán resultados aceptables.

Scheer Industries considera un nuevo programa asistido por computadora destinado a capacitar a los empleados de mantenimiento para reparar las máquinas. Con objeto de evaluar

FIGURA 8.6 Intervalo de confianza de Minitab para el estudio de los saldos en las tarjetas de crédito

Variable	N	Mean	StDev	SE Mean	95% CI
NewBalance	70	9312	4007	479	(8357, 10267)

TABLA 8.4 Duración de la capacitación, en días, para la muestra de 20 empleados de Scheer Industries

WEB archivo
Scheer

52	59	54	42
44	50	42	48
55	54	60	55
44	62	62	57
45	46	43	56

este programa, el director de manufactura solicita una estimación de la media poblacional del tiempo requerido para que los empleados de mantenimiento completen la capacitación asistida por computadora.

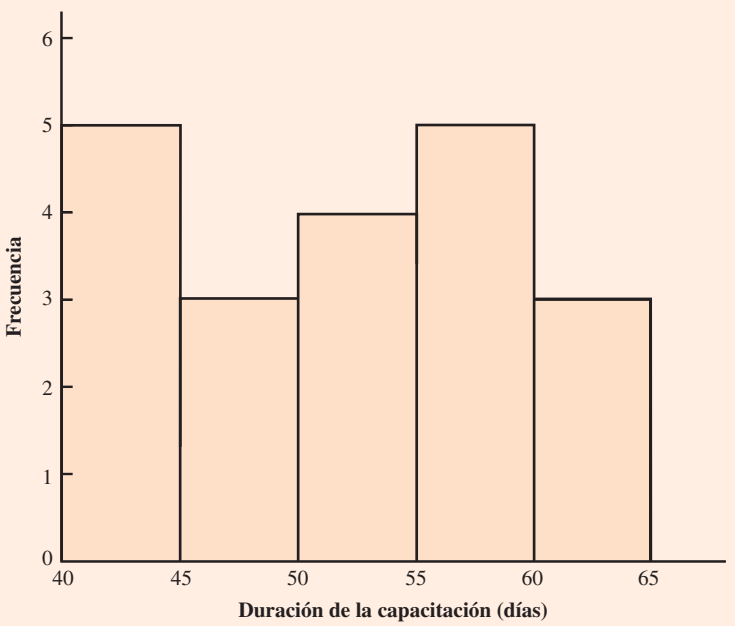
Considere una muestra de 20 individuos que siguen el programa de capacitación. En la tabla 8.4 se muestran los datos del tiempo, en días, que necesitó cada uno para completar el programa. En la figura 8.7 aparece un histograma de los datos. Con base en éste, ¿qué se puede decir de la distribución de la población? Primero, con base en los datos muestrales, no es posible concluir que la población sea normal, si bien no se tienen evidencias de sesgo o de observaciones atípicas. Por tanto, mediante los lineamientos de la subsección anterior, se concluye que una estimación por intervalo basada en la distribución t parece ser aceptable para esta muestra de 20 empleados.

A continuación se calcula la media muestral y la desviación estándar muestral.

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{1\,030}{20} = 51.5 \text{ días}$$

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{889}{20 - 1}} = 6.84 \text{ días}$$

FIGURA 8.7 Histograma sobre la duración de la capacitación en la muestra de Scheer Industries



Para dar un intervalo de confianza de 95%, se usa la tabla 2 del apéndice B y $n - 1 = 19$ grados de libertad y se obtiene $t_{0.025} = 2.093$. La expresión (8.2) suministra la estimación por intervalo de la media poblacional.

$$51.5 \pm 2.093\left(\frac{6.84}{\sqrt{20}}\right)$$

$$51.5 \pm 3.2$$

La estimación puntual de la media poblacional es 51.5 días. El margen de error es 3.2 días y el intervalo de confianza de 95% va de $51.5 - 3.2 = 48.3$ días a $51.5 + 3.2 = 54.7$ días.

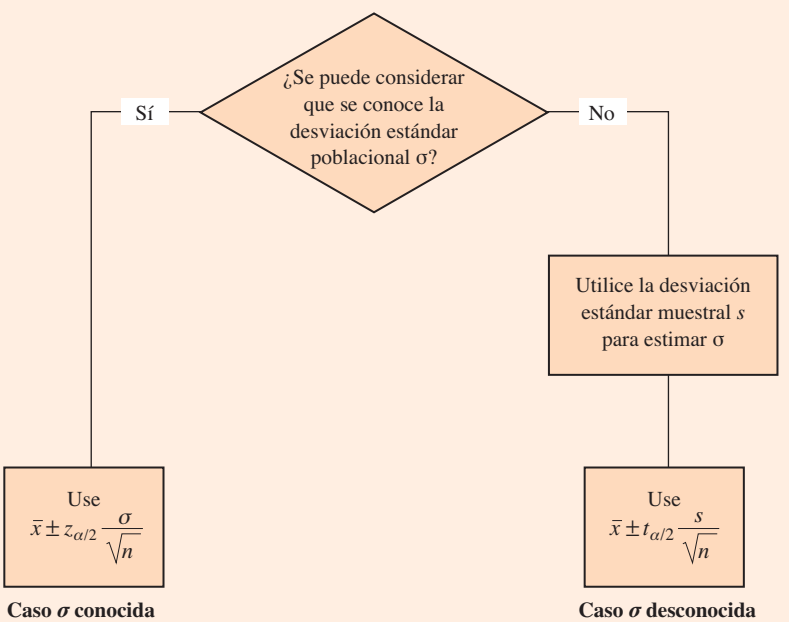
Usar un histograma de los datos muestrales para tener información acerca de la distribución de la población no es siempre concluyente, pero en muchos casos es la única información disponible. El histograma, junto con la opinión del analista, suele utilizarse para decidir si es adecuado usar la expresión (8.2) para obtener una estimación por intervalo.

Resumen de los procedimientos de estimación por intervalo

Se presentaron dos métodos para calcular una estimación por intervalo de la media poblacional. En el caso en que σ es conocida, en la expresión (8.1) se usan σ y la distribución normal estándar para calcular el margen de error y la estimación por intervalo. En el caso en que σ no es conocida, en la expresión (8.2) se utilizan la desviación estándar muestral s y la distribución t para calcular el margen de error y desarrollar la estimación por intervalo.

En la figura 8.8 se presenta un resumen de los procedimientos para la estimación por intervalo de los dos casos. En la mayoría de las aplicaciones, un tamaño de muestra $n \geq 30$ es adecuado. Sin embargo, si la población tiene distribución normal o aproximadamente normal,

FIGURA 8.8 Resumen de los procedimientos para la estimación por intervalo de la media poblacional



se pueden usar tamaños de muestra menores. En caso de que no se conozca σ y si la distribución de la población es muy sesgada o existen observaciones atípicas, se recomienda que el tamaño de la muestra sea $n \geq 50$.

NOTAS Y COMENTARIOS

1. En los casos en que conoce σ, el margen de error, $z_{\alpha/2}(\sigma/\sqrt{n})$, es fijo y es el mismo para todas las muestras de tamaño n. Cuando σ no se conoce, el margen de error, $t_{\alpha/2}(s/\sqrt{n})$, varía de una muestra a otra. Esta variación se debe a que la desviación estándar muestral s cambia de acuerdo con la muestra que se seleccione. Si s es grande, se obtiene un margen de error mayor, mientras que si s es pequeña, se obtiene un margen de error menor.
2. ¿Qué sucede con las estimaciones por intervalo cuando la población es sesgada? Considere una población sesgada a la derecha en la cual los datos con valores grandes jalan la distribución hacia esa dirección. Cuando existe un sesgo así, hay una correlación positiva entre la media muestral $\bar{x}$ y la desviación estándar muestral s. Valores mayores de s tienden a corresponderse con valores mayores de $\bar{x}$. De esta manera, cuando $\bar{x}$ es mayor que la media poblacional, s tiende a ser mayor que σ. Este sesgo hace que el margen de error, $t_{\alpha/2}(s/\sqrt{n})$, sea mayor de lo que sería si se conociera σ. Un intervalo de confianza con un margen de error mayor tenderá a incluir con más frecuencia la media poblacional μ que si se usara el verdadero valor σ. Pero cuando $\bar{x}$ es menor que la media poblacional, la correlación entre $\bar{x}$ y s hace que el margen de error sea más pequeño. En este caso, dichos intervalos de confianza con menor margen de error incluirán la media poblacional menos veces que si se conociera y se usara σ. Por esta razón se recomienda usar tamaños de muestra más grandes cuando la distribución de la población es muy sesgada.

Ejercicios

Métodos

11. En la distribución t con 16 grados de libertad, encuentre el área, o la probabilidad, de cada una de las regiones siguientes.
    a) A la derecha de 2.120
    b) A la izquierda de 1.337
    c) A la izquierda de -1.746
    d) A la derecha de 2.583
    e) Entre -2.120 y 2.120
    f) Entre -1.746 y 1.746

12. Encuentre los valores de t para las situaciones siguientes.
    a) Un área de 0.025 en la cola superior, con 12 grados de libertad.
    b) Un área de 0.05 en la cola inferior, con 50 grados de libertad.
    c) Un área de 0.01 en la cola superior, con 30 grados de libertad.
    d) Entre los que queda 90% del área, con 25 grados de libertad.
    e) Entre los que queda 95% del área, con 45 grados de libertad.

AUTOevaluación

13. Los datos muestrales siguientes provienen de una población normal: 10, 8, 12, 15, 13, 11, 6, 5.
    a) ¿Cuál es la estimación puntual de la media poblacional?
    b) ¿Cuál es la estimación puntual de la desviación estándar poblacional?
    c) Con 95% de confianza, ¿cuál es el margen de error para la estimación de la media poblacional?
    d) ¿Cuál es el intervalo de confianza de 95% para la media poblacional?

14. En una muestra aleatoria simple con $n = 54$, la media muestral es 22.5 y la desviación estándar muestral es 4.4.
    a) Proporcione un intervalo de confianza de 90% para la media poblacional.
    b) Determine un intervalo de confianza de 95% para la media poblacional.

c) Proporcione un intervalo de confianza de 99% para la media poblacional.
d) ¿Qué pasa con el margen de error y con el intervalo de confianza a medida que aumenta el nivel de confianza?

Aplicaciones

AUTO evaluación

15. Los agentes de ventas de Skillings Distributors presentan un informe semanal que enumera a los clientes contactados durante la semana. En una muestra de 65 informes, la media muestral es 19.5 clientes por semana. La desviación estándar muestral es 5.2. Proporcione intervalos de confianza de 90 y 95% para la media poblacional del número de clientes contactados semanalmente por el personal de ventas.

16. El número medio de horas de vuelo de los pilotos de Continental Airlines es 49 horas por mes (*The Wall Street Journal,* 25 de febrero de 2003). Suponga que esta media se basó en las horas de vuelo de una muestra de 100 pilotos de esa empresa y que la desviación estándar muestral es de 8.5 horas.
a) A 95% de confianza, ¿cuál es el margen de error?
b) Proporcione el intervalo de estimación de 95% de confianza para la media poblacional de las horas de vuelo de los pilotos.
c) La media en las horas de vuelo de los pilotos de United Airlines es de 36 horas por mes. Use los resultados del inciso *b*) para analizar la diferencia entre la cantidad de horas de vuelo de los pilotos en las dos líneas aéreas. *The Wall Street Journal* informa que United Airlines tiene el costo laboral más alto de todas las aerolíneas. La información proporcionada en estos ejercicios, ¿sirve para entender por qué se puede esperar que esta empresa tenga los costos más altos?

17. La International Air Transport Association realiza encuestas entre los viajeros de negocios en las que se califica la calidad de los aeropuertos de salida internacional. La calificación máxima es 10. Se seleccionó una muestra aleatoria simple de 50 viajeros de negocios y a cada uno se le solicitó su evaluación para el aeropuerto internacional de Miami. Las calificaciones que proporcionaron estos 50 viajeros se muestran a continuación.

WEB archivo
Miami

6	4	6	8	7	7	6	3	3	8	10	4	8
7	8	7	5	9	5	8	4	3	8	5	5	4
4	4	8	4	5	6	2	5	9	9	8	4	8
9	9	5	9	7	8	3	10	8	9	6		

Proporcione la estimación por intervalo de confianza de 95% para la media poblacional de las calificaciones al aeropuerto de Miami.

WEB archivo
JobSearch

18. Con frecuencia, las personas mayores pasan momentos muy difíciles buscando empleo. La AARP ha reportado el número de semanas que toma a los trabajadores de más de 55 años colocarse en un puesto. Los datos en número de semanas dedicadas a buscar empleo que se encuentran en el archivo JobSearch son congruentes con los hallazgos de la AARP (*AARP Bulletin*, abril de 2008).
a) Proporcione una estimación puntual de la media poblacional del número de semanas que le toma a los trabajadores de más de 55 años encontrar un empleo.
b) ¿Cuál es el margen de error con 95% de confianza?
c) ¿Cuál es la estimación por intervalo de confianza de 95% para la media poblacional?
d) Analice el grado de sesgo que puede encontrarse en los datos muestrales. ¿Qué sugeriría para la repetición de este estudio?

19. El costo promedio por noche de un cuarto de hotel en la ciudad de Nueva York es $273 (*SmartMoney,* marzo de 2009). Suponga que esta estimación se basa en una muestra de 45 hoteles y que la desviación estándar muestral es $65.
a) Con 95% de confianza, ¿cuál es el margen de error?
b) ¿Cuál es la estimación por intervalo de confianza de 95% para la media poblacional?
c) Hace dos años, el costo promedio por noche de un cuarto de hotel en Nueva York era de $229. Analice la variación en el costo en este periodo de dos años.

WEB archivo
Program

20. ¿Los comerciales interrumpen constantemente su programa de televisión favorito? CNBC presentó datos estadísticos sobre la cantidad promedio de minutos de programa en media hora de transmisión (CNBC, 23 de febrero de 2006). Los datos siguientes (en minutos) son representativos de sus hallazgos.

21.06	22.24	20.62
21.66	21.23	23.86
23.82	20.30	21.52
21.52	21.91	23.14
20.02	22.20	21.20
22.37	22.19	22.34
23.36	23.44	

Suponga que la población es aproximadamente normal. Proporcione una estimación puntual y un intervalo de confianza de 95% para la cantidad media de minutos de programa en media hora de transmisión televisiva.

WEB archivo
Alcohol

21. El consumo de bebidas alcohólicas entre mujeres jóvenes en edad de beber se ha incrementado en el Reino Unido, Estados Unidos y Europa (*The Wall Street Journal,* 15 de febrero de 2006). Datos (consumo anual en litros) reportados por este periódico tomados de una muestra de 20 mujeres europeas jóvenes son los siguientes.

266	82	199	174	97
170	222	115	130	169
164	102	113	171	0
93	0	93	110	130

Suponga que la población es más o menos simétrica. Proporcione un intervalo de confianza de 95% para el consumo medio anual de bebidas alcohólicas entre las mujeres europeas jóvenes.

WEB archivo
TicketSales

22. *Hannah Montana: La Película*, de Disney, se estrenó el fin de semana de Pascua en abril de 2009. Durante los tres días del fin de semana, la película se convirtió en la atracción número uno en taquilla (*The Wall Street Journal,* 13 de abril de 2009). Los ingresos de ventas de boletos en dólares de una muestra de 25 salas se listan a continuación.

20 200	10 150	13 000	11 320	9 700
8 350	7 300	14 000	9 940	11 200
10 750	6 240	12 700	7 430	13 500
13 900	4 200	6 750	6 700	9 330
13 185	9 200	21 400	11 380	10 800

a) ¿Cuál es la estimación por intervalo de confianza de 95% para los ingresos medios de las ventas de boletos por sala? Interprete su resultado.
b) Con un precio por boleto de $7.16, ¿cuál es la estimación del número medio de espectadores por sala?
c) La película se exhibió en 3 118 cines. Estime el número total de espectadores que vieron *Hannah Montana: La Película* y el total de las ventas de boletos en taquilla los tres días del fin de semana.

8.3 Determinación del tamaño de la muestra

En esta sección se presenta un procedimiento para determinar el tamaño de muestra que se necesita para tener un margen de error específico establecido antes de tomar la muestra.

En los consejos prácticos de las dos secciones anteriores se habló del papel del tamaño de la muestra para obtener una buena aproximación a los intervalos de confianza en los casos en que la población no tiene una distribución normal. Ahora se enfoca la atención en otro aspecto relacionado con el tamaño de la muestra, y se describe cómo elegir un tamaño suficientemente grande para obtener un margen de error deseado. Para explicar esto, se vuelve al caso de la sección 8.1 en el que se tenía una σ conocida. Con la expresión (8.1), el intervalo de estimación está dado por

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

La cantidad $z_{\alpha/2}(\sigma/\sqrt{n})$ es el margen de error. De manera que, como se ve, $z_{\alpha/2}$, la desviación estándar poblacional σ, y el tamaño de la muestra n se combinan para determinar el margen de error. Una vez que se selecciona el coeficiente de confianza $1 - \alpha$, $z_{\alpha/2}$ puede ser determinado. Por tanto, si se tiene el valor de σ, es posible encontrar el tamaño de muestra n necesario para proporcionar cualquier margen de error deseado. A continuación se presenta el desarrollo de la fórmula utilizada para calcular el tamaño n de muestra deseado.

Sea E = el margen de error deseado:

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Al despejar $\sqrt{n}$ tenemos

$$\sqrt{n} = \frac{z_{\alpha/2}\sigma}{E}$$

Al elevar al cuadrado ambos lados de esta ecuación, se obtiene la expresión siguiente para el tamaño de la muestra.

La ecuación (8.3) proporciona una buena recomendación del tamaño de la muestra. Sin embargo, la opinión del analista cuenta para determinar si el tamaño de muestra final debe ajustarse hacia arriba.

TAMAÑO DE LA MUESTRA PARA UNA ESTIMACIÓN POR INTERVALO DE LA MEDIA POBLACIONAL

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} \qquad \textbf{(8.3)}$$

Este tamaño de muestra proporciona el margen de error deseado al nivel de confianza elegido.

En la ecuación (8.3), E es el margen de error que el usuario está dispuesto a aceptar, y el valor $z_{\alpha/2}$ es consecuencia directa del nivel de confianza que se utilizará para calcular la estimación por intervalo. A reserva de la decisión del usuario, 95% de confianza es el valor más frecuentemente elegido ($z_{0.025} = 1.96$).

Por último, para usar la ecuación (8.3) es necesario contar con el valor de la desviación estándar poblacional σ. Sin embargo, aun cuando este valor no se conozca, puede utilizarse la ecuación (8.3) siempre que se tenga un valor preliminar o un *valor planeado* de σ. En la práctica, suele usarse alguno de los procedimientos siguientes para obtenerlo.

El valor planeado de la desviación estándar poblacional σ debe especificarse antes de determinar el tamaño de la muestra. Aquí se ofrecen tres métodos para obtener este valor planeado de σ.

1. Se utiliza como valor planeado de σ una estimación de la desviación estándar poblacional calculada a partir de datos de estudios anteriores.
2. Se opta por un estudio piloto seleccionando una muestra preliminar. La desviación estándar muestral obtenida de la muestra preliminar puede usarse como valor planeado de σ.
3. Se usa el juicio personal para "adivinar el mejor" valor de σ. Por ejemplo, se puede empezar por estimar el mayor y el menor valor en los datos de la población. La diferencia entre ambos valores proporciona una estimación del rango de los datos. Por último, este valor dividido entre 4 suele considerarse como una aproximación burda a la desviación estándar y tomarse como un valor planeado aceptable de σ.

Se considera el ejemplo siguiente para mostrar el uso de la ecuación (8.3) en la determinación del tamaño de la muestra. En un estudio previo para investigar el costo de la renta de automóviles en Estados Unidos se encontró que el costo medio de rentar un vehículo mediano era aproximadamente de $55 por día. Suponga que la organización que realizó dicho estudio quiere realizar otro para estimar la media poblacional del costo de las rentas por día de automóviles medianos en Estados Unidos. Al diseñar el nuevo estudio, el director del proyecto especificó que la media poblacional de las rentas por día debe estimarse con un margen de error de $2 y que se desea un nivel de 95% de confianza.

El director del proyecto especificó un margen de error deseable de $E = 2$, y el nivel de 95% de confianza indica que $z_{0.025} = 1.96$. Por tanto, sólo falta el valor planeado de la desviación estándar poblacional σ para calcular el tamaño de muestra deseado. En este punto, un analista

La ecuación (8.3) proporciona el tamaño de muestra mínimo necesario para obtener el margen de error deseado. Si el tamaño de muestra calculado no es un número entero, se redondea al siguiente número entero, con lo que se tendrá un margen de error ligeramente menor al requerido.

revisó los datos muestrales del estudio anterior y encontró que la desviación estándar muestral del costo de la renta diaria era \$9.65. Al utilizar \$9.65 como valor planeado de σ, tenemos:

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} = \frac{(1.96)^2(9.65)^2}{2^2} = 89.43$$

De esta manera, el tamaño de la muestra necesario para obtener un margen de error de \$2 debe ser de por lo menos 89.43 rentas de automóviles medianos. En casos como éste, en los que el valor de n no es un número entero, se redondea al siguiente valor entero; así que el tamaño de muestras que se aconseja es 90 rentas de automóviles medianos.

Ejercicios

Métodos

23. ¿Qué tan grande debe seleccionarse una muestra para tener un intervalo de confianza de 95% con un margen de error de 10? Suponga que la desviación estándar poblacional es 40.

AUTOevaluación

24. En un conjunto de datos se estima que el rango es 36.
    a) ¿Cuál es el valor planeado para la desviación estándar poblacional?
    b) ¿De qué tamaño deberá ser la muestra para que el margen de error en un intervalo de confianza de 95% sea 3?
    c) ¿De qué tamaño deberá ser la muestra para que el margen de error sea 2 en un intervalo de confianza de 95%?

Aplicaciones

AUTOevaluación

25. Remítase al ejemplo de Scheer Industries de la sección 8.2. Use 6.84 días como valor planeado para la desviación estándar poblacional.
    a) Asuma 95% de confianza, ¿de qué tamaño deberá ser la muestra para tener un margen de error de 1.5 días?
    b) Si la declaración de precisión se hizo con 90% de confianza, ¿de qué tamaño deberá ser la muestra para tener un margen de error de 2 días?

26. El costo promedio de un galón de gasolina sin plomo en Greater Cincinnati es \$2.41 (*The Cincinnati Enquirer*, 3 de febrero de 2006). En una época de constantes cambios en los precios, un periódico muestrea las gasolineras y presenta un informe sobre los precios del combustible. Suponga que la desviación estándar es \$0.15 en los precios del galón de la gasolina sin plomo y recomiende el tamaño apropiado de muestra n que debe usar este periódico para tener un margen de error con 95% de confianza.
    a) Suponga que el margen de error requerido es \$0.07.
    b) Asuma que el margen de error deseado es \$0.05.
    c) Ahora considere que el margen de error requerido es \$0.03.

27. Los sueldos anuales iniciales para estudiantes graduados en una carrera en administración se espera que estén entre \$30 000 y \$45 000. Suponga que se quiere dar un intervalo de confianza de 95% para estimar la media poblacional anual de los sueldos iniciales. ¿Cuál es el valor planeado de la desviación estándar poblacional? ¿Cuán grande deberá ser la muestra si se quiere que el margen de error sea cualquiera de los siguientes?
    a) \$500.
    b) \$200.
    c) \$100.
    d) ¿Recomendaría usted intentar obtener \$100 como margen de error? Explique.

28. Con base en una encuesta en línea de ShareBuilder, un proveedor de planes de retiro, y Harris Interactive se reportó que 60% de las mujeres propietarias de negocios no están seguras de estar ahorrando lo suficiente para su retiro (*SmallBiz*, invierno de 2006). Suponga que se quiere efectuar un estudio de seguimiento para determinar cuánto están ahorrando las propietarias de negocios cada año con miras a su retiro y se quiere utilizar \$100 como margen de error requerido para un intervalo estimado de la media poblacional. Utilice \$1 100 como un valor planeado para la desviación estándar y recomiende un tamaño muestral para cada una de las siguientes situaciones.
    a) Se requiere un intervalo de confianza de 90% para la cantidad media ahorrada.
    b) Se necesita un intervalo de confianza de 95% para la cantidad media ahorrada.

c) Se requiere un intervalo de confianza de 99% para la cantidad media ahorrada.
d) Cuando se tiene un margen de error fijo, ¿qué sucede con el tamaño de la muestra a medida que el nivel de confianza aumenta? ¿Recomendaría usar en este caso un intervalo de confianza de 99%? Analice su respuesta.

29. Los tiempos requeridos para transportarse al trabajo en las 15 ciudades más grandes de Estados Unidos se consignan en *2003 Information Please Almanac*. Suponga que se usa una muestra aleatoria simple preliminar de los habitantes de San Francisco con el fin de establecer un valor planeado de 6.25 minutos para la desviación estándar poblacional.
a) Si desea estimar la media poblacional del tiempo que necesitan los residentes de San Francisco para transportarse al trabajo, con un margen de error de 2 minutos, ¿cuál debe ser el tamaño de la muestra? Suponga que el nivel de confianza es de 95%.
b) Si desea estimar la media poblacional del tiempo requerido por los habitantes de San Francisco para transportarse al trabajo con un margen de error de 1 minuto, ¿cuál debe ser el tamaño de la muestra? Suponga un nivel de confianza de 95%.

30. Durante el primer trimestre de 2003 la proporción precio/ganancias (P/G) en las acciones de la Bolsa de Nueva York iba de 5 a 60 (*The Wall Street Journal*, 7 de marzo de 2003). Suponga que se desea estimar la media poblacional de esta relación P/G en todas las acciones de la Bolsa de Nueva York, ¿cuántas acciones habrá que tomar en la muestra si se quiere que el margen de error sea 3? Use 95% de confianza.

8.4 Proporción poblacional

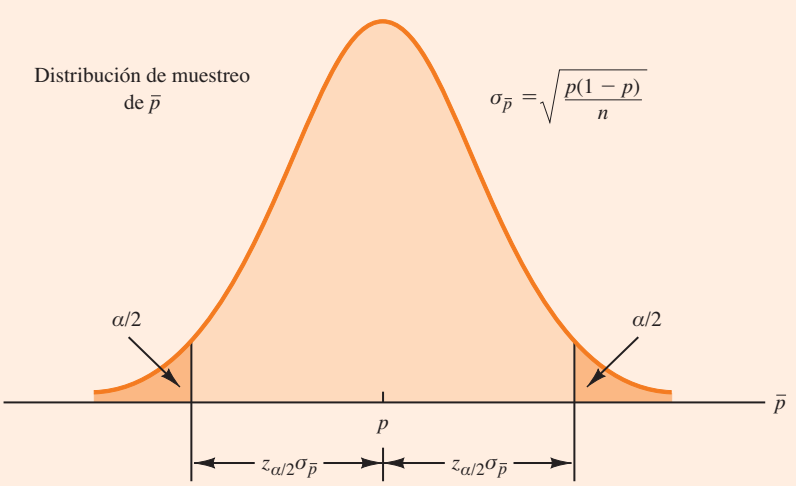
En la introducción a este capítulo se dijo que para obtener una estimación por intervalo de la proporción poblacional p, la fórmula general es:

$$\bar{p} \pm \text{margen de error}$$

La distribución de muestreo de $\bar{p}$ desempeña un papel clave en el cálculo del margen de error de esta estimación por intervalo.

En el capítulo 7 se dijo que la distribución de muestreo de $\bar{p}$ se aproxima mediante una distribución normal siempre que $np \geq 5$ y $n(1 - p) \geq 5$. En la figura 8.9 se presenta una apro-

FIGURA 8.9 Aproximación normal a la distribución de muestreo de $\bar{p}$



ximación normal a la distribución de muestreo de $\bar{p}$. La media de la distribución de muestreo de $\bar{p}$ es la proporción poblacional p, y el error estándar de $\bar{p}$ es

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{8.4}$$

Como la distribución de muestreo de $\bar{p}$ es una distribución normal, si en la estimación por intervalo de la proporción poblacional se elige como margen de error $z_{\alpha/2}\sigma_{\bar{p}}$, entonces $100(1-\alpha)\%$ de los intervalos que se obtengan contendrán la verdadera proporción poblacional. Pero para calcular el margen de error no se puede usar directamente $\sigma_{\bar{p}}$, ya que no se conoce p, pues se está tratando de estimarlo. Lo que se hace es que p se sustituye por $\bar{p}$ y de esta manera el margen de error para la estimación por intervalo de la proporción poblacional queda dado por

$$\text{Margen de error} = z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{8.5}$$

Con este margen de error, la expresión general para la estimación por intervalo de la proporción poblacional es la siguiente.

El margen de error de un intervalo de confianza para la proporción poblacional está dado por la cantidad $z_{\alpha/2}\sqrt{\bar{p}(1-\bar{p})/n}$.

ESTIMACIÓN POR INTERVALO DE UNA PROPORCIÓN POBLACIONAL

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{8.6}$$

donde $1-\alpha$ es el coeficiente de confianza y $z_{\alpha/2}$ es el valor de z que deja un área $\alpha/2$ en la cola superior de la distribución normal estándar.

WEB archivo
TeeTimes

En el siguiente ejemplo se ilustra el cálculo del margen de error y de la estimación por intervalo para una proporción poblacional. Un estudio en Estados Unidos encuestó a 900 mujeres golfistas para conocer su opinión acerca de cómo se les trataba en los cursos de golf. En el estudio se encontró que 396 estaban satisfechas con la disponibilidad de horarios de salida. Por tanto, la estimación puntual de la proporción poblacional de golfistas satisfechas con la disponibilidad de horarios de salida es $396/900 = 0.44$. Utilizando la expresión (8.6) y el nivel de confianza de 95%,

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$0.44 \pm 1.96\sqrt{\frac{0.44(1-0.44)}{900}}$$

$$0.44 \pm 0.0324$$

En consecuencia, el margen de error es 0.0324 y la estimación por intervalo de confianza de 95% de la proporción poblacional es 0.4076 a 0.4724. Empleando porcentajes, los resultados de la investigación permiten decir con 95% de confianza que entre 40.76% y 47.24% de las golfistas están satisfechas con la disponibilidad de horarios de salida.

Determinación del tamaño de la muestra

Ahora se considera cuál debe ser el tamaño de la muestra para obtener una estimación de la proporción poblacional con una precisión determinada. La función que tiene el tamaño de la muestra en la determinación de la estimación por intervalo de p es semejante a la que tiene en la estimación de la media poblacional estudiada en la sección 8.3.

Ya en esa sección se dijo que el margen de error asociado con la estimación por intervalo de la proporción poblacional es $z_{\alpha/2}\sqrt{\bar{p}(1-\bar{p})/n}$. Este margen se basa en el valor de $z_{\alpha/2}$, en la proporción muestral $\bar{p}$ y en el tamaño de la muestra n. Muestras mayores proporcionan márgenes de error menores y mejor precisión.

Sea E el margen de error deseado.

$$E = z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Al despejar n de esta fórmula, se obtiene la fórmula para calcular el tamaño de la muestra con el que se tendrá el margen de error deseado, E.

$$n = \frac{(z_{\alpha/2})^2\bar{p}(1-\bar{p})}{E^2}$$

Sin embargo, debido a que no se conocerá $\bar{p}$ sino hasta que se tome la muestra, no es posible usar esta fórmula para calcular el tamaño de la muestra con el que se obtendrá el margen de error deseado. Se necesita, entonces, un valor planeado de $\bar{p}$ útil para hacer este cálculo. Con p^* como valor planeado de $\bar{p}$, la fórmula para calcular el tamaño de la muestra con el que se obtendrá el error E queda como se presenta a continuación.

TAMAÑO DE LA MUESTRA PARA UNA ESTIMACIÓN POR INTERVALO DE LA PROPORCIÓN POBLACIONAL

$$n = \frac{(z_{\alpha/2})^2 p^*(1-p^*)}{E^2} \qquad \textbf{(8.7)}$$

En la práctica, el valor planeado p^* se determina mediante alguno de los métodos siguientes.

1. Se utiliza la proporción poblacional de una muestra previa de las mismas unidades o de unidades similares.
2. Se toma un estudio piloto y se elige una muestra preliminar. La proporción muestral de esta muestra se usa como valor planeado, p^*.
3. Se utiliza el criterio o una "mejor aproximación" para el valor de p^*.
4. Si no es aplicable ninguna de las alternativas anteriores, se emplea como valor planeado $p^* = 0.50$.

De regreso al estudio de mujeres golfistas, suponga que la empresa desea llevar a cabo otra investigación para determinar la proporción actual en la población de golfistas que está satisfecha con la disponibilidad de horarios de salida. ¿De qué tamaño deberá ser la muestra si se desea que en la estimación de la proporción poblacional el margen de error sea 0.025 a 95% de confianza? Como $E = 0.025$ y $z_{\alpha/2} = 1.96$, se necesita un valor planeado p^* para responder la pregunta. Utilizando como valor planeado p^* el resultado del estudio anterior, $\bar{p} = 0.44$, con la ecuación (8.7) se obtiene

$$n = \frac{(z_{\alpha/2})^2 p^*(1-p^*)}{E^2} = \frac{(1.96)^2(0.44)(1-0.44)}{(0.025)^2} = 1514.5$$

TABLA 8.5 Algunos valores posibles de $p^*(1 - p^*)$

p^*	$p^*(1 - p^*)$	
0.10	(0.10)(0.90) = 0.09	
0.30	(0.30)(0.70) = 0.21	
0.40	(0.40)(0.60) = 0.24	
0.50	(0.50)(0.50) = 0.25	← máximo valor de $p^*(1 - p^*)$
0.60	(0.60)(0.40) = 0.24	
0.70	(0.70)(0.30) = 0.21	
0.90	(0.90)(0.10) = 0.09	

Así, el tamaño de la muestra debe ser por lo menos de 1 514.5 golfistas mujeres para satisfacer el margen de error requerido. Al redondear al valor entero siguiente, tenemos que se necesitan 1 515 golfistas para obtener el margen de error deseado.

La cuarta alternativa sugerida para seleccionar un valor planeado p^* es elegir $p^* = 0.50$. Cuando no se cuenta con ninguna otra información, suele utilizarse este valor. Para entender por qué, observe que el numerador de la ecuación (8.7) indica que el tamaño de la muestra es proporcional a la cantidad $p^*(1 - p^*)$. Si el valor de $p^*(1 - p^*)$ es grande, el tamaño de la muestra también lo será. En la tabla 8.5 se consideran algunos valores que puede tener $p^*(1 - p^*)$. El máximo valor se presenta cuando $p^* = 0.50$. De esta manera, en caso de duda acerca del valor planeado apropiado, sabemos que $p^* = 0.50$ dará el mayor tamaño de muestra que se puede recomendar. En efecto, con el mayor tamaño de muestra posible se va a lo seguro. Si resulta que la proporción muestral es diferente del valor planeado, el margen de error será menor que el anticipado. De manera que al usar $p^* = 0.50$ se garantiza que el tamaño de la muestra será suficiente para obtener el margen de error deseado.

En el ejemplo del estudio de las golfistas, si se usa como valor planeado $p^* = 0.50$, el tamaño de muestra que se obtiene es

$$n = \frac{(z_{\alpha/2})^2 p^*(1 - p^*)}{E^2} = \frac{(1.96)^2(0.50)(1 - 0.50)}{(0.025)^2} = 1\,536.6$$

Es decir, una muestra ligeramente mayor: 1 537 mujeres golfistas.

NOTAS Y COMENTARIOS

El margen de error deseado para calcular una proporción poblacional casi siempre es 0.10 o menos. En las encuestas de opinión pública a nivel nacional en Estados Unidos conducidas por Gallup y Harris, un margen de error de 0.03 o 0.04 es común. Con estos márgenes, la ecuación (8.7) suministra un tamaño de la muestra que es suficiente para satisfacer los requerimientos de $np \geq 5$ y $n(1 - p) \geq 5$ para usar una distribución normal como aproximación de la distribución de muestreo de $\bar{x}$.

Ejercicios

Métodos

AUTOevaluación

31. Una muestra aleatoria simple de 400 individuos proporciona 100 respuestas Sí.
    a) Determine la estimación puntual de la proporción poblacional de individuos cuya respuesta será Sí.
    b) ¿Cuál es la estimación del error estándar de la proporción $\sigma_{\bar{p}}$?
    c) Calcule el intervalo de confianza de 95% para la proporción poblacional.

32. En una muestra aleatoria de 800 elementos se obtiene una proporción muestral, $\bar{p} = 0.70$.
    a) Proporcione un intervalo de 90% de confianza para la proporción poblacional.
    b) Proporcione un intervalo de confianza de 95% para la proporción poblacional.

33. En un estudio, el valor planeado para la proporción poblacional es $p^* = 0.35$. ¿De qué tamaño se debe tomar la muestra para dar un intervalo de confianza de 95% con un margen de error de 0.05?

34. Para 95% de confianza, ¿de qué tamaño se deberá tomar la muestra para obtener un margen de error de 0.03 en la estimación de una proporción poblacional? Suponga que no se cuenta con datos anteriores para obtener un valor planeado de p^*.

Aplicaciones

AUTO evaluación

35. El Consumer Reports National Research Center realizó una encuesta telefónica con 2 000 adultos para conocer sus principales preocupaciones económicas proyectadas al futuro (*Consumer Reports*, enero de 2009). Los resultados mostraron que 1 760 de los encuestados afirmaron que la salud futura es una de sus principales preocupaciones económicas.
    a) ¿Cuál es la estimación puntual de la proporción poblacional de adultos que piensan que la salud futura es una de las principales preocupaciones económicas?
    b) A 90% de confianza, ¿cuál es el margen de error?
    c) Proporcione el intervalo de 90% de confianza para la proporción poblacional de adultos que piensan que la salud futura es una de las principales preocupaciones económicas.
    d) Proporcione el intervalo de 95% de confianza para esta proporción poblacional.

36. Con base en estadísticas publicadas por la CNBC, la cantidad de vehículos que no están asegurados es sorprendente (CNBC, 23 de febrero de 2006). Los resultados muestrales indican que 46 de 200 vehículos no están asegurados.
    a) ¿Cuál es la estimación puntual de la proporción de vehículos no asegurados?
    b) Proporcione un intervalo de confianza de 95% para la proporción poblacional.

37. Towers Perrin, una firma de consultoría de recursos humanos de Nueva York, realizó un estudio con 1 100 empleados de empresas medianas y grandes para determinar qué tan insatisfechos estaban con su trabajo (*The Wall Street Journal,* 29 de enero de 2003). En el archivo JobSatisfaction se muestran datos representativos. Un Sí como respuesta indica que al empleado le desagrada mucho su puesto actual.

WEB archivo
JobSatisfaction

    a) Proporcione la estimación puntual de la proporción poblacional de empleados a quienes les disgusta mucho su puesto actual.
    b) A 95% de confianza, ¿cuál es el margen de error?
    c) ¿Cuál es el intervalo de confianza de 95% para la proporción de la población de empleados a quienes les desagrada mucho su puesto actual?
    d) Towers Perrin estima que a los empleadores les cuesta un tercio de un sueldo anual encontrar a un sucesor y hasta 1.5 veces el sueldo anual encontrar a un sucesor para un empleado que recibe una alta compensación. ¿Cuál es el mensaje de esta investigación para los empleadores?

38. Según Thomson Financial, hasta el 25 de enero de 2006 la mayoría de las empresas que informaban tener utilidades habían superado las estimaciones (*BusinessWeek,* 6 de febrero de 2006). En una muestra de 162 compañías, 104 superaron las estimaciones, 29 coincidieron y 29 se quedaron cortas.
    a) ¿Cuál es la estimación puntual de la proporción de empresas que se quedaron cortas?
    b) Determine el margen de error y proporcione un intervalo de confianza de 95% para la proporción que superó las estimaciones.
    c) ¿De qué tamaño debe de ser la muestra si el margen de error es 0.05?

AUTO evaluación

39. El porcentaje de personas que no tenía un seguro médico en 2003 era de 15.6% (*Statistical Abstract of the United States,* 2006). Se le solicitó a un comité del Congreso realizar un estudio para obtener información actualizada.
    a) ¿Qué tamaño de muestra le recomienda usted al comité si el objetivo es que en la estimación de la proporción actual de individuos que no tienen seguro médico el margen de error sea 0.03? Use 95% de confianza.
    b) Repita el inciso *a*) usando 99% de confianza.

40. Por muchos años, las empresas han luchado con el creciente costo del cuidado de la salud. Recientemente los incrementos han disminuido debido a la menor inflación en los precios del servicio y a los empleados que pagan gran parte de esos beneficios. Una reciente encuesta de Mercer mostró que era probable que 52% de los empleadores estadounidenses requiriera contribuciones más altas de los empleados para la cobertura del cuidado de la salud en 2009 (*BusinessWeek*, 16 de febrero de 2009). Suponga que la encuesta se basó en una muestra de 800 empresas. Calcule el margen de error y un intervalo de confianza de 95% para la proporción de compañías con probabilidad de requerir contribuciones más altas de los empleados para la cobertura del cuidado de la salud en 2009.

41. Los jóvenes de Estados Unidos usan Internet intensamente: 87% de los jóvenes entre 12 y 17 años son usuarios de la red (*The Cincinnati Enquirer*, 1 de febrero de 2006). En una muestra de usuarios de Internet de esta edad, 9% votó por MySpace como el sitio más popular de la Web. Suponga que en este estudio participaron 1400 sujetos. ¿Cuáles son los márgenes de error y la estimación por intervalo de la proporción poblacional de quienes consideran que este sitio es el más popular? Use 95% de nivel de confianza.

42. Una encuesta realizada durante la campaña presidencial tomó en junio una muestra de 491 votantes potenciales. El objetivo consistió en estimar la proporción de votantes potenciales a favor de cada candidato. Suponga que el valor planeado es $p^* = 0.50$, con un nivel de confianza de 95%.
    a) Si $p^* = 0.50$, ¿cuál fue el margen de error planeado en la encuesta de junio?
    b) Al acercarse la elección de noviembre se busca una mejor precisión y un menor margen de error. Suponga que los márgenes de error que se piden son los que se muestran en la tabla siguiente. Calcule el tamaño de muestra que se recomienda para cada estudio.

Estudio	Margen de error
Septiembre	0.04
Octubre	0.03
Inicio de noviembre	0.02
Un día antes de la elección	0.01

43. Phoenix Wealth Management/Harris Interactive realizó un estudio con 1 500 individuos cuyo patrimonio era de un millón o más de dólares, y obtuvo diversos datos estadísticos sobre la gente pudiente (*BusinessWeek*, 22 de septiembre de 2003). Los tres años anteriores habían sido malos para el mercado accionario, lo que motivó algunas de las preguntas planteadas.
    a) En este estudio se encontró que 53% de los encuestados perdió 25% o más del valor de su portafolio en los últimos tres años. Proporcione un intervalo de confianza de 95% para la proporción de personas pudientes que perdieron 25% o más del valor de su portafolio en el periodo de referencia.
    b) El estudio indicó que 31% de los encuestados siente que deberá ahorrar más para su retiro con objeto de compensar lo perdido. Proporcione un intervalo de confianza de 95% para la proporción poblacional.
    c) De los encuestados, 5% donó $25 000 o más para obras de caridad el año anterior. Proporcione un intervalo de confianza de 95% para la proporción de quienes aportaron $25 000 o más para obras caritativas.
    d) Compare los márgenes de error de las estimaciones por intervalo de los incisos *a*), *b*) y *c*). ¿Cuál es la relación entre margen de error y $\bar{p}$? Si usa la misma muestra para obtener varias proporciones, ¿cuál debe usarse para elegir el valor planeado p^*? ¿Por qué considera que en estos casos suela usarse $p^* = 0.50$?

Resumen

En este capítulo se presentaron los métodos para obtener estimaciones por intervalo de la media poblacional y de la proporción poblacional. Un estimador puntual puede o no proporcionar una buena estimación de un parámetro poblacional. Un intervalo de estimación suministra una media de la precisión de una estimación. Tanto la estimación por intervalo de una media poblacional como la de una proporción poblacional tienen la forma: estimación puntual $\pm$ margen de error.

Para la media poblacional se presentaron estimaciones por intervalo en dos casos. En el caso de σ conocida, se usan datos históricos o alguna otra información para obtener una estimación de σ antes de tomar la muestra. Entonces, el análisis de nuevos datos muestrales se realiza bajo el supuesto de que se conoce σ. En el caso de σ desconocida, los datos muestrales se usan para estimar tanto la media poblacional como la desviación estándar poblacional. La decisión final de qué procedimiento de estimación por intervalo utilizar depende de que el analista decida qué método proporciona una mejor estimación de σ.

Para σ conocida, el procedimiento de estimación por intervalo se basa en el valor supuesto de σ y en el uso de la distribución normal estándar. En cuanto a σ desconocida, para el procedimiento de estimación por intervalo se usa la desviación estándar muestral s y la distribución t. En ambos casos, la calidad de la estimación por intervalo depende de la distribución de la población y del tamaño de la muestra. Si la población tiene una distribución normal, la estimación por intervalo será exacta en ambos casos, aun cuando los tamaños de las muestras sean pequeños. Si la población no tiene distribución normal, la estimación por intervalo resultante será aproximada. Tamaños de muestras mayores proporcionarán mejores aproximaciones, pero entre más sesgada sea la población, mayor será el tamaño de la muestra necesario para obtener una buena aproximación. En las secciones 8.1 y 8.2 se proporcionaron consejos prácticos respecto del tamaño de muestra necesario para obtener buenas aproximaciones. En la mayoría de los casos, un tamaño 30 o mayor proporcionará una buena aproximación para el intervalo de confianza.

La forma general de una estimación por intervalo para la proporción poblacional es $\bar{p} \pm$ margen de error. En la práctica, los tamaños de muestra empleados en estimaciones por intervalo de una proporción poblacional suelen ser grandes. Entonces, el procedimiento de estimación por intervalo se basa en la distribución normal estándar.

Algunas veces se especifica un determinado margen de error antes de llevar a cabo el plan de muestreo. También se explicó cómo elegir el tamaño de muestra adecuado para obtener la precisión deseada.

Glosario

σ conocida Caso en el que datos históricos o alguna otra información proporciona un buen valor para ser considerado como desviación estándar poblacional antes de tomar la muestra. Este valor conocido de σ se usa en la estimación por intervalo para calcular el margen de error.

σ desconocida El caso más común cuando no existen bases sólidas para estimar la desviación estándar poblacional antes de tomar la muestra. En la estimación por intervalo se usa la desviación estándar muestral s para calcular el margen de error.

Coeficiente de confianza Nivel de confianza expresado como valor decimal. Por ejemplo 0.95 es el coeficiente de confianza correspondiente al nivel de confianza de 95%.

Distribución *t* Familia de distribuciones de probabilidad utilizada para obtener una estimación por intervalo de la media poblacional cuando la desviación estándar poblacional σ no se conoce y se estima mediante la desviación estándar muestral s.

Estimación por intervalo Estimación de un parámetro poblacional que suministra un intervalo que se cree que contiene el valor del parámetro. Para las estimaciones por intervalo abordadas en este capítulo se adopta la forma: estimación puntual $\pm$ margen de error.

Grados de libertad Parámetro de la distribución t. Cuando se usa esta distribución para calcular una estimación por intervalo de la media poblacional, la distribución t correspondiente tiene $n - 1$ grados de libertad, donde n es el tamaño de la muestra aleatoria simple.

Intervalo de confianza Otro nombre para designar la estimación por intervalo.

Margen de error Valor $\pm$ que se suma y se resta *de* la estimación puntual con objeto de obtener una estimación por intervalo de un parámetro poblacional.

Nivel de confianza Confianza asociada con la estimación por intervalo. Por ejemplo, si un procedimiento de estimación por intervalo proporciona intervalos tales que 95% de ellos contendrá el parámetro poblacional, se dice que esa estimación por intervalo tiene un nivel de confianza de 95%.

Fórmulas clave

Estimación por intervalo de la media poblacional: σ conocida

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{8.1}$$

Estimación por intervalo de la media poblacional: σ desconocida

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{8.2}$$

Tamaño de la muestra para una estimación por intervalo de la media poblacional

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} \tag{8.3}$$

Estimación por intervalo de una proporción poblacional

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{8.6}$$

Tamaño de la muestra para una estimación por intervalo de la proporción poblacional

$$n = \frac{(z_{\alpha/2})^2 p^*(1-p^*)}{E^2} \tag{8.7}$$

Ejercicios complementarios

44. En un estudio realizado con 54 corredores de bolsa con descuento, se encontró que la media de los precios cobrados por una transacción de 100 acciones a $50 la acción, fue $33.77 (*AAII Journal*, febrero de 2006). Este estudio se realiza anualmente. Con base en los datos históricos disponibles, considere que la desviación estándar poblacional conocida es $15.
    a) Según los datos muestrales, ¿cuál es el margen de error asociado con un intervalo de confianza de 95%?
    b) Proporcione un intervalo de confianza de 95% para la media de los precios cobrados por una transacción de 100 acciones a $50 cada una.

45. En una encuesta realizada por la American Automobile Association se encontró que una familia de cuatro miembros gasta en promedio en vacaciones $215.60 por día. Suponga que en una muestra de 64 familias de vacaciones en las cataratas del Niágara la media muestral encontrada fue de $252.45 por día y la desviación estándar muestral fue de $74.50.
    a) Proporcione una estimación, mediante un intervalo de confianza de 95%, para la media de la cantidad que gasta por día una familia de cuatro que está de vacaciones en las cataratas del Niágara.
    b) Con base en el intervalo de confianza del inciso *a*), ¿parece que la media poblacional de la cantidad gastada por día por las familias que visitan las cataratas del Niágara es diferente de la media reportada por la American Automobile Association? Explique.

46. Los 92 millones de estadunidenses de más de 50 años de edad controlan 50% de todos los ingresos discrecionales (*AARP Bulletin*, marzo de 2008). La AARP estima que el gasto promedio anual en restaurantes y comida para llevar fue de $1 873 por individuo de ese grupo de edad. Suponga que tal estimación se basa en una muestra de 80 personas y que la desviación estándar muestral es $550.
    a) ¿Cuál es el margen de error en este estudio? Use 95% de confianza.
    b) ¿Cuál es el intervalo de confianza de 95% de la media poblacional de la cantidad gastada en restaurantes y comida para llevar?
    c) ¿Cuál es su estimación de la cantidad total gastada por los estadunidenses de más de 50 años de edad en restaurantes y comida para llevar?
    d) Si la cantidad gastada en ambos aspectos es sesgada a la derecha, ¿esperaría que la cantidad media gastada sea mayor o menor que $1 873?

47. Numerosos observadores de los mercados bursátiles aseguran que cuando la razón P/E en las acciones es superior a 20, el mercado está sobrevaluado. La razón P/E es el precio de una acción dividido entre las ganancias (earnings) de los últimos 12 meses. Suponga que usted desea saber si actualmente el mercado está sobrevaluado y qué proporción de las empresas pagan dividendos (Dividend). A continuación aparece una muestra aleatoria de 30 firmas que cotizan en la Bolsa de Valores de Nueva York (NYSE) *(Barron's*, 19 de enero de 2004).

WEB archivo
NYSEStocks

Company	Dividend	P/E Ratio	Company	Dividend	P/E Ratio
Albertsons	Yes	14	NY Times A	Yes	25
BRE Prop	Yes	18	Omnicare	Yes	25
CityNtl	Yes	16	PallCp	Yes	23
DelMonte	No	21	PubSvcEnt	Yes	11
EnrgzHldg	No	20	SensientTch	Yes	11
Ford Motor	Yes	22	SmtProp	Yes	12
Gildan A	No	12	TJX Cos	Yes	21
HudsnUtdBcp	Yes	13	Thomson	Yes	30
IBM	Yes	22	USB Hldg	Yes	12
JeffPilot	Yes	16	US Restr	Yes	26
KingswayFin	No	6	Varian Med	No	41
Libbey	Yes	13	Visx	No	72
MasoniteIntl	No	15	Waste Mgt	No	23
Motorola	Yes	68	Wiley A	Yes	21
Ntl City	Yes	10	Yum Brands	No	18

a) Proporcione una estimación puntual para la razón poblacional P/E de las acciones que cotizan en la Bolsa de Valores de Nueva York. Maneje un intervalo de confianza de 95%.
b) Con base en su respuesta del inciso *a*), ¿considera usted que el mercado está sobrevaluado?
c) Proporcione una estimación puntual de la proporción de empresas en la NYSE que pagan dividendos. ¿El tamaño de la muestra es suficientemente grande para justificar el empleo de la distribución normal en el cálculo de un intervalo de confianza para esta proporción? ¿Por qué?

WEB archivo
Flights

48. US Airways llevó a cabo diversos estudios que indican ahorros importantes si los viajeros frecuentes del programa Dividend Miles realizaran en línea el canje de millas y programaran los vuelos ganados (*US Airways Attaché*, febrero de 2003). En un estudio se recabaron datos sobre el tiempo que se requiere para realizar por teléfono el canje de millas y la programación de un vuelo ganado. En el archivo de datos Flights se encuentra una muestra de tiempos en minutos requeridos para programar por teléfono cada uno de los 150 vuelos ganados. Use Minitab o Excel para contestar las preguntas siguientes.
a) ¿Cuál es la media muestral del número de minutos que se requiere para programar por teléfono los vuelos ganados?
b) Proporcione el intervalo de confianza de 95% para la media poblacional del tiempo requerido para programar por teléfono los vuelos.
c) Suponga que un agente de boletos por teléfono trabaja 7.5 horas por día. ¿Cuántos vuelos ganados se espera que atienda en un día?
d) Diga cómo esta información apoya el plan de US Airways de usar un sistema en línea para reducir costos.

WEB archivo
ActTemps

49. En un estudio realizado por Accountemps se le solicitó a 200 ejecutivos de una muestra proporcionar datos sobre la cantidad de minutos por día que pierden los oficinistas tratando de localizar cosas mal guardadas, mal archivadas o mal clasificadas. Los datos congruentes con esta investigación se encuentran en el archivo de datos ActTemps.
a) Use ActTemps para dar una estimación puntual de los minutos por día perdidos por los oficinistas en localizar cosas mal guardadas, mal archivadas o mal clasificadas.
b) ¿Cuál es la desviación estándar muestral?
c) Proporcione un intervalo de confianza de 95% para la cantidad de minutos perdidos por día.

50. Se efectúan pruebas de rendimiento de gasolina con un determinado modelo de automóvil. Si se desea dar un intervalo de confianza de 98% con un margen de error de 1 milla por galón, ¿cuántos automóviles deberán usarse? Suponga que por pruebas anteriores se sabe que la desviación estándar del rendimiento es 2.6 millas por galón.

51. Un centro médico quiere estimar la media del tiempo que se necesita para programar una cita de un paciente. ¿De qué tamaño deberá ser la muestra si se quiere que el margen de error sea de 2 minutos y que el nivel de confianza sea 95%? ¿De qué tamaño deberá tomarse la muestra si se quiere que el nivel de confianza sea 99%? Para la desviación estándar poblacional use 8 minutos como valor planeado.

52. *BusinessWeek* presenta datos sobre el sueldo anual más bonos de presidentes ejecutivos (CEO). En una muestra preliminar la desviación estándar es $675; los datos se dan en miles de dólares. ¿De cuántos CEO deberá constar la muestra si se quiere estimar el sueldo anual más bonos con un margen de error de $100 000? (*Nota*. El margen de error deseado será $E = 100$ si los datos están dados en miles de dólares.) Use 95% de confianza.

53. El National Center for Education Statistics informa que 47% de los estudiantes universitarios trabaja para pagar sus estudios y su sustento. Suponga que se empleó una muestra de 450 estudiantes en la investigación.
    a) Proporcione un intervalo de confianza de 95% para dicha proporción poblacional.
    b) Proporcione un intervalo de confianza de 99% para la proporción poblacional de estudiantes que trabajan para mantenerse y pagar sus estudios.
    c) ¿Qué ocurre con el margen de error cuando el nivel de confianza aumenta de 95% a 99%?

54. En un estudio de *USA Today*/CNN/Gallup realizado con 369 padres que trabajan, se encontró que 200 consideran que pasan muy poco tiempo con sus hijos debido a sus compromisos laborales.
    a) Proporcione una estimación puntual de la proporción poblacional de padres que trabajan y piensan que pasan muy poco tiempo con sus hijos debido a sus compromisos laborales.
    b) ¿Cuál es el margen de error para 95% de confianza?
    c) ¿Cuál es el intervalo de confianza de 95% para la proporción poblacional de padres que trabajan y piensan que pasan muy poco tiempo con sus hijos debido a sus compromisos ocupacionales?

55. ¿De qué le sería más difícil prescindir: de su televisor o de su computadora? En un estudio reciente efectuado con 1 677 usuarios de Internet en Estados Unidos, se encontró que a 74% de la élite tecnológica juvenil (edad promedio de 22 años) le sería más difícil prescindir de su computadora (*PC Magazine*, 3 de febrero de 2004). Sólo para 48% sería más difícil renunciar a su televisor.
    a) Desarrolle un intervalo de confianza de 95% para la proporción de jóvenes a quienes les sería difícil prescindir de su computadora.
    b) Encuentre un intervalo de confianza de 99% para la proporción de jóvenes a quienes les sería difícil renunciar a su televisor.
    c) ¿En cuál de los incisos, *a*) o *b*), es mayor el margen de error? Explique por qué.

56. El aeropuerto internacional Cincinnati/Northern Kentucky obtuvo en 2005 el segundo lugar en puntualidad en la llegada de vuelos entre los aeropuertos con más actividad del país (*The Cincinnati Enquirer*, 3 de febrero de 2003). Suponga que esto se basa en una muestra de 550 vuelos, de los cuales 455 llegaron a tiempo.
    a) Elabore una estimación puntual de la tasa de llegadas puntuales (proporción de vuelos que llegan a tiempo) al aeropuerto.
    b) Construya un intervalo de confianza de 95% para la proporción poblacional de llegadas a tiempo en todos los vuelos del aeropuerto en 2005.

57. El *2003 Statistical Abstract of the United States* proporciona el porcentaje de personas de 18 años o más que fuma. Asuma que en un nuevo estudio para recabar datos sobre los fumadores y no fumadores se usa 0.30 como estimación preliminar de la proporción que fuma.
    a) ¿De qué tamaño deberá tomarse la muestra para estimar la proporción de fumadores con un margen de error de 0.02? Use 95% de confianza.
    b) Suponga que el estudio usa su recomendación para el tamaño de la muestra del inciso *a*) y encuentra 520 fumadores. ¿Cuál es la estimación puntual de la proporción de fumadores en la población?
    c) ¿Cuál es el intervalo de confianza de 95% para la proporción de fumadores en la población?

58. Una firma de tarjetas de crédito de un conocido banco desea estimar la proporción de tarjetahabientes que al final del mes tienen un saldo distinto de cero que ocasiona cargos. Suponga que el margen de error deseado es 0.03 con 98% de confianza.
    a) ¿De qué tamaño deberá tomarse la muestra si se cree que 70% de los tarjetahabientes de la firma tienen un saldo distinto de cero al final del mes?
    b) ¿De qué tamaño deberá tomarse la muestra si no se puede especificar ningún valor planeado para la proporción?
59. En un estudio se le solicitó a 200 personas que indicaran su principal fuente de información de noticias; 110 afirmaron que eran los noticieros de televisión.
    a) Proporcione un intervalo de confianza de 95% para la proporción poblacional de personas que tienen como principal fuente de noticias la televisión.
    b) ¿Cuál será el tamaño de muestra necesario para estimar la proporción poblacional con un margen de error de 0.05 y 95% de confianza?
60. Aunque para los viajeros de negocios, los horarios y los costos son aspectos importantes al elegir una línea aérea, en un estudio realizado por *USA Today* se encontró que para este sector el factor más importante estriba en que la línea tenga un programa de viajero frecuente. En una muestra de $n = 1\,993$ pasajeros que participaron en la encuesta, 618 indicaron como factor más importante un programa de viajero frecuente.
    a) ¿Cuál es la estimación puntual de la proporción poblacional de viajeros de negocios que consideran el programa de viajero frecuente como el factor más importante al elegir una línea aérea?
    b) Proporcione un intervalo de confianza de 95% para estimar la proporción poblacional.
    c) ¿De qué tamaño deberá ser la muestra para un margen de error de 0.01 con 95% de confianza? ¿Aconsejaría que *USA Today* tratara de tener esta precisión? ¿Por qué?

Caso a resolver 1 Revista *Young Professional*

La revista *Young Professional* fue creada para un público formado por personas que se encuentran en los 10 primeros años de su carrera profesional en negocios. En sus dos primeros años de publicación, la revista ha tenido bastante éxito. Ahora el editor está tratando de aumentar su base publicitaria. Los anunciantes potenciales preguntan continuamente sobre los datos demográficos e intereses de los suscriptores de *Young Professional*. Para recabar esta información, la revista realizó un estudio sobre el perfil de sus suscriptores. Los resultados se usarán para ayudar a elegir artículos de interés y proporcionar a los anunciantes un perfil de los suscriptores. Como nuevo empleado de la empresa se le solicita a usted su ayuda para analizar los resultados de la investigación.

WEB archivo
Professional

A continuación se presentan algunas preguntas del estudio.

1. ¿Cuál es su edad?
   (What is your age?)
2. Usted es: Hombre_____ Mujer_____
   Are you: Male_____ Female_____
3. ¿Piensa comprar algún bien inmueble en los próximos dos años? Sí_____ No_____
   Do you plan to make any real estate purchases in the next two years? Yes_____ No_____
4. ¿Cuál es el valor aproximado de las inversiones financieras, excluyendo su casa, que son de su propiedad o de otro miembro de su familia?
   What is the approximate total value of financial investments, exclusive of your home, owned by you or members of your household?
5. ¿Cuántas transacciones de acciones/bonos/fondos de inversión realizó el año pasado?
   How many stock/bond/mutual fund transactions have you made in the past year?
6. ¿Tiene en casa acceso de banda ancha a Internet? Sí_____ No_____
   Do you have broadband access to the Internet at home? Yes_____ No_____
7. Por favor, indique cuál fue el ingreso de su hogar el año pasado.
   Please indicate your total household income last year.
8. ¿Tiene hijos? Sí_____ No_____
   Do you have children? Yes_____ No_____

El archivo denominado Professional contiene las respuestas a estas preguntas. En la tabla 8.6 se muestra la parte de este archivo correspondiente a las respuestas de los primeros cinco entrevistados.

TABLA 8.6 Resultados parciales del estudio de la revista *Young professional*

Age	Gender	Real Estate Purchases	Value of Investments($)	Number of Transactions	Broadband Access	Household Income($)	Children
38	Female	No	12 200	4	Yes	75 200	Yes
30	Male	No	12 400	4	Yes	70 300	Yes
41	Female	No	26 800	5	Yes	48 200	No
28	Female	Yes	19 600	6	No	95 300	No
31	Female	Yes	15 100	5	No	73 300	Yes
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Informe gerencial

Elabore un informe gerencial con los resultados del estudio. Además de los resúmenes estadísticos, analice cómo la revista puede usarlos para atraer más anunciantes. También presente una recomendación a los editores para que empleen los resultados en la elección de los temas de interés para sus suscriptores. Su informe debe contener los siguientes puntos, pero no limite su análisis a estas áreas.

1. Desarrolle la estadística descriptiva adecuada para resumir los datos.
2. Muestre los intervalos de 95% de confianza para la edad promedio y el ingreso promedio por hogar de los suscriptores.
3. Encuentre intervalos de confianza de 95% para la proporción de suscriptores que tienen acceso de banda ancha y para la proporción de éstos que tienen niños.
4. ¿Será *Young Professional* un buen sitio para que los agentes de bolsa en línea contraten publicidad? Justifique su conclusión con datos estadísticos.
5. ¿Será esta revista un buen lugar para la publicidad de empresas que venden software educativo y juegos de computadora para niños?
6. Comente sobre el tipo de artículos que crea usted que son de interés para los lectores de *Young Professional*.

Caso a resolver 2 Gulf Real Estate Properties

Gulf Real Estate Properties, Inc. es una inmobiliaria ubicada en el suroeste de Florida. Esta empresa, que se anuncia como "experta en el mercado de bienes raíces", monitorea las ventas de condominios recabando datos sobre ubicación, precio de lista, precio de venta y días necesarios para vender cada unidad. Los condominios están calificados como *con* o *sin vista al golfo*, dependiendo de su ubicación hacia el golfo de México. Multiple Listing Service en Naples, Florida, proporciona datos muestrales sobre 40 condominios con vista al golfo (Gulf View Condominiums) y 18 sin vista al golfo (No Gulf View Condominiums).* Los precios están dados en miles de dólares. Los datos se presentan en la tabla 8.7.

Informe gerencial

1. Use la estadística descriptiva apropiada para resumir cada una de las tres variables de los 40 condominios con vista al golfo.
2. Aplique la estadística descriptiva adecuada para resumir cada una de las tres variables de los 18 condominios sin vista al golfo.
3. Compare los resultados. Analice cualquier estadístico específico que ayude al agente de ventas inmobiliarias a conocer más sobre el mercado de los condominios.

* Datos sustentados en las ventas de condominios reportadas en el Naples MLS (Coldwell Banker, junio de 2000).

TABLA 8.7 Datos de venta de propiedades vendidas por Gulf Real State Properties

Gulf View Condominiums			No Gulf View Condominiums		
List Price	**Sale Price**	**Days to Sell**	**List Price**	**Sale Price**	**Days to Sell**
495.0	475.0	130	217.0	217.0	182
379.0	350.0	71	148.0	135.5	338
529.0	519.0	85	186.5	179.0	122
552.5	534.5	95	239.0	230.0	150
334.9	334.9	119	279.0	267.5	169
550.0	505.0	92	215.0	214.0	58
169.9	165.0	197	279.0	259.0	110
210.0	210.0	56	179.9	176.5	130
975.0	945.0	73	149.9	144.9	149
314.0	314.0	126	235.0	230.0	114
315.0	305.0	88	199.8	192.0	120
885.0	800.0	282	210.0	195.0	61
975.0	975.0	100	226.0	212.0	146
469.0	445.0	56	149.9	146.5	137
329.0	305.0	49	160.0	160.0	281
365.0	330.0	48	322.0	292.5	63
332.0	312.0	88	187.5	179.0	48
520.0	495.0	161	247.0	227.0	52
425.0	405.0	149			
675.0	669.0	142			
409.0	400.0	28			
649.0	649.0	29			
319.0	305.0	140			
425.0	410.0	85			
359.0	340.0	107			
469.0	449.0	72			
895.0	875.0	129			
439.0	430.0	160			
435.0	400.0	206			
235.0	227.0	91			
638.0	618.0	100			
629.0	600.0	97			
329.0	309.0	114			
595.0	555.0	45			
339.0	315.0	150			
215.0	200.0	48			
395.0	375.0	135			
449.0	425.0	53			
499.0	465.0	86			
439.0	428.5	158			

WEB archivo
GulfProp

4. Proporcione un intervalo de confianza de 95% para estimar las medias poblacionales del precio de venta (Sales Price) y del número de días necesario para vender (Days to Sell) los condominios con vista al golfo. Interprete los resultados.
5. Encuentre un intervalo de confianza de 95% para estimar las medias poblacionales del precio de venta y el número de días necesarios para vender los condominios sin vista al golfo. Interprete los resultados.
6. Suponga que se necesita estimar el precio medio de venta de los condominios con vista al golfo con un margen de error de $40 000 y el precio medio de venta de los condomi-

nios sin vista al golfo con un margen de error de $15 000. Si se usa 95% de confianza, ¿de qué tamaño deberán ser las muestras?

7. Gulf Real Estate Properties firmó contratos para dos nuevos catálogos: un condominio con vista al golfo con un precio de lista de $585 000 y un condominio sin vista al golfo con un precio de $285 000. ¿Cuál es su estimado del precio final de venta y el número de días requerido para vender cada una de estas unidades?

Caso a resolver 3 Metropolitan Research, Inc.

Metropolitan Research, Inc., una organización para la investigación del consumidor, realiza estudios con objeto de evaluar una amplia variedad de bienes y servicios para los consumidores. En uno de sus trabajos, Metropolitan se enfocó en la satisfacción del consumidor respecto del funcionamiento de los automóviles producidos por el principal fabricante de Detroit. En un cuestionario enviado a propietarios de automóviles de esta empresa se encontraron varias quejas relacionadas con problemas prematuros en la transmisión. Para tener más información acerca de estos problemas, Metropolitan empleó una muestra de reparaciones de la transmisión proporcionada por empresas en Detroit dedicadas a esta tarea. Los datos siguientes indican el número de millas recorridas por 50 vehículos hasta el momento en que se presenta-ron los problemas con la transmisión.

WEB archivo
Auto

85 092	32 609	59 465	77 437	32 534	64 090	32 464	59 902
39 323	89 641	94 219	116 803	92 857	63 436	65 605	85 861
64 342	61 978	67 998	59 817	101 769	95 774	121 352	69 568
74 276	66 998	40 001	72 069	25 066	77 098	69 922	35 662
74 425	67 202	118 444	53 500	79 294	64 544	86 813	116 269
37 831	89 341	73 341	85 288	138 114	53 402	85 586	82 256
77 539	88 798						

Informe gerencial

1. Use la estadística descriptiva adecuada para resumir los datos sobre los problemas en la transmisión.
2. Proporcione un intervalo de confianza de 95% para estimar, en la población de automóviles con fallas en la transmisión, el número de millas promedio recorridas hasta que se presenta el problema. Haga una interpretación gerencial del intervalo estimado.
3. Analice las consecuencias de sus hallazgos en términos de la creencia de que algunos propietarios de automóviles tuvieron problemas prematuros con la transmisión.
4. ¿Cuántos registros de reparación deben tomarse en la muestra si se desea estimar la media poblacional del número de millas recorridas hasta la aparición de problemas en la transmisión con un margen de error de 5 000 millas? Use 95% de confianza.
5. ¿Qué otra información desearía recolectar para evaluar mejor los problemas con la transmisión?

Apéndice 8.1 Estimación por intervalo con Minitab

A continuación se describe cómo usar Minitab para obtener intervalos de confianza de la media poblacional y la proporción poblacional.

Media poblacional: σ conocida

WEB archivo
Lloyd's

La estimación por intervalo se ilustra mediante el ejemplo de Lloyd's de la sección 8.1. En una muestra de 100 clientes, las cantidades gastadas en cada visita a la tienda están en la columna Cl de la hoja de cálculo de Minitab. Se supone que la desviación estándar poblacional se conoce y es $\sigma = 20$. Los pasos siguientes permiten calcular un intervalo de confianza de 95% para estimar la media poblacional.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Basic Statistics**.
Paso 3. Seleccione **1-Sample Z**.
Paso 4. Cuando aparezca el cuadro de diálogo 1-Sample Z:
Ingrese C1 en el cuadro **Samples in columns**.
Ingrese 20 en el cuadro **Standard deviation**.
Paso 5. Haga clic en **OK**.

Por omisión, Minitab emplea 95% como nivel de confianza. Para especificar otro nivel de confianza, por ejemplo 90%, al paso 4 hay que agregar lo siguiente.

Seleccione **Options**.
Cuando el cuadro de diálogo 1-Sample Z-Options aparezca:
Ingrese 90 en el cuadro **Confidence level**.
Haga clic en **OK**.

Media poblacional: σ desconocida

WEB archivo
NewBalance

La estimación por intervalo se ilustra empleando los datos de la tabla 8.3 que proporcionan los saldos en las tarjetas de crédito en una muestra de 70 hogares. Los datos están en la columna Cl de la hoja de cálculo de Minitab. En este caso se estima la desviación estándar poblacional σ mediante la desviación estándar muestral s. Con los pasos siguientes se obtiene un intervalo de confianza de 95% para estimar la media poblacional.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Basic Statistics**.
Paso 3. Escoja **1-Sample t**.
Paso 4. Cuando el cuadro de diálogo 1-Sample t aparezca:
Ingrese C1 en el cuadro **Samples in columns**.
Paso 5. Haga clic en **OK**.

Por omisión, Minitab emplea 95% como nivel de confianza. Para especificar otro nivel, por ejemplo 90%, hay que agregar al paso 4 lo siguiente.

Seleccione **Options**.
Cuando el cuadro de diálogo 1-Sample t-Options aparezca:
Ingrese 90 en el cuadro **Confidence level**.
Haga clic en **OK**.

Proporción poblacional

WEB archivo
TeeTimes

La estimación por intervalo se ilustra utilizando los datos de las mujeres golfistas presentados en la sección 8.4. Los datos aparecen en la columna C1 de la hoja de cálculo de Minitab. Las respuestas individuales se registraron como Yes (Sí) cuando la golfista está satisfecha con la disponibilidad de horarios de salida y No, en caso contrario. Usando los pasos siguientes se calcula un intervalo de confianza de 95% para estimar la proporción de golfistas satisfechas con la disponibilidad de los horarios de salida.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Basic Statistics**.
Paso 3. Elija **1 Proportion**.
Paso 4. Cuando el cuadro de diálogo 1 Proportion aparezca:
Ingrese C1 en el cuadro **Samples in columns**.
Paso 5. Elija **Options**.
Paso 6. Cuando el cuadro de diálogo 1 Proportion-Options aparezca:
Seleccione **Use test and interval based on normal distribution**.
Haga clic en **OK**.
Paso 7. Haga clic en **OK**.

Por omisión, Minitab emplea 95% como nivel de confianza. Para especificar otro nivel, como 90%, cuando aparezca el cuadro de diálogo 1 Proportion-Options en el paso 6, ingrese 90 en el cuadro **Confidence Level.**

Nota. La rutina 1 Proportion de Minitab usa un ordenamiento alfabético de las respuestas y selecciona la *segunda respuesta* como la proporción poblacional de interés. En el ejemplo de las mujeres golfistas, Minitab maneja el orden alfabético No-Yes y de esta manera da el intervalo de confianza para la proporción de las respuestas Yes. Como Yes era la respuesta de interés, los resultados de Minitab fueron los adecuados. Sin embargo, si el orden alfabético no da la respuesta de interés, se selecciona cualquier celda de la columna y se usa la secuencia: Editor > Column > Value Order. Minitab le proporcionará la opción de usar un orden especificado por el usuario, pero usted debe ubicar en segundo lugar de la lista la respuesta de interés en el cuadro define-an-order.

Apéndice 8.2 Estimación por intervalo usando Excel

A continuación se describe el uso de Excel para calcular intervalos de confianza para la media poblacional y la proporción poblacional.

Media poblacional: σ conocida

WEB archivo
Lloyd's

La estimación por intervalo se ilustra con el ejemplo de Lloyd's de la sección 8.1. Se supone que se conoce la desviación estándar poblacional y que $\sigma = 20$. Las cantidades gastadas por la muestra de 100 clientes se encuentran en la columna A de la hoja de cálculo de Excel. En el cálculo del margen de error para estimar la media poblacional se aplican los pasos que se indican a continuación. Se empieza usando la herramienta para estadística descriptiva de Excel descrita en el capítulo 3.

Paso 1. Haga clic en la ficha **Data** en la cinta de opciones.
Paso 2. En el grupo **Analysis**, haga clic en **Data Analysis**.
Paso 3. Elija **Descriptive Statistics** de la lista Analysis Tools.
Paso 4. Cuando aparezca el cuadro de diálogo Descriptive Statistics:
Ingrese A1:A101 en el cuadro **Input Range**.
Seleccione **Grouped by Columns**.
Elija **Labels in First Row**.
Seleccione **Output Range**.
Ingrese C1 en el cuadro **Output Range**.
Seleccione **Summary Statistics**.
Haga clic en **OK**.

El resumen de estadísticas aparecerá en las columnas C y D. Continúe con el cálculo del margen de error usando la función Confidence de Excel como sigue:

Paso 5. Seleccione la celda C16 e ingrese el título Margin of error.
Paso 6. Elija la celda D16 e ingrese la fórmula de Excel =CONFIDENCE(0.5,20,100).

Los tres parámetros de esta función son:

Alfa = 1 − coeficiente de confianza = 1 − 0.95 = 0.05.
Desviación estándar poblacional = 20.
Tamaño de la muestra = 100 (*Nota*. Este parámetro aparece como Count en la celda D15.)

La estimación puntual de la media poblacional se encuentra en la celda D3 y el margen de error en la celda DI6. La estimación puntual (82) y el margen de error (3.92) permiten calcular con facilidad el intervalo de confianza para la media poblacional.

Media poblacional: σ desconocida

WEB archivo
NewBalance

La estimación por intervalo se ilustra con los datos de la tabla 8.2 en la que se registran los saldos en las tarjetas de crédito de 70 hogares. Los datos se encuentran en la columna A de la hoja de cálculo de Excel. Para calcular una estimación puntual y el margen de error de una estimación por intervalo de la media poblacional se siguen los pasos que se indican a continuación. Se emplea la herramienta para estadística descriptiva estudiada en el capítulo 3.

Paso 1. Haga clic en la ficha **Data** en la cinta de opciones.
Paso 2. En el grupo **Analysis**, haga clic en **Data Analysis**.
Paso 3. Elija **Descriptive Statistics** de la lista Analysis Tools.
Paso 4. Cuando aparezca el cuadro de diálogo Descriptive Statistics:
Ingrese A1:A71 en el cuadro **Input Range**.
Seleccione **Grouped by Columns**.
Elija **Labels in First Row**.
Seleccione **Output Range**.
Ingrese C1 en el cuadro Output Range.
Seleccione **Summary Statistics**.
Elija **Confidence Level for Mean**.
Ingrese 95 en el cuadro Confidence Level for Mean.
Haga clic en **OK**.

El resumen estadístico aparecerá en las columnas C y D. La estimación puntual de la media poblacional se presenta en la celda D3. El margen de error aparecerá como "Confidence Level(95.0%)" en la celda DI6. La estimación puntual ($9 312) y el margen de error ($955) permiten estimar con facilidad el intervalo de confianza para la media poblacional. La figura 8.10 ilustra el resultado de este procedimiento de Excel.

FIGURA 8.10 Estimación por intervalo de la media poblacional de saldos en tarjetas de crédito usando Excel

	A	B	C	D	E	F
1	NewBalance		*NewBalance*			
2	9 430					
3	7 535		Mean	9 312	Estimaciôm puntual	
4	4 078		Standard Error	478.9281		
5	5 604		Median	9 466		
6	5 179		Mode	13 627		
7	4 416		Standard Deviation	4 007		
8	10 676		Sample Variance	16 056 048		
9	1 627		Kurtosis	−0.296		
10	10 112		Skewness	0.18792		
11	6 567		Range	18 648		
12	13 627		Minimum	615		
13	18 719		Maximum	19 263		
14	14 661		Sum	651 840		
15	12 195		Count	70		
16	10 544		Confidence Level(95.0%)	955.4354	Margen de error	
17	13 659					
70	9 743					
71	10 324					
71						

Nota. Las filas 18 a 69 están ocultas.

Proporción poblacional

WEB archivo
Interval p

Esta estimación por intervalo se ilustra usando los datos del estudio de las mujeres golfistas presentado en la sección 8.4. Los datos se encuentran en la columna A de la hoja de cálculo de Excel. En la información recabada, una respuesta Yes (Sí) implica que la golfista está satisfecha con los horarios de salida disponibles y No cuando no es el caso. Excel no proporciona una rutina ya elaborada para la estimación de una proporción poblacional; sin embargo, es relativamente fácil disponer una plantilla para usarla con tal propósito. La plantilla de la figura 8.11 proporciona un intervalo de confianza de 95% para la estimación de la proporción de golfistas satisfechas con los horarios de salida disponibles. Observe que en la figura 8.11, en las

FIGURA 8.11 Plantilla de Excel para la estimación por intervalo de una proporción poblacional

	A	B	C	D
1	**Response**		**Interval Estimate of a Population Proportion**	
2	Yes			
3	No		**Sample Size**	=COUNTA(A2:A901)
4	Yes		**Response of Interest**	Yes
5	Yes		**Count for Response**	=COUNTIF(A2:A901,D4)
6	No		**Sample Proportion**	=D5/D3
7	No			
8	No		**Confidence Coefficient**	0.95
9	Yes		***z* Value**	=NORMSINV(0.5+D8/2)
10	Yes			
11	Yes		**Standard Error**	=SQRT(D6*(1-D6)/D3)
12	No		**Margin of Error**	=D9*D11
13	No			
14	Yes		**Point Estimate**	=D6
15	No		**Lower Limit**	=D14-D12
16	No		**Upper Limit**	=D14+D12
17	Yes			
18	No			
901	Yes			
902				

	A	B	C		
1	**Response**		**Interval Estimate of a Population Proportion**		
2	Yes				
3	No		**Sample Size**	900	
4	Yes		**Response of Interest**	Yes	Ingrese la respuesta de interés
5	Yes		**Count for Response**	396	
6	No		**Sample Proportion**	0.4400	
7	No				
8	No		**Confidence Coefficient**	0.95	Ingrese el coeficiente de confianza
9	Yes		***z* Value**	1.960	
10	Yes				
11	Yes		**Standard Error**	0.0165	
12	No		**Margin of Error**	0.0324	
13	No				
14	Yes		**Point Estimate**	0.4400	
15	No		**Lower Limit**	0.4076	
16	No		**Upper Limit**	0.4724	
17	Yes				
18	No				
901	Yes				
902					

Nota. Las filas 19 a 900 están ocultas.

celdas de la hoja de cálculo que aparece en segundo plano, se presentan las fórmulas que proporcionan los resultados de la hoja de cálculo que aparece en primer plano. Los siguientes son los pasos para usar la plantilla con este archivo de datos.

Paso 1. Ingrese el rango de datos A2:A901 en la fórmula =COUNTA de la celda D3.
Paso 2. Introduzca Sí como respuesta de interés en la celda D4.
Paso 3. Ingrese el rango de datos A2:A901 en la fórmula =COUNTIF de la celda D5.
Paso 4. Incorpore 0.95 como coeficiente de confianza en la celda D8.

Esta plantilla proporciona automáticamente los límires inferior y superior del intervalo de confianza en las celdas D15 y D16, y se usa para calcular un intervalo de confianza para la proporción poblacional en otras aplicaciones. Por ejemplo, para calcular la estimación por intervalo de un nuevo archivo de datos, se ingresan los nuevos datos muestrales en la columna A de la hoja de cálculo y después se modifican las cuatro celdas indicadas en los anteriores pasos. Si la nueva muestra de datos ya ha sido resumida, no es necesario ingresar los datos muestrales en la hoja de cálculo. En este caso se ingresa el tamaño de la muestra en la celda D3 y la proporción muestral en la celda D6; la plantilla proporcionará el intervalo de confianza para la proporción poblacional. La hoja de cálculo de la figura 8.11 se encuentra en el archivo *Interval p* del sitio web del libro.

Apéndice 8.3 Estimación por intervalo con StatTools

En este apéndice se muestra el uso de StatTools para establecer una estimación por intervalo de una media poblacional cuando se desconoce σ, y determinar el tamaño de la muestra necesario para obtener el margen de error deseado.

Estimación por intervalo de la media poblacional: caso de σ desconocida

En este caso se estimará la desviación estándar poblacional σ mediante la desviación estándar muestral s. Se emplearán los datos de los saldos en las tarjetas de crédito de la tabla 8.3 para ilustrarlo. Se inicia con el uso del Data Set Manager para crear un archivo de datos de StatTools con esos datos utilizando el procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes pueden usarse para calcular un intervalo de confianza estimado de 95% de la media poblacional.

WEB archivo
NewBalance

Paso 1. Haga clic en la ficha **StatTools** en la cinta de opciones.
Paso 2. En el grupo **Analyses**, dé clic en **Statistical Inference**.
Paso 3. Elija la opción **Confidence Interval**.
Paso 4. Seleccione Mean/Std. Deviation.
Paso 5. Cuando aparezca el cuadro de diálogo StatTools-Confidence Interval for Mean/Std. Deviation:
En **Analysis Type**, elija **One-Sample Analysis**.
En la sección **Variables**, seleccione **NewBalance**.
En la sección **Confidence Intervals to Calculate**:
Elija la opción **Fort the Mean**.
Seleccione 95% en **Confidence Level**.
Haga clic en **OK**.

Aparecerán algunos estadísticos descriptivos y el intervalo de confianza.

Determinación del tamaño de la muestra

En la sección 8.3 se mostró cómo determinar el tamaño de la muestra necesario para proporcionar un margen de error deseado. El ejemplo utilizado involucra un estudio diseñado para

estimar la media poblacional del costo diario del alquiler de automóviles medianos en Estados Unidos. El director del proyecto especificó que la media poblacional del costo del alquiler por día debe estimarse con un margen de error de $2 y un nivel de confianza de 95%. Los datos muestrales de un estudio anterior proporcionaron una desviación estándar muestral de $9.65; esta cifra se utilizará como el valor planeado de la desviación estándar poblacional. Los pasos siguientes pueden usarse para calcular el tamaño de la muestra recomendado para proporcionar un intervalo de confianza estimado de 95% de la media poblacional con un margen de error de $2.

Paso 1. Haga clic en la ficha **StatTools** en la cinta de opciones.
Paso 2. En el grupo **Analyses**, dé clic en **Statistical Inference**.
Paso 3. Elija la opción **Sample Size Selection**.
Paso 4. Cuando aparezca el cuadro de diálogo StatTools-Sample Size Selection:
En la sección **Parameter to Estimate**, elija **Mean**.
En la sección **Confidence Interval Specification**:
Elija **95%** en **Confidence Level**.
Ingrese **2** en el cuadro **Half-Length of Interval**.
Ingrese **9.65** en el cuadro **Estimated Std Dev**.
Haga clic en **OK**.

El valor en half-length of Interval es el margen de error.

Aparecerá el resultado presentando un tamaño de muestra recomendado de 90.

CAPÍTULO 9

Pruebas de hipótesis

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: JOHN MORRELL & COMPANY

9.1 FORMULACIÓN DE LAS HIPÓTESIS NULA Y ALTERNATIVA
- La hipótesis alternativa como hipótesis de investigación
- La hipótesis nula como un supuesto para ser rebatido
- Resumen de las formas para las hipótesis nula y alternativa

9.2 ERRORES TIPO I Y TIPO II

9.3 MEDIA POBLACIONAL: σ CONOCIDA
- Prueba de una cola
- Prueba de dos colas
- Resumen y consejo práctico
- Relación entre estimación por intervalo y prueba de hipótesis

9.4 MEDIA POBLACIONAL: σ DESCONOCIDA
- Prueba de una cola
- Prueba de dos colas
- Resumen y consejo práctico

9.5 PROPORCIÓN POBLACIONAL
- Resumen

9.6 PRUEBA DE HIPÓTESIS Y TOMA DE DECISIONES

9.7 CÁLCULO DE LA PROBABILIDAD DE LOS ERRORES TIPO II

9.8 DETERMINACIÓN DEL TAMAÑO DE LA MUESTRA EN UNA PRUEBA DE HIPÓTESIS PARA LA MEDIA POBLACIONAL

ESTADÍSTICA *en* LA PRÁCTICA

JOHN MORRELL & COMPANY*

CINCINNATI, OHIO

John Morrell & Company inició en Inglaterra en 1827 y es considerado el fabricante de productos cárnicos con operación continua más antiguo de Estados Unidos. Es una subsidiaria de propiedad absoluta y administrada independientemente de Smithfield Foods, Smithfield, Virginia. John Morrell & Company ofrece a los consumidores una amplia línea de productos de carne de puerco procesada y fresca de 13 marcas regionales que comprenden John Morrell, E-Z-Cut, Tobin's First Prize, Dinner Bell, Hunter, Kretschmar, Rath, Rodeo, Shenson, Farmers Hickory Brand, Iowa Quality y Peyton's. Cada marca regional disfruta del reconocimiento y la lealtad de sus consumidores.

Las investigaciones de mercado de Morrell proporcionan a los directivos información actualizada acerca de los diversos productos de la empresa y su posición en relación con las otras marcas competidoras de productos similares. En un estudio reciente se comparó uno de los productos de Morrell, Beef Pot Roast, con otros similares de dos de sus competidores principales. En esta prueba de comparación de los tres productos se empleó una muestra de consumidores para que indicaran cómo calificaban los productos en términos de sabor, apariencia, aroma y preferencia en general.

Una de las cuestiones que se deseaba investigar era si el producto de Morrell era la elección preferente de más de 50% de la población de consumidores. Si p representa la proporción poblacional que prefiere tal producto, la prueba de hipótesis para la cuestión que se investiga es la siguiente.

$$H_0: p \leq 0.50$$
$$H_a: p > 0.50$$

La hipótesis nula H_0 indica que la preferencia por el producto de Morrell es menor o igual que 50%. Si los datos muestrales respaldan el rechazo de H_0 en favor de la hipótesis alternativa H_a, la empresa concluirá que en una comparación de los tres productos, el suyo es preferido por más de 50% de la población de consumidores.



Platillos totalmente listos para que el consumidor los caliente y sirva en una charola incluida para horno de microondas.
© Cortesía de John Morrell's Convenient Cuisine Products.

En un estudio independiente se efectuó una prueba de degustación empleando una muestra de 224 consumidores de Cincinnati, Milwaukee y Los Ángeles, en la que 150 eligieron el producto de Morrell como el de su preferencia. A partir del procedimiento estadístico de prueba de hipótesis, la hipótesis nula fue rechazada. Mediante el estudio se encontraron evidencias estadísticas que favorecían la H_a y se llegó a la conclusión de que el producto de Morrell es preferido por más de 50% de la población de consumidores.

La estimación puntual de la proporción poblacional es $\bar{p} = 150/224 = 0.67$. De este modo, los datos muestrales sirvieron para hacer publicidad en una revista de alimentos en la cual se mostraba que en una comparación del sabor de los tres productos, el de Morrell era "preferido en una relación 2 a 1 sobre los de la competencia".

En este capítulo se estudiará cómo formular hipótesis y la forma de elaborar pruebas como la utilizada por Morrell. Mediante el análisis de datos muestrales se podrá determinar si una hipótesis debe o no ser rechazada.

* Los autores agradecen a Marty Butler, vicepresidente de Marketing de John Morrell, por proporcionar este artículo para *Estadística en la práctica*.

En los capítulos 7 y 8 se describió cómo usar una muestra para calcular estimaciones puntuales y por intervalo de parámetros poblacionales. En este capítulo se continúa con el estudio de la inferencia estadística mostrando cómo usar la prueba de hipótesis para determinar si una afirmación acerca del valor de un parámetro poblacional debe o no ser rechazada.

En las pruebas de hipótesis se empieza por hacer un supuesto tentativo acerca del parámetro poblacional. A este supuesto tentativo se le llama **hipótesis nula**, y se denota por H_0. Después se define otra hipótesis, llamada **hipótesis alternativa**, que contradice lo que establece

la hipótesis nula y se denota como H_a. En el procedimiento de pruebas de hipótesis se usan datos de una muestra para probar dos afirmaciones contrarias indicadas por H_0 y H_a.

En este capítulo se describe el modo de realizar pruebas de hipótesis acerca de una media poblacional y una proporción poblacional. Para empezar, se facilitan ejemplos que ilustran los métodos para desarrollar las hipótesis nula y alternativa.

9.1 Formulación de las hipótesis nula y alternativa

Para aprender a formular correctamente las hipótesis se necesita práctica. Se debe esperar al principio cierta confusión en la elección apropiada de la hipótesis nula y la hipótesis alternativa. Los ejemplos de esta sección tienen el propósito de proporcionar algunas directrices.

No siempre es obvio cómo formular las hipótesis nula y alternativa. Se debe tener cuidado en estructurarlas de manera apropiada para que la conclusión de la prueba de hipótesis proporcione la información que el investigador o la persona que toma las decisiones desea. El contexto de la situación es muy importante para determinar cómo deben establecerse las hipótesis. Todas las aplicaciones de prueba de hipótesis involucran la recolección de una muestra y el uso de resultados muestrales para proporcionar evidencias y emitir conclusiones. Algunas buenas preguntas a considerar al formular las hipótesis nula y alternativa son: ¿cuál es el propósito de recolectar la muestra? ¿Qué conclusiones se espera formular?

En la introducción del capítulo se establece que la hipótesis nula H_0 es un supuesto tentativo acerca de un parámetro poblacional tal como una media poblacional o una proporción poblacional. La hipótesis alternativa H_a es una declaración que contradice lo que establece la hipótesis nula. En algunas situaciones es más fácil identificar la hipótesis alternativa primero y luego desarrollar la nula. En otras es más fácil identificar la hipótesis nula primero y luego desarrollar la alternativa. En los siguientes ejemplos se ilustrarán esas situaciones.

La hipótesis alternativa como hipótesis de investigación

Numerosas aplicaciones de prueba de hipótesis involucran un intento de obtener evidencia en apoyo de una hipótesis de investigación. En tales situaciones, con frecuencia es mejor empezar con la hipótesis alternativa y convertirla en la conclusión que el investigador espera sustentar. Considere un modelo de automóvil determinado que actualmente alcanza un rendimiento de gasolina de 24 millas por galón en manejo urbano. Un grupo de investigación de productos desarrolló un nuevo sistema de inyección de combustible diseñado para dar un mejor rendimiento en millas por galón de gasolina. El grupo realizará pruebas controladas con el nuevo sistema de inyección de combustible en busca de un sustento estadístico para concluir que proporciona más millas por galón que el sistema actual.

Se fabricarán varias unidades del nuevo sistema de inyección de combustible, se instalarán en automóviles de prueba y se someterán a condiciones de manejo bajo investigación controlada. Se calculará la media muestral de millas por galón para esos autos y se utilizará en una prueba de hipótesis para determinar si se puede concluir que el nuevo sistema de inyección de combustible proporciona más de 24 millas por galón. En términos de la media poblacional de millas por galón μ, la hipótesis de investigación $\mu > 24$ se convierte en la hipótesis alternativa. El sistema actual proporciona un promedio o media de 24 millas por galón, por lo que se hace el supuesto tentativo de que el nuevo sistema no es de ninguna manera mejor que el actual y se escoge $\mu \leq 24$ como la hipótesis nula. Las hipótesis nula y alternativa adecuadas son

$$H_0: \mu \leq 24$$
$$H_a: \mu > 24$$

Si los resultados muestrales llevan a la conclusión de rechazar H_0, se puede hacer la inferencia de que $\mu > 24$ es verdadera. Los investigadores tendrían el sustento estadístico necesario para afirmar que el nuevo sistema de inyección de combustible aumenta el rendimiento medio en millas por galón. Debería considerarse por tanto la producción de automóviles con el nuevo sistema de inyección de combustible. Pero si los resultados obtenidos indican que no se puede

La conclusión de que la hipótesis de investigación es verdadera se formula si los datos muestrales proporcionan suficiente evidencia para demostrar que se puede rechazar la hipótesis nula.

rechazar H_0, los investigadores no pueden concluir que el nuevo sistema es mejor que el actual. La producción de automóviles con el nuevo diseño no se puede justificar sobre la base de un millaje mayor por gasolina. Quizá será necesario investigar más y realizar futuras pruebas.

Las empresas exitosas se mantienen en la competencia desarrollando nuevos productos, métodos, marcas, sistemas y similares, que son lo mejor de lo que se dispone en la actualidad. Antes de adoptar algo nuevo, es deseable realizar investigación para determinar si hay sustento estadístico para la conclusión de que el nuevo enfoque es en efecto mejor. En tales casos, la hipótesis de investigación se establece como la hipótesis alternativa. Por ejemplo, se desarrolla un método nuevo de enseñanza que se considera mejor que el actual. La hipótesis alternativa indica que el método nuevo es mejor. La hipótesis nula establece que el método nuevo no es mejor que el antiguo. Se desarrolla un nuevo plan de bono para la fuerza de ventas en un intento por aumentar estas últimas. La hipótesis alternativa es que el nuevo plan de bono aumentará las ventas. La hipótesis nula es que el nuevo plan de bono no aumentará las ventas. Se desarrolla un medicamento con el objetivo de reducir la presión arterial con mayor eficacia que un medicamento ya existente. La hipótesis alternativa es que el nuevo fármaco reducirá la presión arterial más que el anterior. La hipótesis nula indica que el nuevo medicamento no reducirá la presión arterial más que la medicina existente. En cada caso, el rechazo de la hipótesis nula H_0 proporciona el sustento estadístico para la hipótesis de investigación. Se verán muchos ejemplos de pruebas de hipótesis en situaciones de investigación como éstas a lo largo de este capítulo y en lo que resta en el libro.

La hipótesis nula como un supuesto para ser rebatido

Naturalmente, no todas las pruebas de hipótesis involucran hipótesis de investigación. En el siguiente análisis veremos aplicaciones de pruebas de hipótesis donde se inicia con la creencia o supuesto de que una declaración acerca del valor de un parámetro poblacional es verdadero. Luego se usará una prueba de hipótesis para rebatir el supuesto y determinar si hay evidencia estadística para concluir que no es correcto. En tales situaciones, resulta útil establecer primero la hipótesis nula. La H_0 expresa la creencia o supuesto acerca del valor del parámetro poblacional. La hipótesis alternativa H_a establece que la creencia o supuesto no es correcto.

Como ejemplo, considere la situación de un fabricante de bebidas refrescantes. La etiqueta en los envases de bebida asegura que contienen 67.6 onzas de líquido. Se considera correcta la leyenda toda vez que la media poblacional de peso de llenado de los envases es *por lo menos* de 67.6 onzas de líquido. Sin razón alguna para creer otra cosa, se le da al fabricante el beneficio de la duda y se asume que la información proporcionada en la etiqueta es correcta. Así, en una prueba de hipótesis acerca de la media poblacional de peso de líquido por botella, se debería comenzar con el supuesto de que la leyenda es correcta y se establece la hipótesis nula como $\mu \geq 67.6$. El desafío para este supuesto implicaría que la leyenda no es correcta y que los envases se llenan de forma insuficiente. Este reto al supuesto deberá establecerse como la hipótesis alternativa $\mu < 67.6$. Así, las hipótesis nula y alternativa son:

$$H_0: \mu \geq 67.6$$
$$H_a: \mu < 67.6$$

Usualmente se asume como cierta la información que proporciona un fabricante acerca de su producto y se establece como hipótesis nula. Puede formularse la conclusión de que la información no es correcta si la hipótesis nula es rechazada.

Una agencia gubernamental responsable de validar las etiquetas de fabricación podría seleccionar una muestra de envases con bebida refrescante, calcular la media muestral del peso de llenado y usar los resultados para probar las hipótesis anteriores. Si los resultados muestrales llevan a la conclusión de rechazar H_0, se puede hacer la inferencia de que H_a: $\mu < 67.6$ es verdadera. Con este sustento estadístico, la agencia tiene justificada la conclusión de que la leyenda no es correcta y se está realizando un llenado insuficiente de los envases. Se podrán considerar acciones para obligar al fabricante a cumplir con los estándares del etiquetado. Pero si los resultados muestrales indican que no se puede rechazar H_0, no es apropiado rechazar el supuesto de que el etiquetado del fabricante es correcto. Con esta conclusión no se puede realizar ninguna acción.

Analicemos ahora una variación del ejemplo de las bebidas refrescantes viendo la misma situación desde la perspectiva del fabricante. La operación de llenado de los envases está diseñada para completarlos con 67.6 onzas de líquido como se declara en la etiqueta. La empresa no quiere llenar de manera incompleta los contenedores porque podría terminar en una queja de los clientes por llenado insuficiente, o quizás hasta de una agencia gubernamental. Sin embargo, tampoco quiere sobrellenar los contenedores, pues agregar más bebida refrescante de la apropiada podría resultar un costo innecesario. La meta de la empresa sería ajustar la operación de forma tal que la media poblacional del peso de llenado por envase sea 67.6 onzas de líquido como se declara en la etiqueta.

Aunque ésta es la meta de la empresa, de tiempo en tiempo cualquier proceso de producción puede salirse del ajuste. Si esto ocurre en el ejemplo, podría presentarse un llenado insuficiente o en exceso de la bebida refrescante. En ambos casos la empresa quisiera saberlo a fin de corregir la situación reajustando la operación de llenado a las 67.6 onzas de líquido programadas. En una aplicación de prueba de hipótesis, se empezaría de nuevo con el supuesto de que el proceso de producción opera de forma correcta y establecer la hipótesis nula como $\mu = 67.6$ onzas de líquido. La hipótesis alternativa que rebate este supuesto sostiene que $\mu \neq 67.6$, la cual indica que está ocurriendo llenado insuficiente o en demasía. Las hipótesis nula y alternativa de la prueba de hipótesis del fabricante son:

$$H_0: \mu = 67.6$$
$$H_a: \mu \neq 67.6$$

Suponga que el fabricante utiliza un procedimiento de control de calidad para seleccionar periódicamente una muestra de envases de la operación de llenado y calcular la media muestral del peso de llenado por botella. Si los resultados muestrales llevan a la conclusión de rechazar H_0, se puede hacer la inferencia de que $H_a: \mu \neq 67.6$ es verdadera. Concluimos que los contenedores no se están llenando de manera apropiada y el proceso de producción debe ajustarse para restaurar la media poblacional a 67.6 onzas de líquido por envase. Pero si los resultados muestrales indican que no se puede rechazar H_0, no es posible descartar el supuesto de que la operación de llenado de los envases del fabricante funciona de manera apropiada. En este caso no se tomaría ninguna acción adicional y la producción continuaría adelante.

Las dos formas anteriores de pruebas de hipótesis del fabricante de bebidas refrescantes muestran que las hipótesis nula y alternativa varían dependiendo del punto de vista del investigador o de quien toma las decisiones. Para formular hipótesis correctamente, es importante comprender el contexto de la situación y estructurarlas a efecto de proporcionar la información que requiere el investigador o quien toma la decisión.

Resumen de las formas para las hipótesis nula y alternativa

Las pruebas de hipótesis de este capítulo se refieren a dos parámetros poblacionales: la media poblacional y la proporción poblacional. A partir de la situación, las pruebas de hipótesis para un parámetro poblacional asumen una de estas tres formas: en dos se emplean desigualdades en la hipótesis nula, y en la tercera se aplica una igualdad en la hipótesis nula. En las pruebas de hipótesis para la media poblacional, μ_0 denota el valor hipotético, y hay que escoger una de las formas siguientes.

$$H_0: \mu \geq \mu_0 \qquad H_0: \mu \leq \mu_0 \qquad H_0: \mu = \mu_0$$
$$H_a: \mu < \mu_0 \qquad H_a: \mu > \mu_0 \qquad H_a: \mu \neq \mu_0$$

Aquí se muestran las tres formas que pueden tener H_0 y H_a. Observe que en la hipótesis nula H_0 siempre aparece la igualdad.

Por razones que se aclararán más adelante, a las dos primeras formas se les llama *pruebas de una cola*. A la tercera se le llama *prueba de dos colas*.

En muchas situaciones no es obvio cómo elegir H_0 y H_a, y resulta necesario el criterio para elegirlas en forma adecuada. Sin embargo, como se observa en las formas anteriores, la

igualdad (ya sea $\geq$, $\leq$ o $=$) debe aparecer *siempre* en la hipótesis nula. Al elegir la forma adecuada para H_0 y H_a hay que tener en mente que la hipótesis alternativa a menudo es lo que la prueba trata de demostrar. Por tanto, preguntarse si el usuario busca evidencias en apoyo de $\mu < \mu_0$, $\mu > \mu_0$, o $\mu \neq \mu_0$ ayudará a determinar H_a. Los ejercicios siguientes tienen por objeto aportar práctica en la elección de la forma adecuada de una prueba de hipótesis para la media poblacional.

Ejercicios

1. El gerente del Danvers-Hilton Resort Hotel afirma que la cantidad media que gastan los huéspedes en un fin de semana es de $600 o menos. Un miembro del equipo de contadores observó que en los últimos meses habían aumentado tales cantidades. El contador emplea una muestra de las cuentas de fin de semana de los huéspedes para probar la afirmación del gerente.
   a) ¿Qué forma de hipótesis deberá usar para probar la afirmación del gerente? Explique.

$$H_0: \mu \geq 600 \qquad H_0: \mu \leq 600 \qquad H_0: \mu = 600$$
$$H_a: \mu < 600 \qquad H_a: \mu > 600 \qquad H_a: \mu \neq 600$$

   b) ¿Cuál es la conclusión apropiada cuando no se puede rechazar la hipótesis nula H_0?
   c) ¿Qué conclusión es adecuada cuando se puede rechazar la hipótesis nula H_0?

AUTOevaluación

2. El gerente de un negocio de venta de automóviles piensa en un nuevo plan de bono diseñado para incrementar el volumen de ventas. En el momento actual, el volumen medio de ventas es 14 automóviles por mes. El gerente desea realizar un estudio para ver si el plan de bono incrementa el volumen de ventas. Para recolectar los datos, se le permitirá a una muestra de vendedores vender bajo el nuevo plan de bono durante un mes.
   a) Desarrolle las hipótesis nula y alternativa más adecuadas para esta situación.
   b) Comente la conclusión en caso de que no pueda rechazarse H_0.
   c) Comente la conclusión en caso de que pueda rechazarse H_0.
3. Una operación de la línea de producción está diseñada para llenar cajas con un peso medio de 32 onzas de detergente para lavar. Con periodicidad se selecciona una muestra de los empaques y se pesan para determinar si se están llenando de manera insuficiente o en demasía. Si los datos muestrales llevan a la conclusión de que hay llenado insuficiente o excesivo, la producción se suspende y se ajusta al llenado correcto.
   a) Formule las hipótesis nula y alternativa que ayudarán a determinar si se debe detener la producción y ajustar el peso.
   b) Comente sobre la conclusión y la decisión en caso de que H_0 no se pueda rechazar.
   c) Comente acerca de la conclusión y la decisión en caso de que H_0 se pueda rechazar.
4. Antes de implantar un método de fabricación propuesto, y debido a los costos y al tiempo de adaptación de la producción, un director de manufactura debe convencer a la dirección de que ese método nuevo reducirá los costos. El costo medio del actual método de producción es $220 por hora. Un estudio de investigación medirá el costo del método nuevo durante un periodo muestral de producción.
   a) Formule las hipótesis nula y alternativa más adecuadas para este estudio.
   b) Comente acerca de la conclusión cuando H_0 no pueda rechazarse.
   c) Comente acerca de la conclusión cuando H_0 pueda rechazarse.

9.2 Errores tipo I y tipo II

Las hipótesis nula y alternativa son afirmaciones opuestas acerca de la población. Una de las dos, ya sea la hipótesis nula H_0 o la alternativa H_a, es verdadera, pero no ambas. Lo ideal es que la prueba de hipótesis lleve a la aceptación de H_0 cuando sea verdadera y a su rechazo en

TABLA 9.1 Errores y conclusiones correctas en las pruebas de hipótesis

		Condición poblacional	
		H_0 verdadera	H_a verdadera
Conclusión	H_0 es aceptada	Conclusión correcta	Error tipo II
	H_0 es rechazada	Error tipo I	Conclusión correcta

caso de que H_a sea verdadera. Desafortunadamente, las conclusiones correctas no siempre son posibles. Como la prueba de hipótesis se basa en una información muestral, debe considerarse que existe la posibilidad de error. La tabla 9.1 ilustra las dos clases de errores comunes en una prueba de hipótesis.

En la primera fila se ilustra qué sucede cuando H_0 es aceptada. Si H_0 es verdadera, la conclusión es correcta. Pero si H_a es verdadera, se comete un **error tipo II**; es decir, H_0 es aceptada cuando es falsa. En la segunda fila de la tabla 9.1 se muestra qué sucede si la conclusión es rechazar H_0. Si H_0 es verdadera, se comete un **error tipo I**; es decir, H_0 es rechazada cuando es verdadera. Pero si H_a es verdadera, es correcto rechazar H_0.

Recuerde la prueba de hipótesis analizada en la sección 9.1 en la cual un grupo de investigación desarrolló un nuevo sistema de inyección de combustible con objeto de aumentar el rendimiento del hidrocarburo en un determinado modelo de automóvil. Como con el sistema actual el rendimiento promedio es 24 millas por galón, la prueba de hipótesis se formuló como sigue.

$$H_0: \mu \leq 24$$
$$H_a: \mu > 24$$

La hipótesis alternativa, H_a: $\mu > 24$, indica que los investigadores buscan evidencias muestrales que apoyen la conclusión de que con el nuevo sistema de inyección de combustible la media poblacional del rendimiento es mayor que 24.

En esta aplicación, el error tipo I de rechazar H_0 cuando es verdadera implica que los investigadores afirmen que el nuevo sistema mejora el rendimiento de millas por galón ($\mu > 24$) cuando en realidad no es nada mejor que el actual. En cambio, el error tipo II de aceptar H_0 cuando es falsa corresponde a la conclusión de los investigadores de que el nuevo sistema no es mejor que el actual ($\mu \leq 24$) cuando en realidad sí mejora el rendimiento de millas por galón.

En la prueba de hipótesis del rendimiento de millas por galón, la hipótesis nula es H_0: $\mu \leq 24$. Admita que la hipótesis nula es verdadera como una igualdad; es decir $\mu = 24$. A la probabilidad de cometer un error tipo I cuando la hipótesis nula es verdadera como igualdad se le conoce como **nivel de significancia**. Por tanto, en la prueba de hipótesis del rendimiento de combustible, el nivel de significancia es la probabilidad de rechazar H_0: $\mu \leq 24$ cuando $\mu = 24$. Dada la importancia de este concepto, se redacta otra vez la definición de nivel de significancia.

NIVEL DE SIGNIFICANCIA

Consiste en la probabilidad de cometer un error tipo I cuando la hipótesis nula es verdadera como igualdad.

Para denotar el nivel de significancia se usa la letra griega α (alfa), y los valores que suelen utilizarse para α son 0.05 y 0.01.

En la práctica, el responsable de la prueba de hipótesis especifica el nivel de significancia. Al elegir α controla la probabilidad de cometer un error tipo I. Si el costo de cometer este error es alto, los valores pequeños de α son preferibles. Si el costo no es demasiado alto, entonces usa valores mayores para α. A las aplicaciones de la prueba de hipótesis en que sólo se controla el error tipo I se les llama *pruebas de significancia.* Muchas aplicaciones de las pruebas de hipótesis son de este tipo.

Si los datos muestrales son consistentes con la hipótesis nula H_0, se seguirá la práctica de concluir que "no es rechazada H_0". Esta conclusión es preferible a la de "H_0 es aceptada", porque al aceptarla se corre el riesgo de cometer un error tipo II.

Aunque en la mayoría de las aplicaciones de las pruebas de hipótesis se controla la probabilidad de cometer un error tipo I, no siempre sucede lo mismo con uno tipo II. Por tanto, si se decide aceptar H_0, no es posible establecer el nivel de confianza en esa decisión. Debido a la incertidumbre asociada con el hecho de cometer un error tipo II al realizar una prueba de significancia, los profesionales de la estadística suelen recomendar que se diga "H_0 no es rechazada" en lugar de "H_0 es aceptada". Decir "H_0 no es rechazada" implica la recomendación de reservarse tanto el juicio como la acción. En efecto, al no aceptar directamente H_0, se evita el riesgo de cometer un error tipo II. Siempre que no se determine y controle la probabilidad de cometerlo, no se dirá "H_0 es aceptada". En esos casos sólo son posibles dos conclusiones: H_0 *no es rechazada* o H_0 *es rechazada*.

Aunque es poco común controlar el error tipo II en una prueba de hipótesis, es posible. En las secciones 9.7 y 9.8 se ilustra el procedimiento para controlar y determinar la probabilidad de cometer este tipo de error. Si se ha establecido un control adecuado del mismo, las medidas basadas en la conclusión "H_0 es aceptada" pueden ser adecuadas.

NOTAS Y COMENTARIOS

Walter Williams, columnista y profesor de economía de la Universidad George Mason, indica que existe siempre la posibilidad de cometer un error tipo I o un error tipo II al tomar cualquier decisión (*The Cincinnati Enquirer*, 14 de agosto de 2005). Hace notar que la Food and Drug Administration (FDA) corre el riesgo de cometer estos errores en sus procedimientos para la aprobación de medicamentos. Cuando incurre en un error tipo I, la FDA no aprueba un medicamento que es seguro y efectivo. Al cometer en un error tipo II, aprueba un fármaco que presenta efectos secundarios imprevistos. Sin importar la decisión que se tome, la probabilidad de cometer un error costoso no se puede eliminar.

Ejercicios

AUTOevaluación

5. Nielsen informó que los hombres jóvenes estadounidenses ven diariamente 56.2 minutos de televisión en las horas de mayor audiencia (*The Wall Street Journal Europe*, 18 de noviembre de 2003). Un investigador cree que en Alemania los jóvenes ven más tiempo la televisión en las horas de mayor audiencia. Este investigador toma una muestra de hombres jóvenes alemanes y registra el tiempo que ven televisión en un día. Los resultados muestrales se usan para probar las siguientes hipótesis nula y alternativa.

$$H_0: \mu \leq 56.2$$
$$H_a: \mu > 56.2$$

   a) ¿Cuál es el error tipo I en esta situación? ¿Qué consecuencia tiene cometerlo?
   b) ¿Cuál es el error tipo II en esta situación? ¿Qué consecuencia tiene cometerlo?

6. En la etiqueta de una botella de jugo de naranja de 3 cuartos de galón se afirma que el jugo contiene en promedio 1 gramo o menos de grasa. Responda las preguntas siguientes relacionadas con una prueba de hipótesis para probar lo que se asegura en la etiqueta.
   a) Desarrolle las hipótesis nula y alternativa adecuadas.

b) ¿Cuál es el error tipo I en esta situación? ¿Qué consecuencias tiene cometerlo?
c) ¿Cuál es el error tipo II en esta situación? ¿Qué consecuencias tiene cometerlo?

7. El personal de ventas de Carpetland tiene un promedio de $8 000 semanales en ventas. Steve Contois, vicepresidente de la empresa, propone un plan de compensación con nuevos incentivos. Steve espera que los resultados de un periodo de prueba permitirán concluir que el plan de compensación aumenta el promedio de ventas de los vendedores.
   a) Establezca las hipótesis nula y alternativa adecuadas.
   b) ¿Cuál es el error tipo I en esta situación? ¿Qué consecuencias tiene cometerlo?
   c) ¿Cuál es el error tipo II en esta situación? ¿Qué consecuencias tiene cometerlo?

8. Suponga que se implementará un nuevo método de producción si mediante una prueba de hipótesis se confirma la conclusión de que el nuevo método reduce el costo medio de operación por hora.
   a) Proporcione las hipótesis nula y alternativa adecuadas si el costo medio de producción actual por hora es $220.
   b) ¿Cuál es el error tipo I en esta situación? ¿Qué consecuencias tiene cometerlo?
   c) ¿Cuál es el error tipo II en esta situación? ¿Qué consecuencias tiene cometerlo?

9.3 Media poblacional: σ conocida

En el capítulo 8 se dijo que el caso de σ conocida se refiere a aplicaciones en las que se cuenta con datos históricos o con alguna información que permita obtener buenas estimaciones de la desviación estándar poblacional antes de tomar la muestra. En tales casos, para propósitos prácticos se considera que se conoce la desviación estándar poblacional. En esta sección se muestra cómo realizar una prueba de hipótesis para la media poblacional en el caso en que σ es conocida.

Los métodos que se presentan dan resultados exactos si la población de la que se selecciona la muestra tiene distribución normal. En los casos en los que no sea razonable suponer que la población tiene esta distribución, se pueden aplicar estos métodos siempre y cuando el tamaño de la muestra sea suficientemente grande. Al final de esta sección se proporcionan algunos consejos prácticos en relación con la distribución poblacional y el tamaño de la muestra.

Prueba de una cola

La **prueba de una cola** para la media poblacional toma una de las dos formas siguientes.

Prueba de cola inferior (o izquierda)	Prueba de cola superior (o derecha)
$H_0: \mu \geq \mu_0$	$H_0: \mu \leq \mu_0$
$H_a: \mu < \mu_0$	$H_a: \mu > \mu_0$

A continuación se presenta un ejemplo de una prueba para la cola inferior.

La Federal Trade Commission (FTC) de Estados Unidos realiza periódicamente estudios estadísticos con objeto de comprobar las afirmaciones de los fabricantes acerca de sus productos. Por ejemplo, en la etiqueta de una lata grande de Hilltop Coffee se dice que contiene 3 libras de café. La FTC sabe que el proceso de producción de Hilltop no permite llenar las latas con 3 libras exactas de café, incluso si la media poblacional del peso de llenado de todas las latas es esa cantidad por unidad. Sin embargo, mientras la media poblacional del peso de llenado sea por lo menos de 3 libras por lata, los derechos del consumidor estarán protegidos. Por tanto, la FTC interpreta que la información de la etiqueta de un contenedor grande de café Hilltop tiene una media poblacional del peso de llenado de por lo menos 3 libras por lata. Se mostrará cómo verificar esto realizando una prueba de hipótesis de cola inferior.

El primer paso consiste en desarrollar las hipótesis nula y alternativa para la prueba. Si la media poblacional del peso de llenado es por lo menos de 3 libras por lata, lo que afirma Hilltop es correcto. Esto establece la hipótesis nula de la prueba. No obstante, si la media poblacional del peso de llenado es menor que 3 libras por lata, la afirmación de Hilltop es incorrecta. Así,

se establece la hipótesis alternativa. Si μ denota la media poblacional del peso de llenado, las hipótesis nula y alternativa son las siguientes.

$$H_0: \mu \geq 3$$
$$H_a: \mu < 3$$

Observe que el valor hipotético de la media poblacional es $\mu_0 = 3$.

Si los datos muestrales indican que H_0 no puede ser rechazada, la evidencia estadística no conducirá a concluir que ha habido una violación en lo que se afirma en la etiqueta. Luego, no se tomará ninguna acción en contra de Hilltop. Pero si los datos muestrales indican que H_0 puede ser rechazada, se concluirá que la hipótesis alternativa H_a: $\mu < 3$ es verdadera. En este caso la conclusión de que hay falta de peso y un cargo por violación a lo que se establece en la etiqueta estarán justificados.

Suponga que se selecciona una muestra de 36 latas de café y se calcula la media muestral $\bar{x}$ como una estimación de la media poblacional μ. Si el valor de la media muestral $\bar{x}$ es menor de 3 libras, los resultados muestrales despertarán dudas sobre lo que establece la hipótesis nula. Lo que se busca saber es cuánto menos de 3 libras tiene que ser $\bar{x}$ para declarar que la diferencia es significativa y se esté dispuesto a correr el riesgo de cometer un error tipo I al acusar indebidamente a Hilltop de violar lo que establece en la etiqueta. Aquí el factor clave es el valor elegido como nivel de significancia por quien tomará la decisión.

Como se hizo notar en la sección anterior, el nivel de significancia, que se denota como α, es la probabilidad de cometer un error tipo I al rechazar la hipótesis nula cuando ésta, considerada en forma de una igualdad, es verdadera. La persona que tomará la decisión debe especificar el nivel de significancia. Si el costo de cometer un error tipo I es alto, se deberá elegir un valor pequeño para el nivel de significancia. Si el costo no es alto, es más apropiado seleccionar un valor grande. En el caso del café Hilltop, el director del programa de pruebas de la FTC afirma: "Si la empresa satisface sus especificaciones de peso en $\mu = 3$, no tomaré ninguna medida en su contra. Pero estoy dispuesto a asumir un riesgo de 1% de cometer tal error." De acuerdo con lo establecido por el director, el nivel de significancia en esta prueba de hipótesis se establece en $\alpha = 0.01$. Así, la prueba de hipótesis deberá diseñarse de manera que la probabilidad de cometer un error tipo I cuando $\mu = 3$ sea 0.01.

En este estudio sobre Hilltop Coffee, al proponer las hipótesis nula y alternativa y especificar el nivel de significancia para la prueba se han dado los dos primeros pasos requeridos en cualquier prueba de hipótesis. Con esto estamos listos para el tercer paso en una prueba de hipótesis: recabar los datos muestrales y calcular el valor de lo que se conoce como el *estadístico de prueba*.

Estadístico de prueba En el estudio de Hilltop Coffee las pruebas realizadas con anterioridad por la FTC indican que la desviación estándar poblacional se considera conocida, siendo su valor $\sigma = 0.18$. Estas pruebas muestran también que se puede sostener que la población de los pesos de llenado tiene una distribución normal. Según lo estudiado en el capítulo 7 sobre distribuciones de muestreo, sabemos que si la población de la que se toma la muestra tiene una distribución normal, la distribución de muestreo de $\bar{x}$ también es normal. En consecuencia, en el estudio de Hilltop Coffee, la distribución de muestreo de $\bar{x}$ será normal. Con un valor conocido de $\sigma = 0.18$ y un tamaño de muestra de $n = 36$, en la figura 9.1 se ilustra la distribución de muestreo de $\bar{x}$ si la hipótesis nula, considerada como igualdad, es verdadera; es decir, cuando $\mu = \mu_0 = 3$.[1] Observe que el error estándar de $\bar{x}$ está dado por $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 0.18/\sqrt{36} = 0.3$.

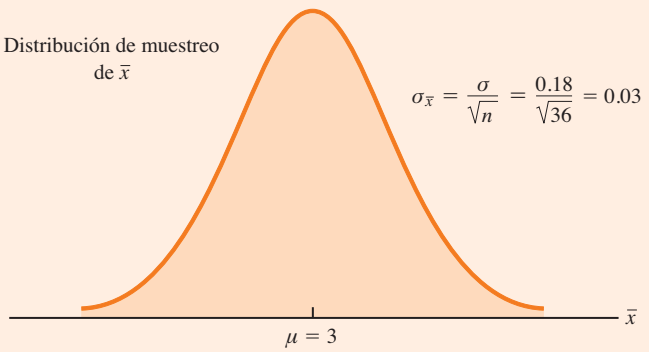
El error estándar de $\bar{x}$ es la desviación estándar de la distribución de muestreo de $\bar{x}$.

Como la distribución de muestreo de $\bar{x}$ está distribuida normalmente, la distribución de muestreo de

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{\bar{x} - 3}{0.03}$$

[1] Cuando se elaboran distribuciones de muestreo para una prueba de hipótesis, se asume que H_0 es satisfecha como igualdad.

FIGURA 9.1 Distribución de muestreo de $\bar{x}$ en el estudio de Hilltop Coffee cuando la hipótesis nula es verdadera como igualdad ($\mu = 3$)



es una distribución normal estándar. Si el valor de $z = -1$, esto significa que el valor de $\bar{x}$ es un error estándar menor que el valor hipotético de la media; si el valor de $z = -2$, esto significa que el valor de $\bar{x}$ es dos errores estándar menor que el valor hipotético de la media, y así sucesivamente. Para determinar la probabilidad que corresponde a cualquier valor de z en la cola inferior se usa la tabla de probabilidad normal estándar. Por ejemplo, el área en la cola inferior para $z = -3.00$ es 0.0013. Así, la probabilidad de obtener un valor de z que sea tres o más errores estándar menor que la media es 0.0013. Como resultado, la probabilidad de registrar un valor de $\bar{x}$ que sea 3 o más errores estándar menor que la media poblacional hipotética $\mu_0 = 3$ también es 0.0013. Si la hipótesis nula es verdadera, un resultado así es poco probable.

En una prueba de hipótesis para la media poblacional en el caso de σ conocida, se emplea la variable aleatoria normal estándar z como **estadístico de prueba** para determinar si $\bar{x}$ se desvía lo suficiente del valor hipotético de μ como para justificar el rechazo de la hipótesis nula. Como $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, el estadístico de prueba es el siguiente.

ESTADÍSTICO DE PRUEBA EN LAS PRUEBAS DE HIPÓTESIS PARA LA MEDIA POBLACIONAL: σ CONOCIDA

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \qquad \textbf{(9.1)}$$

La cuestión clave en una prueba de cola inferior es: ¿qué tan pequeño debe ser el estadístico de prueba z para que se decida rechazar la hipótesis nula? Para responder esta pregunta se usan dos métodos: el método del valor-p y el método del valor crítico.

Método del valor-p En este enfoque se usa el valor del estadístico de prueba z para calcular una probabilidad llamada **valor-*p***.

Un valor-p *pequeño indica que el valor del estadístico de prueba es inusual bajo el supuesto de que H_0 es verdadera.*

VALOR-p

Es una probabilidad que aporta una medida de la evidencia suministrada por la muestra contra la hipótesis nula. Valores-p pequeños indican una evidencia mayor contra H_0.

El valor p se utiliza para determinar si la hipótesis nula debe ser rechazada.

Ahora se verá cómo se calcula y utiliza el valor-*p*. Para el cálculo, se usa el valor del estadístico de prueba. El método a seguir depende de si se trata de una prueba de cola inferior, de cola superior o de dos colas. En la primera, el valor-*p* es la probabilidad de conseguir un valor del estadístico de prueba tan pequeño o menor que el obtenido con la muestra. Por ende, para calcular el valor-*p* en una prueba de cola inferior, en el caso de σ conocida, se debe determinar el área bajo la curva normal estándar para valores de $z \leq$ que el valor del estadístico de prueba. Una vez calculado el valor-*p*, se debe decidir si es lo suficientemente pequeño para rechazar la hipótesis nula. Como se verá más adelante, para esta decisión hay que comparar el valor-*p* con el nivel de significancia.

WEB archivo
Coffee

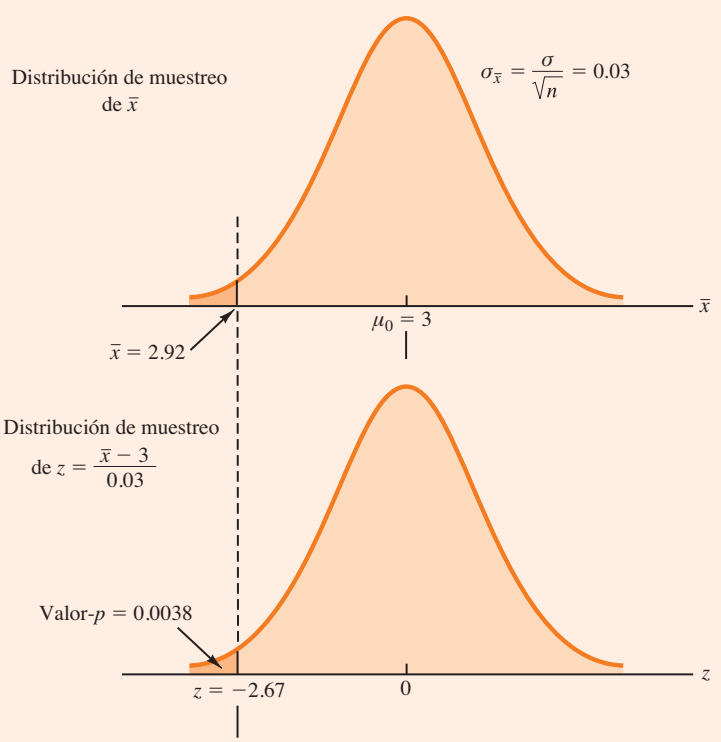
Ahora calculamos el valor-*p* para la prueba de cola inferior del estudio de Hilltop Coffee. Suponga que en la muestra de las 36 latas de café, la media muestral obtenida es $\bar{x} = 2.92$ libras. ¿Es $\bar{x} = 2.92$ lo suficientemente pequeña para que se rechace H_0? Como es una prueba de cola inferior, el valor-*p* es el área bajo la curva normal estándar para valores de $z \leq$ que el valor del estadístico de prueba. Al usar $\bar{x} = 2.92$, $\sigma = 0.18$ y $n = 36$, se determina el valor del estadístico de prueba *z*.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{2.92 - 3}{0.18/\sqrt{36}} = -2.67$$

Por consiguiente, el valor-*p* es la probabilidad de que el estadístico de prueba *z* sea menor o igual que -2.67 (el área bajo la curva normal estándar a la izquierda del estadístico de prueba).

En la tabla de probabilidad normal estándar aparece que el área en la cola inferior para $z = -2.67$ es 0.0038. En la figura 9.2 se muestra que a $\bar{x} = 2.92$ le corresponde $z = -2.67$ y el valor-$p = 0.0038$. Este último indica que si la muestra se ha tomado de una población con $\mu = 3$, la probabilidad de obtener una media muestral $\bar{x} = 2.92$ (y un estadístico de prueba de -2.67), o menor, es pequeña. Este valor-*p* no favorece mucho la hipótesis nula, pero, ¿es lo

FIGURA 9.2 Valor-*p* en el estudio de Hilltop Coffee, en el que $\bar{x} = 2.92$ y $z = -2.67$



suficientemente pequeño como para que H_0 sea rechazada? La respuesta depende del nivel de significancia de la prueba.

Como se indicó antes, el director del programa de pruebas de la FTC eligió como nivel de significancia un valor de 0.01. Seleccionar $\alpha = 0.01$ significa que él está dispuesto a tolerar una probabilidad de 0.01 para rechazar la hipótesis nula cuando sea verdadera como igualdad ($\mu_0 = 3$). La muestra de 36 latas de Hilltop Coffee dio como resultado un valor-$p = 0.0038$, lo cual significa que la probabilidad de obtener $\bar{x} = 2.92$ o menor, si la hipótesis nula considerada como igualdad es verdadera, es 0.0038. Como 0.0038 es menor o igual que $\alpha = 0.01$, H_0 es rechazada. De manera que para el nivel de significancia 0.01 se encontró evidencia estadística suficiente para rechazar la hipótesis nula.

Ahora se puede establecer ya la regla general para determinar cuándo rechazar la hipótesis nula al usar el método del valor-*p*. Dado un nivel de significancia α, la regla para el rechazo utilizando el método del valor-*p* es la siguiente.

REGLA PARA EL RECHAZO USANDO EL VALOR-*p*

Rechazar H_0 si el valor-$p \leq \alpha$

En la prueba para Hilltop Coffee, el valor-*p* de 0.0038 llevó a que la hipótesis nula fuera rechazada. Aunque la base para tomar la decisión del rechazo fue comparar el valor-*p* con el nivel de significancia especificado por el director de la FTC, el valor-*p* observado de 0.0038 indica que H_0 hubiera sido rechazada para cualquier valor de $\alpha \geq 0.0038$. Debido a esto, el valor-*p* se conoce también como *nivel de significancia observado*.

Quienes toman decisiones pueden expresar opiniones distintas respecto del costo de cometer un error tipo I y elegir niveles de significancia distintos. Al proporcionar el valor-*p* como parte de los resultados de la prueba de hipótesis, alguien que toma decisiones puede comparar el valor-*p* con su propio nivel de significancia y posiblemente tome otra decisión respecto de rechazar o no H_0.

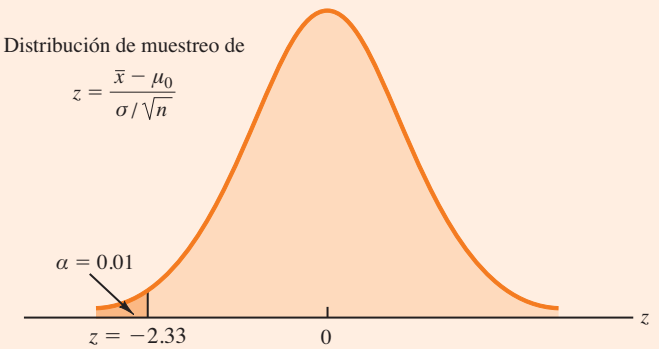
Método del valor crítico En este método primero se determina un valor para el estadístico de prueba llamado **valor crítico**. En una prueba de cola inferior éste sirve como punto de referencia para determinar si el valor del estadístico de prueba es lo suficientemente pequeño para rechazar la hipótesis nula. El valor crítico es el valor del estadístico de prueba que corresponde a un área de α (nivel de significancia) en la cola inferior de la distribución de muestreo del estadístico. En otras palabras, es el mayor valor del estadístico de prueba que hará que se rechace la hipótesis nula. A continuación, de nuevo con el ejemplo de Hilltop Coffee, se verá cómo funciona este método.

En el caso de σ conocida, la distribución de muestreo del estadístico de prueba z es la distribución normal estándar. Por tanto, el valor crítico es el valor del estadístico de prueba que corresponde a un área de $\alpha = 0.01$ en la cola inferior de la distribución normal estándar. En la tabla de probabilidad normal estándar aparece que $z = -2.33$ proporciona un área de 0.01 en la cola inferior (figura 9.3). De manera que si con la muestra se obtiene un valor del estadístico de prueba menor o igual a -2.33, el valor-*p* correspondiente será menor o igual a 0.01; en este caso la hipótesis nula deberá ser rechazada. Entonces, en el estudio de Hilltop Coffee la regla para el rechazo usando el valor crítico para un nivel de significancia de 0.01 es

Rechazar H_0 si $z \leq -2.33$

En nuestro ejemplo, $\bar{x} = 2.92$ y el estadístico de prueba es $z = -2.67$. Como $z = -2.67 < -2.33$, H_0 puede ser rechazada y concluir que Hilltop Coffee está llenando las latas de manera deficiente.

FIGURA 9.3 Valor crítico $= -2.33$ en la prueba de hipótesis de Hilltop Coffee



La regla de rechazo se puede generalizar empleando el método del valor crítico para cualquier nivel de significancia. La regla de rechazo en una prueba de cola inferior es la siguiente.

REGLA PARA EL RECHAZO EN UNA PRUEBA DE COLA INFERIOR: MÉTODO DEL VALOR CRÍTICO

$$\text{Rechazar } H_0 \text{ si } z \leq -z_\alpha$$

donde $-z_\alpha$ es el valor crítico; es decir, el valor z que proporciona un área de α en la cola inferior de la distribución normal estándar.

En las pruebas de hipótesis, el método del valor-p y el método del valor crítico llevarán siempre a la misma decisión de rechazo; esto es, siempre que el valor-p sea menor o igual que α, el valor del estadístico de prueba será menor o igual al valor crítico. La ventaja del método del valor-p radica en que dice *cuán* significativos son los resultados (el nivel de significancia observado). Si se usa el método del valor crítico, sólo se sabe que los resultados son significativos al nivel de significancia establecido.

Al principio de esta sección se dijo que las pruebas de una cola, para la media poblacional, toman una de las dos formas siguientes.

Prueba de cola inferior	**Prueba de cola superior**
$H_0: \mu \geq \mu_0$	$H_0: \mu \leq \mu_0$
$H_a: \mu < \mu_0$	$H_a: \mu > \mu_0$

El estudio de Hilltop Coffee sirvió para ilustrar cómo realizar una prueba de cola inferior. El mismo método general se usa para realizar una prueba de cola superior. Para ésta también se calcula el estadístico de prueba z usando la ecuación (9.1). Pero en una prueba de cola superior el valor-p es la probabilidad de obtener un valor para el estadístico de prueba tan grande o mayor que el obtenido con la muestra. Por tanto, para calcular el valor-p de una prueba de cola superior en el caso de α conocida, es necesario determinar el área bajo la curva normal estándar a la derecha del estadístico de prueba. Utilizando el método del valor crítico, la hipótesis nula es rechazada si el valor del estadístico de prueba es mayor o igual al valor crítico z_α; en otras palabras, H_0 es rechazada si $z \geq z_\alpha$.

Prueba de dos colas

En las pruebas de hipótesis, la forma general de una **prueba de dos colas** es la siguiente.

$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

En esta subsección se muestra cómo realizar una prueba de dos colas para la media poblacional en el caso en que se conoce σ. Como ilustración, se considera el caso de una prueba de hipótesis en la empresa MaxFlight, Inc.

La U.S. Golf Association (USGA), establece reglas que deben satisfacer los fabricantes de equipos de golf si quieren que sus productos sean aceptados en los eventos de la organización. MaxFlight emplea procesos de manufactura de alta tecnología para fabricar pelotas de golf que tienen una distancia media de recorrido de 295 yardas. Sin embargo, algunas veces el proceso se desajusta y se fabrican pelotas que tienen una distancia media de recorrido diferente. Cuando la distancia media es menor que 295 yardas, a la empresa le preocupa perder clientes porque las pelotas de golf no proporcionen la medida anunciada. Cuando es mayor de 295 yardas, las pelotas de MaxFlight pueden ser rechazadas por la USGA por exceder los estándares respecto de distancia de recorrido y rodaje.

El programa de control de calidad de la empresa consiste en tomar muestras periódicas de 50 pelotas de golf para monitorear el proceso de manufactura. Con cada muestra se realiza una prueba de hipótesis para determinar si el proceso se ha desajustado. Para elaborar las hipótesis nula y alternativa se empieza por suponer que el proceso está funcionando correctamente; es decir, las pelotas de golf que se fabrican alcanzan una distancia media de 295 yardas. Este es el supuesto que establece la hipótesis nula. La hipótesis alternativa indica que la distancia media no es igual 295 yardas. Como el valor hipotético es $\mu_0 = 295$, las hipótesis nula y alternativa en el caso de la prueba de hipótesis de MaxFlight son las siguientes.

$$H_0: \mu = 295$$
$$H_a: \mu \neq 295$$

Si la media muestral $\bar{x}$ es significativamente menor o significativamente mayor que 295 yardas, H_0 será rechazada. En este caso, se tomarán medidas para ajustar el proceso de manufactura. Por otro lado, si $\bar{x}$ no se desvía una cantidad significativa de la media hipotética $\mu_0 = 295$, H_o no será rechazada, y no se tomará medida alguna para ajustar el proceso de manufactura.

El equipo de control de calidad elige $\alpha = 0.05$ como nivel de significancia para esta prueba. Datos de pruebas anteriores realizadas sabiendo que el proceso está ajustado, indican que se puede suponer que la desviación estándar poblacional se conoce y que su valor es $\sigma = 12$. Por ende, con un tamaño de muestra $n = 50$, el error estándar $\bar{x}$ es

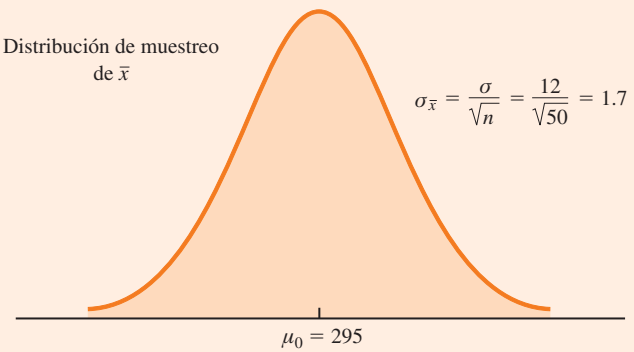
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{50}} = 1.7$$

Como el tamaño de la muestra es grande, el teorema del límite central (capítulo 7) permite concluir que la distribución de muestreo de $\bar{x}$ puede aproximarse mediante una distribución normal. En la figura 9.4 se ilustra la distribución de muestreo de $\bar{x}$ para la prueba de hipótesis de MaxFlight con una media poblacional hipotética de $\mu_0 = 295$.

WEB archivo
GolfTest

Suponga que se toma una muestra de 50 pelotas de golf y que la media muestral es $\bar{x} = 297.6$ yardas. Esta media muestral favorece la conclusión de que la media poblacional es mayor de 295 yardas. ¿Este valor de $\bar{x}$ es suficientemente mayor que 295 para hacer que H_0 sea rechazada a un nivel de significancia de 0.05? En la sección anterior se describieron dos métodos que pueden utilizarse para responder esta pregunta: el método del valor-p y el método del valor crítico.

FIGURA 9.4 Distribución de muestreo de $\bar{x}$ en la prueba de hipótesis de MaxFlight



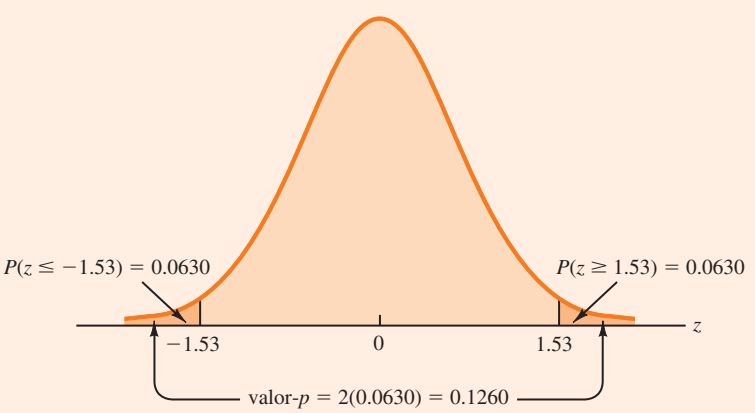
Método del valor-p Recuerde que el valor-p es la probabilidad que sirve para determinar si la hipótesis nula es rechazada. En una prueba de dos colas, los valores del estadístico de prueba en *ambas* colas proporcionan evidencias contra la hipótesis nula. En este tipo de prueba el valor-p es la probabilidad de obtener un valor para el estadístico de prueba *tan improbable o más improbable que* el obtenido con la muestra. A continuación veremos cómo se calcula el valor-p en la prueba de hipótesis de MaxFlight.

Primero calculamos el valor del estadístico de prueba. En el caso en que se conoce σ, el estadístico de prueba z es la variable aleatoria normal estándar. Empleando la ecuación (9.1) con $\bar{x} = 297.6$, el valor del estadístico de prueba es

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{297.6 - 295}{12/\sqrt{50}} = 1.53$$

Ahora, para calcular el valor-p hay que encontrar la probabilidad de obtener, para el estadístico de prueba, un valor por *lo menos tan improbable como* $z = 1.53$. Es claro que los valores $z \geq 1.53$ son *por lo menos igual de improbables.* Pero como ésta es una prueba de dos colas, los valores $z \leq -1.53$ también *son al menos tan improbables como* el valor del estadístico de prueba obtenido con la muestra. En la figura 9.5 vemos que el valor-p para dos colas está dado,

FIGURA 9.5 Valor-p en la prueba de hipótesis de MaxFlight



en este caso, por $P(z \leq -1.53) + P(z \geq 1.53)$. Como la curva normal es simétrica, calculamos la probabilidad determinando el área bajo la curva normal estándar a la derecha de $z = 1.53$ y la duplicamos. La tabla de la distribución normal estándar indica que el área a la izquierda de $z = 1.53$ es 0.9370. Entonces, el área bajo la curva normal estándar a la derecha de $z = 1.53$ es $1.0000 - 0.9370 = 0.0630$. Al duplicar esta cantidad, encontramos que en la prueba de hipótesis de dos colas de MaxFlight el valor-$p = 2(0.0630) = 0.1260$.

Ahora se compara el valor-p con el nivel de significancia para ver si la hipótesis nula es rechazada. Como el nivel de significancia es de $\alpha = 0.05$, la hipótesis nula no es rechazada, porque el valor-$p = 0.1260 > 0.05$. Como no hay rechazo, no es necesario tomar medidas para ajustar el proceso de manufactura de MaxFlight.

El cálculo del valor-p en una prueba de dos colas puede parecer un poco complicado en comparación con el cálculo del valor-p en las pruebas de una cola, pero se simplifica mediante los siguientes tres pasos.

CÁLCULO DEL VALOR-p EN UNA PRUEBA DE DOS COLAS

1. Determine el valor del estadístico de prueba z.
2. Si el valor del estadístico de prueba está en la cola superior ($z > 0$), encuentre el área bajo la curva normal estándar a la derecha de z; si está en la cola inferior ($z < 0$), localice el área bajo la curva normal estándar a la izquierda de z.
3. Duplique el área, o probabilidad, en la cola, obtenida en el paso 2 y determine el valor-p.

Método del valor crítico Antes de dejar esta sección, se verá la forma de comparar el valor del estadístico de prueba z con un valor crítico para tomar la decisión en una prueba de dos colas. En la figura 9.6 se aprecia que los valores críticos en esta prueba se encuentran tanto en la cola superior como en la cola inferior de la distribución normal estándar. Si el nivel de significancia es $\alpha = 0.05$, en cada cola, el área más allá del valor crítico es $\alpha/2 = 0.05/2 = 0.025$. En la tabla de probabilidad normal estándar se encuentra que los valores críticos para el estadístico de prueba son $-z_{0.025} = -1.96$ y $z_{0.025} = 1.96$. Entonces, al utilizar el método del valor crítico, la regla de rechazo para dos colas es:

$$\text{Rechazar } H_0 \text{ si } z \leq -1.96 \text{ o si } z \geq 1.96$$

Como en el estudio de MaxFlight el valor del estadístico de prueba es $z = 1.53$, la evidencia estadística no permitirá rechazar la hipótesis nula a un nivel de significancia de 0.05.

FIGURA 9.6 Valores críticos en la prueba de hipótesis de MaxFlight

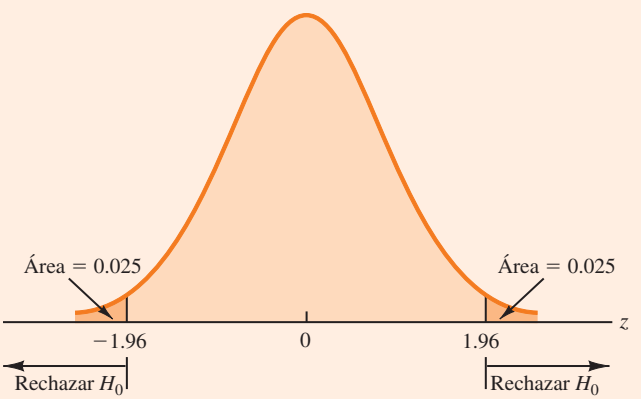


TABLA 9.2 Resumen de las pruebas de hipótesis para la media poblacional: caso con σ conocida

	Prueba de cola inferior	Prueba de cola superior	Prueba de dos colas
Hipótesis	$H_0: \mu \geq \mu_0$ $H_a: \mu < \mu_0$	$H_0: \mu \leq \mu_0$ $H_a: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$
Estadístico de prueba	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
Regla de rechazo: método del valor-*p*	Rechazar H_0 si el valor-$p \leq \alpha$	Rechazar H_0 si el valor-$p \leq \alpha$	Rechazar H_0 si el valor-$p \leq \alpha$
Regla de rechazo: método del valor crítico	Rechazar H_0 si $z \leq -z_\alpha$	Rechazar H_0 si $z \geq z_\alpha$	Rechazar H_0 si $z \leq -z_{\alpha/2}$ o si $z \geq z_{\alpha/2}$

Resumen y consejo práctico

Se presentaron ejemplos de una prueba de cola inferior y de una prueba de dos colas para la media poblacional. Con base en estos ejemplos es posible resumir ahora, como se muestra en la tabla 9.2, los procedimientos de prueba de hipótesis para la media poblacional en el caso de σ conocida. Observe que μ_0 es el valor hipotético de la media poblacional.

Los pasos en las pruebas de hipótesis seguidos en los dos ejemplos presentados en esta sección son comunes a toda prueba de hipótesis.

PASOS EN LAS PRUEBAS DE HIPÓTESIS

Paso 1. Establecer la hipótesis nula y la hipótesis alternativa.
Paso 2. Especificar el nivel de significancia.
Paso 3. Recabar los datos muestrales y calcular el valor del estadístico de prueba.

Método del valor-p

Paso 4. Emplear el valor del estadístico de prueba para calcular el valor-*p*.
Paso 5. Rechazar H_0 si el valor-$p \leq \alpha$.

Método del valor crítico

Paso 4. Utilizar el nivel de significancia para determinar el valor crítico y la regla de rechazo.
Paso 5. Emplear el valor del estadístico de prueba y la regla de rechazo para determinar si H_0 es rechazada.

El consejo práctico acerca del tamaño de la muestra para pruebas de hipótesis es semejante a la recomendación sugerida en el capítulo 8 con respecto a la estimación por intervalo. En la mayor parte de las aplicaciones, para el procedimiento de prueba de hipótesis revisado en esta sección, un tamaño de muestra $n \geq 30$ es adecuado. En los casos en los que el tamaño sea menor de 30, la distribución de la población de la cual se toma la muestra se vuelve una consideración importante. Si la población tiene una distribución normal, el procedimiento de prueba de hipótesis descrito es exacto y puede utilizarse con cualquier tamaño de muestra. Si la población no tiene una distribución normal, pero es por lo menos aproximadamente simétrica, con tamaños de muestra hasta de 15 pueden esperarse resultados aceptables.

Relación entre estimación por intervalo y prueba de hipótesis

En el capítulo 8 se explicó la forma de obtener una estimación de la media poblacional mediante un intervalo de confianza. En el caso en que σ es conocida, esta estimación mediante un intervalo de $(1 - \alpha)\%$ de confianza está dada por

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

En este capítulo se mostró que una prueba de hipótesis de dos colas para la media poblacional tiene la siguiente forma.

$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

donde μ_0 es el valor hipotético de la media poblacional.

Suponga que se sigue el procedimiento descrito en el capítulo 8 para construir un intervalo de $100(1 - \alpha)\%$ de confianza para la media poblacional. Sabemos que $100(1 - \alpha)\%$ de los intervalos de confianza generados contendrán la media poblacional y $100\alpha\%$ de los intervalos generados no la contendrán. En consecuencia, si H_0 es rechazada, cuando el intervalo de confianza no contenga μ_0, la probabilidad de rechazar la hipótesis nula cuando sea verdadera ($\mu = \mu_0$) será α. Recuerde que el nivel de significancia es la probabilidad de rechazar la hipótesis nula cuando es verdadera. Entonces, construir un intervalo de $100(1 - \alpha)\%$ de confianza y rechazar H_0 cuando el intervalo no contenga μ_0 es equivalente a realizar una prueba de hipótesis de dos colas con α como nivel de significancia. El procedimiento para usar un intervalo de confianza para efectuar una prueba de hipótesis de dos colas se resume a continuación.

MÉTODO DEL INTERVALO DE CONFIANZA PARA PROBAR UNA HIPÓTESIS DE LA FORMA

$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

1. Seleccionar de la población una muestra aleatoria simple y emplear el valor de la media muestral $\bar{x}$ para obtener un intervalo de confianza de la media poblacional μ.

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

2. Si el intervalo de confianza contiene el valor hipotético μ_0, H_0 no es rechazada. En caso contrario, H_0 es rechazada.[2]

En una prueba de hipótesis de dos colas, la hipótesis nula es rechazada si el intervalo de confianza no contiene μ_0.

El uso del método del intervalo de confianza para realizar una prueba de hipótesis se ilustrará empleando el ejemplo de MaxFlight. Esta prueba de hipótesis tiene la forma siguiente.

$$H_0: \mu = 295$$
$$H_a: \mu \neq 295$$

[2] Para ser congruentes con la regla para rechazar H_0 cuando el valor-$p \leq \alpha$, se debe rechazar H_0 utilizando el método del intervalo de confianza si ocurre que μ_0 es igual a uno de los puntos finales del intervalo de $100(1 - \alpha)\%$.

Para probar esta hipótesis con un nivel de confianza de $\alpha = 0.05$, se tomó una muestra de 50 pelotas de golf y se encontró una distancia media muestral de $\bar{x} = 297.6$ yardas. Recuerde que la desviación estándar poblacional es $\sigma = 12$. Al aplicar estos resultados a $z_{0.025} = 1.96$, obtenemos que el intervalo de 95% de confianza para estimar la media poblacional es

$$\bar{x} \pm z_{0.025}\frac{\sigma}{\sqrt{n}}$$

$$297.6 \pm 1.96\frac{12}{\sqrt{50}}$$

$$297.6 \pm 3.3$$

o

$$294.3 \text{ a } 300.9$$

Este hallazgo permite al gerente de control de calidad concluir que con 95% de confianza la distancia media para la población de pelotas de golf está entre 294.3 y 300.9 yardas. Como el valor hipotético de la media poblacional $\mu_0 = 295$ está en dicho intervalo, la conclusión de la prueba de hipótesis es que no se puede rechazar la hipótesis nula, H_0: $\mu = 295$.

Preste atención a que estos análisis y ejemplo pertenecen a pruebas de hipótesis de dos colas para la media poblacional. Sin embargo, la misma relación entre intervalo de confianza y prueba de hipótesis de dos colas existe para otros parámetros poblacionales. Esta relación también se extiende a pruebas de hipótesis de una cola para parámetros poblacionales; sin embargo, para ello se pide elaborar intervalos de confianza unilaterales que son muy poco utilizados en la práctica.

NOTAS Y COMENTARIOS

Se mostró cómo usar el valor-*p*. Entre menor sea éste, mayor es la evidencia en contra de H_0 y a favor de H_a. A continuación se listan algunos lineamientos que los expertos en estadística recomiendan para interpretar valores-*p* pequeños.

- Menor que 0.01: evidencia terminante para concluir que H_a es verdadera.
- Entre 0.01 y 0.05: fuerte evidencia para concluir que H_a es verdadera.
- Entre 0.05 y 0.10: evidencia débil para concluir que H_a es verdadera.
- Mayor que 0.10: evidencia insuficiente para concluir que H_a es verdadera.

Ejercicios

Nota para el estudiante. En algunos ejercicios que siguen se pide usar el método del valor-*p* y en otros el método del valor crítico. Ambos llevarán a la misma conclusión en una prueba de hipótesis. Se presentan ejercicios con ambos métodos para que el lector adquiera práctica en su uso. En las secciones y capítulos posteriores se preferirá usar el enfoque del valor-*p*, pero el estudiante puede elegir el que prefiera.

Métodos

9. Considere la prueba de hipótesis siguiente.

$$H_0: \mu \geq 20$$
$$H_a: \mu < 20$$

En una muestra de 50, la media muestral es 19.4 y la desviación estándar poblacional es 2.

a) Calcule el valor del estadístico de prueba.
b) ¿Cuál es el valor-*p*?
c) Use $\alpha = 0.05$, ¿qué concluye usted?
d) ¿Cuál es la regla de rechazo si se usa el método del valor crítico? ¿Cuál es su conclusión?

AUTO evaluación

10. Considere la prueba de hipótesis siguiente.

$$H_0: \mu \leq 25$$
$$H_a: \mu > 25$$

En una muestra de 40, la media muestral es 26.4 y la desviación estándar poblacional es 6.

a) Calcule el valor del estadístico de prueba.
b) ¿Cuál es el valor-*p*?
c) Use $\alpha = 0.01$, ¿cuál es su conclusión?
d) ¿Cuál es la regla de rechazo si se usa el método del valor crítico? ¿Qué concluye?

AUTO evaluación

11. Considere la prueba de hipótesis siguiente.

$$H_0: \mu = 15$$
$$H_a: \mu \neq 15$$

En una muestra de 50, la media muestral es 14.15 y la desviación estándar poblacional es 3.

a) Calcule el valor del estadístico de prueba.
b) ¿Cuál es el valor-*p*?
c) Use $\alpha = 0.05$, ¿cuál es su conclusión?
d) ¿Cuál es la regla de rechazo si se usa el método del valor crítico? ¿Cuál es su conclusión?

12. Considere la prueba de hipótesis siguiente.

$$H_0: \mu \geq 80$$
$$H_a: \mu < 80$$

Se utilizó una muestra de 100 y la desviación estándar poblacional es 12. Calcule el valor-*p* y establezca su conclusión para cada uno de los resultados muestrales siguientes. Use $\alpha = 0.01$.

a) $\bar{x} = 78.5$
b) $\bar{x} = 77$
c) $\bar{x} = 75.5$
d) $\bar{x} = 81$

13. Considere la prueba de hipótesis siguiente.

$$H_0: \mu \leq 50$$
$$H_a: \mu > 50$$

Se utilizó una muestra de 60 y la desviación estándar poblacional es 8. Use el método de valor crítico y fije sus conclusiones para cada uno de los resultados muestrales siguientes. Use $\alpha = 0.05$.

a) $\bar{x} = 52.5$
b) $\bar{x} = 51$
c) $\bar{x} = 51.8$

14. Considere la prueba de hipótesis siguiente:

$$H_0: \mu = 22$$
$$H_a: \mu \neq 22$$

Con una muestra de 75, la desviación estándar poblacional es 10. Calcule el valor-*p* y establezca sus conclusiones para cada uno de los resultados muestrales siguientes. Use $\alpha = 0.01$.

a) $\bar{x} = 23$
b) $\bar{x} = 25.1$
c) $\bar{x} = 20$

Aplicaciones

AUTOevaluación

15. Las declaraciones de impuestos presentadas antes del 31 de marzo obtienen un reembolso que en promedio es de $1 056. Considere la población de los contribuyentes de "última hora" que presentan su declaración en los últimos cinco días del periodo para este trámite (normalmente del 10 al 15 de abril).
    a) Un investigador sugiere que la razón por la que estos declarantes esperan hasta los últimos días se debe a que en promedio obtienen un reembolso menor que los que declaran antes del 31 de marzo. Establezca las hipótesis apropiadas de manera que el rechazo de H_0 favorezca la sugerencia de este investigador.
    b) En una muestra de 400 personas que presentaron su declaración entre el 10 y el 15 de abril, la media muestral de los reembolsos fue $910. Por experiencia se sabe que es posible considerar que la desviación estándar poblacional es $\sigma = \$1\,600$. ¿Cuál es el valor-*p*?
    c) Con $\sigma = 0.05$, ¿cuál es su conclusión?
    d) Repita la prueba de hipótesis anterior usando el método del valor crítico.
16. En un estudio acerca de cómo utilizan las tarjetas de crédito los estudiantes no titulados, se reportó que este tipo de población tiene un saldo medio en su tarjeta de crédito de $3 173 (*Sallie Mae*, abril de 2009). Tal cifra fue la más alta de todos los tiempos y tuvo un incremento de 44% sobre la de los cinco años previos. Suponga que se realiza un estudio actual para determinar si es posible concluir que el saldo medio en la tarjeta de crédito de estudiantes no titulados ha continuado en aumento comparado con el informe de abril de 2009. Con base en reportes previos, utilice una desviación estándar poblacional de $\sigma = \$1\,000$.
    a) Establezca las hipótesis nula y alternativa.
    b) ¿Cuál es el valor-*p* de una muestra de 180 estudiantes no titulados con un saldo medio muestral en su tarjeta de crédito de $3 325?
    c) Usando un nivel de significancia de 0.05, ¿cuál es su conclusión?
17. Las sociedades de valores de Wall Street pagaron en 2005 gratificaciones de fin de año de $125 500 por empleado (*Fortune*, 6 de febrero de 2006). Suponga que se desea tomar una muestra de los empleados de la empresa de valores Jones & Ryan para ver si la media de la gratificación de fin de año es diferente de la media de $125 500 reportada para la población.
    a) Establezca las hipótesis nula y alternativa que se usarían para probar si las gratificaciones de fin de año de Jones & Ryan difieren de la media poblacional.
    b) Suponga que una muestra de 40 empleados de Jones & Ryan exhibió una media muestral de las gratificaciones de $118 000. Suponga que la desviación estándar poblacional es $\sigma = \$30\,000$ y calcule el valor-*p*.
    c) Con $\alpha = 0.05$ como nivel de significancia, ¿cuál es su conclusión?
    d) Repita esta prueba de hipótesis usando el método del valor crítico.
18. La rentabilidad total anual promedio de los fondos de inversión de U.S. Diversified Equity de 1999 a 2003 fue de 4.1% (*BusinessWeek*, 26 de enero de 2004). Un investigador desea realizar una prueba de hipótesis para saber si los rendimientos de determinados fondos de crecimiento (mid-cap growth funds) difieren de manera significativa del promedio de los fondos de U.S. Diversified Equity.
    a) Establezca las hipótesis que se pueden usar para determinar si la rentabilidad anual media de estos fondos difiere de la media de los fondos de U.S. Diversified Equity.
    b) En una muestra de 40 fondos el rendimiento medio fue de $\bar{x} = 3.4\%$. Suponga que por estudios anteriores se sabe que la desviación estándar poblacional de estos fondos es $\sigma = 2\%$. Use los resultados muestrales para calcular el estadístico de prueba y el valor-*p* para la prueba de hipótesis.
    c) Con $\alpha = 0.05$, ¿cuál es su conclusión?

19. El U.S. Department of Labor informó que los ingresos promedio para los trabajadores estadounidenses de la producción en 2001 era $14.32 por hora (*The World Almanac*, 2003). En una muestra de 75 trabajadores tomada en 2003, la media muestral fue $14.68 por hora. Si la desviación estándar poblacional es $\sigma = \$1.45$, ¿se puede concluir que ha habido un aumento en la media de las ganancias por hora? Use $\alpha = 0.05$.

20. En Estados Unidos, un hogar paga en promedio $32.79 mensuales por el servicio de Internet (CNBC, 18 de enero de 2006). En una muestra de 50 hogares de un estado del sur la media muestral fue $30.63. Use la desviación estándar poblacional de $\sigma = \$5.60$.
    a) Formule las hipótesis para una prueba en la que se quiere determinar si los datos muestrales favorecen la conclusión de que la cantidad media mensual pagada por el servicio de Internet en este estado del sur es menor a la media de todo el país, que es de $32.79.
    b) ¿Cuál es el valor del estadístico de prueba?
    c) ¿Cuál es el valor-*p*?
    d) Con $\alpha = 0.01$, ¿qué concluye?

WEB archivo
Fowle

21. Fowle Marketing Research, Inc. tasa la tarifa que cobra a sus clientes en el supuesto de que una encuesta por teléfono se realiza en un promedio de 15 minutos o menos. Si se requiere más tiempo en promedio, se cobra una cantidad adicional. La duración de las encuestas en una muestra de 35 de ellas se presentan en el archivo Fowle. Por estudios anteriores se puede considerar que la desviación estándar poblacional es conocida y que es $\sigma = 4$ minutos. ¿El cobro de la cantidad adicional está justificado?
    a) Formule las hipótesis nula y alternativa para esta aplicación.
    b) Calcule el valor del estadístico de prueba.
    c) ¿Cuál es el valor-*p*?
    d) Con $\alpha = 0.01$, ¿cuál es su conclusión?

22. CNN y ActMedia presentaron un canal de televisión dirigido a las personas que esperan en las colas de los supermercados. En este canal se transmitían noticias, reportajes cortos y publicidad. La duración de la programación se basaba en el supuesto de que la media poblacional del tiempo que los clientes esperan en la fila de la caja es 8 minutos. Se utilizará una muestra de tiempos de espera reales para probar ese supuesto y determinar si el tiempo medio de espera difiere de ese estándar.
    a) Formule las hipótesis para esta aplicación.
    b) En una muestra de 120 clientes, la media muestral de tiempo de espera fue 8.5 minutos. Suponga que la desviación estándar poblacional es $\sigma = 3.2$ minutos. ¿Cuál es el valor-*p*?
    c) Con $\alpha = 0.05$, ¿cuál es su conclusión?
    d) Calcule un intervalo de 95% de confianza para la media poblacional. ¿Esto sustenta su conclusión?

9.4 Media poblacional: σ desconocida

En esta sección se describe cómo realizar pruebas de hipótesis para la media poblacional en el caso de σ desconocida. Como ésta corresponde a la situación en que no se tiene una estimación de la desviación estándar poblacional antes de tomar la muestra, esta última se usa para obtener una estimación tanto de μ como de σ. Por tanto, para realizar una prueba sobre la media poblacional en el caso en que σ no se conoce, la media muestral $\bar{x}$ se utiliza como estimación de μ y la desviación estándar muestral s, como estimación de σ.

Los pasos a seguir para las pruebas de hipótesis en el caso en que σ no se conoce son los mismos que cuando σ se conoce, descritos en la sección 9.3. Pero como σ no se conoce, los cálculos del estadístico de prueba y del valor-*p* son ligeramente distintos. Recuerde que en el caso de σ conocida la distribución de muestreo del estadístico de prueba tiene distribución normal estándar. Sin embargo, en el caso de σ desconocida la distribución de muestreo del estadístico de prueba sigue la distribución *t*; tiene ligeramente más variabilidad debido a que la muestra se usa para obtener estimaciones tanto de μ como de σ.

En la sección 8.2 se vio que una estimación por intervalo de la media poblacional en el caso de σ desconocida se basa en una distribución de probabilidad llamada distribución *t*. Las pruebas de hipótesis para la media poblacional cuando σ no se conoce también se basan en la distribución *t*. Para σ desconocida, el estadístico de prueba tiene distribución *t* con $n - 1$ grados de libertad.

ESTADÍSTICO DE PRUEBA EN LAS PRUEBAS DE HIPÓTESIS PARA LA MEDIA POBLACIONAL: σ DESCONOCIDA

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{9.2}$$

En el capítulo 8 también se dijo que la distribución *t* se basa en el supuesto de que la población de la que se toma la muestra tiene distribución normal. Sin embargo, las investigaciones demuestran que este supuesto no es muy fuerte si el tamaño de la muestra es suficientemente grande. Al final de esta sección se proporciona una recomendación práctica acerca de la distribución de la población y del tamaño de la muestra.

Prueba de una cola

A continuación se considera un ejemplo de prueba de una cola para la media poblacional en el caso de σ desconocida. Una revista de viajes de negocios desea clasificar los aeropuertos internacionales con base en una evaluación externada por la población de viajeros de negocios. Se utiliza una escala de evaluación que va desde un mínimo de 0 hasta un máximo de 10, y aquellos aeropuertos que obtengan una media mayor de 7 serán considerados de servicio superior. Para obtener los datos de evaluación, el personal de la revista entrevista una muestra de 60 viajeros de negocios de cada terminal aeroportuaria. En la muestra tomada en el aeropuerto Heathrow de Londres la media muestral es $\bar{x} = 7.25$ y la desviación estándar muestral es $s = 1.052$. Con base en estos datos muestrales, ¿Heathrow deberá ser designado como un aeropuerto de servicio superior?

WEB archivo
AirRating

Se quiere realizar una prueba de hipótesis para que la decisión de rechazar H_0 permita concluir que la media poblacional en la evaluación de la terminal de Heathrow es *mayor* de 7. Entonces se requiere una prueba de cola superior en la que H_a: $\mu > 7$. Las hipótesis nula y alternativa en esta prueba son las siguientes.

$$H_0: \mu \leq 7$$
$$H_a: \mu > 7$$

Se usa como nivel de significancia $\alpha = 0.05$.

Al aplicar la ecuación (9.2) con $\bar{x} = 7.25$, $\mu_0 = 7$, $s = 1.052$ y $n = 60$, el valor del estadístico de prueba es

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{7.25 - 7}{1.052/\sqrt{50}} = 1.84$$

La distribución de muestreo de *t* tiene $n - 1 = 60 - 1 = 59$ grados de libertad. Como es una prueba de cola superior, el valor-*p* es el área bajo la curva de la distribución *t* a la derecha de $t = 1.84$.

Las tablas de distribución *t* proporcionadas en la mayor parte de los libros de texto no son suficientemente detalladas para determinar el valor-*p* exacto, como es el caso del valor-*p*

correspondiente a $t = 1.84$. Por ejemplo, en la tabla 2 del apéndice B, la distribución t con 59 grados de libertad proporciona la información siguiente.

Área en la cola superior	0.20	0.10	0.05	0.025	0.01	0.005
Valor t (59 gl)	0.848	1.296	1.671	2.001	2.391	2.662

$t = 1.84$

Como se ve, $t = 1.84$ está entre 1.671 y 2.001. Aunque esta tabla no proporciona el valor exacto de t, los valores en la fila "Área en la cola superior" indican que el valor-p debe ser menor que 0.05 y mayor que 0.025. Con un nivel de significancia $\alpha = 0.05$, esto es todo lo que se necesita saber para rechazar la hipótesis nula y concluir que Heathrow debe ser considerado un aeropuerto de servicio superior.

El apéndice F indica cómo calcular los valores-p usando Excel o Minitab.

Debido a que es engorroso usar una tabla t para calcular los valores-p, y puesto que sólo se pueden obtener valores-p aproximados, se mostrará cómo calcular valores-p exactos usando Excel o Minitab. Estas instrucciones se encuentran al final del libro, en el apéndice F. Usando ambos programas con $t = 1.84$, el valor-p que se obtiene en la cola superior es 0.0354 para la prueba de hipótesis del aeropuerto de Heathrow. Como $0.0354 < 0.05$, la hipótesis nula es rechazada y se concluye que éste se debe considerar un aeropuerto de servicio superior.

Prueba de dos colas

Con objeto de ilustrar cómo realizar una prueba de dos colas para la media poblacional en el caso de σ desconocida, se considerará la situación de la prueba de hipótesis de Holiday Toys. Esta empresa distribuye sus productos a través de más de 1 000 puntos de venta. Al planear sus niveles de producción para la temporada de invierno siguiente, debe decidir cuántas unidades de cada producto fabricar antes de saber cuál será la verdadera demanda en cada tienda. El gerente de Marketing de Holiday espera que su juguete de novedad más importante de este año tenga una demanda de 40 unidades en promedio por punto de venta. Antes de tomar la decisión final de producción con base en dicha estimación, la empresa decide levantar una encuesta en una muestra de 25 puntos de venta con objeto de obtener más información acerca de la demanda del nuevo producto. A cada uno de estos puntos de venta se le proporciona información sobre las características del nuevo juguete e información sobre el costo y el precio de venta sugerido. Después se le pide que anticipe la cantidad que solicitará.

Siendo μ la media poblacional de las cantidades ordenadas por punto de venta, los datos muestrales se usan para realizar la siguiente prueba de hipótesis de dos colas:

$$H_0: \mu = 40$$
$$H_a: \mu \neq 40$$

Si H_0 no puede ser rechazada, Holiday continuará con la producción planeada con base en la estimación del director de Marketing de que la media poblacional de la cantidad solicitada por punto de venta será $\mu = 40$ unidades. Pero si H_0 es rechazada, Holiday reevaluará de inmediato su plan de producción de este juguete. Se usa una prueba de dos colas porque la empresa quiere reevaluar su plan de producción si la media poblacional de la cantidad demandada por punto de venta es menor o mayor a la prevista. Como no se cuenta con datos históricos (se trata de un producto nuevo), la media poblacional μ y la desviación estándar poblacional deben estimarse usando los valores $\bar{x}$ y s que se obtengan con los datos muestrales.

WEB archivo
Orders

En la muestra de 25 puntos de venta la media que se obtiene es $\bar{x} = 37.4$ y la desviación estándar $s = 11.79$ unidades. Antes de usar la distribución t, el analista elabora un histograma con los datos muestrales con objeto de ver cuál es la forma de la distribución poblacional. El histograma no indica evidencias de sesgo ni de valores atípicos, de manera que el analista

concluye que es adecuado usar la distribución t con $n - 1 = 24$ grados de libertad. Usando la ecuación (9.2) con $\bar{x} = 37.4$, $\mu_0 = 40$, $s = 11.79$ y $n = 25$, el valor que se obtiene para el estadístico de prueba es

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{37.4 - 40}{11.79/\sqrt{25}} = -1.10$$

Como se trata de una prueba de dos colas, el valor-p es el doble del área bajo la curva de la distribución t para $t \leq -1.10$. En la tabla 2 del apéndice B, la fila de la distribución t para 24 grados de libertad proporciona la información siguiente.

Área en la cola superior	**0.20**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
Valor *t* (24 *gl*)	0.857	1.318	1.711	2.064	2.492	2.797

$t = 1.10$

La tabla de distribución t sólo contiene valores t positivos. Sin embargo, como la distribución t es simétrica, el área bajo la curva a la derecha de $t = 1.10$ es igual al área bajo la curva a la izquierda de $t = -1.10$. Se encuentra así que $t = 1.10$ está entre 0.857 y 1.318. En la fila "Área en la cola superior" se ve que el área en la cola a la derecha de $t = 1.10$ está entre 0.20 y 0.10. Duplicando estas cantidades, el valor-p debe estar entre 0.40 y 0.20. Como el nivel de significancia es $\alpha = 0.05$, se ve que el valor-p es mayor que α. Por tanto, H_0 no puede ser rechazada. No hay evidencia suficiente para concluir que Holiday deba modificar su plan de producción para la temporada siguiente.

En el apéndice F se indica cómo calcular el valor-p para esta prueba usando Minitab o Excel. El valor-p que se obtiene es 0.2822. Con el nivel de significancia $\alpha = 0.05$, H_0 no puede ser rechazada, dado que $0.2822 > 0.05$.

Para tomar la decisión en esta prueba de dos colas también se puede comparar el estadístico de prueba con el valor crítico. Usando $\alpha = 0.05$ y la distribución t con 24 grados de libertad, $-t_{0.025} = -2.064$ y $t_{0.025} = 2.064$ son los valores críticos para la prueba de dos colas. La regla de rechazo usando el estadístico de prueba es

$$\text{Rechazar } H_0 \text{ si } t \leq -2.064 \text{ o si } t \geq 2.064$$

Con base en el estadístico de prueba $t = -1.10$, H_0 no puede ser rechazada. Este resultado indica que Holyday puede continuar con su plan de producción para la temporada próxima con base en la expectativa de $\mu = 40$.

Resumen y consejo práctico

En la tabla 9.3 se proporciona un resumen de los procedimientos de prueba de hipótesis en los casos de σ desconocida. La diferencia principal entre estos procedimientos y el del caso de σ conocida estriba en que para calcular el estadístico de prueba se usa s en lugar de σ. A esto se debe que el estadístico de prueba siga la distribución t.

La aplicabilidad de los procedimientos de prueba de hipótesis de esta sección depende de la distribución de la población de donde se toma la muestra y del tamaño de ésta. Si la población tiene una distribución normal, las pruebas de hipótesis descritas en esta sección dan resultados exactos con cualquier tamaño de muestra. Si la población no está distribuida normalmente, los procedimientos son aproximaciones. De cualquier manera, se encuentra que tamaños de muestra de 30 o mayores proporcionan buenos resultados en la mayor parte de los casos. Si la población es aproximadamente normal, muestras pequeñas (por ejemplo, $n < 15$) pueden ofrecer resultados aceptables. Si la población es muy sesgada o si contiene observaciones atípicas, se recomiendan tamaños de alrededor de 50.

TABLA 9.3 Resumen de las pruebas de hipótesis para la media poblacional: caso de σ desconocida

	Prueba de cola inferior	Prueba de cola superior	Prueba de dos colas
Hipótesis	$H_0: \mu \geq \mu_0$ $H_a: \mu < \mu_0$	$H_0: \mu \leq \mu_0$ $H_a: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$
Estadístico de prueba	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$z = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$z = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
Regla de rechazo: método del valor-*p*	Rechazar H_0 si el valor-$p \leq \alpha$	Rechazar H_0 si el valor-$p \leq \alpha$	Rechazar H_0 si el valor-$p \leq \alpha$
Regla de rechazo: método del valor crítico	Rechazar H_0 si $t \leq -t_\alpha$	Rechazar H_0 si $t \geq t_\alpha$	Rechazar H_0 si $t \leq -t_{\alpha/2}$ o si $t \geq t_{\alpha/2}$

Ejercicios

Métodos

23. Considere la prueba de hipótesis siguiente.

$$H_0: \mu \leq 12$$
$$H_a: \mu > 12$$

En una muestra de 25, la media muestral es $\bar{x} = 14$ y la desviación estándar $s = 4.32$.

a) Calcule el valor del estadístico de prueba.
b) Use la tabla de distribución *t* (tabla 2 del apéndice B) a fin de calcular un intervalo para el valor-*p*.
c) Con $\alpha = 0.05$, ¿cuál es su conclusión?
d) ¿Cuál es la regla de rechazo usando el valor crítico? ¿Qué concluye?

AUTOevaluación

24. Considere la prueba de hipótesis siguiente.

$$H_0: \mu = 18$$
$$H_a: \mu \neq 18$$

En una muestra de 48, la media muestral es $\bar{x} = 17$ y la desviación estándar muestral $s = 4.5$.

a) Calcule el valor del estadístico de prueba.
b) Use la tabla de distribución *t* (tabla 2 del apéndice B) con objeto de calcular un intervalo para el valor-*p*.
c) Con $\alpha = 0.05$, ¿cuál es su conclusión?
d) ¿Cuál es la regla de rechazo usando el valor crítico? ¿Qué concluye?

25. Considere la prueba de hipótesis siguiente.

$$H_0: \mu \geq 45$$
$$H_a: \mu < 45$$

Se usa una muestra de 36. Identifique el valor-*p* y establezca su conclusión para cada uno de los siguientes resultados muestrales. Use $\alpha = 0.01$.

a) $\bar{x} = 44$ y $s = 5.2$
b) $\bar{x} = 43$ y $s = 4.6$
c) $\bar{x} = 46$ y $s = 5.0$

26. Considere la prueba de hipótesis siguiente.

$$H_0: \mu = 100$$
$$H_a: \mu \neq 100$$

Se utiliza una muestra de 65. Identifique el valor-p y establezca su conclusión para cada uno de los siguientes resultados muestrales. Use $\alpha = 0.05$.

a) $\bar{x} = 103$ y $s = 11.5$
b) $\bar{x} = 96.5$ y $s = 11$
c) $\bar{x} = 102$ y $s = 10.5$

Aplicaciones

AUTOevaluación

27. La Employment and Training Administration informó que la prestación media del seguro de desempleo es de $238 por semana (*The World Almanac*, 2003). Un investigador del estado de Virginia anticipó que datos muestrales indicarán que la prestación media semanal del seguro de desempleo en ese estado es menor que la media de todo el país.
a) Establezca las hipótesis adecuadas de manera que el rechazo de H_0 favorezca la afirmación del investigador.
b) En una muestra de 100 individuos, la media muestral semanal del seguro de desempleo encontrada fue $231, con una desviación estándar muestral de $80. ¿Cuál es el valor-$p$?
c) Si $\alpha = 0.05$, ¿cuál es su conclusión?
d) Repita la prueba de hipótesis anterior utilizando el método del valor crítico.

28. Al presentar una protesta, un grupo de accionistas exige que la media de permanencia de un presidente ejecutivo (CEO) sea de por lo menos nueve años. En una encuesta de empresas publicada en *The Wall Street Journal* se encontró una media muestral de permanencia de los CEO de $\bar{x} = 7.27$ años, con una desviación estándar de $s = 6.38$ años (*The Wall Street Journal*, 2 de enero de 2007).
a) Formule las hipótesis que se usarán para rebatir la validez de la exigencia del grupo de accionistas.
b) Suponga que en la muestra se incluyeron 85 empresas. ¿Cuál es el valor-p para la prueba de hipótesis?
c) Con $\alpha = 0.01$, ¿cuál es su conclusión?

WEB archivo
Diamonds

29. El precio de un diamante de un quilate de color H y pureza VS2 de Diamond Source USA es $5 600 (sitio web de Diamond Source, marzo de 2003). Un joyero del medio oeste llama a sus contactos en el distrito de los diamantes de Nueva York para saber si el precio medio de los que venden en ese lugar difiere de $5 600.
a) Formule las hipótesis que se usarán para determinar si el precio medio en Nueva York difiere de $5 600.
b) Los precios en una muestra de 25 contactos en la ciudad de Nueva York se presentan en el archivo *Diamonds*. ¿Cuál es el valor-p?
c) Con $\alpha = 0.05$, ¿es posible rechazar la hipótesis nula? ¿Cuál es su conclusión?
d) Repita la prueba de hipótesis anterior usando el método del valor crítico.

30. CNN, compañía de AOL Time Warner Inc., tiene el liderazgo de noticias en televisión por cable. Nielsen Media Research indica que en 2002 la media de la audiencia de CNN fue de 600 000 espectadores por día (*The Wall Street Journal*, 10 de marzo de 2003). Suponga que en una muestra de 40 días durante la primera mitad de 2003, la cantidad diaria de espectadores haya sido de 612 000, con una desviación estándar muestral de 65 000 sujetos.
a) ¿Cuáles son las hipótesis si el director de CNN desea información sobre cualquier cambio en la cantidad de espectadores de la empresa?
b) ¿Cuál es el valor-p?
c) Elija su propio nivel de significancia. ¿Cuál es su conclusión?
d) ¿Qué recomendación le haría al director de CNN en esta aplicación?

31. The Coca-Cola Company reportó que la media de ventas anuales per cápita de sus bebidas en Estados Unidos fue de 423 botellas de 8 onzas (sitio web de Coca-Cola Company, 3 de febrero

de 2009). Suponga que se tiene la curiosidad de verificar si el consumo de estas bebidas es más alto en Atlanta, Georgia, donde se ubican las oficinas corporativas de la empresa. Una muestra de 36 individuos del área de Atlanta mostró un consumo anual medio muestral de 460.4 botellas de 8 onzas, con una desviación estándar de $s = 101.9$ onzas. Utilizando $\alpha = 0.05$, ¿los resultados muestrales sustentan la conclusión de que el consumo anual medio de las bebidas de Coca-Cola es más alto en Atlanta?

WEB archivo
UsedCars

32. Según la National Automobile Dealers Association, el precio medio de un automóvil usado es de $10 192. El gerente de una distribuidora de la ciudad de Kansas revisó una muestra de 50 automóviles usados vendidos recientemente en ese establecimiento, con objeto de determinar si la media poblacional de sus precios difería del precio medio en todo el país. Los precios de los 50 automóviles se encuentran en el archivo denominado *UsedCars.*
   a) Formule las hipótesis que se usarán para determinar si existe diferencia en el precio medio de los automóviles usados de la distribuidora.
   b) ¿Cuál es el valor-*p*?
   c) Con $\alpha = 0.05$, ¿cuál es su conclusión?

33. El consumo anual per cápita de leche en Estados Unidos es de 21.6 galones (*Statistical Abstract of the United States: 2006*). Usted cree que en el oeste medio el consumo de leche es mayor y desea fundamentar su opinión. En una muestra de 16 personas de Webster City, pueblo del oeste medio, la media muestral del consumo anual es de 24.1 galones y la desviación estándar es $s = 4.8$.
   a) Elabore una prueba de hipótesis que se pueda usar para determinar si el consumo medio anual en Webster City es mayor que la media nacional.
   b) ¿Cuál sería una estimación puntual de la diferencia entre el consumo medio anual en Webster City y la media nacional?
   c) Con $\alpha = 0.05$ pruebe si hay una diferencia significativa. ¿Cuál es su conclusión?

34. Joan's Nursery se especializa en jardines de zonas residenciales diseñados según el gusto del cliente. La estimación del precio de un proyecto se basa en el número de árboles, arbustos, etc., a emplear en el proyecto. Para propósitos de estimación de costos, los gerentes consideran que se requieren dos horas de trabajo para plantar un árbol mediano. A continuación se presentan los tiempos (en horas) realmente requeridos en una muestra de 10 árboles plantados durante el mes pasado.

1.7	1.5	2.6	2.2	2.4	2.3	2.6	3.0	1.4	2.3

Con un nivel de significancia $\alpha = 0.05$, realice una prueba para ver si el tiempo necesario promedio para plantar los árboles difiere de 2 horas.
   a) Establezca las hipótesis nula y alternativa.
   b) Calcule la media muestral.
   c) Calcule la desviación estándar muestral.
   d) ¿Cuál es el valor-*p*?
   e) ¿Cuál es su conclusión?

9.5 Proporción poblacional

En esta sección se describe cómo realizar una prueba de hipótesis para la proporción poblacional p si mediante p_0 se denota el valor hipotético para la proporción poblacional. Las tres formas de una prueba de hipótesis para la proporción poblacional son las siguientes.

$$H_0: p \geq p_0 \qquad H_0: p \leq p_0 \qquad H_0: p = p_0$$
$$H_a: p < p_0 \qquad H_a: p > p_0 \qquad H_a: p \neq p_0$$

La primera forma es una prueba de cola inferior, la segunda es de cola superior y la tercera es de dos colas.

Las pruebas de hipótesis para la proporción poblacional se basan en la diferencia entre la proporción muestral $\bar{p}$ y la proporción poblacional hipotética p_0. Los métodos para realizarlas son semejantes a los usados para las pruebas de hipótesis de la media poblacional. La única diferencia radica en que para calcular el estadístico de prueba se usa la proporción muestral y su error estándar. Después, para determinar si la hipótesis nula es rechazada, se utiliza el método del valor-p o el método del valor crítico.

Para ver un ejemplo, considere el caso del campo de golf Pine Creek. En los años anteriores, 20% de los jugadores del campo eran mujeres. Para aumentar la proporción del sector femenino, Pine Creek realizó una promoción especial diseñada para atraer a mujeres golfistas. Un mes después de realizada la promoción, el directivo del campo solicitó un estudio estadístico para determinar si la proporción de jugadoras había aumentado. Como el objetivo es determinar si la proporción de jugadoras se incrementó, lo apropiado es una prueba de cola superior en la que H_a: $p > 0.20$. Las hipótesis nula y alternativa para esta prueba son:

$$H_0: p \leq 0.20$$
$$H_a: p > 0.20$$

Si H_0 se puede rechazar, los resultados de la prueba darán sustento estadístico a la conclusión de que la proporción de golfistas aumentó y que la promoción fue efectiva. El directivo del campo especificó que se usara $\alpha = 0.05$ como nivel de significancia para realizar esta prueba de hipótesis.

El paso siguiente en el procedimiento de prueba de hipótesis es seleccionar una muestra y calcular el valor del estadístico de prueba adecuado. Para demostrar cómo se realiza este paso en la prueba de cola superior, se comienza por calcular el valor del estadístico de prueba en cualquiera de las formas de prueba de hipótesis para la proporción poblacional. La distribución de muestreo de $\bar{p}$, el estimador puntual del parámetro poblacional p, es la base para desarrollar el estadístico de prueba.

Si la hipótesis nula es verdadera como igualdad, el valor esperado de $\bar{p}$ es igual al valor hipotético p_0; es decir, $E(\bar{p}) = p_0$. El error estándar de $\bar{p}$ está dado por

$$\sigma_{\bar{p}} = \sqrt{\frac{p_0(1 - p_0)}{n}}$$

En el capítulo 7 se dijo que si $np \geq 5$ y $n(1 - p) \geq 5$, la distribución de muestreo de $\bar{p}$ puede aproximarse mediante una distribución normal.[3] Bajo estas condiciones que generalmente se pueden aplicar en la práctica, el estadístico

$$z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}} \tag{9.3}$$

tiene una distribución de probabilidad normal estándar. Con $\sigma_{\bar{p}} = \sqrt{p_0(1 - p_0)/n}$, la variable aleatoria normal estándar z es el estadístico de prueba empleado para realizar las pruebas de hipótesis acerca de la proporción poblacional.

[3] En la mayor parte de las aplicaciones de pruebas de hipótesis para la proporción poblacional, los tamaños de las muestras son suficientemente grandes para usar la aproximación a la distribución normal. La distribución de muestreo exacta de $\bar{p}$ es discreta y la probabilidad para cada valor de $\bar{p}$ está dada por la distribución binomial. En consecuencia, las prue-bas de hipótesis son un poco más complicadas cuando las muestras son pequeñas y no se puede usar la aproximación a la distribución normal.

ESTADÍSTICO DE PRUEBA EN LAS PRUEBAS DE HIPÓTESIS PARA LA PROPORCIÓN POBLACIONAL

$$z = \frac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} \tag{9.4}$$

WEB archivo
WomenGolf

Ahora es posible calcular el estadístico de prueba para la prueba de hipótesis del campo de golf Pine Creek. Considere una muestra aleatoria de 400 jugadores en la que 100 de ellos son mujeres. La proporción de las golfistas en la muestra es

$$\bar{p} = \frac{100}{400} = 0.25$$

Al aplicar la ecuación (9.4) el valor del estadístico de prueba es

$$z = \frac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} = \frac{0.25 - 0.20}{\sqrt{\dfrac{0.20(1 - 0.20)}{400}}} = \frac{0.05}{0.02} = 2.50$$

Como la prueba de hipótesis para el campo de golf es una prueba de cola superior, el valor-p es la probabilidad de que z sea mayor o igual que $z = 2.50$; esto es, es el área bajo la curva normal estándar para $z \geq 2.50$. En la tabla de probabilidad normal estándar aparece que el área a la izquierda de $z = 2.50$ es 0.9938. Por tanto, el valor-p en la prueba de Pine Creek es $1.0000 = 0.9938 = 0.0062$. En la figura 9.7 se ilustra el cálculo de este valor-p.

Recuerde que el administrador del campo especificó $\alpha = 0.05$ como nivel de significancia. Un valor-$p = 0.0062 < 0.05$ proporciona evidencia estadística suficiente para rechazar H_0 al nivel de significancia 0.05. Así, la prueba proporciona apoyo estadístico suficiente para concluir que la promoción especial incrementó la proporción de jugadoras en el campo de golf.

La decisión de rechazar o no la hipótesis nula también se toma utilizando el método del valor crítico. El valor crítico que corresponde a un área de 0.05 en la cola superior de una distribución de probabilidad normal es $z_{0.05} = 1.645$. Entonces, la regla de rechazo usando el método del valor crítico exige descartar H_0 si $z \geq 1.645$. Como $z = 2.50 > 1.645$, H_0 es rechazada.

Una vez más, los métodos del valor-p y del valor crítico llevan a la misma conclusión en una prueba de hipótesis, pero el primero proporciona más información. Para un valor-$p = 0.0062$,

FIGURA 9.7 Cálculo del valor-p para la prueba de hipótesis de Pine Creek

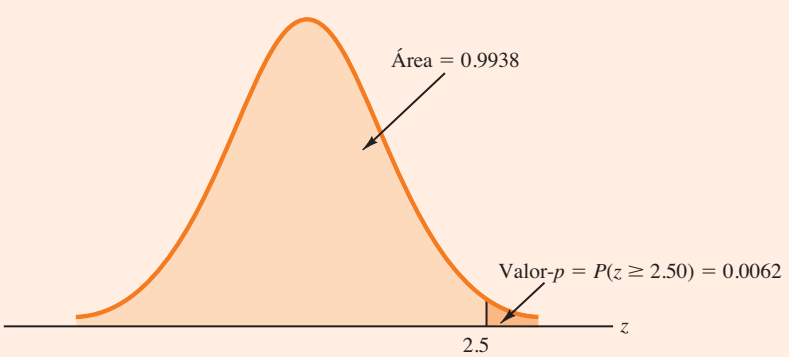


TABLA 9.4 Resumen de las pruebas de hipótesis para la proporción poblacional

	Prueba de cola inferior	Prueba de cola superior	Prueba de dos colas
Hipótesis	$H_0: p \geq p_0$ $H_a: p < p_0$	$H_0: p \leq p_0$ $H_a: p > p_0$	$H_0: p = p_0$ $H_a: p \neq p_0$
Estadístico de prueba	$z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	$z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	$z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$
Regla de rechazo: método del valor-*p*	Rechazar H_0 si el valor-$p \leq \alpha$	Rechazar H_0 si el valor-$p \leq \alpha$	Rechazar H_0 si el valor-$p \leq \alpha$
Regla de rechazo: método del valor crítico	Rechazar H_0 si $z \leq -z_\alpha$	Rechazar H_0 si $z \geq z_\alpha$	Rechazar H_0 si $z \leq -z_{\alpha/2}$ o si $z \geq z_{\alpha/2}$

la hipótesis nula será rechazada para cualquier nivel de significancia mayor o igual que 0.0062.

Resumen

El procedimiento empleado en una prueba de hipótesis para la proporción poblacional es semejante al método usado en una prueba de hipótesis para la media poblacional. Aunque sólo se ilustró cómo realizar una prueba de hipótesis de cola superior para la proporción poblacional, en el caso de pruebas de cola inferior o de dos colas se recurre a procedimientos similares. En la tabla 9.4 se presenta una síntesis de las pruebas de hipótesis para la proporción poblacional. Se supone que $np \geq 5$ y $n(1 - p) \geq 5$, con lo cual se puede usar una distribución normal como aproximación a la distribución de muestreo de $\bar{p}$.

Ejercicios

Métodos

35. Considere la prueba de hipótesis siguiente:

$$H_0: p = 0.20$$
$$H_a: p \neq 0.20$$

En una muestra de 400 se encontró una proporción muestral de $\bar{p} = 0.175$.

a) Calcule el valor del estadístico de prueba.
b) ¿Cuál es el valor-*p*?
c) Con $\alpha = 0.05$, ¿cuál es su conclusión?
d) ¿Cuál es la regla de rechazo usando el valor crítico? ¿Qué concluye?

AUTOevaluación

36. Considere la prueba de hipótesis siguiente:

$$H_0: p \geq 0.75$$
$$H_a: p < 0.75$$

Se seleccionó una muestra de 300 elementos. Calcule el valor-*p* y establezca su conclusión para cada uno de los resultados muestrales siguientes. Use $\alpha = 0.05$.

a) $\bar{p} = 0.68$
b) $\bar{p} = 0.72$
c) $\bar{p} = 0.70$
d) $\bar{p} = 0.77$

Aplicaciones

37. En un estudio se encontró que, en 2005, el 12.5% de los trabajadores estadounidenses pertenecía a un sindicato (*The Wall Street Journal*, 21 de enero de 2006). El caso es que en 2006 se toma una muestra de 400 trabajadores para ver si el esfuerzo realizado por los sindicatos por organizarse ha hecho que aumente el número de sus miembros.
    a) Formule las hipótesis que puedan ser usadas para determinar si la afiliación a los sindicatos ha aumentado en 2006.
    b) Si los resultados muestrales indican que 52 de los trabajadores pertenecen a los sindicatos, ¿cuál es el valor-*p* de esta prueba de hipótesis?
    c) Con $\alpha = 0.05$, ¿cuál es su conclusión?

AUTO evaluación

38. Un estudio realizado por *Consumer Reports* indica que 64% de los clientes de los supermercados piensa que las marcas de esos establecimientos son tan buenas como las marcas nacionales. Para investigar si estos resultados aplican a sus propios productos, un fabricante de salsa de tomate de una marca nacional preguntó a los integrantes de una muestra si consideraban las salsas de tomate de marca propia de los supermercados tan buenas como la de marca nacional.
    a) Formule las hipótesis para determinar si el porcentaje de clientes de los supermercados que considera las salsas de tomate de marca propia de estos establecimientos tan buenas como la de marca nacional difiere de 64%.
    b) Si en una muestra de 100 clientes, 52 opinan que las marcas de los supermercados son tan buenas como las nacionales, ¿cuál es el valor-*p*?
    c) Con $\alpha = 0.05$, ¿cuál es la conclusión?
    d) ¿Le dará gusto esta conclusión al fabricante de la marca nacional de salsa de tomate? Explique.

WEB archivo
AgeGroup

39. Según el Pew Internet & American Life Project, 75% de los estadounidenses adultos usa Internet (sitio web de Pew Internet, 19 de abril de 2008). Los autores del projecto Pew también reportaron el porcentaje de estadounidenses que usa Internet por grupo de edad. Los datos en el archivo AgeGroup son congruentes con sus hallazgos. Esos datos fueron obtenidos de una muestra de 100 usuarios en el grupo de edad de 30 a 49 años y 200 usuarios en el grupo de edad de 50 a 64 años. Un Yes (Sí) indica que el encuestado usa Internet; un No indica que el encuestado no lo hace.
    a) Establezca las hipótesis que pueden utilizarse para determinar si el porcentaje de usuarios de Internet en ambos grupos de edad difieren del promedio general de 75%.
    b) Estime la proporción de usuarios en el grupo de edad de 30 a 49 años. ¿Esta proporción difiere significativamente de la proporción general de 0.75? Utilice $\alpha = 0.05$.
    c) Determine la proporción de usuarios en el grupo de edad de 50 a 64 años. ¿Esta proporción difiere significativamente de la proporción general de 0.75? Utilice $\alpha = 0.05$.
    d) ¿Esperaría que la proporción de usuarios en el grupo de 18 a 29 años sea más grande o más pequeña que la proporción del grupo de edad de 30 a 49 años? Sustente su conclusión con los resultados obtenidos en los incisos *b*) y *c*).

40. Antes del Super Bowl de 2003, la ABC pronosticó que 22% de la audiencia por televisión expresaría interés por ver uno de sus programas por estrenar, entre ellos: *8 Simple Rules, Are You Hot?* y *Dragnet.* Durante el Super Bowl, la ABC pasó anuncios sobre estos programas de televisión. Al día siguiente del evento, una firma de publicidad tomó una muestra de 1 532 espectadores que los vieron, de los cuales 414 afirmaron que verían alguna de las series promovidas por la ABC (*The Wall Street Journal*, 30 de enero de 2003).
    a) ¿Cuál es la estimación puntual de la proporción de espectadores que después de ver los anuncios dijeron que verían los programas de televisión?
    b) Con $\alpha = 0.05$, determine si la intención de ver los programas de la ABC aumentó significativamente después de ver los anuncios. Establezca las hipótesis apropiadas, calcule el valor-*p* y defina su conclusión.
    c) ¿Por qué tales estudios son valiosos para las empresas y los negocios de publicidad?

41. En una conferencia en 2006, un ejecutivo de una empresa de correduría en el mercado de dinero dijo a un grupo de analistas que por lo menos 70% de los inversionistas confiaba en lograr

sus objetivos de inversión. UBS Investor Optimism Survey realizó un estudio, del 2 al 15 de enero, y encontró que 67% de los inversionistas confiaba en lograr sus objetivos de inversión (CNCB, 20 de enero de 2006).

a) Formule las hipótesis para probar la validez de lo expresado por el ejecutivo de la empresa mediadora en el mercado de dinero.

b) Suponga que para este estudio, UBS Investor Optimism Survey recabó información de 300 inversionistas. ¿Cuál es el valor-*p* en esta prueba de hipótesis?

c) Con $\alpha = 0.50$, ¿debe rechazarse lo que afirma el ejecutivo?

42. Según el Center for Logistics Management de la Universidad de Nevada, 6% de todas las mercancías vendidas en Estados Unidos son devueltas (*BusinessWeek*, 15 de enero de 2007). Una tienda departamental en Houston tomó una muestra de 80 artículos vendidos en enero y encontró que 12 de ellos fueron devueltos.

a) Calcule una estimación puntual de la proporción de artículos devueltos para la población de transacciones de ventas en el almacén de Houston.

b) Calcule un intervalo de 95% de confianza para la proporción de devoluciones en el almacén de Houston.

c) ¿La proporción de devoluciones es significativamente distinta de las devoluciones de toda la nación en conjunto? Proporcione sustento estadístico para su respuesta.

WEB archivo
Eagle

43. Eagle Outfitters es una cadena de tiendas que se especializa en ropa de invierno y equipo para excursionismo. Esta empresa planea una promoción con envío de cupones de descuento para todos sus clientes con tarjeta de crédito. La promoción se considerará un éxito si más de 10% de los que reciban el cupón lo utilizan. Antes de realizar la promoción a nivel nacional, se envía cupones a una muestra de 100 clientes con tarjeta de crédito.

a) Desarrolle las hipótesis que pueden utilizarse para probar si la proporción poblacional de aquellos que usarán el cupón es suficiente como para hacer la promoción en todo el país.

b) El archivo Eagle contiene los datos muestrales. Obtenga una estimación puntual de la proporción poblacional.

c) Use $\alpha = 0.05$ y realice la prueba de hipótesis. ¿La empresa debe realizar esta promoción en todo el país?

WEB archivo
Drowsy

44. En un artículo anunciado en su portada, *BusinessWeek* publicó información acerca de los hábitos de sueño de los estadounidenses (*BusinessWeek*, 26 de enero de 2004). El artículo señalaba que la privación del sueño ocasiona diversos problemas, entre ellos muertes en las autopistas. El 51% de los conductores admitió manejar sintiéndose somnoliento. Un investigador planteó la hipótesis de que este problema es aún mayor entre los trabajadores de los turnos nocturnos.

a) Formule las hipótesis que ayuden a determinar si más de 51% de la población de trabajadores de los turnos nocturnos admite conducir somnoliento.

b) En una muestra de 400 trabajadores de turnos nocturnos se identificó a quienes admitían conducir somnolientos. Consulte el archivo Drowsy. ¿Cuál es la proporción muestral? ¿Cuál es el valor-*p*?

c) Con $\alpha = 0.1$, ¿cuál es su conclusión?

45. Numerosos inversionistas y analistas financieros piensan que el promedio industrial Dow Jones (DJIA) es un buen barómetro del mercado de acciones. El 31 de enero de 2006, de las 30 acciones que constituyen el DJIA, 9 aumentaron de precio (*The Wall Street Journal*, 1 de febrero de 2006). A partir de este hecho, un analista bursátil afirmó que 30% de las acciones de la Bolsa de Nueva York se incrementarían ese mismo día.

a) Formule las hipótesis nula y alternativa para probar lo que afirma el analista.

b) En una muestra de 50 acciones de la bolsa de Nueva York, 24 aumentaron. Establezca la estimación puntual de la proporción poblacional de las acciones se incrementaron.

c) Realice una prueba de hipótesis usando $\alpha = 0.01$ como nivel de significancia. ¿Cuál es la conclusión?

9.6 Prueba de hipótesis y toma de decisiones

En las secciones previas de este capítulo se estudiaron aplicaciones de pruebas de hipótesis consideradas pruebas de significancia. Después de formular las hipótesis nula y alternativa, se selecciona una muestra y se calcula el valor de un estadístico de prueba y el valor-*p* asociado.

Se compara, entonces, el valor-p con una probabilidad controlada de cometer un error tipo I, α, que se conoce como *nivel de significancia para la prueba*. Si el valor-$p \leq \alpha$, se concluye "rechazar H_0", y los resultados se declaran significantes; de otra manera, se concluye "no rechazar H_0". Con una prueba de significancia se controla la probabilidad de cometer un error tipo I, pero no uno tipo II. Por tanto, se recomienda la conclusión "no rechazar H_0" más que "aceptar H_0", po que esta última nos expone al riesgo de cometer un error tipo II de aceptar H_0 cuando es falsa. Con la conclusión de "no rechazar H_0" la evidencia estadística se considera no concluyente y es por lo general un indicador para postergar una decisión o una acción hasta que se pueda realizar mayor investigación y pruebas.

Pero si el propósito de una prueba de hipótesis es tomar cierta decisión cuando H_0 es verdadera y una decisión diferente cuando H_a es verdadera, quien debe tomarla deseará, y en muchos casos tendrá que actuar tanto en el caso en que la conclusión sea *no rechazar* H_0 como en el caso en que sea *rechazar* H_0. Si se da esta situación, los expertos en estadística recomiendan controlar la probabilidad de cometer un error tipo II. Con las probabilidades controladas de cometer tanto un error tipo I como tipo II, la conclusión de la prueba de hipótesis es ya sea *aceptar* H_0 o *rechazar* H_0. En el primer caso, se concluye que H_0 es verdadera, mientras que en el segundo, que H_a es verdadera. Así, se puede tomar una decisión y emprender una acción apropiada cuando se llegó a una conclusión.

Una buena ilustración de una prueba de hipótesis para tomar decisiones es el muestreo de aceptación de lotes, un tema que se discutirá con más detalle en el capítulo 20. Por ejemplo, un director de control de calidad tiene que decidir si acepta un pedido de baterías de un proveedor o si lo rechaza por ser de mala calidad. Suponga que las especificaciones de diseño indican que se requieren baterías con una vida útil promedio de por lo menos 120 horas. Para evaluar si el pedido recibido satisface esta especificación, se selecciona una muestra de 36 baterías y se prueban. Con base en esta muestra, se deberá tomar la decisión de aceptar el pedido o devolverlo al proveedor por no tener la calidad adecuada. Sea μ el número medio de horas de vida útil que tienen las baterías del envío. Las hipótesis nula y alternativa para la media poblacional se presentan a continuación.

$$H_0: \mu \geq 120$$
$$H_a: \mu < 120$$

Si H_0 es rechazada, se concluye que la hipótesis alternativa es verdadera. Esta conclusión indica que lo adecuado es devolver el pedido al proveedor. Pero si H_0 no es rechazada, la persona que toma la decisión deberá determinar qué medidas tomar. Así, sin haber concluido que H_0 es verdadera, sino sólo por no haberla rechazado, dicha persona tendrá que aceptar el envío y considerarlo de la calidad adecuada.

En tales situaciones es recomendable que el procedimiento de prueba de hipótesis se amplíe para controlar la probabilidad de cometer un error tipo II. Como se tomará una decisión y alguna medida cuando H_0 no sea rechazada, será útil conocer la probabilidad de cometer un error de este tipo. En las secciones 9.7 y 9.8 se explica cómo calcular la probabilidad de cometer un error tipo II y ajustar el tamaño de la muestra para controlar esta probabilidad.

9.7 Cálculo de la probabilidad de los errores tipo II

En esta sección se describe cómo calcular la probabilidad de cometer un error tipo II en una prueba de hipótesis para la media poblacional. Este procedimiento se ilustra usando el ejemplo del muestreo de aceptación de lotes descrito en la sección 9.6. Las hipótesis nula y alternativa para el número medio de horas de vida útil de un pedido de baterías son: H_0: $\mu \geq 120$ y H_a: $\mu < 120$. Si H_0 es rechazada, la decisión será devolver el producto al proveedor, debido a que

la media del número de horas de vida útil es menor que la especificada de 120 horas. Si H_0 no se rechaza, la decisión será aceptar el pedido.

Suponga que se usa el nivel de significancia de $\alpha = 0.05$ para realizar la prueba de hipótesis. El estadístico de prueba en el caso de σ conocida es

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 120}{\sigma/\sqrt{n}}$$

Con base en el método del valor crítico y $z_{0.05} = 1.645$, la regla de rechazo en esta prueba de cola inferior es

$$\text{rechazar } H_0 \text{ si } z \leq -1.645$$

Asuma que se seleccionará una muestra de 36 baterías y que por pruebas anteriores se puede considerar que se conoce la desviación estándar poblacional y que su valor es $\sigma = 12$ horas. La regla de rechazo indica que H_0 será descartada si

$$z = \frac{\bar{x} - 120}{12/\sqrt{36}} \leq -1.645$$

Al despejar $\bar{x}$ de la expresión anterior, tenemos que H_0 será rechazada si

$$\bar{x} \leq 120 - 1.645\left(\frac{12}{\sqrt{36}}\right) = 116.71$$

Rechazar H_0 siempre que $\bar{x} \leq 116.71$ significa que se tomará la decisión de aceptar el pedido siempre que

$$\bar{x} > 116.71$$

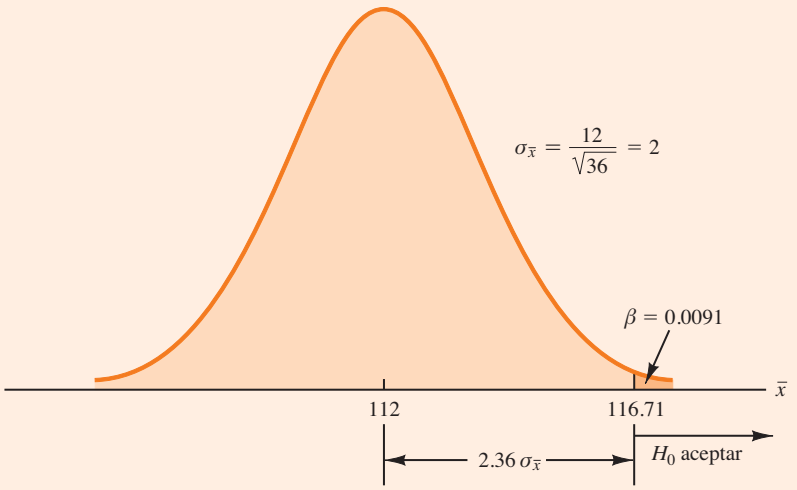
Con esta información se pueden calcular ya las probabilidades asociadas con cometer un error tipo II. Primero, recuerde que se comete este error cuando la verdadera media del pedido es menor de 120 horas y se decida aceptar H_0: $\mu \geq 120$. Por tanto, para calcular la probabilidad de cometerlo, se debe elegir un valor de μ menor que 120 horas. Por ejemplo, suponga que la calidad del envío es pobre si la vida promedio de las baterías es $\mu = 112$ horas. Si en realidad es verdad que $\mu = 112$, ¿cuál es la probabilidad de aceptar H_0: $\mu \geq 120$ y cometer así un error tipo II? Observe que es la probabilidad de que la media muestral $\bar{x}$ sea mayor de 116.71 cuando $\mu = 112$.

En la figura 9.8 se presenta la distribución de muestreo de $\bar{x}$ si la media es $\mu = 112$. El área sombreada en la cola superior da la probabilidad de obtener $\bar{x} > 116.71$. Utilizando la distribución normal estándar vemos que para $\bar{x} = 116.71$.

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{116.71 - 112}{12/\sqrt{36}} = 2.36$$

La tabla de probabilidad normal estándar indica que para $z = 2.36$, el área en la cola superior es $1.0000 - 0.9909 = 0.0091$. Entonces, 0.0091 es la probabilidad de cometer un error tipo II cuando $\mu = 112$. Si se usa β para denotar la probabilidad de cometer este error, tenemos que si $\mu = 112$, $\beta = 0.0091$. Podemos concluir que si la media de la población es 112 horas, la probabilidad de incurrir en un error tipo II es de sólo 0.0091.

FIGURA 9.8 Probabilidad de un error tipo II cuando $\mu = 112$



Estos cálculos se repiten con otros valores de μ menores de 120. Para cada valor de μ se obtendrán diferentes probabilidades de cometer un error tipo II. Por ejemplo, suponga que en el pedido de baterías la media de vida útil es $\mu = 115$ horas. Como H_0 será aceptada siempre que $\bar{x} > 116.71$, el valor z obtenido con $\mu = 115$ está dado por

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{116.71 - 115}{12/\sqrt{36}} = 0.86$$

En la tabla de probabilidad normal estándar vemos que el área en la cola superior de la distribución normal estándar que corresponde a $z = 0.86$ es $1.0000 - 0.8051 = 0.1949$. Si la verdadera media es $\mu = 115$, la probabilidad de incurrir en un error tipo II es $\beta = 0.1949$.

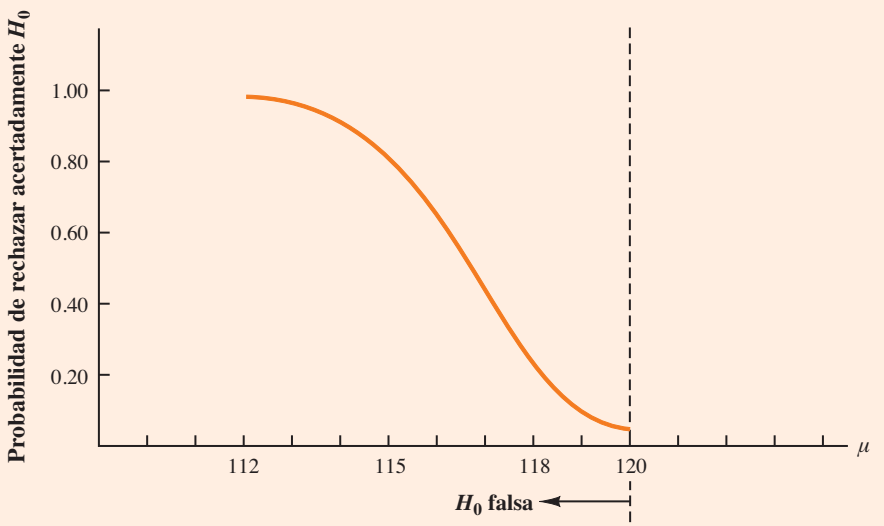
Como se muestra en la tabla 9.5, la probabilidad de cometer un error tipo II depende del valor de la media poblacional μ. Si los valores de μ son cercanos a μ_0, la probabilidad de cometer un error tipo II puede ser alta.

En la tabla 9.5 se muestran las probabilidades de cometer un error tipo II para varios valores de μ menores de 120. Observe que si μ aumenta y se acerca a 120, la probabilidad aumenta hacia un límite superior de 0.95. Pero a medida que μ disminuye y se aleja de 120, la probabilidad de cometer el error disminuye. Este es el patrón que se debe esperar. Cuando la verdadera media poblacional está cerca del valor de la hipótesis nula, $\mu = 120$, la probabilidad de cometer un error tipo II es alta. Pero cuando la verdadera media poblacional está muy por debajo del valor $\mu = 120$ de la hipótesis nula, la probabilidad que se menciona es baja.

TABLA 9.5 Probabilidad de cometer un error tipo II en la prueba de hipótesis del muestreo de aceptación de lotes

Valor de μ	$z = \frac{116.71 - \mu}{12/\sqrt{36}}$	Probabilidad de un error tipo II (β)	Potencia ($1 - \beta$)
112	2.36	0.0091	0.9909
114	1.36	0.0869	0.9131
115	0.86	0.1949	0.8051
116.71	0.00	0.5000	0.5000
117	−0.15	0.5596	0.4404
118	−0.65	0.7422	0.2578
119.999	−1.645	0.9500	0.0500

FIGURA 9.9 Curva de potencia para la prueba de hipótesis del muestreo de aceptación de lotes



A la probabilidad de rechazar acertadamente H_0 cuando es falsa se le llama **potencia** de la prueba. Para cada valor específico de μ la potencia es $1 - \beta$; es decir, la probabilidad de rechazar acertadamente la hipótesis nula es 1 menos la probabilidad de cometer un error tipo II. En la tabla 9.5 se listan también los valores de la potencia. Con base en estos valores, en la figura 9.9 se presentan gráficamente las potencias correspondientes a cada valor μ. A este tipo de gráficas se les conoce como **curva de potencia**. Observe que esta curva se extiende sobre los valores de μ para los que la hipótesis nula es falsa. La altura en la curva de potencia para cualquier valor de μ indica la probabilidad de rechazar acertadamente H_0 cuando es falsa.[4]

En resumen, para calcular la probabilidad de cometer un error tipo II en una prueba de hipótesis para la media poblacional se puede seguir, paso a paso, el procedimiento siguiente.

1. Formular las hipótesis nula y alternativa.
2. Usar el nivel de significancia α y el método del valor crítico para determinar el valor crítico y la regla de rechazo para la prueba.
3. Usar la regla de rechazo para encontrar el valor de la media muestral que corresponde al valor crítico del estadístico de prueba.
4. Utilizar el resultado del paso 3 para determinar el valor de la media muestral que llevará a la aceptación de H_0. Este valor define la región de aceptación de la prueba.
5. Usar la distribución de muestreo de $\bar{x}$ para un valor de μ que satisfaga la hipótesis alternativa y la región de aceptación del paso 4 para calcular la probabilidad de que la media muestral se encuentre en la región de aceptación. Ésta es la probabilidad de cometer un error tipo II dado el valor de μ elegido.

Ejercicios

Métodos

AUTOevaluación

46. Considere la prueba de hipótesis siguiente.

$$H_0: \mu \geq 10$$
$$H_a: \mu < 10$$

[4] Algunas veces, para proporcionar información acerca de la probabilidad de cometer un error tipo II, se usa otra gráfica denominada *curva característica de operación*, la cual muestra la probabilidad de aceptar H_0 y por tanto proporciona β para los valores de μ en los que la hipótesis nula es falsa. Con esta gráfica se puede leer directamente la probabilidad de cometer un error tipo II.

El tamaño de la muestra es 120 y la desviación estándar poblacional es conocida con $\sigma = 5$. Use $\alpha = 0.05$.

a) Si la media poblacional es 9, ¿cuál es la probabilidad de que la media muestral lleve a la conclusión de *no rechazar* H_0?

b) ¿Qué tipo de error se comete si la verdadera media poblacional es 9 y se concluye que H_0: $\mu \geq 10$ es verdadera?

c) ¿Cuál es la probabilidad de cometer un error tipo II si la verdadera media poblacional es 8?

47. Considere la prueba de hipótesis siguiente.

$$H_0: \mu = 20$$
$$H_a: \mu \neq 20$$

Se toma una muestra de 200 elementos y la desviación estándar poblacional es $\sigma = 10$. Utilice $\alpha = 0.05$. Calcule la probabilidad de cometer un error tipo II si la media poblacional es:

a) $\mu = 18.0$

b) $\mu = 22.5$

c) $\mu = 21.0$

Aplicaciones

48. Fowle Marketing Research, Inc. basa la tarifa que cobra a sus clientes en el supuesto de que una encuesta por teléfono se puede realizar en un promedio de 15 minutos o menos. Si se requiere más tiempo en promedio, se cobra una cantidad adicional. Con una muestra de 35 encuestas, una desviación estándar poblacional de 4 minutos y 0.01 como nivel de significancia, se usará la media muestral para probar la hipótesis nula H_0: $\mu \leq 15$.

a) ¿Cuál es su interpretación del error tipo II en este problema? ¿Qué impacto tiene en la empresa?

b) ¿Cuál es la probabilidad de cometer un error tipo II si la verdadera media de los tiempos es $\mu = 17$ minutos?

c) ¿Cuál es la probabilidad de cometerlo si la verdadera media de los tiempos es $\mu = 18$ minutos?

d) Dibuje la forma general de la curva de potencia de esta prueba.

AUTOevaluación

49. Un grupo de investigación del consumidor está interesado en probar la afirmación de un fabricante de automóviles de que un nuevo modelo económico recorrerá por lo menos 25 millas por cada galón de gasolina (H_0: $\mu \geq 25$).

a) Con 0.02 como nivel de significancia y una muestra de 30 automóviles, ¿cuál es la regla de rechazo basada en el valor de $\bar{x}$ en la prueba para determinar si debe rechazarse la afirmación del fabricante? Suponga que σ es 3 millas por galón.

b) ¿Cuál es la probabilidad de cometer un error tipo II si el verdadero rendimiento es 23 millas por galón?

c) ¿Cuál es la probabilidad de cometerlo si el verdadero rendimiento es 24 millas por galón?

d) ¿Cuál es la probabilidad si el verdadero rendimiento es 25.5 millas por galón?

50. La revista *Young Adult* establece la hipótesis siguiente acerca de la edad de sus suscriptores.

$$H_0: \mu = 28$$
$$H_a: \mu \neq 28$$

a) En esta situación, ¿qué significa cometer un error tipo II?

b) Se supone que la desviación estándar poblacional es conocida como de $\sigma = 6$ años, y que el tamaño de la muestra es 100. Si $\alpha = 0.05$, ¿cuál es la probabilidad de aceptar H_0 si μ es igual a 26, 27, 29 y 30?

c) ¿Cuál es la potencia si $\mu = 26$? ¿Qué le dice este resultado?

51. En la operación de una línea de producción se prueba la exactitud del peso de llenado mediante la prueba de hipótesis siguiente.

Hipótesis	Conclusión y acción
H_0: $\mu = 16$	Llenado correcto; puede continuar
H_a: $\mu \neq 16$	Llenado fuera del estándar; detener y ajustar la máquina

El tamaño de la muestra es 30 y la desviación estándar poblacional es $\sigma = 0.8$. Use $\alpha = 0.05$.

a) En esta situación, ¿qué significa un error tipo II?

b) ¿Cuál es la probabilidad de cometer un error de tipo II si la máquina está llenando con 0.5 onzas de exceso?

c) Si se está llenando con 0.5 onzas de exceso, ¿cuál es la potencia de la prueba estadística?

d) Muestre la curva de potencia para esta prueba de hipótesis. ¿Qué información aporta al gerente de producción?

52. Vaya al ejercicio 48. Suponga que la empresa toma una muestra de 50 encuestas y repita los incisos *b*) y *c*). ¿Qué observación se puede hacer sobre cómo el incremento del tamaño de la muestra afecta la probabilidad de cometer un error tipo II?

53. Sparr Investments, Inc. se especializa en oportunidades de inversión para sus clientes con pago de impuestos diferido. Hace poco, Sparr ofreció un programa de inversión con deducción vía nómina para los empleados de una determinada empresa. Estimó que en este momento los empleados tienen en promedio $100 o menos por mes en inversiones con impuestos diferidos. Para probar la hipótesis de Sparr acerca del nivel actual de las inversiones entre la población de empleados, se toma una muestra de 40 sujetos. Suponga que las cantidades invertidas mensualmente por éstos en inversiones con impuestos diferidos tienen una desviación estándar de $75 y que en esta prueba de hipótesis se usará 0.05 como nivel de significancia.

a) En esta situación, ¿cuál es el error tipo II?

b) ¿Cuál es la probabilidad de cometer un error tipo II si la media de la inversión mensual de los empleados es $120?

c) ¿Cuál es la probabilidad de cometer un error de tipo II si la media de la inversión mensual de los empleados es $130?

d) Suponiendo que se usa un tamaño muestral de 80 empleados, repita los incisos *b*) y *c*).

9.8 Determinación del tamaño de la muestra en una prueba de hipótesis para la media poblacional

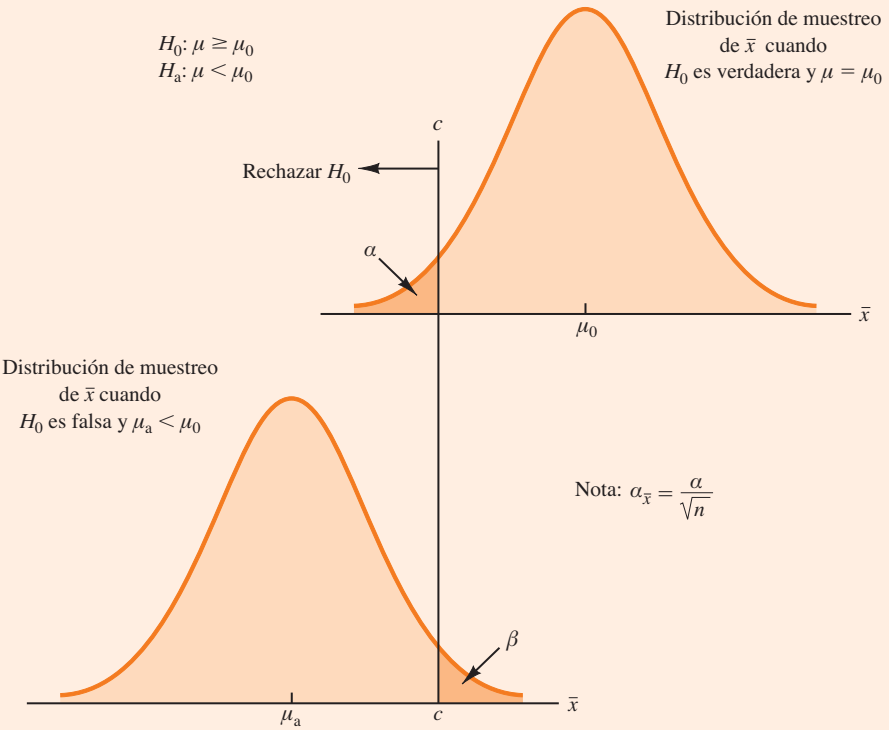
Considere realizar una prueba de hipótesis para el valor de la media poblacional. El nivel de significancia elegido por el usuario determina la probabilidad de cometer un error tipo I en esta prueba. Al controlar el tamaño de la muestra, el usuario también controla la probabilidad de cometer un error tipo II. Enseguida se muestra cómo determinar el tamaño de la muestra en la prueba de hipótesis de cola inferior para la media poblacional que se da a continuación.

$$H_0: \mu \geq \mu_0$$
$$H_a: \mu < \mu_0$$

En la figura 9.10, la gráfica superior presenta la distribución de muestreo de $\bar{x}$ cuando H_0 es verdadera y $\mu = \mu_0$. En una prueba de cola inferior el valor crítico del estadístico de prueba se denota como $-z_\alpha$. La línea vertical, c, en la gráfica superior de la figura, señala el valor correspondiente de $\bar{x}$. Observe que si H_0 es rechazada cuando $\bar{x} \leq c$, la probabilidad de cometer un error tipo I será α. Si z_α representa el valor de z que corresponde al área α en la cola superior de la distribución normal estándar, la fórmula siguiente se emplea para calcular c.

$$c = \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}} \qquad \textbf{(9.5)}$$

FIGURA 9.10 Determinación del tamaño de la muestra para valores específicos de las probabilidades de cometer un error tipo I (α) y un error tipo II (β)



La gráfica inferior es la distribución de muestreo de $\bar{x}$ cuando la hipótesis alternativa es verdadera siendo $\mu = \mu_a < \mu_0$. La región sombreada muestra β, la probabilidad de cometer un error tipo II a la cual está expuesta la persona que toma la decisión de aceptar la hipótesis nula cuando $\bar{x} > c$. Si z_β representa el valor z que corresponde al área de β en la cola superior de la distribución normal estándar, c se calcula empleando la fórmula siguiente.

$$c = \mu_a - z_\beta \frac{\sigma}{\sqrt{n}} \tag{9.6}$$

Ahora lo que buscamos es elegir un valor para c, de manera que cuando H_0 sea rechazada y H_a aceptada, la probabilidad de cometer un error tipo I sea igual a la probabilidad elegida para α, y la probabilidad de cometer un error tipo II sea igual al valor elegido para β. Por consiguiente, con ambas ecuaciones (9.5) y (9.6) se debe obtener el mismo valor de c y la ecuación siguiente debe satisfacerse.

$$\mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}} = \mu_a + z_\beta \frac{\sigma}{\sqrt{n}}$$

Para determinar el tamaño de muestra que se necesita, primero se despeja $\sqrt{n}$ como sigue.

$$\mu_0 - \mu_a = z_\alpha \frac{\sigma}{\sqrt{n}} + z_\beta \frac{\sigma}{\sqrt{n}}$$

$$\mu_0 - \mu_a = \frac{(z_\alpha + z_\beta)\sigma}{\sqrt{n}}$$

y

$$\sqrt{n} = \frac{(z_\alpha + z_\beta)\sigma}{(\mu_0 - \mu_a)}$$

Al elevar al cuadrado ambos lados de la expresión, obtenemos la fórmula siguiente para el tamaño de la muestra necesario en una prueba de hipótesis de una cola para la media poblacional.

TAMAÑO DE LA MUESTRA EN UNA PRUEBA DE HIPÓTESIS DE UNA COLA PARA LA MEDIA POBLACIONAL

$$n = \frac{(z_\alpha + z_\beta)^2\sigma^2}{(\mu_0 - \mu_a)^2} \qquad \textbf{(9.7)}$$

donde

z_α = valor de z que proporciona un área de α en la cola superior de la distribución normal estándar.

z_β = valor de z que proporciona un área de β en la cola superior de la distribución normal estándar.

σ = desviación estándar poblacional.

μ_0 = valor de la media poblacional en la hipótesis nula.

μ_a = valor de la media poblacional utilizada para el error tipo II.

Nota. Para una prueba de hipótesis de dos colas, en la ecuación (9.7) se usa $z_{\alpha/2}$ en lugar de z_α.

Aunque la lógica de la ecuación (9.7) se desarrolló para la prueba de hipótesis mostrada en la figura 9.10, también es válida en cualquier prueba de hipótesis de una cola para la media poblacional. En una prueba de hipótesis de dos colas para la media poblacional se usa $z_{\alpha/2}$ en lugar de z_α en la misma ecuación.

Volvamos al ejemplo del muestreo de aceptación de lotes presentado en las secciones 9.6 y 9.7. Las especificaciones de diseño para el embarque de las baterías indican una vida media útil de por lo menos 120 horas. Los pedidos se regresan si H_0 es rechazada: $\mu \geq 120$. Suponga que el gerente de control de calidad establece las siguientes declaraciones acerca de las probabilidades admisibles de cometer los errores tipo I y tipo II:

Declaración para el error tipo I. Si la vida media de las baterías del pedido es $\mu = 120$, estoy dispuesto a asumir el riesgo de que la probabilidad de rechazar el embarque sea $\alpha = 0.05$.

Declaración para el error tipo II. Si la vida media de las baterías del pedido es 5 horas por debajo de lo que indican las especificaciones (es decir, $\mu = 115$), estoy dispuesto a asumir el riesgo de que la probabilidad de aceptar el embarque sea $\beta = 0.10$.

Estas declaraciones se basan en el criterio del gerente. Otra persona podría establecer diferentes restricciones para las probabilidades. Sin embargo, las declaraciones acerca de las probabilidades admisibles de ambos errores deben establecerse antes de determinar el tamaño de la muestra.

En el ejemplo, $\alpha = 0.05$ y $\beta = 0.10$. Mediante la distribución de probabilidad normal estándar, se tiene $z_{0.05} = 1.645$ y $z_{0.10} = 1.28$. De acuerdo con lo dicho al especificar las probabilidades para los errores, observamos que $\mu_0 = 120$ y $\mu_a = 115$. Por último, supusimos que la desviación estándar poblacional se conocía y era $\sigma = 12$. Mediante la ecuación (9.7) encontramos que el tamaño de muestra recomendado para el ejemplo del muestreo de aceptación de lotes es

$$n = \frac{(1.645 + 1.28)^2(12)^2}{(120 - 115)^2} = 49.3$$

Al redondear hacia arriba, el tamaño de muestra recomendado es 50.

Como las probabilidades de los dos errores tipo I y tipo II se han controlado usando $n = 50$, queda justificado que el gerente de control de calidad utilice las declaraciones H_0 *es aceptada* o H_0 *es rechazada* en esta prueba de hipótesis. Las inferencias correspondientes se hacen teniendo probabilidades admisibles de cometer un error de cualquiera de ambos tipos.

Acerca de la relación entre α, β y el tamaño n de la muestra caben tres observaciones.

1. Una vez que se tienen dos de estos tres valores, el tercero puede calcularse.
2. Dado un nivel de significancia α, aumentando el tamaño de la muestra se reduce β.
3. Dado un tamaño de muestra, al reducirse α aumenta β y al incrementarse α, disminuye β.

La tercera observación debe tenerse en cuenta cuando no se controla la probabilidad de cometer un error tipo II. Dicha observación indica que no se deben elegir niveles de significancia α innecesariamente pequeños. Para un tamaño de muestra dado, elegir un nivel de significancia pequeño implica más riesgo de cometer un error tipo II. Personas con poca experiencia piensan que al realizar una prueba de hipótesis es mejor usar siempre valores pequeños de α, lo cual es cierto si la única preocupación es cometer un error tipo I. Sin embargo, los valores pequeños de α tienen la desventaja de incrementar la probabilidad de cometer un error tipo II.

Ejercicios

Métodos

AUTOevaluación

54. Considere la prueba de hipótesis siguiente.

$$H_0: \mu \geq 10$$
$$H_a: \mu < 10$$

El tamaño de la muestra es 120 y la desviación estándar poblacional 5. Use $\alpha = 0.05$. Si la media poblacional real es 9, la probabilidad de cometer un error tipo II es 0.2912. Suponga que el investigador desea reducir a 0.10 la probabilidad de cometer este tipo de error si la media poblacional verdadera es 9. ¿Qué tamaño de muestra se recomienda?

55. Considere la prueba de hipótesis siguiente.

$$H_0: \mu = 20$$
$$H_a: \mu \neq 20$$

La desviación estándar poblacional es 10. Use $\alpha = 0.05$. ¿De qué tamaño deberá tomarse la muestra si el investigador está dispuesto a aceptar una probabilidad de 0.05 de cometer un error tipo II cuando la media poblacional actual sea 22?

Aplicaciones

56. Suponga que el director del proyecto de estudio de Hilltop Coffee (vea la sección 9.3) solicita una probabilidad de 0.10 de declarar que la empresa no comete ninguna violación si en realidad está llenando con 1 onza de menos ($\mu_a = 2.9375$ libras), ¿Cuál será el tamaño de muestra recomendado?

AUTOevaluación

57. Una batería industrial especial debe tener una vida de por lo menos 400 horas. Considere una prueba de hipótesis con 0.02 como nivel de significancia. Si en las baterías de un determinado lote de producción la media verdadera de vida útil es 385 horas, el gerente de producción desea un procedimiento de muestreo que sólo 10% de las veces indique de manera errónea que el lote es aceptable. ¿Qué tamaño de muestra se recomienda para esta prueba de hipótesis? Use 30 horas como estimación de la desviación estándar poblacional.

58. La revista *Young Adult* plantea la hipótesis siguiente acerca de la edad de sus suscriptores.

$$H_0: \mu = 28$$
$$H_a: \mu \neq 28$$

Si el gerente que realiza la prueba admite una probabilidad de 0.15 de cometer un error tipo II cuando la verdadera edad promedio es 29 años, ¿de qué tamaño debe tomarse la muestra? Suponga que $\sigma = 6$ y que el nivel de significancia es 0.05.

59. En un estudio sobre el rendimiento de la gasolina en los automóviles, medido como millas por galón, se probaron las hipótesis siguientes.

Hipótesis	Conclusión
$H_0: \mu \geq 25$ mpg	Confirma lo que sostiene el fabricante
$H_a: \mu < 25$ mpg	Refuta lo que sostiene el fabricante; el rendimiento es menor de lo afirmado

Para $\sigma = 3$ y un nivel de significancia de 0.02, ¿qué tamaño de muestra se recomienda si el investigador desea tener 80% de probabilidad de detectar que μ es menor que 25 millas por galón cuando en realidad es 24?

Resumen

Las pruebas de hipótesis constituyen un procedimiento estadístico que utiliza datos muestrales para determinar si una afirmación acerca del valor de un parámetro poblacional debe o no rechazarse. Como hipótesis se tienen dos afirmaciones opuestas acerca de un parámetro poblacional. A una se le llama *hipótesis nula* (H_0) y a la otra *hipótesis alternativa* (H_a). En la sección 9.1 se proporcionaron los lineamientos para elaborar estas hipótesis en tres situaciones encontradas a menudo en la práctica.

Si se tienen datos históricos o alguna otra información que proporcione una base para suponer que se conoce la desviación estándar poblacional, el procedimiento de prueba de hipótesis para la media poblacional se sustenta en la distribución normal estándar. Si no se conoce σ, se usa la desviación estándar muestral s para estimarlo, y el procedimiento de la prueba de hipótesis se basa en la distribución t. En ambos casos, la calidad de los resultados depende tanto de la forma de la distribución de la población como del tamaño de la muestra. Si la población tiene distribución normal, los dos procedimientos para la prueba de hipótesis son aplicables, aun con tamaños de muestra pequeños. Si la población no está distribuida normalmente, se necesitan tamaños de muestra mayores. En las secciones 9.3 y 9.4 se proporcionaron los lineamientos generales para determinar este tamaño. En el caso de pruebas de hipótesis para la proporción poblacional, en el procedimiento de la prueba de hipótesis se usa un estadístico de prueba sustentado en la distribución normal estándar.

En todos los casos el valor del estadístico de prueba se utiliza para calcular un valor-p para la prueba. Éste es una probabilidad que se usa para determinar si la hipótesis nula es rechazada o no. Si el valor-p es menor o igual que el nivel de significancia α, la hipótesis nula puede ser rechazada.

Las conclusiones de una prueba de hipótesis también pueden obtenerse al comparar el valor del estadístico de prueba con el valor crítico. En pruebas de cola inferior, la hipótesis nula es rechazada si el valor del estadístico de prueba es menor o igual que el valor crítico. En pruebas de cola superior, la hipótesis nula es rechazada si el valor del estadístico de prueba es mayor o igual al valor crítico. En pruebas de dos colas hay dos valores críticos: uno en la cola inferior de la distribución de muestreo y otro en la cola superior. En este caso, la hipótesis nula es rechazada si el valor del estadístico de prueba es menor o igual al valor crítico de la cola inferior, o bien, mayor o igual que el valor crítico de la cola superior.

También se presentaron extensiones de los procedimientos de prueba de hipótesis para incluir un análisis del error tipo II. En la sección 9.7 se mostró la forma de calcular la probabilidad de cometerlo. En la sección 9.8 se explicó cómo determinar el tamaño de la muestra de manera que se controlen tanto la probabilidad de cometer un error tipo I como un error tipo II.

Glosario

Curva de potencia Gráfica que da la probabilidad de rechazar H_0 para cada uno de los posibles valores del parámetro poblacional que no satisfaga la hipótesis nula. La curva de potencia proporciona las probabilidades de rechazar correctamente la hipótesis nula.
Error tipo I Error de rechazar H_0 cuando es verdadera.
Error tipo II Error de aceptar H_0 cuando es falsa.
Estadístico de prueba Un estadístico cuyo valor ayuda a determinar si la hipótesis nula es rechazada.
Hipótesis alternativa Hipótesis que se concluye como verdadera cuando la hipótesis nula es rechazada.
Hipótesis nula Hipótesis que se supone tentativamente verdadera en una prueba de hipótesis.
Nivel de significancia Probabilidad de cometer un error tipo I cuando la hipótesis nula es verdadera como igualdad.
Potencia Probabilidad de rechazar adecuadamente H_0 cuando ésta es falsa.
Prueba de dos colas Prueba de hipótesis en la que la hipótesis nula es rechazada debido a un valor del estadístico de prueba que se encuentra en cualquiera de las dos colas de la distribución muestral.
Prueba de una cola Prueba de hipótesis en la que la hipótesis nula es rechazada para valores del estadístico de prueba en una de las colas de la distribución de muestreo.
Valor crítico Valor que se compara con el estadístico de prueba para determinar si H_0 es rechazada.
Valor-*p* Probabilidad que proporciona una medida de la evidencia, dada por la muestra, contra la hipótesis nula. Entre menor sea un valor-*p*, mayor será la evidencia contra H_0. En una prueba de cola inferior, el valor-*p* es la probabilidad de obtener un valor para el estadístico de prueba tan pequeño o menor que el aportado por la muestra. En una prueba de cola superior, el valor-*p* es la probabilidad de obtener un valor para el estadístico de prueba tan grande o mayor que el proporcionado por la muestra. En una prueba de dos colas, el valor-*p* es la probabilidad de obtener un valor para el estadístico de prueba tan poco probable, o aun menos probable, que el aportado por la muestra.

Fórmulas clave

Estadístico de prueba en las pruebas de hipótesis para la media poblacional: σ conocida

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{9.1}$$

Estadístico de prueba en las pruebas de hipótesis para la media poblacional: σ desconocida

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{9.2}$$

Estadístico de prueba en las pruebas de hipótesis para la proporción poblacional

$$z = \frac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} \tag{9.4}$$

Tamaño de la muestra en una prueba de hipótesis de una cola para la media poblacional

$$n = \frac{(z_\alpha + z_\beta)^2\sigma^2}{(\mu_0 - \mu_a)^2} \tag{9.7}$$

En una prueba de dos colas se sustituye z_α por $z_{\alpha/2}$.

Ejercicios complementarios

60. En una línea de producción, el peso promedio con que se llena cada recipiente es 16 onzas. Un exceso o una insuficiente de llenado ocasionan problemas serios y, cuando son detectados, es necesario que el operador detenga la línea de producción para reajustar el mecanismo de llenado. Con base en datos anteriores, se supone que la desviación estándar poblacional es $\sigma = 0.8$ onzas. Cada hora, un inspector de control de calidad toma una muestra de 30 recipientes y decide si es necesario detener la producción y hacer un reajuste. El nivel de significancia es $\alpha = 0.05$.
    a) Establezca la prueba de hipótesis para esta aplicación al control de calidad.
    b) Si se encuentra que la media muestral es $\bar{x} = 16.32$ onzas, ¿cuál es el valor-*p*? ¿Qué medidas recomendaría usted tomar?
    c) Si se encuentra que la media muestral es $\bar{x} = 15.82$ onzas, ¿cuál es el valor-*p*? ¿Qué medidas sería preferible tomar?
    d) Use el método del valor crítico. ¿Cuál es la regla de rechazo en la prueba de hipótesis anterior? Repita los incisos *b*) y *c*). ¿Llega a la misma conclusión?

61. En Western University, la media histórica en las puntuaciones de los solicitantes de una beca es 900. La desviación estándar poblacional histórica que se considera conocida es $\sigma = 180$. Cada año, el decano asistente utiliza una muestra de las solicitudes para determinar si la puntuación media ha cambiado entre los solicitantes de becas.
    a) Establezca las hipótesis.
    b) ¿Cuál es el intervalo de 95% de confianza para la estimación de la media poblacional de las puntuaciones en el examen si en una muestra de 200 estudiantes la media muestral es $\bar{x} = 935$?
    c) Use el intervalo de confianza para realizar una prueba de hipótesis. Manejando $\alpha = 0.05$, ¿a qué conclusión llega?
    d) ¿Cuál es el valor-*p*?

62. *Playbill* es una revista que se distribuye entre las personas que asisten a obras musicales y otro tipo de producciones teatrales. El ingreso medio anual por familia de la población de lectores de *Playbill* es de $119 155 (*Playbill*, enero de 2006). Suponga que la desviación estándar es $\sigma = $20 700. Un grupo cívico de San Francisco asegura que entre las personas de la zona de la Bahía que van al teatro el ingreso medio es más alto. En una muestra de 60 personas de la Bahía que suelen acudir al teatro se encontró que el ingreso medio por hogar es de $126 100.
    a) Establezca las hipótesis que sean útiles para determinar si los datos muestrales apoyan la conclusión de que las personas de la zona de la Bahía que suelen asistir al teatro tienen un ingreso medio por familia más alto que los demás lectores de *Playbill.*
    b) ¿Cuál es el valor-*p* a partir de la muestra de las 60 personas de la Bahía que suelen acudir al teatro?
    c) Use $\alpha = 0.01$ como nivel de significancia. ¿A qué conclusión llega?

63. El viernes los corredores de bolsa de Wall Street esperaban ansiosos la publicación del gobierno federal sobre el aumento de nóminas no agrícolas en enero. El primer consenso estimado entre los economistas fue que se esperaba un aumento de 250 000 nuevos empleos (CNBC, 3 de febrero de 2006). Sin embargo, en una muestra de 20 economistas tomada el jueves en la tarde, la media muestral fue 266 000, con una desviación estándar muestral de 24 000. Los analistas financieros suelen llamar a tales medias muestrales, basadas en las últimas informaciones, "*whisper number*". Trate la "estimación del consenso" como la media poblacional. Realice una prueba de hipótesis para determinar si el "whisper number" justifica la conclusión de un aumento estadísticamente significativo en la estimación del consenso de los economistas. Use $\alpha = 0.01$ como nivel de significancia.

WEB archivo
FirstBirth

64. Datos emitidos por el National Center for Health Statistics muestran que la edad media a la que las mujeres tienen su primer hijo fue 25.0 en 2006 (*The Wall Street Journal*, 4 de febrero de 2009). La reportera Sue Shellenbarger señaló que, en varios años, éste fue el primer descenso en el indicador de referencia. Una muestra reciente de 42 mujeres proporcionó los datos del archivo FirstBirth en el sitio web acerca de la edad a la que tuvieron su primer hijo. ¿Estos datos indican un cambio con respecto a 2006 en la edad media a la que las mujeres tienen su primer hijo? Use $\alpha = 0.05$.

65. En un amplio estudio sobre los costos de atención a la salud en Estados Unidos se presentaron datos que mostraban un gasto medio de Medicare por derechohabiente de $6 883 en 2003 (*Money*, otoño de 2003). Para investigar las diferencias en todo el país, un investigador tomó una muestra de 40 derechohabientes en Indianápolis. En la muestra, el gasto medio de Medicare en 2003 fue de $5 980 y la desviación estándar de $2 518.
   a) Establezca las hipótesis a usar para determinar si el gasto anual medio de Medicare en Indianápolis es menor a la media nacional.
   b) Use los resultados muestrales anteriores para calcular el estadístico de prueba y el valor-*p*.
   c) Use $\alpha = 0.05$. ¿Cuál es su conclusión?
   d) Repita la prueba de hipótesis usando el método del valor crítico.

66. La cámara de comercio de una comunidad de la costa del Golfo en Florida anuncia en su publicidad que hay disponibilidad de propiedades en el área residencial a un costo medio de $125 000 o menos por lote. Suponga que en una muestra de 32 propiedades se encuentra una media muestral de $130 000 por terreno y una desviación estándar muestral es $12 500. Use 0.05 como nivel de significancia para probar la validez de lo que se dice en la publicidad.

WEB archivo
Gasoline

67. La U.S. Energy Administration informó que en Estados Unidos el precio medio del galón de gasolina era de $2.357 (U.S. Energy Administration, 30 de enero de 2006). En el archivo de datos llamado *Gasoline* se encuentran los precios de gasolina normal encontrados en una muestra de 50 estaciones de servicio en estados del Atlántico sur. Realice una prueba de hipótesis para determinar si el precio medio del galón de combustible en los estados del Atlántico sur es diferente a la media nacional. Use $\alpha = 0.05$ como nivel de significancia y formule su conclusión.

68. En un estudio del Center for Disease Control (CDC) se encontró que 23% de los adultos son fumadores y de éstos, 70% indicó que quiere dejar este hábito (Associated Press, 26 de julio de 2002). El CDC informó que, de las personas que fumaron en algún momento de su vida, 50% habían podido dejar de hacerlo. Parte del estudio indicó que el éxito en prescindir del cigarro aumenta con el nivel de estudios. Suponga que en una muestra de 100 personas con título universitario que han fumado en algún momento de su vida, 64 lograron renunciar a su hábito.
   a) Especifique las hipótesis a usar para determinar si la población de personas con título universitario tiene más éxito para dejar de fumar que la población general.
   b) Dados los datos muestrales, ¿cuál es la proporción de personas con título universitario que, habiendo fumado en algún momento de su vida, pudieran dejar de hacerlo?
   c) ¿Cuál es el valor-*p*? Con $\alpha = 0.01$, ¿cuál es la conclusión de la prueba de hipótesis?

69. La promoción de una línea aérea se sustenta en el supuesto de que dos terceras partes de los viajeros de negocios usan una computadora portátil en sus viajes durante la noche.
   a) Establezca las hipótesis a usar para probar este supuesto.
   b) ¿Cuál es la proporción muestral encontrada en un estudio patrocinado por American Express, en que 355 de 546 viajeros de negocios utilizaban su computadora portátil en sus viajes de negocios durante la noche?
   c) ¿Cuál es el valor-*p*?
   d) Use $\alpha = 0.05$. ¿A qué conclusión llega?

70. Los call centers (centros de atención telefónica) virtuales son atendidos por personas que trabajan desde su hogar. La mayoría de los agentes en casa gana de $10 a $15 por hora sin beneficios de seguridad social frente a $7 a $9 por hora con beneficios en un call center tradicional (*BusinessWeek*, 23 de enero de 2006). Regional Airways considera emplear agentes en casa, pero sólo si conservan una satisfacción del cliente mayor de 80%. Se realizó una prueba con agentes de este tipo. En una muestra de 300 clientes, 252 indicaron estar satisfechos con el servicio.
   a) Elabore las hipótesis de prueba para determinar si los datos muestrales apoyan la conclusión de que el servicio al cliente con agentes en casa satisface el criterio de Regional Airways.
   b) ¿Cuál es la estimación puntual del porcentaje de clientes satisfechos?
   c) ¿Cuál es el valor-*p* proporcionado por los datos muestrales?
   d) ¿Cuál es la conclusión en esta prueba de hipótesis? Use como nivel de significancia $\alpha = 0.05$.

71. Durante el año electoral 2004 se publicaban a diario los resultados de los nuevos sondeos. En una consulta de IBD/TIPP a 910 adultos, 503 encuestados dijeron sentirse optimistas ante las

perspectivas nacionales y el índice de liderazgo del presidente Bush aumentó 4.7 puntos, a 55.3 puntos (*Investor's Business Daily*, 14 de enero de 2004).

a) ¿Cuál es la proporción muestral de encuestados optimistas ante las perspectivas nacionales?

b) Un director de campaña quiere afirmar que el sondeo indica que la mayoría de los adultos se sienten optimistas ante las perspectivas nacionales. Elabore una prueba de hipótesis de manera que el rechazo de la hipótesis nula permita concluir que la proporción de optimistas es mayor de 50%.

c) Use los datos del sondeo para calcular el valor-*p* en la prueba de hipótesis del inciso *b*). Explique al director lo que dice este valor-*p* acerca del nivel de significancia de los resultados.

72. Una estación de radio de Myrtle Beach anuncia que, por lo menos, 90% de los hoteles y moteles estarán llenos el fin de semana en que se conmemora el Día de los Caídos. La radiodifusora aconseja a sus oyentes hacer sus reservaciones con anticipación si piensan pasar ese fin de semana en esa localidad vacacional. La noche del sábado, una muestra de 58 hoteles y moteles, indicó que 49 estaban completamente llenos y 9 aún tenían habitaciones libres. ¿Cuál es su reacción ante lo anunciado por la estación de radio después de ver la evidencia muestral? Use $\alpha = 0.05$ al realizar el estadístico de prueba. ¿Cuál es el valor-*p*?

73. En Estados Unidos, según el gobierno federal, 24% de los trabajadores amparados por el plan de atención a la salud no tuvieron que contribuir a la prima en su empresa (*Statistical Abstract of the United States*: *2006*). En un estudio reciente se encontró que a 81 de los 400 trabajadores muestreados no se les pidió que contribuyeran para el plan de atención a la salud en su empresa.

a) Elabore las hipótesis para probar si ha disminuido el porcentaje de trabajadores a quienes no se les pide que contribuyan con su empresa para el plan de atención a la salud.

b) ¿Cuál es la estimación puntual de la proporción que tiene un seguro de salud financiado totalmente por su empresa?

c) ¿Se ha registrado una disminución estadísticamente significativa en la proporción de trabajadores que tienen un seguro de salud financiado totalmente por su empresa? Use $\alpha = 0.05$.

74. Shorney Construction Company licita proyectos suponiendo que la media del tiempo desperdiciado por trabajador es de 72 minutos o menos por día. Para probar este supuesto se usa una muestra de 30 trabajadores de la construcción. Suponga que la desviación estándar poblacional es 20 minutos.

a) Establezca las hipótesis para esta prueba.

b) ¿Cuál es la probabilidad de cometer un error tipo II si la media poblacional del tiempo desperdiciado fueran 80 minutos?

c) ¿Cuál es la probabilidad de cometer este error si la media poblacional del tiempo desperdiciado fueran 75 minutos?

d) ¿Cuál es la probabilidad de cometerlo si la media poblacional del tiempo desperdiciado fueran 70 minutos?

e) Trace la curva de potencia para este problema.

75. Existe un programa de ayuda federal para las zonas de bajos ingresos. Para recibirla, el ingreso medio de la localidad debe ser menor de $15 000 anuales. Aquellas con ingreso medio anual de $15 000 o más no son elegibles. La decisión de asignación de la ayuda se basa en una muestra de los habitantes de la zona. Se realiza una prueba de hipótesis con 0.02 como nivel de significancia. Si los lineamientos establecen una probabilidad máxima de 0.05 de no otorgar esta ayuda a una zona en la que el ingreso medio anual sea de $14 000, ¿qué tamaño de muestra deberá utilizarse en el estudio? Use $\sigma = \$4\,000$ para desarrollar su plan.

76. Para probar si en el proceso de fabricación de un jabón de baño se satisface el estándar de producir 120 barras por lote se usan las hipótesis H_0: $\mu = 120$ y H_a: $\mu \neq 120$. Use 0.05 como nivel de significancia en esta prueba y 5 para la desviación estándar.

a) Si la media de producción disminuye a 117 barras por lote, la empresa desea tener 98% de oportunidad de concluir que no se está satisfaciendo el estándar de producción. ¿De qué tamaño deberá tomarse la muestra?

b) Con el tamaño de muestra del inciso *a*), ¿cuál es la probabilidad de concluir que el proceso opera satisfactoriamente para cada una de las siguientes medias de producción verdaderas: 117, 118, 119, 121, 122 y 123 barras por lote? Es decir, ¿cuál es, en cada caso, la probabilidad de cometer un error tipo II?

Caso a resolver 1

Quality Associates, Inc.

Quality Associates, Inc., una firma de consultoría, asesora a sus clientes sobre procedimientos estadísticos y de muestreo para el control de sus procesos de manufactura. En una determinada asesoría, el cliente dio a Quality Associates una muestra de 800 observaciones tomadas mientras el proceso operaba satisfactoriamente. La desviación estándar de estos datos fue 0.21; al ser tantos los datos, se consideró que la desviación estándar poblacional era 0.21. Quality Associates recomendó que, para monitorear el proceso, se tomaran de manera periódica muestras aleatorias de tamaño 30. Al analizar las nuevas muestras, el cliente sabrá pronto si el proceso operaba de manera satisfactoria. Si el proceso no operaba de manera adecuada, se podían emprender las acciones correctivas para eliminar el problema. Las especificaciones de diseño indicaban que la media en el proceso debería ser 12. A continuación, la prueba de hipótesis sugerida por Quality Associates.

$$H_0: \mu = 12$$
$$H_a: \mu \neq 12$$

Siempre que H_0 fuera rechazada, deberían emprenderse las acciones correctivas.

Durante el primer día en que se realizó este nuevo procedimiento de control estadístico de procesos se tomaron las siguientes muestras (Sample) a intervalos de una hora. Esta información se encuentra en el conjunto de datos Quality.

WEB archivo

Quality

Sample 1	Sample 2	Sample 3	Sample 4
11.55	11.62	11.91	12.02
11.62	11.69	11.36	12.02
11.52	11.59	11.75	12.05
11.75	11.82	11.95	12.18
11.90	11.97	12.14	12.11
11.64	11.71	11.72	12.07
11.80	11.87	11.61	12.05
12.03	12.10	11.85	11.64
11.94	12.01	12.16	12.39
11.92	11.99	11.91	11.65
12.13	12.20	12.12	12.11
12.09	12.16	11.61	11.90
11.93	12.00	12.21	12.22
12.21	12.28	11.56	11.88
12.32	12.39	11.95	12.03
11.93	12.00	12.01	12.35
11.85	11.92	12.06	12.09
11.76	11.83	11.76	11.77
12.16	12.23	11.82	12.20
11.77	11.84	12.12	11.79
12.00	12.07	11.60	12.30
12.04	12.11	11.95	12.27
11.98	12.05	11.96	12.29
12.30	12.37	12.22	12.47
12.18	12.25	11.75	12.03
11.97	12.04	11.96	12.17
12.17	12.24	11.95	11.94
11.85	11.92	11.89	11.97
12.30	12.37	11.88	12.23
12.15	12.22	11.93	12.25

Informe gerencial

1. Con cada una de las muestras realice una prueba de hipótesis usando 0.01 como nivel de significancia. Determine las acciones a emprender si resulta necesario. Proporcione el estadístico de prueba y el valor-p de cada prueba.
2. Calcule la desviación estándar de cada una de las cuatro muestras. ¿Parece razonable el supuesto de 0.21 para la desviación estándar poblacional?
3. Calcule límites de alrededor de $\mu = 12$ para la media muestral $\bar{x}$ de manera que, en tanto las medias muestrales se encuentren dentro de estos límites, pueda considerarse que el proceso opera de manera satisfactoria. Pero si $\bar{x}$ excede el límite superior, o es menor al límite inferior, será necesario emprender las acciones correctivas. Estos límites se conocen en el control de calidad como *límites de control superior e inferior*.
4. Analice las consecuencias de modificar el nivel de significancia por un valor mayor. ¿Qué falla o error crece si se aumenta el valor del nivel de significancia?

Caso a resolver 2

Comportamiento ético de los estudiantes de negocios en la Universidad de Bayview

Durante la recesión global de 2008 y 2009 se fincaron muchos cargos por comportamiento no ético de ejecutivos de Wall Street, directores de finanzas y otros funcionarios corporativos. Al mismo tiempo se publicó un artículo que sugería que parte del motivo de ese comportamiento no ético podía provenir del hecho de que el engaño se ha vuelto más frecuente entre los estudiantes de negocios (*Chronicle of Higher Education*, 10 de febrero de 2009). El artículo menciona que 56% de los estudiantes de negocios admitió haber hecho trampa alguna vez durante su carrera académica en comparación con 47% de estudiantes de otras carreras.

El engaño se ha vuelto una preocupación para el decano de la Facultad de Negocios de la Universidad de Bayview por varios años. Algunos académicos de la facultad consideran que es más generalizado en Bayview que en otras universidades, en tanto que otros piensan que el embuste no es uno de los principales problemas de la institución. Para resolver algo de estos temas, el decano encargó un estudio que permitiera evaluar el comportamiento ético actual de los estudiantes de negocios en Bayview. Como parte del estudio, se aplicó una encuesta de salida anónima a una muestra de 90 estudiantes de negocios de la clase de graduados de este año. Para obtener datos de tres tipos de engaños, se utilizaron las respuestas a las siguientes preguntas.

> Durante su estancia en Bayview, ¿alguna vez presentó algún trabajó copiado de Internet como propio?
>
> Sí _____ No _____
>
> Durante su estancia en Bayview, ¿alguna vez copió las respuestas del examen de otro estudiante?
>
> Sí _____ No _____
>
> Durante su estancia en Bayview, ¿alguna vez colaboró con otros estudiantes para realizar proyectos que se suponía que debían ser efectuados de manera individual?
>
> Sí _____ No _____

Cualquier estudiante que contestó Sí a una o más de esas preguntas se considera que estuvo involucrado en algún tipo de engaño. Parte de los datos recolectados se presentan a continuación con las siguientes etiquetas: Student (estudiante); Copied from Internet (copió de Internet); Copied on Exam (copió de otro examen); Collaborated on Individual Project (colaboró en proyectos individuales) y Gender (género). El conjunto completo está en el archivo llamado *Bayview*.

WEB archivo
Bayview

Student	Copied from Internet	Copied on Exam	Collaborated on Individual Project	Gender
1	No	No	No	Female
2	No	No	No	Male
3	Yes	No	Yes	Male
4	Yes	Yes	No	Male
5	No	No	Yes	Male
6	Yes	No	No	Female
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
88	No	No	No	Male
89	No	Yes	Yes	Male
90	No	No	No	Female

Informe gerencial

Prepare un informe para el decano de la universidad que resuma su evaluación de la naturaleza del engaño entre los estudiantes de negocios de la Universidad de Bayview. Asegúrese de incluir los siguientes puntos en su reporte.

1. Use la estadística descriptiva para resumir los datos y comente sus hallazgos.
2. Elabore un intervalo de 95% de confianza para la proporción de todos los estudiantes, la proporción de los estudiantes hombres y la proporción de las estudiantes mujeres que estuvieron involucrados en algún tipo de trampa.
3. Realice una prueba de hipótesis para determinar si la proporción de estudiantes de negocios en Bayview que estuvieron involucrados en algún tipo de engaño es menor que la de alumnos en la misma disciplina de otras instituciones, como lo reportó el *Chronicle of Higher Education*.
4. Realice una prueba de hipótesis para determinar si la proporción de estudiantes de negocios en Bayview que estuvieron involucrados en algún tipo de engaño es menor que la de alumnos de otras carreras en otras instituciones, como lo reportó el *Chronicle of Higher Education*.
5. Con base en su análisis de los datos, ¿qué consejo podría dar al decano?

Apéndice 9.1 Pruebas de hipótesis con Minitab

Se describe el uso de Minitab para realizar pruebas de hipótesis para la media y la proporción poblacionales.

Media poblacional: σ conocida

Se ilustra con el ejemplo presentado en la sección 9.3 acerca de la distancia recorrida por las pelotas de golf de MaxFlight. Los datos están en la columna Cl de la hoja de cálculo de Minitab. Se asume que se conoce la desviación estándar poblacional como $\sigma = 12$ y el nivel de significancia es $\alpha = 0.05$. Se pueden utilizar los siguientes pasos para probar la hipótesis H_0: $\mu = 295$ frente a H_a: $\mu \neq 295$.

WEB archivo
GolfTest

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Basic Statistics**.
Paso 3. Elija **1-Sample Z**.

Paso 4. Cuando el cuadro de diálogo 1-Sample Z aparezca:
Ingrese Cl en el cuadro **Samples in columns**.
Ingrese 12 en el cuadro **Standard deviation**.
Seleccione **Perform Hypothesis Test**.
Ingrese 295 en el cuadro **Hypothesized mean**.
Seleccione **Options**.

Paso 5. Cuando el cuadro de diálogo 1-Sample Z-Options aparezca:
Ingrese 95 en el cuadro **Confidence level***.
Seleccione **not equal** en el cuadro **Alternative**.
Haga clic en **OK**.

Paso 6. Dé clic en **OK**.

Además de los resultados de la prueba de hipótesis, Minitab proporciona un intervalo de 95% de confianza para la media poblacional.

Este procedimiento se modifica con facilidad para una prueba de hipótesis de una cola seleccionando la opción menor que (less than), o mayor que (greater than) en el cuadro **Alternative** del paso 5.

Media poblacional: σ desconocida

WEB archivo
AirRating

Las puntuaciones proporcionadas por 60 viajeros de negocios al aeropuerto de Heathrow se han ingresado en la columna Cl de la hoja de cálculo de Minitab. El nivel de significancia para esta prueba es $\alpha = 0.05$, y la desviación estándar poblacional σ se estimará mediante la desviación estándar muestral s. Los pasos siguientes se usan para probar la hipótesis H_0: $\mu \leq 7$ frente a H_a: $\mu > 7$.

Paso 1. Seleccione el menú **Stat**.

Paso 2. Elija **Basic Statistics**.

Paso 3. Elija **1-Sample t**.

Paso 4. Cuando el cuadro de diálogo 1-Sample t aparezca:
Ingrese C1 en el cuadro **Samples in columns**.
Seleccione **Perform Hypothesis Test**.
Ingrese 7 en el cuadro **Hypothesized mean**.
Seleccione **Options**.

Paso 5. Cuando el cuadro de diálogo 1-Sample t-options aparezca:
Ingrese 95 en el cuadro **Confidence level**.
Seleccione **greater than** en el cuadro **Alternative**.
Haga clic en **OK**.

Paso 6. Haga clic en **OK**.

En el estudio de las puntuaciones para el aeropuerto de Heathrow se tiene una hipótesis alternativa del estilo mayor que. Los pasos anteriores se modifican con facilidad para otras pruebas de hipótesis al seleccionar las opciones menor que (less than) o no igual (not equal) en el cuadro **Alternative** del paso 5.

Proporción poblacional

WEB archivo
WomenGolf

Se ilustra con el ejemplo del campo de golf Pine Creek presentado en la sección 9.5. Los datos con las respuestas mujer (Female) y hombre (Male) están en la columna Cl de la hoja de cálculo de Minitab. Este software usa un orden alfabético de las respuestas y selecciona la *segunda respuesta* para la proporción poblacional de interés. En este caso Minitab utiliza el orden alfabético Female-Male (mujer-hombre) y entrega la proporción poblacional de las respuestas Male (hombre). Como Female (mujer) es la respuesta de interés, hay que modificar el orden

* Minitab proporciona simultáneamente los resultados de la prueba de hipótesis y la estimación por intervalo. El usuario debe seleccionar cualquier nivel de confianza para la estimación por intervalo de la media poblacional: aquí se sugiere 95% de confianza.

alfabético como sigue. Seleccione cualquier celda de la columna y use la secuencia Editor > Column > Value Order. Después elija la opción de ingresar un orden especificado por el usuario. Ingrese Male-Female en el cuadro **Define-an-order** y dé clic en OK. La rutina 1 Proportion de Minitab suministrará los resultados de la prueba de hipótesis para la proporción poblacional de golfistas. Proceda como sigue:

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Basic Statistics**.
Paso 3. Elija **1 Proportion**.
Paso 4. Cuando el cuadro de diálogo 1 Proportion aparezca:
Ingrese C1 en el cuadro **Samples in Columns**.
Seleccione **Perform Hypothesis Test**.
Ingrese 0.20 en el cuadro **Hypothesized proportion**.
Seleccione **Options**.
Paso 5. Cuando el cuadro de diálogo 1 Proportion-Options aparezca:
Ingrese 95 en el cuadro **Confidence level**.
Seleccione **greater than** en el cuadro **Alternative**.
Elija **Use test and interval based on normal distribution**.
Haga clic en **OK**.
Paso 6. Haga clic en **OK**.

Apéndice 9.2 Pruebas de hipótesis con Excel

Excel no cuenta con rutinas predefinidas para las pruebas de hipótesis presentadas en este capítulo. Para resolver esas situaciones se presentan hojas de cálculo de Excel, diseñadas por los autores de este libro, para usarse como plantillas en pruebas de hipótesis acerca de una media poblacional y una proporción poblacional. Utilizar estas hojas de cálculo es sencillo y también pueden modificarse para cualesquiera datos muestrales. Las hojas están disponibles en el sitio web del libro.

Media poblacional: σ conocida

Se ilustra con el ejemplo presentado en la sección 9.3 de la distancia de las pelotas de golf de MaxFlight. Los datos están en la columna A de la hoja de cálculo de Excel. Se asume que se conoce la desviación estándar poblacional como $\sigma = 12$, y que el nivel de significancia es $\alpha = 0.05$. Los pasos siguientes se usan para probar la hipótesis H_0: $\mu = 295$ frente a la hipótesis H_a: $\mu \neq 295$.

A medida que se describe este procedimiento, consulte la figura 9.11. En la hoja de cálculo que aparece en segundo plano se presentan las celdas con las fórmulas usadas para calcular los resultados que figuran en la hoja de cálculo en primer plano. Los datos se han introducido en las celdas A2:A51. Para usar la plantilla con este conjunto de datos son necesarios los pasos siguientes.

WEB archivo
Hyp Sigma Known

Paso 1. Ingrese el rango de datos A2:A51 en la fórmula =COUNT de la celda D4 para obtener el tamaño de la muestra.
Paso 2. Ingrese el rango de datos A2:A51 en la fórmula =AVERAGE de la celda D5 para obtener la media muestral.
Paso 3. Incorpore la desviación estándar poblacional $\sigma = 12$ en la celda D6.
Paso 4. Ingrese el valor hipotético de la media poblacional 295 en la celda D8.

Las fórmulas de las celdas restantes proporcionarán en automático el error estándar en la celda D10; el valor del estadístico de prueba z en la D11, y tres valores-p. Como la hipótesis alternativa ($\mu \neq 295$) indica que se trata de una prueba de dos colas, para tomar la decisión de rechazar o no, se usa el valor-p (Two Tail) de la celda D15. Como el valor-$p = 0.1255 > \alpha = 0.05$, la hipótesis nula no puede ser rechazada. Los valores-p de las celdas D13 o D14 se usarían si se tratara de una prueba de hipótesis de una sola cola, inferior o superior, respectivamente.

Esta plantilla se utiliza para los cálculos de pruebas de hipótesis de otras aplicaciones. Por ejemplo, para realizar una prueba de hipótesis con otro conjunto de datos, ingresélos en

FIGURA 9.11 Hoja de cálculo de Excel para pruebas de hipótesis sobre la media poblacional con σ conocida

	A	B	C	D	E
1	Yards		Hypothesis Test About a Population Mean		
2	303		With σ Known		
3	282				
4	289		Sample Size	=COUNT(A2:A51)	
5	298		Sample Mean	=AVERAGE(A2:A51)	
6	283		Population Std. Deviation	12	
7	317				
8	297		Hypothesized Value	295	
9	308				
10	317		Standard Error	=D6/SQRT(D4)	
11	293		Test Statistic z	=(D5-D8)/D10	
12	284				
13	290		p-value (Lower Tail)	=NORMSDIST(D11)	
14	304		p-value (Upper Tail)	=1-D13	
15	290		p-value (Two Tail)	=2*MIN(D13,D14)	
16	311				
17	305				
49	303				
50	301				
51	292				
52					

	A	B	C	D	E
1	Yards		Hypothesis Test About a Population Mean		
2	303		With σ Known		
3	282				
4	289		Sample Size	50	
5	298		Sample Mean	297.6	
6	283		Population Std. Deviation	12	
7	317				
8	297		Hypothesized Value	295	
9	308				
10	317		Standard Error	1.70	
11	293		Test Statistic z	1.53	
12	284				
13	290		p-value (Lower Tail)	0.9372	
14	304		p-value (Upper Tail)	0.0628	
15	290		p-value (Two Tail)	0.1255	
16	311				
17	305				
49	303				
50	301				
51	292				
52					

Nota. Las filas 18 a 48 están ocultas.

la columna A de la hoja de cálculo. Modifique las fórmulas de las celdas D4 y D5 para que correspondan al nuevo rango de datos. Para obtener los resultados, ingrese la desviación estándar poblacional en la celda D6, y en la celda D8 ingrese el valor hipotético de la media poblacional. Si los nuevos datos muestrales ya han sido resumidos, no es necesario ingresarlos en la hoja de cálculo. En este caso, para obtener los resultados se ingresa el tamaño de la muestra en la celda D4, la media muestral en la celda D5, la desviación estándar poblacional en la celda D6 y el valor hipotético de la media poblacional en la celda D8. La hoja de cálculo que se presenta en la figura 9.11 está disponible en el archivo Hyp Sigma Known en el sitio web del libro.

Media poblacional: σ desconocida

WEB archivo
Hyp Sigma Unknown

Se ilustra con el ejemplo presentado en la sección 9.4 de las puntuaciones sobre el aeropuerto de Heathrow. Los datos están en la columna A de la hoja de cálculo de Excel. La desviación estándar poblacional σ no se conoce y se estimará a partir de la desviación estándar muestral s. El nivel de significancia es $\alpha = 0.05$. Los pasos siguientes se utilizan para probar la hipótesis H_0: $\mu \leq 7$ frente a la hipótesis H_a: $\mu > 7$.

Consulte la figura 9.12 a medida que se describe este procedimiento. La hoja de cálculo que aparece en segundo plano indica las fórmulas usadas para obtener los resultados en la

FIGURA 9.12 Hoja de cálculo de Excel para pruebas de hipótesis sobre una media poblacional con σ desconocida

	A	B	C	D	E
1	Rating		Hypothesis Test About a Population Mean		
2	5		With σ Unknown		
3	7				
4	8		Sample Size	=COUNT(A2:A61)	
5	7		Sample Mean	=AVERAGE(A2:A61)	
6	8		Sample Std. Deviation	=STDEV(A2:A61)	
7	8				
8	8		Hypothesized Value	7	
9	7				
10	8		Standard Error	=D6/SQRT(D4)	
11	10		Test Statistic *t*	=(D5-D8)/D10	
12	6		Degrees of Freedom	=D4-1	
13	7				
14	8		*p*-value (Lower Tail)	=IF(D11<0,TDIST(-D11,D12,1),1-TDIST(D11,D12,1))	
15	8		*p*-value (Upper Tail)	=1-D14	
16	9		*p*-value (Two Tail)	=2*MIN(D14,D15)	
17	7				
59	7				
60	7				
61	8				
62					

	A	B	C	D	E
1	Rating		Hypothesis Test About a Population Mean		
2	5		With σ Unknown		
3	7				
4	8		Sample Size	60	
5	7		Sample Mean	7.25	
6	8		Sample Std. Deviation	1.05	
7	8				
8	8		Hypothesized Value	7	
9	7				
10	8		Standard Error	0.136	
11	10		Test Statistic *t*	1.841	
12	6		Degrees of Freedom	59	
13	7				
14	8		*p*-value (Lower Tail)	0.9647	
15	8		*p*-value (Upper Tail)	0.0353	
16	9		*p*-value (Two Tail)	0.0706	
17	7				
59	7				
60	7				
61	8				
62					

Nota. Las filas 18 a 58 están ocultas.

versión de la hoja de cálculo presentada en primer plano. Los datos se ingresan en las celdas A2:A61. Para usar la plantilla con estos datos son necesarios los pasos siguientes.

Paso 1. Ingrese el rango de datos A2:A61 en la fórmula =COUNT de la celda D4 para obtener el tamaño de la muestra.

Paso 2. Introduzca el rango de datos A2:A61 en la fórmula =AVERAGE de la celda D5 para obtener la media muestral.

Paso 3. Ingrese el rango de datos A2:A61 en la fórmula =STDEV de la celda D6 para obtener la desviación estándar muestral.

Paso 4. Ingrese el valor hipotético 7 de la media poblacional en la celda D8.

Las fórmulas de las celdas restantes proporcionarán automáticamente el error estándar en la celda D8, el valor del estadístico de prueba t en la D11, el número de grados de libertad en la D12, y tres valores-p. Como la hipótesis alternativa ($\mu > 7$) indica que se trata de una prueba de cola superior, para tomar la decisión de rechazar o no, se usa el valor-p (Upper Tail) de la celda D15. Como el valor-$p = 0.0353 < \alpha = 0.05$, la hipótesis nula es rechazada. Los valores-p de las celdas D14 y D16 se usarían si se tratara de una prueba de hipótesis de cola inferior o de dos colas.

Esta plantilla se utiliza para los cálculos de pruebas de hipótesis de otras aplicaciones. Por ejemplo, para realizar una prueba de hipótesis con un nuevo conjunto de datos, éstos se ingresan en la columna A de la hoja de cálculo y se modifican las fórmulas de las celdas D4, D5 y D6 para que correspondan al nuevo rango de datos. Para obtener los resultados, se ingresa en la celda D8 el valor hipotético de la media poblacional. Si los datos muestrales ya han sido resumidos, no es necesario incorporarlos en la hoja de cálculo. En este caso, para obtener los resultados se ingresa el tamaño de la muestra en la celda D4, la media muestral en la celda D5, la desviación estándar muestral en D6 y el valor hipotético de la media poblacional en la celda D8. La hoja de cálculo que se presenta en la figura 9.12 se encuentra con el nombre Hyp Sigma Unknown en el sitio web del libro.

Proporción poblacional

WEB archivo
Hypothesis p

Se ilustra con el ejemplo del campo de golf Pine Creek presentado en la sección 9.5. Los datos con las respuestas golfista Mujer (Female) y Hombre (Male) están en la columna A de la hoja de cálculo de Excel. Consulte la figura 9.13 a medida que se describe este procedimiento. La hoja de cálculo que aparece en segundo plano indica las fórmulas usadas para obtener los resultados que figuran en la hoja que está en primer plano. Los datos están en las celdas A2:A401. Los pasos siguientes se usan para probar la hipótesis H_0: $p \leq 0.20$ frente a H_a: $p > 0.20$.

Paso 1. Ingrese el rango de datos A2:A401 en la fórmula =COUNT de la celda D3 para obtener el tamaño de la muestra.

Paso 2. Ingrese Female como respuesta de interés en la celda D4.

Paso 3. Incorpore el rango de datos A2:A401 en la fórmula =COUNTIF de la celda D5 para contar el número de respuestas de interés.

Paso 4. Ingrese el valor hipotético 0.20 de la proporción poblacional en la celda D8.

Las fórmulas de las celdas restantes proporcionarán automáticamente el error estándar en la celda D10, el valor del estadístico de prueba z en la D11, y tres valores-p. Como la hipótesis alternativa ($p > 0.20$) indica que se trata de una prueba de cola superior, para tomar la decisión de rechazar o no se usa el valor-p (Upper Tail) de la celda D14. Como el valor-$p = 0.0062 < \alpha = 0.05$, la hipótesis nula es rechazada. Los valores-p de las celdas D13 o D15 se usarían si se tratara de una prueba de hipótesis de cola inferior o de dos colas, respectivamente.

Esta planilla se puede utilizar para los cálculos de pruebas de hipótesis con otras aplicaciones. Por ejemplo, para realizar una prueba de hipótesis con otro conjunto de datos, ingrésélos en la columna A de la hoja de cálculo. Se modifican las fórmulas de las celdas D3 y D5 para que correspondan al nuevo rango de datos. Para obtener los resultados, se ingresa en la celda D4 la respuesta de interés y en la D8 el valor hipotético de la proporción poblacional. Si los nuevos datos muestrales ya han sido resumidos, no es necesario ingresarlos en la hoja de cálculo. En este caso, para obtener los resultados se ingresa el tamaño de la muestra en la celda D3, la proporción muestral en D6 y el valor hipotético de la proporción poblacional en la D8. La hoja de cálculo que se presenta en la figura 9.13 se encuentra bajo el nombre de Hypothesis p en el sitio web del libro.

FIGURA 9.13 Hoja de cálculo de Excel para pruebas de hipótesis de una proporción poblacional

	A	B	C	D	E
1	**Golfer**		**Hypothesis Test About a Population Proportion**		
2	Female				
3	Male		**Sample Size**	=COUNTA(A2:A401)	
4	Female		**Response of Interest**	Female	
5	Male		**Count for Response**	=COUNTIF(A2:A401,D4)	
6	Male		**Sample Proportion**	=D5/D3	
7	Female				
8	Male		**Hypothesized Value**	0.20	
9	Male				
10	Female		**Standard Error**	=SQRT(D8*(1-D8)/D3)	
11	Male		**Test Statistic *z***	=(D6-D8)/D10	
12	Male				
13	Male		***p*-value (Lower Tail)**	=NORMSDIST(D11)	
14	Male		***p*-value (Upper Tail)**	=1-D13	
15	Male		***p*-value (Two Tail)**	=2*MIN(D13,D14)	
16	Female				
400	Male				
401	Male				
402					

	A	B	C	D	E
1	**Golfer**		**Hypothesis Test About a Population Proportion**		
2	Female				
3	Male		**Sample Size**	400	
4	Female		**Response of Interest**	Female	
5	Male		**Count for Response**	100	
6	Male		**Sample Proportion**	0.2500	
7	Female				
8	Male		**Hypothesized Value**	0.20	
9	Male				
10	Female		**Standard Error**	0.0200	
11	Male		**Test Statistic *z***	2.50	
12	Male				
13	Male		***p*-value (Lower Tail)**	0.9938	
14	Male		***p*-value (Upper Tail)**	0.0062	
15	Male		***p*-value (Two Tail)**	0.0124	
16	Female				
400	Male				
401	Male				
402					

Nota. Las filas 17 a 399 están ocultas.

Apéndice 9.3 Pruebas de hipótesis con StatTools

En este apéndice se muestra el uso de StatTools para realizar pruebas de hipótesis sobre la media poblacional en el caso en que se desconoce σ.

Media poblacional: σ desconocida

WEB archivo
AirRating

En este caso, la desviación estándar poblacional σ será estimada a partir de la desviación estándar muestral s. Se usará el ejemplo presentado en la sección 9.4 de las puntuaciones proporcionadas por 60 viajeros de negocios al aeropuerto de Heathrow.

Se inicia con Data Set Manager para crear un conjunto de datos de StatTools utilizando el procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes se utilizan para probar la hipótesis de H_0: $\mu \leq 7$ contra H_a: $\mu > 7$.

Paso 1. Haga clic en la ficha **StatTools** en la cinta.
Paso 2. En el grupo **Analyses**, haga clic en **Statistical Inference**.
Paso 3. Seleccione la opción **Hypothesis Test**.
Paso 4. Elija Mean/Std. Deviation.
Paso 5. Cuando el cuadro de diálogo StatTools-Hypothesis Test for Mean/Std. Deviation aparezca:
Para **Analysis Type**, elija **One-Sample Analysis**.
En la sección **Variables,** seleccione **Rating**.
En la sección **Hypothesis Tests to Perform:**
Elija la opción **Mean**.
Ingrese 7 en el cuadro **Null Hypothesis Value**.
Seleccione **Greater Than Null Value (One-Tailed Test)** en el cuadro **Alternative Hypothesis**.
Si está seleccionado, suprima la marca en el cuadro **Standard Deviation**.
Haga clic en **OK**.

Aparecerán los resultados de la prueba de hipótesis. Ahí se incluyen el valor-*p* y el valor del estadístico de prueba.

CAPÍTULO 10

Inferencia estadística acerca de medias y proporciones con dos poblaciones

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: U.S. FOOD AND DRUG ADMINISTRATION

10.1 INFERENCIAS ACERCA DE LA DIFERENCIA ENTRE DOS MEDIAS POBLACIONALES: σ_1 Y σ_2 CONOCIDAS
- Estimación por intervalo para $\mu_1 - \mu_2$
- Pruebas de hipótesis acerca de $\mu_1 - \mu_2$
- Consejo práctico

10.2 INFERENCIAS ACERCA DE LA DIFERENCIA ENTRE DOS MEDIAS POBLACIONALES: σ_1 Y σ_2 DESCONOCIDAS
- Estimación por intervalo para $\mu_1 - \mu_2$
- Pruebas de hipótesis acerca de $\mu_1 - \mu_2$
- Consejo práctico

10.3 INFERENCIAS ACERCA DE LA DIFERENCIA ENTRE DOS MEDIAS POBLACIONALES: MUESTRAS PAREADAS

10.4 INFERENCIAS ACERCA DE LA DIFERENCIA ENTRE DOS PROPORCIONES POBLACIONALES
- Estimación por intervalo para $p_1 - p_2$
- Prueba de hipótesis acerca de $p_1 - p_2$

ESTADÍSTICA *en* LA PRÁCTICA

U.S. FOOD AND DRUG ADMINISTRATION
WASHINGTON, D.C.

La U.S. Food and Drug Administration (Oficina de Alimentos y Medicamentos de Estados Unidos, FDA), a través del Centro de Evaluación e Investigación sobre Medicamentos (CDER, por sus siglas en inglés), garantiza que estos últimos sean confiables y efectivos. Pero el CDER no se encarga de realizar las pruebas necesarias a los productos nuevos. La empresa interesada en comercializarlos es la responsable de presentar las evidencias de que son confiables y efectivos. Después, en el CDER, científicos y especialistas en estadística revisan las evidencias presentadas.

Las empresas interesadas en que se apruebe un nuevo medicamento realizan extensos estudios estadísticos para apoyar su solicitud. Las pruebas que se efectúan en la industria farmacéutica suelen comprender tres etapas: 1) pruebas preclínicas; 2) pruebas de uso prolongado y confiabilidad, y 3) pruebas de eficiencia clínica. En cada una de las etapas sucesivas disminuye la posibilidad de que el medicamento pase las rigurosas pruebas; en cambio, el costo de las pruebas subsiguientes aumenta enormemente. Los estudios realizados informan que el costo promedio de la investigación y desarrollo de un nuevo fármaco es de $250 millones y dura 12 años. De esta manera, es importante descartar ya en las pruebas de las primeras etapas aquellos medicamentos que no resultarán útiles e identificar los que parecen prometedores para continuar sometiéndolos a las distintas pruebas.

La estadística desempeña un papel muy importante en la investigación farmacéutica, para la cual existen regulaciones gubernamentales estrictas y rigurosas. En las pruebas preclínicas suelen emplearse pruebas estadísticas que implican dos o tres estudios de poblaciones para determinar si se debe continuar con las pruebas de uso prolongado y confiabilidad del nuevo producto. Las poblaciones son: una para el nuevo medicamento, una de control, y otra para un medicamento estándar. Los estudios preclínicos inician con el envío del fármaco al departamento de farmacología para que evalúe su eficacia; esto es, su capacidad para producir los efectos esperados. Como parte de este proceso, se le pide a un especialista en estadística que diseñe un experimento para probar el nuevo producto. En este diseño se especifica el tamaño de la muestra y los métodos estadísticos de análisis. En los estudios en los que intervienen dos poblaciones se usa una muestra para obtener datos sobre la eficacia del nuevo medicamento (población 1) y otra para obtener datos acerca de la eficacia de un producto similar estándar (población 2). Dependiendo del uso que tendrá el nuevo ejemplar en disciplinas como neurología, cardiología e inmunología, se probarán tanto éste como el fármaco estándar. En la mayor parte de los estudios se emplea una prueba de hipótesis para determinar la diferencia entre las medias de las poblaciones del nuevo medicamento y del medicamento estándar. Si el primero es menos eficaz que el segundo o tiene efectos indeseables, se rechaza y queda eliminado del programa de pruebas. Sólo los nuevos fármacos que muestran comparaciones prometedoras respecto de los de tipo estándar se mantienen en pruebas de uso prolongado y confiabilidad.



Los métodos estadísticos se utilizan para desarrollar y probar medicamentos nuevos. © Lester Lefkowitz/CORBIS.

En las etapas de pruebas de uso prolongado y confiabilidad y de eficiencia clínica también se recolectan más datos y se realizan otros estudios multipoblacionales. La FDA requiere que los métodos estadísticos se definan antes de que se realicen las pruebas para evitar sesgos. Además, para no incurrir en sesgos humanos, algunos estudios clínicos son de doble o de triple ciego. Es decir, ni el paciente ni el investigador saben quién recibe cuál medicamento. Si el nuevo satisface todos los requerimientos en comparación con el estándar, se presenta una solicitud de aceptación del nuevo producto ante la FDA. La solicitud es rigurosamente analizada por los científicos y especialistas en estadística de esta oficina.

En este capítulo aprenderá cómo calcular intervalos de confianza y realizar pruebas de hipótesis para medias y proporciones cuando se comparan dos poblaciones. Se presentarán las técnicas para analizar tanto muestras aleatorias independientes como muestras pareadas.

En los capítulos 8 y 9 se mostró cómo obtener una estimación por intervalo y conducir una prueba de hipótesis cuando se trata de una sola media poblacional o de una sola proporción poblacional. En este capítulo se continúa con el estudio de la inferencia estadística mostrando la forma de estimar por intervalo y realizar pruebas de hipótesis cuando se tienen dos poblaciones y lo que interesa es la diferencia entre dos medias poblacionales o entre dos proporciones poblacionales. Por ejemplo, quizá desee obtener una estimación por intervalo para la diferencia entre el sueldo inicial medio de una población de hombres y el sueldo inicial medio de una población de mujeres, o necesite realizar una prueba de hipótesis para determinar si hay alguna diferencia entre la proporción de piezas defectuosas producidas por el proveedor A y la proporción de partes defectuosas producidas por el proveedor B. El estudio de inferencia estadística para dos poblaciones se inicia mostrando cómo obtener una estimación mediante un intervalo de confianza y cómo realizar una prueba de hipótesis para la diferencia entre las medias de dos poblaciones en el caso en que se conozcan sus desviaciones estándar.

10.1 Inferencias acerca de la diferencia entre dos medias poblacionales: σ_1 y σ_2 conocidas

Sean μ_1 la media de la población 1 y μ_2 la media de la población 2. Lo que interesa aquí son las inferencias acerca de la diferencia entre las medias: $\mu_1 - \mu_2$. Para hacer la inferencia respectiva se elige una muestra aleatoria simple de n_1 unidades de la población 1 y otra muestra aleatoria simple de n_2 unidades de la población 2. A estos dos conjuntos que se toman separada e independientemente se les conoce como **muestras aleatorias simples independientes**. En esta sección se supondrá que se cuenta con información que permite considerar que las dos desviaciones estándar σ_1 y σ_2 se conocen antes de tomar las muestras. Esta situación se identifica como el caso de σ_1 y σ_2 conocidas. En el ejemplo siguiente se describe cómo calcular el margen de error y obtener una estimación por intervalo sobre la diferencia entre las dos medias poblacionales cuando se conocen σ_1 y σ_2.

Estimación por intervalo para $\mu_1 - \mu_2$

Greystone Department Stores, Inc. tiene dos tiendas en Buffalo, Nueva York: una en el centro de la ciudad y otra en un centro comercial suburbano. El gerente regional ha observado que los productos que se venden bien en una tienda no siempre se venden bien en la otra. Cree que esto se debe a diferencias demográficas entre los clientes de las dos localidades. Debe haber diferencias de edad, educación, ingreso, etc., entre los compradores de una y otra tienda. Suponga que el gerente pide que se investigue la diferencia entre las medias de las edades de los clientes de los dos establecimientos.

Si la población 1 es de clientes que compran en la tienda del centro de la ciudad y la población 2 es de clientes que compran en la sucursal del centro comercial suburbano, tenemos:

μ_1 = media de la población 1 (es decir, media de las edades de los clientes que compran en la tienda del centro de la ciudad)

μ_2 = media de la población 2 (es decir, media de las edades de los clientes que compran en la tienda del centro comercial suburbano)

La diferencia entre las dos medias poblacionales es $\mu_1 - \mu_2$.

Para estimar $\mu_1 - \mu_2$ se toma una muestra aleatoria simple de n_1 clientes de la población 1 y una muestra aleatoria simple de n_2 clientes de la población 2, y se calculan las dos medias muestrales.

$\bar{x}_1$ = media muestral de las edades para la muestra aleatoria simple de n_1 clientes del centro de la ciudad

$\bar{x}_2$ = media muestral de las edades para la muestra aleatoria simple de n_2 clientes suburbanos

La estimación puntual de la diferencia entre las dos medias poblacionales es la diferencia entre las dos medias muestrales.

ESTIMADOR PUNTUAL DE LA DIFERENCIA ENTRE DOS MEDIAS POBLACIONALES

$$\bar{x}_1 - \bar{x}_2 \qquad (10.1)$$

El error estándar de $\bar{x}_1 - \bar{x}_2$ es la desviación estándar de la distribución de muestreo de $\bar{x}_1 - \bar{x}_2$.

En la figura 10.1 se proporciona una visión esquemática del proceso que se sigue para estimar la diferencia entre dos medias poblacionales empleando dos muestras aleatorias simples independientes.

Como otros estimadores puntuales, el estimador puntual $\bar{x}_1 - \bar{x}_2$ tiene un error estándar que describe la variación en la distribución de muestreo del estimador. Cuando se tienen dos muestras aleatorias simples independientes, el error estándar de $\bar{x}_1 - \bar{x}_2$ es el siguiente.

ERROR ESTÁNDAR DE $\bar{x}_1 - \bar{x}_2$

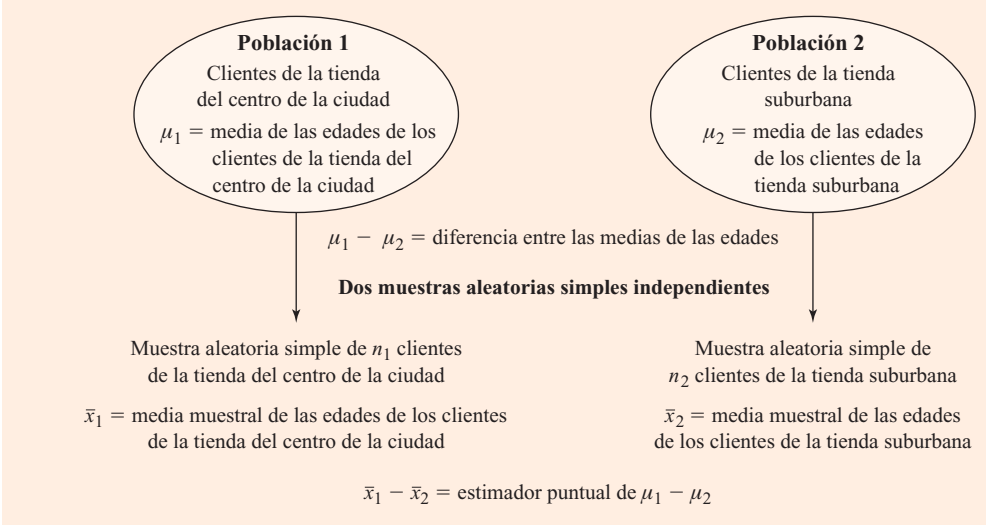
$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad (10.2)$$

Si ambas poblaciones tienen distribución normal o si los tamaños de las muestras son suficientemente grandes para que el teorema del límite central permita concluir que las distribuciones de muestreo de $\bar{x}_1 - \bar{x}_2$ puedan ser aproximadas mediante una distribución normal, la distribución muestral de $\bar{x}_1 - \bar{x}_2$ tendrá una distribución normal cuya media es $\mu_1 - \mu_2$.

Como se ilustra en el capítulo 8, una estimación por intervalo está dada por una estimación puntual $\pm$ un margen de error. En el caso de la estimación de la diferencia entre dos medias poblacionales, una estimación por intervalo tendrá la forma siguiente.

$$\bar{x}_1 - \bar{x}_2 \pm \text{margen de error}$$

FIGURA 10.1 Estimación de la diferencia entre dos medias poblacionales



Como la distribución de muestreo de $\bar{x}_1 - \bar{x}_2$ tiene una distribución normal, el margen de error se expresa de la manera siguiente.

El margen de error se obtiene multiplicando el error estándar por $z_{\alpha/2}$.

$$\text{Margen de error} = z_{\alpha/2}\sigma_{\bar{x}_1-\bar{x}_2} = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad \textbf{(10.3)}$$

En consecuencia, la estimación por intervalo de la diferencia entre las dos medias poblacionales es la que se indica enseguida:

ESTIMACIÓN POR INTERVALO DE LA DIFERENCIA ENTRE DOS MEDIAS POBLACIONALES: σ_1 Y σ_2 CONOCIDAS

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad \textbf{(10.4)}$$

donde $1 - \alpha$ es el coeficiente de confianza.

Volvamos al ejemplo de la tienda departamental de Greystone. Con base en datos de anteriores estudios demográficos sobre los clientes, las dos desviaciones estándar poblacionales se conocen y son $\sigma_1 = 9$ años y $\sigma_2 = 10$ años. De los datos de las dos muestras aleatorias simples independientes de los clientes de Greystone se obtuvieron los resultados siguientes.

	Tienda del centro de la ciudad	Tienda suburbana
Tamaño de la muestra	$n_1 = 36$	$n_2 = 49$
Media muestral	$\bar{x}_1 = 40$ años	$\bar{x}_2 = 35$ años

Mediante la expresión (10.1) se encuentra que la estimación puntual de la diferencia entre las medias de las edades de las dos poblaciones es $\bar{x}_1 - \bar{x}_2 = 40 - 35 = 5$ años. Por tanto, se estima que, en promedio, los clientes de la tienda en el centro de la ciudad son cinco años mayores que los de la tienda suburbana. Ahora, con la expresión (10.4) se calcula el margen de error y se proporciona una estimación por intervalo de $\mu_1 - \mu_2$. Utilizando 95% de confianza y $z_{\alpha/2} = z_{0.025} = 1.96$, tenemos:

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$40 - 35 \pm 1.96\sqrt{\frac{9^2}{36} + \frac{10^2}{49}}$$

$$5 \pm 4.06$$

Así pues, el margen de error es 4.06 años y la estimación por intervalo de 95% de confianza de la diferencia entre las dos medias poblacionales va de $5 - 4.06 = 0.94$ años a $5 + 4.06 = 9.06$ años.

Pruebas de hipótesis acerca de $\mu_1 - \mu_2$

Ahora se verán las pruebas de hipótesis acerca de la diferencia entre dos medias poblacionales. D_0 denota la diferencia hipotética entre μ_1 y μ_2. Las tres formas que puede adoptar una prueba de hipótesis son las siguientes:

$$H_0: \mu_1 - \mu_2 \geq D_0 \qquad H_0: \mu_1 - \mu_2 \leq D_0 \qquad H_0: \mu_1 - \mu_2 = D_0$$
$$H_a: \mu_1 - \mu_2 < D_0 \qquad H_a: \mu_1 - \mu_2 > D_0 \qquad H_a: \mu_1 - \mu_2 \neq D_0$$

En muchas aplicaciones, $D_0 = 0$. Con un ejemplo de prueba de hipótesis de dos colas, cuando $D_0 = 0$, la hipótesis nula es H_0: $\mu_1 - \mu_2 = 0$. En este caso, la hipótesis nula indica que μ_1 y μ_2 son iguales. Rechazar H_0 lleva a la conclusión de que H_a: $\mu_1 - \mu_2 \neq 0$ es verdadera; es decir, μ_1 y μ_2 no son iguales.

Los pasos presentados en el capítulo 9 para realizar una prueba de hipótesis también son aplicables aquí. Hay que elegir el nivel de significancia, calcular el valor del estadístico de prueba y encontrar el valor-p para determinar si se rechaza la hipótesis nula. En el caso de dos muestras aleatorias independientes, se mostró que el estimador puntual $\bar{x}_1 - \bar{x}_2$ tiene un error estándar $\sigma_{\bar{x}_1 - \bar{x}_2}$ dado por la expresión (10.2), y cuando los tamaños de las muestras son suficientemente grandes, la distribución de $\bar{x}_1 - \bar{x}_2$ se puede considerar normal. En este caso, el estadístico de prueba para la diferencia entre dos medias poblacionales cuando se conocen σ_1 y σ_2 es el que se da a continuación.

ESTADÍSTICO DE PRUEBA PARA PRUEBAS DE HIPÓTESIS ACERCA DE $\mu_1 - \mu_2$: σ_1 Y σ_2 CONOCIDAS

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \qquad \textbf{(10.5)}$$

El uso de dicho estadístico de prueba se muestra en el siguiente ejemplo de prueba de hipótesis.

Como parte de un estudio para evaluar las diferencias en la calidad educativa entre dos centros de enseñanza, se aplica un examen estandarizado a los individuos de ambos centros. La diferencia de calidad se evalúa al comparar las medias de las puntuaciones obtenidas en el examen. Las medias poblacionales en cada uno de los centros son:

μ_1 = media de las puntuaciones en los exámenes para la población de individuos del centro A

μ_2 = media de las puntuaciones en los exámenes para la población de individuos del centro B

Se parte del supuesto tentativo de que no hay diferencia entre la calidad de la educación en uno y otro centro de enseñanza. Entonces, en términos de las puntuaciones medias obtenidas en el examen, la hipótesis nula es que $\mu_1 - \mu_2 = 0$. Si las evidencias muestrales llevan a rechazar esta hipótesis, se concluirá que sí hay diferencia entre las medias de las puntuaciones de examen en las dos poblaciones. Esta conclusión indicará que hay diferencia en la calidad de los dos centros y sugerirá la necesidad de realizar un estudio para investigar las razones de estas discrepancias. Las hipótesis nula y alternativa en esta prueba de dos colas se expresan como se indica a continuación.

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 \neq 0$$

En exámenes estandarizados aplicados ya en diversas ocasiones, siempre se ha obtenido una desviación estándar cercana a 10 puntos. Por tanto, se usará esta información y se considerará que las desviaciones estándar poblacionales se conocen y que son $\sigma_1 = 10$ y $\sigma_1 = 10$. Para este estudio se especifica un nivel de significancia $\alpha = 0.05$.

WEB archivo
ExamScores

Se tienen muestras aleatorias simples independientes de $n_1 = 30$ individuos del centro de enseñanza A y $n_2 = 40$ individuos del centro de enseñanza B. Las medias muestrales correspondientes son $\bar{x}_1 = 82$ y $\bar{x}_2 = 78$. ¿Estos datos indican que existe una diferencia significativa

entre las medias poblacionales de los dos centros de enseñanza? Para responder esta pregunta se calcula el estadístico de prueba empleando la ecuación (10.5).

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(82 - 78) - 0}{\sqrt{\frac{10^2}{30} + \frac{10^2}{40}}} = 1.66$$

A continuación se calcula el valor-p de esta prueba de dos colas. Como el estadístico de prueba z se encuentra en la cola superior, se calcula primero el área bajo la curva a la derecha de $z = 1.66$. En la tabla de distribución normal estándar, el área a la izquierda de $z = 1.66$ es 0.9515. Por ende, el área en la cola superior de la distribución es $1.000 - 0.9515 = 0.0485$. Como es una prueba de dos colas, se debe duplicar el área que queda a la cola: valor-$p = 2(0.0485) = 0.0970$. Dado que la regla es rechazar H_0 si el valor-$p \leq \alpha$, como el valor-p es 0.0970, no se puede rechazar H_0 al nivel de significancia 0.05. Los resultados muestrales no proporcionan suficiente evidencia para concluir que hay una diferencia de calidad entre los dos centros de enseñanza.

En este capítulo, para las pruebas de hipótesis se utilizará el método del valor-p descrito en el capítulo 9. Sin embargo, si así se prefiere, se puede utilizar el estadístico de prueba y la regla de rechazo del valor crítico. Para $\alpha = 0.05$ y $z_{\alpha/2} = z_{0.025} = 1.96$, la regla de rechazo empleando el método del valor crítico será rechazar H_0 si $z \leq -1.96$ o si $z \geq 1.96$. Como $z = 1.66$, se llega a la misma conclusión de no rechazar H_0.

En el ejemplo anterior se demostró la prueba de hipótesis de dos colas acerca de la diferencia entre dos medias poblacionales. También se consideraron las pruebas de cola superior e inferior, en las cuales se aplica el mismo estadístico de prueba dado en la ecuación (10.5). El procedimiento para calcular el valor-p y las reglas de rechazo para pruebas de una cola se presentaron en el capítulo 9.

Consejo práctico

En la mayor parte de las aplicaciones de estimaciones por intervalo y de pruebas de hipótesis presentadas en esta sección, las muestras aleatorias con $n_1 \geq 30$ y $n_2 \geq 30$ se consideran adecuadas. En el caso en que una o las dos muestras sea menor que 30, las distribuciones de las poblaciones son importantes. En general, cuando esto ocurre, es importante que el analista se convenza de que es razonable suponer que las distribuciones de las dos poblaciones son al menos aproximadamente normales.

Ejercicios

Métodos

AUTOevaluación

1. Los resultados siguientes provienen de muestras aleatorias simples independientes tomadas de dos poblaciones.

Muestra 1	Muestra 2
$n_1 = 50$	$n_2 = 35$
$\bar{x}_1 = 13.6$	$\bar{x}_2 = 11.6$
$\sigma_1 = 2.2$	$\sigma_2 = 3.0$

   a) ¿Cuál es la estimación puntual de la diferencia entre las dos medias poblacionales?
   b) Proporcione un intervalo de 90% de confianza para la diferencia entre las dos medias poblacionales.
   c) Proporcione un intervalo de 95% de confianza para la diferencia entre las dos medias poblacionales.

AUTOevaluación

2. Considere la prueba de hipótesis que se da a continuación.

$$H_0: \mu_1 - \mu_2 \leq 0$$
$$H_a: \mu_1 - \mu_2 > 0$$

Los resultados siguientes se obtuvieron de dos muestras independientes tomadas de dos poblaciones.

Muestra 1	Muestra 2
$n_1 = 40$	$n_2 = 50$
$\bar{x}_1 = 25.2$	$\bar{x}_2 = 22.8$
$\sigma_1 = 5.2$	$\sigma_2 = 6.0$

a) ¿Cuál es el valor del estadístico de prueba?
b) ¿Cuál es el valor-*p*?
c) Si $\alpha = 0.05$, ¿cuál es la conclusión de la prueba de hipótesis?

3. Considere la prueba de hipótesis:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 \neq 0$$

Los siguientes son los resultados de dos muestras independientes tomadas de dos poblaciones.

Muestra 1	Muestra 2
$n_1 = 80$	$n_2 = 70$
$\bar{x}_1 = 104$	$\bar{x}_2 = 106$
$\sigma_1 = 8.4$	$\sigma_2 = 7.6$

a) ¿Cuál es el valor del estadístico de prueba?
b) ¿Cuál es el valor-*p*?
c) Si $\alpha = 0.05$, ¿cuál es la conclusión de la prueba de hipótesis?

Aplicaciones

AUTOevaluación

4. *Condé Nast Traveler* realiza una encuesta anual en la que los lectores califican su crucero favorito. Los navíos se califican en una escala de 100 puntos, donde los valores más altos indican un mejor servicio. Una muestra de 37 cruceros que transportan menos de 500 pasajeros resultó con una calificación promedio de 85.36 y una muestra de 44 navíos que transportan 500 o más pasajeros recibió una calificación promedio de 81.40 (*Condé Nast Traveler*, febrero de 2008). Suponga que la desviación estándar poblacional es 4.55 para los cruceros que transportan a menos de 500 pasajeros y 3.97 para los que transportan a 500 o más paseantes.
   a) ¿Cuál es la estimación puntual de la diferencia entre la calificación media poblacional de los navíos que transportan menos de 500 pasajeros y la calificación media poblacional de los que transportan 500 o más personas.
   b) ¿Cuál es el margen de error con 95% de confianza?
   c) ¿Cuál es la estimación por intervalo de 95% de confianza para la diferencia entre las calificaciones medias poblacionales de ambos tamaños de cruceros?

5. Se esperaba que el Día de San Valentín el gasto promedio fuera de $100.89 (*USA Today,* 13 de febrero de 2006). ¿Hay diferencia en las cantidades que desembolsan los hombres y las mujeres? El gasto promedio en una muestra de 40 hombres fue de $135.67 y en una muestra de 30 mujeres fue de $68.64. Por estudios anteriores se sabe que la desviación estándar poblacional en el consumo de los hombres es $35 y en el de las mujeres es $20.

a) ¿Cuál es la estimación puntual de la diferencia entre el gasto medio poblacional de los hombres y el gasto medio poblacional de las mujeres?
b) Con 99% de confianza, ¿cuál es el margen de error?
c) Elabore un intervalo de confianza de 99% para la diferencia entre las dos medias poblacionales.

6. Suponga que usted es el responsable de realizar las gestiones de una convención de negocios. Debido a recortes en el presupuesto por la reciente recesión, se le ha encargado escoger la ciudad con las habitaciones de hotel menos caras para el evento. La elección se ha reducido a optar entre Atlanta y Houston. El archivo nombrado *Hotel* contiene muestras de precios de habitaciones en Atlanta y Houston que son congruentes con los resultados reportados por Smith Travel Research (*SmartMoney*, marzo de 2009). Debido a que está disponible una considerable cantidad de datos históricos acerca de las tarifas de hospedaje en ambas ciudades, se puede asumir que la desviación estándar poblacional para los precios sea $20 en Atlanta y $25 en Houston. Con base en los datos muestrales, ¿puede concluir que la media de precio de una habitación de hotel en Atlanta es menor que la de una en Houston?

WEB archivo
Hotel

7. Durante la temporada 2003, las grandes ligas de beisbol tomaron medidas para acelerar el juego en los partidos con objeto de mantener el interés de los aficionados (CNN Headline News, 30 de septiembre de 2003). Los resultados siguientes se obtuvieron de una muestra de 60 partidos jugados en el verano de 2002 y de una muestra de 50 celebrados en el verano de 2003. La media muestral proporciona la duración media de los juegos que formaron parte de cada muestra.

Temporada 2002	Temporada 2003
$n_1 = 60$	$n_2 = 50$
$\bar{x}_1 =$ 2 horas, 52 minutos	$\bar{x}_2 =$ 2 horas, 46 minutos

a) La hipótesis de investigación consistió en plantear que las disposiciones tomadas en la temporada 2003 reducirían la duración media poblacional de los juegos de beisbol. Formule las hipótesis nula y alternativa.
b) ¿Cuál es la estimación puntual de la reducción de la duración media de los juegos en la temporada 2003?
c) Datos de estudios anteriores indican que, para ambos años, la desviación estándar poblacional fue 12 minutos. Realice la prueba de hipótesis y proporcione el valor-*p*. Utilizando como nivel de significancia 0.05, ¿cuál es su conclusión?
d) Proporcione una estimación por intervalo de 95% de confianza de la reducción en la duración media de los partidos en la temporada 2003.
e) ¿Cuál es la reducción porcentual en la duración media de los partidos de beisbol en la temporada 2003? ¿Estarán satisfechos los directivos con los resultados del análisis estadístico? Analice. En los años por venir, ¿seguirá siendo un problema la duración de los juegos de beisbol? Explique.

8. ¿Optimizar el servicio al cliente redunda en precios más altos de las acciones de las empresas que ofrecen el mejor servicio? "Los estudios revelan que, cuando mejora el nivel de satisfacción de una compañía con respecto al de años anteriores y queda por arriba del promedio nacional (actualmente 75.7), sus acciones tienen una buena posibilidad de superar al grueso del mercado de valores a largo plazo" (*BusinessWeek*, 2 de marzo de 2009). Los siguientes niveles de satisfacción de tres empresas en el 4o. trimestre de 2007 y 2008 se obtuvieron del índice estadounidense de satisfacción del cliente. Asuma que esas calificaciones se basan en un estudio de 60 consumidores de cada firma. Como el estudio se realizó por varios años, se puede asumir que la desviación estándar es igual a 6 puntos en cada caso.

Empresa	Calificación 2007	Calificación 2008
Rite Aid	73	76
Expedia	75	77
J.C. Penney	77	78

a) En el caso de Rite Aid, ¿el incremento en la calificación de la satisfacción de 2007 a 2008 es estadísticamente significativo? Utilice $\alpha = 0.05$. ¿Cuál es su conclusión?
b) ¿Puede concluir que la calificación de 2008 en el caso de Rite Aid está por arriba del promedio nacional de 75.7? Utilice $\alpha = 0.05$.
c) Para Expedia, ¿el incremento de 2007 a 2008 es estadísticamente significativo? Utilice $\alpha = 0.05$.
d) Al realizar una prueba de hipótesis con valores dados para la desviación estándar, el tamaño muestral y α, ¿cuán grande debe ser el incremento de 2007 a 2008 para que sea estadísticamente significativo?
e) Utilice el resultado del inciso *d*) para establecer si el incremento de JC Penney de uno a otro año es estadísticamente significativo.

10.2 Inferencias acerca de la diferencia entre dos medias poblacionales: σ_1 y σ_2 desconocidas

En esta sección el estudio de las inferencias sobre la diferencia entre dos medias poblacionales se extiende al caso en el cual las dos desviaciones estándar poblacionales, σ_1 y σ_2, no se conocen. En este caso, para estimar las desviaciones estándar poblacionales desconocidas se emplean las desviaciones estándar muestrales s_1 y s_2. Cuando se usan estas últimas desviaciones, las estimaciones por intervalo y las pruebas de hipótesis se basan en la distribución t en lugar de en la distribución normal estándar.

Estimación por intervalo para $\mu_1 - \mu_2$

En el ejemplo siguiente se muestra cómo calcular el margen de error y obtener una estimación por intervalo para la diferencia entre dos medias poblacionales cuando σ_1 y σ_2 no se conocen. Clearwater National Bank realiza un estudio para identificar diferencias entre las cuentas de cheques de sus clientes en dos de sus sucursales. Toma una muestra aleatoria simple de 28 cuentas de la sucursal Cherry Grove y otra muestra aleatoria simple e independiente de 22 cuentas de cheques de la sucursal Beechmont. El saldo actual de las cuentas de cheques se registra para cada cuenta. A continuación se presenta un resumen de los saldos en estas cuentas de cheques.

WEB archivo
CheckAcct

	Cherry Grove	Beechmont
Tamaño de la muestra	$n_1 = 28$	$n_2 = 22$
Media muestral	$\bar{x}_1 = \$1025$	$\bar{x}_2 = \$910$
Desviación estándar muestral	$s_1 = \$150$	$s_2 = \$125$

El banco desea estimar la diferencia entre el saldo medio en las cuentas de cheques de la población de clientes de Cherry Grove y el saldo medio en las cuentas de cheques de la población de clientes de Becchmont. Enseguida se calculará el margen de error y una estimación por intervalo para la diferencia entre estas dos medias poblacionales.

En la sección 10.1 se proporciona la estimación por intervalo siguiente para el caso en que se conocen las desviaciones estándar poblacionales, σ_1 y σ_2.

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Cuando se estiman σ_1 y σ_2 mediante s_1 y s_2, se usa la distribución t para hacer inferencias sobre la diferencia entre dos medias poblacionales.

Cuando no se conocen σ_1 y σ_2 se emplean las desviaciones estándar muestrales s_1 y s_2 para estimar σ_1 y σ_2, y $z_{\alpha/2}$ se sustituye por $t_{\alpha/2}$. Entonces la estimación por intervalo para la diferencia entre dos medias poblacionales queda dada por la expresión siguiente.

ESTIMACIÓN POR INTERVALO PARA LA DIFERENCIA ENTRE DOS MEDIAS POBLACIONALES: σ_1 Y σ_2 DESCONOCIDAS

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad \textbf{(10.6)}$$

donde $1 - \alpha$ es el coeficiente de confianza.

En esta expresión el uso de la distribución t es una aproximación, pero proporciona resultados excelentes y es relativamente fácil de usar. La única dificultad que se enfrenta al emplear la expresión (10.6) consiste en determinar los grados de libertad apropiados para $t_{\alpha/2}$. El software estadístico los calcula automáticamente. La fórmula que se usa es la siguiente.

GRADOS DE LIBERTAD: DISTRIBUCIÓN t CON DOS MUESTRAS ALEATORIAS INDEPENDIENTES

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} \qquad \textbf{(10.7)}$$

De vuelta al ejemplo del Clearwater National Bank, se mostrará cómo usar la expresión (10.6) para obtener una estimación por intervalo de 95% de confianza de la diferencia entre las medias poblacionales de los saldos en las cuentas de cheques de sus dos sucursales. Los datos muestrales de Cherry Grove son $n_1 = 28$, $\bar{x}_1 = \$1\,025$ y $s_1 = \$150$, y los de la sucursal Beechmont son $n_2 = 22$, $\bar{x}_2 = \$910$ y $s_2 = \$125$. El cálculo de los grados de libertad (gl) para $t_{\alpha/2}$ es:

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{150^2}{28} + \frac{125^2}{22}\right)^2}{\frac{1}{28 - 1}\left(\frac{150^2}{28}\right)^2 + \frac{1}{22 - 1}\left(\frac{125^2}{22}\right)^2} = 47.8$$

Como el resultado no es un número entero, se redondea hacia *abajo* a 47 para obtener un valor t mayor y una estimación por intervalo más prudente. En la tabla de la distribución t para 47 grados de libertad se encuentra que $t_{0.025} = 2.012$. Según la expresión (10.6), el intervalo de 95% de confianza para la diferencia entre las dos medias poblacionales se calcula como sigue.

$$\bar{x}_1 - \bar{x}_2 \pm t_{0.025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$1\,025 - 910 \pm 2.012\sqrt{\frac{150^2}{28} + \frac{125^2}{22}}$$

$$115 \pm 78$$

La estimación puntual de la diferencia entre las dos medias poblacionales de los saldos en las cuentas de cheques es \$115. El margen de error es \$78 y la estimación por intervalo de 95% de

confianza para la diferencia entre las dos medias poblacionales es el que va de $115 - 78 = \$37$ a $115 + 78 = \$193$.

Esta sugerencia es útil cuando se usa la ecuación (10.7) para calcular manualmente los grados de libertad.

El cálculo manual de los grados de libertad usando la expresión (10.7) es laborioso, pero muy sencillo si se usa software. Sin embargo, observe que las expresiones s_1^2/n_1 y s_2^2/n_2 aparecen tanto en la expresión (10.6) como en la (10.7). Por tanto, sólo habrá que calcular estos valores una vez para usarlos en ambas ecuaciones, (10.6) y (10.7).

Pruebas de hipótesis acerca de $\mu_1 - \mu_2$

Ahora se estudiarán las pruebas de hipótesis acerca de la diferencia entre las medias de dos poblaciones cuando no se conocen las desviaciones estándar poblacionales σ_1 y σ_2. Sea D_0 la diferencia hipotética entre μ_1 y μ_2. En la sección 10.1 se mostró que el estadístico de prueba utilizado cuando se conocen σ_1 y σ_2 es el siguiente.

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

El estadístico de prueba z sigue la distribución normal estándar.

Cuando no se conocen σ_1 y σ_2, se usa s_1 para estimar σ_1 y s_2 para estimar σ_2. Al sustituir σ_1 y σ_2 por estas desviaciones estándar muestrales se obtiene el siguiente estadístico de prueba para el caso en que no se conocen σ_1 y σ_2.

ESTADÍSTICO DE PRUEBA PARA PRUEBAS DE HIPÓTESIS ACERCA DE $\mu_1 - \mu_2$: σ_1 Y σ_2 DESCONOCIDAS

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad \textbf{(10.8)}$$

Los grados de libertad para la distribución t se obtienen con la ecuación (10.7).

Ahora se demostrará el uso del estadístico de prueba en el ejemplo siguiente.

Considere un nuevo software que ayuda a los analistas de sistemas a reducir el tiempo requerido para diseñar, elaborar y poner en marcha un sistema de información. Para evaluar las ventajas del nuevo programa, se toma una muestra aleatoria de 24 analistas de sistemas. A cada analista se le proporciona información sobre un sistema de información hipotético. A 12 de ellos se les pide que elaboren el sistema de información usando la tecnología actual. A los otros 12 se les capacita para usar el nuevo software y se les instruye para que lo empleen en el desarrollo del sistema de información.

En el estudio participan dos poblaciones: una de analistas de sistemas que usan la tecnología actual y otra de analistas de sistemas que aplican el nuevo software. En términos del tiempo necesario para completar el proyecto del sistema de información, las medias poblacionales son las siguientes.

μ_1 = media del tiempo que necesitan para completar el proyecto los analistas que emplean la tecnología actual.

μ_2 = media del tiempo que necesitan para completar el proyecto los analistas que emplean el nuevo software.

El investigador encargado de la evaluación del nuevo software espera demostrar que con el nuevo programa se requiere menos tiempo para completar el proyecto del sistema de información. De manera que tratará de hallar evidencias que le permitan concluir que μ_2 es menor que μ_1,

TABLA 10.1 Datos y resumen estadísticos del tiempo requerido en el estudio de la prueba de software

WEB archivo
SoftwareTest

	Tecnología actual	Nuevo software
	300	274
	280	220
	344	308
	385	336
	372	198
	360	300
	288	315
	321	258
	376	318
	290	310
	301	332
	283	263
Resumen estadístico		
Tamaño de la muestra	$n_1 = 12$	$n_2 = 12$
Media muestral	$\bar{x}_1 = 325$ horas	$\bar{x}_2 = 286$ horas
Desviación estándar muestral	$s_1 = 40$	$s_2 = 44$

caso en el que la diferencia entre las dos muestras poblacionales $\mu_1 - \mu_2$ será mayor que cero. La hipótesis de investigación $\mu_1 - \mu_2 > 0$ se establece como hipótesis alternativa. Por ende, la prueba de hipótesis será:

$$H_0: \mu_1 - \mu_2 \leq 0$$
$$H_a: \mu_1 - \mu_2 > 0$$

Como nivel de significancia se usará $\alpha = 0.05$.

Suponga que los resultados de los 24 analistas son los que se presentan en la tabla 10.1. Con el estadístico de prueba establecido en la ecuación (10.8) tenemos:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(325 - 286) - 0}{\sqrt{\frac{40^2}{12} + \frac{44^2}{12}}} = 2.27$$

Con base en la ecuación (10.7) los grados de libertad son:

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{40^2}{12} + \frac{44^2}{12}\right)^2}{\frac{1}{12 - 1}\left(\frac{40^2}{12}\right)^2 + \frac{1}{12 - 1}\left(\frac{44^2}{12}\right)^2} = 21.8$$

Al redondear hacia abajo se usará una distribución t con 21 grados de libertad. La fila correspondiente de la tabla de distribución t es la siguiente.

Área en la cola superior	0.20	0.10	0.05	0.025	0.01	0.005
Valor-t (21 gl)	0.859	1.323	1.721	2.080	2.518	2.831

$t = 2.27$

FIGURA 10.2 Resultado de Minitab para la prueba de hipótesis usando la tecnología actual y el nuevo software

```
Two-sample T for Current vs New

            N      Mean     StDev    SE Mean
Current    12     325.0      40.0         12
New        12     286.0      44.0         13

Difference = mu Current - mu New
Estimate for difference:  39.0000
95% lower bound for difference = 9.5
T-Test of difference = 0 (vs >):  T-Value = 2.27  P-Value = 0.017  DF = 21
```

*Sólo es posible determinar un rango para el valor-*p *cuando se utiliza la tabla de distribución* t. *Excel o Minitab muestran el valor-*p *exacto = 0.017.*

En una prueba de cola superior el valor-p es el área en la cola superior a la derecha de $t = 2.27$. Con base en este resultado, se ve que el valor-p está entre 0.025 y 0.01. Por tanto, es menor que $\alpha = 0.05$ y se rechaza H_0. Los resultados muestrales permiten al investigador concluir que $\mu_1 - \mu_2 > 0$, o que $\mu_1 > \mu_2$. Por consiguiente, el estudio de investigación apoya la conclusión de que el nuevo software proporciona una media poblacional menor de tiempo requerido.

Para las pruebas de hipótesis acerca de la diferencia entre dos medias poblacionales se usan Excel o Minitab. En la figura 10.2 se presentan los resultados que proporciona Minitab para la comparación de la tecnología actual y el nuevo software. En la última línea se ve que $t = 2.27$ y el valor-$p = 0.017$. Observe que Minitab usa la ecuación (10.7) para calcular los 21 grados de libertad de este análisis.

Consejo práctico

Se recomienda, siempre que sea posible, usar muestras del mismo tamaño, $n_1 = n_2$.

Los procedimientos aquí presentados para estimaciones por intervalo y pruebas de hipótesis son sólidos y pueden usarse con muestras relativamente pequeñas. En la mayor parte de las aplicaciones con muestras iguales o casi del mismo tamaño, y de manera que el tamaño total de la muestra, $n_1 + n_2$, sea por lo menos 20, se esperan muy buenos resultados aun cuando las poblaciones no sean normales. Si las distribuciones de las poblaciones son muy sesgadas o contienen valores atípicos, se recomiendan muestras más grandes. Las pequeñas sólo deben usarse cuando el analista está convencido de que las distribuciones de las poblaciones son aproximadamente normales.

NOTAS Y COMENTARIOS

Otro método que se utiliza para hacer inferencias acerca de la diferencia entre dos medias poblacionales cuando no se conocen σ_1 y σ_2 se basa en el supuesto de que las dos desviaciones estándar poblacionales son *iguales* ($\sigma_1 = \sigma_2 = \sigma$). Bajo este supuesto, las dos desviaciones estándar muestrales se combinan para obtener la siguiente *varianza muestral combinada*.

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

El estadístico de prueba t es:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}},$$

el cual tiene $n_1 + n_2 - 2$ grados de libertad. A partir de aquí el cálculo del valor-p y la interpretación de los resultados muestrales se efectúan de manera idéntica a lo indicado en esta sección.

El problema con este procedimiento es la dificultad que suele existir para verificar que las dos desviaciones estándar poblacionales son iguales. Lo más frecuente es encontrar que no son iguales. Con el procedimiento de la varianza combinada pueden no obtenerse resultados satisfactorios, en especial si los tamaños de las muestras, n_1 y n_2, son muy distintos.

El procedimiento t presentado en esta sección no requiere el supuesto de que las dos desviaciones estándar poblacionales son iguales y puede usarse cuando éstas son o no iguales. Es un procedimiento más general y se recomienda para la mayoría de las aplicaciones.

Ejercicios

Métodos

AUTOevaluación

9. Los resultados siguientes se obtuvieron de muestras aleatorias independientes tomadas de dos poblaciones.

Muestra 1	Muestra 2
$n_1 = 20$	$n_2 = 30$
$\bar{x}_1 = 22.5$	$\bar{x}_2 = 20.1$
$s_1 = 2.5$	$s_2 = 4.8$

a) ¿Cuál es la estimación puntual de la diferencia entre las dos medias poblacionales?
b) ¿Cuáles son los grados de libertad para la distribución *t*?
c) Con 95% de confianza, ¿cuál es el margen de error?
d) ¿Cuál es el intervalo de 95% de confianza para la diferencia entre las dos medias poblacionales?

AUTOevaluación

10. Considere la prueba de hipótesis siguiente.

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 \neq 0$$

Los resultados que se listan enseguida se obtuvieron de muestras independientes tomadas de dos poblaciones

Muestra 1	Muestra 2
$n_1 = 35$	$n_2 = 40$
$\bar{x}_1 = 13.6$	$\bar{x}_2 = 10.1$
$s_1 = 5.2$	$s_2 = 8.5$

a) ¿Cuál es el valor del estadístico de prueba?
b) ¿Cuáles son los grados de libertad para la distribución *t*?
c) ¿Cuál es el valor-*p*?
d) Con $\alpha = 0.05$, qué concluye?

11. Considere los datos siguientes de dos muestras aleatorias independientes que se obtuvieron de dos poblaciones normales.

Muestra 1	10	7	13	7	9	8
Muestra 2	8	7	8	4	6	9

a) Calcule las dos medias muestrales.
b) Determine las dos desviaciones estándar muestrales.
c) ¿Cuál es la estimación puntual de la diferencia entre las dos medias poblacionales?
d) ¿Cuál es la estimación por intervalo de 95% de confianza para la diferencia entre las dos medias poblacionales?

Aplicaciones

AUTOevaluación

12. El Departamento de Transporte de Estados Unidos informa sobre la cantidad de millas que recorren en automóvil los habitantes de las 75 principales áreas metropolitanas del país. Suponga que en una muestra aleatoria simple de 50 residentes de Buffalo, la media es 22.5 millas

por día y la desviación estándar es 8.4 millas por día, mientras que en una muestra aleatoria simple independiente de 40 habitantes de Boston la media es 18.6 millas por día y la desviación estándar es 7.4 millas por día.

a) ¿Cuál es la estimación puntual de la diferencia entre la media de las millas por día que recorre un habitante de Buffalo y la media de las millas por día que recorre un habitante Boston?

b) ¿Cuál es el intervalo de 95% de confianza para la diferencia entre las dos medias poblacionales?

WEB archivo
Cargo

13. FedEx y United Parcel Service (UPS) son las dos empresas de paquetería líderes en el mundo en cuanto a volumen e ingresos (*The Wall Street Journal,* 27 de enero de 2004). Según el Consejo Internacional de Aeropuertos, las terminales áereas internacionales de Memphis (FedEx) y de Louisville (UPS) son dos de los 10 mayores aeropuertos de carga del mundo. Las muestras aleatorias siguientes describen las toneladas de carga por día que pasan por estas terminales. Los datos se registran en miles de toneladas.

Memphis

9.1	15.1	8.8	10.0	7.5	10.5
8.3	9.1	6.0	5.8	12.1	9.3

Louisville

4.7	5.0	4.2	3.3	5.5
2.2	4.1	2.6	3.4	7.0

a) Calcule la media muestral y la desviación estándar muestral para cada uno de los aeropuertos.

b) ¿Cuál es la estimación puntual de la diferencia entre las dos medias poblacionales? Interprete este valor en términos del aeropuerto de mayor volumen y compare la diferencia de volúmenes entre las dos terminales.

c) Proporcione un intervalo de 95% de confianza para la diferencia entre las medias poblacionales diarias de los dos aeropuertos.

14. ¿Los sueldos de las enfermeras en Tampa, Florida, son más bajos que en Dallas, Texas? La información de sueldos muestra que el personal de enfermería en Tampa gana menos que el de Dallas (*The Tampa Tribune*, 15 de enero de 2007). Suponga que en un estudio de seguimiento de 40 enfermeras en Tampa y 50 en Dallas se obtuvieron los siguientes resultados.

Tampa	Dallas
$n_1 = 40$	$n_2 = 50$
$\bar{x}_1 = \$56\,100$	$\bar{x}_2 = \$59\,400$
$s_1 = \$6\,000$	$s_2 = \$7\,000$

a) Formule la hipótesis de tal forma que, si se rechaza la hipótesis nula, se pueda concluir que los sueldos de las enfermeras en Tampa son significativamente más bajos que en Dallas. Use $\alpha = 0.05$.

b) ¿Cuál es el valor del estadístico de prueba?

c) ¿Cuál es el valor-*p*?

d) ¿A qué conclusión llega?

15. Las lesiones entre los jugadores de las grandes ligas de beisbol han aumentado en los últimos años. La expansión de las ligas, de 1992 a 2001, hizo que la lista de jugadores se incrementará 15%. Sin embargo, la cantidad de beisbolistas en la lista de inhabilitados a causa de una lesión aumentó 32% en ese mismo periodo (*USA Today,* 8 de julio de 2002). Una cuestión a investigar es si los jugadores de las grandes ligas permanecían en la lista de lesionados más tiempo en 2001 que quienes estaban en esa lista una década antes.

a) Con la media poblacional de la cantidad de días que permanecía un jugador en la lista de inhabilitados, formule las hipótesis nula y alternativa que se pueden usar para probar la cuestión a investigar.

b) Tome los siguientes datos.

	Temporada 2001	Temporada 1992
Tamaño de la muestra	$n_1 = 45$	$n_2 = 38$
Media muestral	$\bar{x}_1 = 60$ días	$\bar{x}_2 = 51$ días
Desviación estándar muestral	$s_1 = 18$ días	$s_2 = 15$ días

¿Cuál es la estimación puntual de la diferencia entre las medias poblacionales de la cantidad de días en la lista de inhabilitados de 2001 comparado con la de 1992? ¿Cuál es el porcentaje de incremento en el número de días en esta lista?

c) Use $\alpha = 0.01$. ¿Cuál es su conclusión acerca de la cantidad de días en la lista de inhabilitados? ¿Cuál es el valor-*p*?

d) ¿Estos datos sugieren que las Grandes Ligas deben preocuparse por la situación?

WEB archivo
SATVerbal

16. El Consejo Universitario compara las puntuaciones obtenidas en el examen de aptitudes escolares (SAT, por sus siglas en inglés) con base en el nivel máximo de estudios de los padres de los sustentantes. La hipótesis de investigación indica que los estudiantes cuyos padres tienen un nivel educativo más alto obtendrán mejores puntuaciones en el SAT. En 2003, la media general en el examen oral fue 507 (*The World Almanac*, 2004). A continuación se presentan las puntuaciones obtenidas en el examen verbal en dos muestras independientes de estudiantes. La primera corresponde a las puntuaciones de estudiantes cuyos padres tienen una licenciatura. La segunda, a las puntuaciones de sustentantes cuyos padres terminaron el bachillerato pero no cursaron una licenciatura.

Padres de los estudiantes			
Con licenciatura		**Con bachillerato**	
485	487	442	492
534	533	580	478
650	526	479	425
554	410	486	485
550	515	528	390
572	578	524	535
497	448		
592	469		

a) Formule las hipótesis pertinentes para determinar si los datos muestrales confirman la hipótesis de que los estudiantes cuyos padres tienen un nivel de estudios más alto obtienen mejores puntuaciones en el SAT.

b) ¿Cuál es la estimación puntual de la diferencia entre las medias de las dos poblaciones?

c) Calcule el valor-*p* en esta prueba de hipótesis.

d) Con $\alpha = 0.05$, ¿cuál es su conclusión?

17. Con cierta periodicidad, Merrill Lynch solicita a sus clientes evaluaciones sobre los consultores y los servicios financieros que les proporciona. Las puntuaciones más altas en la encuesta de satisfacción del cliente indican mejor servicio con 7 como la puntuación más alta. A continuación se presentan en forma resumida las puntuaciones otorgadas a dos consultores financieros por los miembros de dos muestras aleatorias independientes. El consultor A tiene 10 años de experiencia, mientras que el consultor B tiene sólo 1 año. Use $\alpha = 0.05$ y realice una prueba para determinar si el consultor con más experiencia tiene la media poblacional más alta en la evaluación del servicio.

Consultor A	Consultor B
$n_1 = 16$	$n_2 = 10$
$\bar{x}_1 = 6.82$	$\bar{x}_2 = 6.25$
$s_1 = 0.64$	$s_2 = 0.75$

a) Establezca las hipótesis nula y alternativa.
b) Calcule el valor del estadístico de prueba.
c) ¿Cuál es el valor-*p*?
d) ¿A qué conclusión llega?

WEB archivo
SAT

18. Las empresas de preparación para exámenes ofrecen asesorías, clases y pruebas simuladas con el fin de ayudar a los estudiantes a obtener mejores resultados en exámenes como el de aptitudes escolares (SAT). Estas empresas aseguran que sus cursos ayudan a los estudiantes a mejorar sus puntuaciones hasta en un promedio de 120 puntos (*The Wall Street Journal,* 23 de enero de 2003). Un investigador duda de esta aseveración y cree que 120 puntos es una exageración de las empresas para motivar a los aspirantes a tomar los cursos de preparación. En un estudio para evaluar un curso para dicho examen, los investigadores recabaron datos de las puntuaciones de 35 estudiantes que tomaron el curso de preparación y de 48 que no lo tomaron. El archivo SAT contiene los datos de este estudio.
a) Formule las hipótesis apropiadas para probar el supuesto de los investigadores de que la mejora en la puntuación del SAT debe ser menor que 120 puntos.
b) Use $\alpha = 0.05$. ¿Cuál es su conclusión?
c) ¿Cuál es la estimación puntual de la mejora en la puntuación promedio del SAT obtenida con los cursos de preparación? Proporcione un intervalo de 95% de confianza para la estimación de la mejora.
d) ¿Qué consejo daría al investigador después de ver el intervalo de confianza?

10.3 Inferencias acerca de la diferencia entre dos medias poblacionales: muestras pareadas

Suponga que los empleados de una empresa de manufactura utilizan dos métodos distintos para realizar una tarea de producción. Con el fin de maximizar la producción, la empresa desea identificar el método con la menor media poblacional del tiempo necesario para completar esta tarea. Sea μ_1 la media poblacional del tiempo empleando el método 1, y μ_2 la media poblacional del tiempo requerido para completar la tarea con el método 2. Puesto que no hay ninguna indicación de cuál es el mejor método, se empieza por suponer que con los dos se obtiene la misma media poblacional del tiempo requerido para completar la tarea. De esta manera, la hipótesis nula es H_0: $\mu_1 - \mu_2 = 0$. Si se rechaza esta hipótesis, se podrá concluir que las medias poblacionales de los tiempos requeridos para realizar la tarea son diferentes con los dos métodos. En tal caso se recomendará el que proporcione la menor media del tiempo para la realización de la tarea. Las hipótesis nula y alternativa se expresan como sigue.

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 \neq 0$$

En la elección del procedimiento de muestreo para obtener los datos y probar las hipótesis se consideran dos diseños alternos. Uno se basa en muestras independientes y el otro en **muestras pareadas** (o muestras por pares).

1. *Diseño de muestras independientes.* Se toma una muestra aleatoria simple de trabajadores y cada uno de ellos usa el método 1. Se extrae una segunda muestra aleatoria simple independiente de trabajadores y cada uno usa el método 2. El procedimiento

que se utiliza para probar la diferencia entre las dos medias es el presentado en la sección 10.2.

2. *Diseño de muestras pareadas*. Se toma una muestra aleatoria simple de trabajadores. Cada trabajador usa primero uno de los métodos y después el otro. A cada sujeto se le asigna en forma aleatoria el orden en que aplicará los dos métodos; algunos primero usarán el método 1 y otros el método 2. Cada trabajador proporcionará un par de valores de datos, un valor para el método 1 y otro para el método 2.

En el diseño de muestras pareadas los dos métodos se prueban bajo condiciones similares (es decir, con los mismos trabajadores); por tanto, este diseño suele conducir a errores muestrales más pequeños que el de muestras independientes. La razón principal se debe a que en el diseño de muestras pareadas se elimina la variación entre los trabajadores, ya que los mismos sujetos prueban los dos métodos.

A continuación, con el empleo del diseño de muestras pareadas se demostrará la diferencia entre las medias poblacionales para los dos métodos de producción. Se emplea una muestra aleatoria de seis trabajadores. En la tabla 10.2 se indican los tiempos que requirieron los seis sujetos para realizar la tarea. Observe que de cada trabajador se obtuvieron dos datos, uno con cada método de producción. También observe que en la última columna se proporciona, para cada sujeto de la muestra, la diferencia d_i entre los tiempos para completar la tarea.

Lo principal en el análisis de muestras pareadas consiste en darse cuenta de que únicamente hay que considerar la columna de las diferencias. De manera que se tienen seis valores de datos (0.6, −0.2, 0.5, 0.3, 0 y 0.6) que se usarán para analizar la discrepancia entre las medias poblacionales de los dos métodos de producción.

Sea μ_d = la media de la *diferencia* en los valores de la población de trabajadores. Con esta notación, las hipótesis nula y alternativa se expresan como sigue.

$$H_0: \mu_d = 0$$
$$H_a: \mu_d \neq 0$$

Si H_0 es rechazada, se concluye que las medias poblacionales difieren en los tiempos requeridos para completar la tarea con los dos métodos.

Fuera del uso de la notación d, *las fórmulas para la media muestral y la desviación estándar muestral son las mismas que se han manejado antes en este libro.*

La notación d sirve para recordar que las muestras pareadas proporcionan datos que son *diferencias.* A continuación se calcula la media y la desviación estándar muestrales de las seis diferencias en los valores que se presentan en la tabla 10.2.

$$\bar{d} = \frac{\Sigma d_i}{n} = \frac{1.8}{6} = 0.30$$

$$s_d = \sqrt{\frac{\Sigma(d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{0.56}{5}} = 0.335$$

TABLA 10.2 Tiempos para completar una tarea en un diseño de muestras pareadas

Trabajador	Tiempo de realización con el método 1 (minutos)	Tiempo de realización con el método 2 (minutos)	Diferencia en los tiempos de realización (d_i)
1	6.0	5.4	0.6
2	5.0	5.2	− 0.2
3	7.0	6.5	0.5
4	6.2	5.9	0.3
5	6.0	6.0	0
6	6.4	5.8	0.6

WEB archivo

Matched

Si la muestra es grande, no se necesita el supuesto de que la población tiene una distribución normal. En los capítulos 8 y 9 se presentan los lineamientos sobre el tamaño de la muestra para usar la distribución t.

Como la muestra es pequeña, $n = 6$, es preciso suponer que la población de las diferencias tiene una distribución normal. Este supuesto es necesario para usar la distribución t en la prueba de hipótesis y para calcular la estimación por intervalo. Con esta presunción, el estadístico de prueba siguiente tiene una distribución t con $n - 1$ grados de libertad.

ESTADÍSTICO DE PRUEBA PARA PRUEBAS DE HIPÓTESIS CON MUESTRAS PAREADAS

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} \tag{10.9}$$

Una vez que los datos de las diferencias son calculados, el procedimiento para la distribución t de las muestras pareadas es el mismo que para la estimación y la prueba de hipótesis de una sola población descritas en los capítulos 8 y 9.

A continuación se usará la ecuación (10.9) para probar las hipótesis H_0: $\mu_d = 0$ y H_a: $\mu_d \neq 0$, usando $\alpha = 0.05$. El estadístico de prueba se calcula sustituyendo en la ecuación (10.9) los resultados muestrales, $\bar{d} = 0.30$, $s_d = 0.335$ y $n = 6$.

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} = \frac{0.30 - 0}{0.335/\sqrt{6}} = 2.20$$

Ahora se calculará el valor-p para esta prueba de dos colas. Como $t = 2.20 > 0$, el estadístico de prueba se encuentra en la cola superior de la distribución t. Como $t = 2.20$, el área en la cola superior a la derecha del estadístico de prueba se identifica usando la tabla de distribución t con grados de libertad $= n - 1 = 6 - 1 = 5$. A continuación se copia la información correspondiente a la fila de la tabla de distribución t para 5 grados de libertad.

Área en la cola superior	**0.20**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
Valor-*t* (5 *gl*)	0.920	1.476	2.015	2.571	3.365	4.032

$t = 2.20$

Como se ve, el área en la cola superior está entre 0.05 y 0.025. Por tratarse de una prueba de dos colas, se duplica este valor y se concluye que el valor-p se ubica entre 0.10 y 0.05. Este valor-p es mayor que $\alpha = 0.05$. Por ende, no se rechaza la hipótesis nula H_0: $\mu_d = 0$. Con Excel o Minitab y los datos de la tabla 10.2 se encuentra el valor-p exacto $= 0.080$.

Además, también se puede obtener un intervalo de confianza para estimar la diferencia entre las dos medias poblacionales usando la metodología para una sola población presentada en el capítulo 8. A continuación se presenta el cálculo para obtener un intervalo de 95% de confianza.

$$\bar{d} \pm t_{0.025}\frac{s_d}{\sqrt{n}}$$

$$0.3 \pm 2.571\left(\frac{0.335}{\sqrt{6}}\right)$$

$$0.3 \pm 0.35$$

Por tanto, el margen de error es 0.35 y el intervalo de 95% de confianza para estimar la diferencia entre las medias poblacionales de los dos métodos de producción va de -0.05 minutos a 0.65 minutos.

NOTAS Y COMENTARIOS

1. En el ejemplo presentado en esta sección, los trabajadores realizan la tarea primero con un método y luego con el otro. Este ejemplo ilustra un diseño de muestras pareadas en el que cada elemento (trabajador) de la muestra arroja dos datos. Para obtener el par de datos también se emplean elementos diferentes pero "similares". Por ejemplo, un trabajador en una ubicación forma pareja con otro en diferente ubicación (con similitud en edad, género, experiencia, nivel de estudio, etc.) De las parejas de sujetos se obtendrán los datos de las diferencias a ser usados en el análisis de muestras pareadas.
2. Con el método de muestras pareadas para obtener inferencias sobre dos medias poblacionales, por lo general se logran mejores resultados que con el método de muestras independientes; por tanto, es el más recomendable. Sin embargo, en muchas aplicaciones no se logran formar pares o el tiempo y el costo requeridos son excesivos. En tales casos se deberá recurrir al método de muestras independientes.

Ejercicios

Métodos

AUTOevaluación

19. Considere la prueba de hipótesis siguiente.

$$H_0: \mu_d \leq 0$$
$$H_a: \mu_d > 0$$

Los datos siguientes provienen de muestras pareadas tomadas de dos poblaciones.

	Población	
Elemento	1	2
1	21	20
2	28	26
3	18	18
4	20	20
5	26	24

a) Calcule la diferencia en el valor de cada elemento.
b) Determine $\bar{d}$.
c) Calcule la desviación estándar s_d.
d) Realice una prueba de hipótesis usando $\alpha = 0.05$. ¿Cuál es su conclusión?

20. Los datos siguientes provienen de muestras pareadas tomadas de dos poblaciones.

	Población	
Elemento	1	2
1	11	8
2	7	8
3	9	6
4	12	7
5	13	10
6	15	15
7	15	14

a) Estime la diferencia de valores para cada elemento.
b) Calcule $\overline{d}$.
c) Determine la desviación estándar s_d.
d) ¿Cuál es la estimación puntual de la diferencia entre las dos medias poblacionales?
e) Proporcione un intervalo de 95% de confianza para estimar la diferencia entre las dos medias poblacionales.

Aplicaciones

AUTOevaluación

21. Una firma de investigación de mercados usa una muestra de individuos para calificar el potencial de compra de un determinado producto antes y después de que los individuos vean un comercial de televisión que lo promociona. La calificación del potencial de compra se efectúa con una escala del 0 al 10, con los valores más altos indicando un mayor potencial. En la hipótesis nula se establece que la media de las calificaciones de "después" será menor o igual a la media de las calificaciones "antes". El rechazo de esta hipótesis indica que el comercial mejora la media de la calificación del potencial de compra. Use $\alpha = 0.05$ y los datos de la tabla siguiente para probar esta hipótesis y exprese un comentario sobre la utilidad del comercial.

	Calificación de compra			Calificación de compra	
Individuos	Después	Antes	Individuos	Después	Antes
1	6	5	5	3	5
2	6	4	6	9	8
3	7	7	7	7	5
4	4	3	8	6	6

WEB archivo
Earnings2005

22. En el archivo titulado *Earnings2005* se encuentran datos sobre las utilidades por acción en los que se comparan las obtenidas en un trimestre con las del trimestre anterior. Proporcione un intervalo de 95% de confianza para estimar la diferencia entre las medias poblacionales del trimestre dado frente a las del trimestre anterior. ¿Las utilidades aumentaron?

23. En un estudio del Bank of America sobre el gasto de los consumidores, se recabaron datos sobre las cantidades pagadas con tarjetas de crédito en seis categorías diferentes: transporte, supermercado, restaurantes, gastos domésticos, mobiliario, vestido y diversión (*US. Airways Attaché,* diciembre de 2003). Suponga que con datos de 43 tarjetas de crédito se identifican las cantidades anuales que se gastaron en supermercado (población 1) y en restaurantes (población 2). A partir de las diferencias, la media muestral de éstas fue $\overline{d} = \$850$, y la desviación estándar muestral fue $s_d = \$1\,123$.
a) Formule las hipótesis nula y alternativa para probar que no hay diferencia entre las medias poblacionales de los gastos en supermercado y en restaurantes pagados ambos con tarjeta de crédito.
b) Con 0.05 como nivel de significancia, ¿se puede concluir que hay diferencia entre las medias poblacionales? ¿Cuál es el valor-*p*?
c) ¿En qué categoría, supermercado o restaurantes, es mayor la media poblacional de los gastos anuales pagados con tarjeta de crédito? Proporcione la estimación puntual de la diferencia entre las medias poblacionales. Considere un intervalo de 95% de confianza para estimar la diferencia entre estas medias poblacionales.

WEB archivo
AirFare

24. Las personas que viajan en avión suelen elegir de qué aeropuerto salir con base en el costo del vuelo. Para determinar de qué aeropuerto es más costoso salir, considerando el de Dayton, Ohio, y el de Louisville, Kentucky, se recolectan datos (en dólares) de una muestra de vuelos a ocho ciudades (*The Cincinnati Enquirer,* 19 de febrero de 2006). Un investigador sostiene que es mucho más costoso partir de Dayton que de Louisville. Use los datos muestrales para ver si sustentan tal afirmación. Como nivel de significancia use $\alpha = 0.05$.

Destination	Dayton	Louisville
Chicago O'Hare	$319	$142
Grand Rapids, Michigan	192	213
Portland, Oregon	503	317
Atlanta	256	387
Seattle	339	317
South Bend, Indiana	379	167
Miami	268	273
Dallas–Ft. Worth	288	274

25. En los últimos años prolifera una cantidad cada vez mayor de opciones de entretenimiento que compiten por el tiempo de los consumidores. En 2004 la televisión por cable y la radio superaron a la televisión abierta, la música grabada y los periódicos, convirtiéndose en los medios de entretenimiento más usados (*The Wall Street Journal*, 26 de enero de 2004). Con una muestra de 15 individuos, los investigadores obtienen los datos de las horas por semana que destinan a ver televisión por cable y de las horas por semana en que escuchan la radio.

WEB archivo
TVRadio

Individual	Television	Radio	Individual	Television	Radio
1	22	25	9	21	21
2	8	10	10	23	23
3	25	29	11	14	15
4	22	19	12	14	18
5	12	13	13	14	17
6	26	28	14	16	15
7	22	23	15	24	23
8	19	21			

a) Use como nivel de significancia 0.05 y haga una prueba para la diferencia entre las medias poblacionales de la cantidad de horas destinadas a la televisión por cable y la cantidad de horas destinadas a la radio. ¿Cuál es el valor-*p*?

b) ¿Cuál es la media muestral de la cantidad de horas por semana empleadas en ver televisión por cable? ¿Cuál es la media muestral de la cantidad de horas por semana destinadas a escuchar radio? ¿Cuál de estos medios tiene mayor uso?

26. En la tabla siguiente se presentan las puntuaciones en las rondas primera y cuarta (final) de una muestra de 20 golfistas (Player) que compitieron en los torneos de la PGA (*Golfweek*, 14 y 28 de febrero de 2009). Suponga que se desea determinar si la media de las puntuaciones en la primera ronda (First Round) de un evento del PGA Tour es significativamente diferente de la media de las puntuaciones en la cuarta y última ronda (Final Round). ¿La presión del juego en la ronda final causa que las puntuaciones aumenten? ¿O el incremento en la concentración del golfista ocasiona que las puntuaciones disminuyan?

WEB archivo
GolfScores

Player	First Round	Final Round	Player	First Round	Final Round
Michael Letzig	70	72	Aron Price	72	72
Scott Verplank	71	72	Charles Howell	72	70
D. A. Points	70	75	Jason Dufner	70	73
Jerry Kelly	72	71	Mike Weir	70	77
Soren Hansen	70	69	Carl Pettersson	68	70
D. J. Trahan	67	67	Bo Van Pelt	68	65
Bubba Watson	71	67	Ernie Els	71	70
Reteif Goosen	68	75	Cameron Beckman	70	68
Jeff Klauk	67	73	Nick Watney	69	68
Kenny Perry	70	69	Tommy Armour III	67	71

a) Utilice $\alpha = 0.10$ para probar si existe una diferencia estadísticamente significativa entre las medias poblacionales de las puntuaciones de la primera y cuarta rondas. ¿Cuál es el valor-*p*? ¿A qué conclusión llega?

b) ¿Cuál es la estimación puntual de la diferencia entre las dos medias poblacionales? ¿Para cuál ronda es menor la media poblacional de la puntuación?

c) ¿Cuál es el margen de error para la estimación por intervalo de 95% de confianza para la diferencia entre las medias poblacionales? ¿Podría utilizarse este intervalo de confianza para probar la hipótesis del inciso *a*)? Explique.

27. Un fabricante produce dos modelos de una lijadora automática, uno de lujo y otro estándar, diseñado para uso doméstico. Los precios de venta de una muestra de distribuidores minoristas se presentan a continuación.

	Precio del modelo ($)			Precio del modelo ($)	
Minorista	Lujo	Estándar	Minorista	Lujo	Estándar
1	39	27	5	40	30
2	39	28	6	39	34
3	45	35	7	35	29
4	38	30			

a) En los precios minoristas sugeridos por el fabricante para los dos modelos, la diferencia es de $10. Use como nivel de significancia 0.05 y pruebe que la diferencia media entre los precios de los dos modelos es realmente de $10.

b) ¿Cuál es el intervalo de 95% de confianza para la diferencia entre la media de los precios de ambos modelos?

10.4 Inferencias acerca de la diferencia entre dos proporciones poblacionales

Siendo p_1 la proporción de la población 1 y p_2 la proporción de la población 2, a continuación se considerarán inferencias acerca de la diferencia entre dos proporciones poblacionales: $p_1 - p_2$. Para determinar las inferencias acerca de estas diferencias, se seleccionan dos muestras aleatorias independientes: una de n_1 unidades de la población 1 y otra de n_2 unidades de la población 2.

Estimación por intervalo para $p_1 - p_2$

En el ejemplo siguiente se mostrará cómo calcular un margen de error y una estimación por intervalo para la diferencia entre dos proporciones poblacionales.

Una empresa que se dedica a elaborar declaraciones de impuestos desea comparar la calidad del trabajo que se realiza en dos de sus oficinas regionales. Con muestras aleatorias de las declaraciones de impuestos elaboradas en dichas oficinas y verificando la exactitud de los reportes, la empresa podrá estimar la proporción de declaraciones con errores en que incurrió cada una de estas oficinas. Interesa conocer la diferencia entre las proporciones siguientes:

p_1 = proporción de declaraciones erróneas en la población 1 (oficina 1)

p_2 = proporción de declaraciones erróneas en la población 2 (oficina 2)

$\bar{p}_1$ = proporción muestral en una muestra aleatoria simple de la población 1

$\bar{p}_2$ = proporción muestral en una muestra aleatoria simple de la población 2

La diferencia entre las dos proporciones poblacionales está dada por $p_1 - p_2$. La estimación puntual de $p_1 - p_2$ se indica enseguida.

ESTIMADOR PUNTUAL DE LA DIFERENCIA ENTRE DOS PROPORCIONES POBLACIONALES

$$\bar{p}_1 - \bar{p}_2 \tag{10.10}$$

Por ende, el estimador puntual de la diferencia entre dos proporciones poblacionales es la diferencia entre las proporciones muestrales de dos muestras aleatorias simples independientes.

Como ocurre con otros estimadores puntuales, el estimador puntual $\bar{p}_1 - \bar{p}_2$ tiene una distribución de muestreo que refleja los valores que podría asumir $\bar{p}_1 - \bar{p}_2$ si se tomaran repetidamente dos muestras aleatorias simples independientes. La media de esta distribución de muestreo es $p_1 - p_2$ y el error estándar de $\bar{p}_1 - \bar{p}_2$ es el siguiente.

ERROR ESTÁNDAR DE $\bar{p}_1 - \bar{p}_2$

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}} \tag{10.11}$$

Si los tamaños de las muestras son suficientemente grandes para que n_1p_1, $n_1(1 - p_1)$, n_2p_2 y $n_2(1 - p_2)$ sean todos mayores o iguales que 5, la distribución de muestreo de $\bar{p}_1 - \bar{p}_2$ puede ser aproximada por una distribución normal.

Como ya se indicó antes, una estimación por intervalo está dada por una estimación puntual $\pm$ un margen de error. En el cálculo de la diferencia entre dos proporciones poblacionales, una estimación por intervalo toma la forma siguiente.

$$\bar{p}_1 - \bar{p}_2 \pm \text{margen de error}$$

Al aproximar la distribución de muestreo de $\bar{p}_1 - \bar{p}_2$ mediante una distribución normal, se podrá usar como margen de error $z_{\alpha/2}\sigma_{\bar{p}_1 - \bar{p}_2}$. Sin embargo, como $\sigma_{\bar{p}_1 - \bar{p}_2}$ está dada por la ecuación (10.11) no se puede usar directamente porque no se conoce ninguna de las dos proporciones poblacionales p_1 y p_2. Al usar la proporción muestral $\bar{p}_1$ para estimar p_1 y la proporción muestral $\bar{p}_2$ para estimar p_2, el margen de error queda como sigue.

$$\text{Margen de error} = z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2(1 - \bar{p}_2)}{n_2}} \tag{10.12}$$

La forma general de una estimación por intervalo para la diferencia entre dos proporciones poblacionales es la siguiente.

ESTIMACIÓN POR INTERVALO DE LA DIFERENCIA ENTRE DOS PROPORCIONES POBLACIONALES

$$\bar{p}_1 - \bar{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2(1 - \bar{p}_2)}{n_2}} \tag{10.13}$$

donde $1 - \alpha$ es el coeficiente de confianza.

De regreso al ejemplo de elaboración de declaraciones de impuestos, se encuentra que de las muestras aleatorias simples independientes de las dos oficinas se obtienen los datos siguientes.

Oficina 1	Oficina 2
$n_1 = 250$	$n_2 = 300$
Número de declaraciones con errores = 35	Número de declaraciones con errores = 27

WEB archivo
TaxPrep

Las proporciones muestrales en cada una de las oficinas son las siguientes.

$$\bar{p}_1 = \frac{35}{250} = 0.14$$

$$\bar{p}_2 = \frac{27}{300} = 0.09$$

La estimación puntual de la diferencia entre las proporciones de declaraciones con errores en las dos poblaciones es $\bar{p}_1 - \bar{p}_2 = 0.14 - 0.09 = 0.05$. Por tanto, se estima que la oficina 1 comete 0.05 o 5% más errores que la oficina 2.

Ahora se puede usar la expresión (10.13) para calcular el margen de error y la estimación por intervalo para la diferencia entre las dos proporciones poblacionales. Utilizando un intervalo de 90% de confianza con $z_{\alpha/2} - z_{0.05} = 1.645$, tenemos

$$\bar{p}_1 - \bar{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$$

$$0.14 - 0.09 \pm 1.645\sqrt{\frac{0.14(1-0.14)}{250} + \frac{0.09(1-0.09)}{300}}$$

$$0.05 \pm 0.045$$

El margen de error es 0.045 y el intervalo de 90% de confianza va de 0.005 a 0.095.

Prueba de hipótesis acerca de $p_1 - p_2$

Ahora se considerarán las pruebas de hipótesis acerca de la diferencia entre las proporciones de dos poblaciones. Se analizarán pruebas que comprenden el caso en que no hay diferencia entre las dos proporciones poblacionales. En tal situación, las tres formas que adoptan las pruebas de hipótesis son las siguientes.

$$H_0: p_1 - p_2 \geq 0 \qquad H_0: p_1 - p_2 \leq 0 \qquad H_0: p_1 - p_2 = 0$$
$$H_a: p_1 - p_2 < 0 \qquad H_a: p_1 - p_2 > 0 \qquad H_a: p_1 - p_2 \neq 0$$

En todas las hipótesis consideradas se usa 0 como la diferencia de interés.

Si se supone que H_0, considerada como igualdad, es verdadera, se tiene $p_1 - p_2 = 0$, lo cual equivale a decir que dichas proporciones poblacionales son iguales, $p_1 = p_2$.

El estadístico de prueba se basará en la distribución de muestreo del estimador puntual $\bar{p}_1 - \bar{p}_2$. En la ecuación (10.11) se mostró que el error estándar de $\bar{p}_1 - \bar{p}_2$ está dado por:

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Si se supone que H_0 es verdadera como igualdad, las proporciones poblacionales son iguales y $p_1 = p_2 = p$. En este caso, $\sigma_{\bar{p}_1 - \bar{p}_2}$ se convierte en la expresión que se presenta enseguida.

ERROR ESTÁNDAR DE $\bar{p}_1 - \bar{p}_2$ CUANDO $p_1 = p_2 = p$

$$\sigma_{\bar{p}_1-\bar{p}_2} = \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \qquad \textbf{(10.14)}$$

Como no se conoce p, se combinan los estimadores puntuales de las dos muestras ($\bar{p}_1$ y $\bar{p}_2$) con objeto de obtener un solo estimador puntual de p como se indica a continuación.

ESTIMADOR COMBINADO DE p CUANDO $p_1 = p_2 = p$

$$\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} \qquad \textbf{(10.15)}$$

El **estimador combinado de *p*** es un promedio ponderado de $\bar{p}_1$ y $\bar{p}_2$.

Al sustituir $\bar{p}$ por p en la ecuación (10.14), se obtiene una estimación del error estándar de $\bar{p}_1 - \bar{p}_2$. Dicha estimación se usa en el estadístico de prueba. La fórmula general del estadístico de prueba para una prueba de hipótesis acerca de la diferencia entre dos proporciones poblacionales es el estimador puntual dividido entre la estimación de $\sigma_{\bar{p}_1-\bar{p}_2}$.

ESTADÍSTICO DE PRUEBA PARA PRUEBAS DE HIPÓTESIS ACERCA DE $p_1 - p_2$

$$z = \frac{(\bar{p}_1 - \bar{p}_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \textbf{(10.16)}$$

Tal estadístico de prueba se aplica en situaciones de muestras grandes en las que n_1p_1, $n_1(1 - p_1)$, n_2p_2 y $n_2(1 - p_2)$, son todos mayores o iguales que 5.

En el ejemplo de la empresa que se dedica a elaborar declaraciones de impuestos, suponga que ésta desea realizar una prueba de hipótesis para determinar si las proporciones de errores en las dos oficinas son diferentes. Para esto se requiere una prueba de dos colas. Las hipótesis nula y alternativa son las siguientes.

$$H_0: p_1 - p_2 = 0$$
$$H_a: p_1 - p_2 \neq 0$$

Si H_0 es rechazada, la empresa concluiría que la proporción de errores que se cometen en las dos oficinas es distinta. Como nivel de significancia se usará $\alpha = 0.10$.

En los datos muestrales recabados previamente se encuentra que $\bar{p}_1 = 0.14$ en la muestra de $n_1 = 250$ declaraciones de la oficina 1, y $\bar{p}_2 = 0.09$ en la muestra de $n_2 = 300$ declaraciones en la muestra de la oficina 2. A continuación se calculará la estimación combinada de p.

$$\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} = \frac{250(0.14) + 300(0.09)}{250 + 300} = 0.1127$$

Con la estimación combinada y la diferencia entre las proporciones muestrales, se obtiene el valor del estadístico de prueba como se indica a continuación.

$$z = \frac{(\bar{p}_1 - \bar{p}_2)}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(0.14 - 0.09)}{\sqrt{0.1127(1 - 0.1127)\left(\frac{1}{250} + \frac{1}{300}\right)}} = 1.85$$

Al calcular el valor-p para esta prueba de dos colas se observa, primero, que $z = 1.85$ se encuentra en la cola superior de la distribución normal estándar. Considerando $z = 1.85$ y la tabla de distribución normal estándar, se encuentra que el área en la cola superior es $1.0000 - 0.9678 = 0.0322$. Al multiplicar esta área por 2, dado que se trata de una prueba de dos colas, se tiene que el valor-p es $2(0.0322) = 0.0644$. Como este valor es menor que $\alpha = 0.10$, se rechaza H_0 para el nivel de significancia 0.10. La empresa concluye que las proporciones de errores de las dos oficinas difieren. La conclusión de esta prueba de hipótesis es consistente con los resultados de la estimación por intervalo calculada antes, los cuales indicaban que la diferencia entre las proporciones poblacionales de errores en las dos oficinas estaba entre 0.005 y 0.095, siendo la oficina 1 la que arrojaba una mayor tasa de errores.

Ejercicios

Métodos

AUTOevaluación

28. Considere los resultados siguientes obtenidos de muestras independientes tomadas de dos poblaciones.

Muestra 1	Muestra 2
$n_1 = 400$	$n_2 = 300$
$\bar{p}_1 = 0.48$	$\bar{p}_2 = 0.36$

a) ¿Cuál es la estimación puntual de la diferencia entre las dos proporciones poblacionales?
b) Calcule un intervalo de 90% de confianza para la diferencia entre las dos proporciones poblacionales.
c) Proporcione un intervalo de 95% de confianza para la diferencia entre las dos proporciones poblacionales.

AUTOevaluación

29. Considere la prueba de hipótesis:

$$H_0: p_1 - p_2 \leq 0$$
$$H_a: p_1 - p_2 > 0$$

Los siguientes resultados se obtuvieron de muestras independientes tomadas de las dos poblaciones.

Muestra 1	Muestra 2
$n_1 = 200$	$n_2 = 300$
$\bar{p}_1 = 0.22$	$\bar{p}_2 = 0.16$

a) ¿Cuál es el valor-p?
b) Usando $\alpha = 0.05$, ¿cuál es la conclusión en esta prueba de hipótesis?

Aplicaciones

30. En una encuesta de *BusinessWeek*/Harris se pidió a los ejecutivos de empresas grandes su opinión acerca de cómo veían las perspectivas económicas para el futuro. Una de las preguntas era: ¿Piensa usted que en los próximos 12 meses aumentará en su empresa el número de empleados de tiempo completo? En la encuesta actual, 220 de 400 ejecutivos respondieron Sí, mientras que en la realizada el año anterior, 192 de 400 respondieron en el mismo sentido. Encuentre un intervalo de 95% de confianza para estimar la diferencia entre las proporciones en estas dos encuestas. ¿Cuál es su interpretación de la estimación por intervalo?

31. La Asociación Profesional de Golf (PGA, por sus siglas en inglés) midió la precisión en golpes cortos de los golfistas profesionales que jugaron en el PGA Tour, y de los mejores golfistas amateurs que participaron en el World Amateur Championship (*Golf Magazine*, enero de 2007). En una muestra de 1 075 golpes cortos de 6 pies de golfistas profesionales se encontró que 688 fueron efectivos. En una muestra de 1 200 golpes cortos de 6 pies de golfistas amateurs se encontró que 696 fueron efectivos.
    a) Calcule la proporción de golpes cortos efectivos de 6 pies de golfistas profesionales. Estime la proporción de golpes cortos efectivos de 6 pies de los golfistas amateurs. ¿Qué grupo tiene mayor precisión?
    b) ¿Cuál es la estimación puntual de la diferencia entre las proporciones de las dos poblaciones? ¿Qué indica tal estimación acerca del porcentaje de golpes cortos efectivos de ambos grupos de golfistas?
    c) ¿Cuál es el intervalo de 95% de confianza para la diferencia entre las dos proporciones poblacionales? Interprete este intervalo de confianza en términos del porcentaje de golpes cortos efectivos de ambos grupos.

32. En un estudio de la Asociación Estadounidense de Automovilistas (AAA, por sus siglas en inglés) se investigó si era más probable que conductores de género masculino o femenino se detuvieran para solicitar indicaciones sobre cómo llegar a una dirección (AAA, enero de 2006). Se preguntaba a los conductores: "Si usted y su cónyuge van en su automóvil y se pierden, ¿se detiene para preguntar por el domicilio que busca?" En una muestra representativa se encontró que 300 de 811 mujeres dijeron que sí se detenían para preguntar, mientras que 255 de 750 hombres dijeron que también lo hacían.
    a) La hipótesis de investigación de AAA afirmaba que era más probable que las mujeres se detuvieran para preguntar por el domicilio. Formule las hipótesis nula y alternativa para este estudio.
    b) ¿Cuál es el porcentaje de mujeres que dijo que se detendría para preguntar por el domicilio?
    c) ¿Cuál es el porcentaje de hombres que se manifestó en el mismo sentido?
    d) Pruebe la hipótesis usando $\alpha = 0.05$. ¿Cuál es el valor-p y cuál es la conclusión a la que se esperaría que llegara AAA?

33. Los aeropuertos O'Hare de Chicago y Hartsfield-Jackson de Atlanta son dos de las terminales aéreas más saturadas en Estados Unidos. A menudo la congestión ocasiona retrasos en las llegadas y salidas de los vuelos. La Oficina de Transporte monitoreó el desempeño a tiempo y con demora en los principales aeropuertos (*Travel & Leisure*, noviembre de 2006). Se considera que un vuelo está retrasado si tiene más de 15 minutos de dilación respecto de su horario. A continuación se presentan datos de las salidas retrasadas en los aeropuertos O'Hare de Chicago y Hartsfield-Jackson de Atlanta.

	O'Hare de Chicago	Hartsfield-Jackson de Atlanta
Vuelos	900	1 200
Salidas retrasadas	252	312

    a) Establezca la hipótesis pertinente para determinar las dos proporciones poblacionales de salidas retrasadas que difieren en ambos aeropuertos.
    b) ¿Cuál es la estimación puntual de la proporción de vuelos con salidas retrasadas en el O'Hare de Chicago?

c) ¿Cuál es la estimación puntual de la proporción de vuelos con salidas demoradas en el Hartsfield-Jackson de Atlanta?

d) ¿Cuál es el valor-*p* de la prueba de hipótesis? ¿Cuál es su conclusión?

34. *BusinessWeek* reportó que, al parecer, existen diferencias por grupos de edad en lo referente a la gente que disfruta vivir en Rusia (*BusinessWeek*, 10 de marzo de 2008). Los datos muestrales listados a continuación son congruentes con los hallazgos de la revista y representan las respuestas por grupos de edad a la pregunta: "¿Le gusta vivir en Rusia?"

	Grupo de edad de rusos	
	17–26	**40 y más**
Muestra	300	260
Respondió "Sí"	192	117

a) ¿Cuál es la estimación puntual de la proporción de rusos de entre 17 y 26 años a quienes les gusta vivir en Rusia?

b) ¿Cuál es la estimación puntual de la proporción de rusos de 40 y más años a quienes les gusta vivir en Rusia?

c) Proporcione una estimación por intervalo de 95% de confianza para la diferencia entre la proporción de jóvenes rusos de entre 17 y 26 años y adultos de 40 y más años a quienes les agrada vivir en su país.

35. En una prueba de calidad de dos comerciales de televisión, cada anuncio se trasmitió, en áreas separadas de prueba, seis veces en una semana. A la semana siguiente se realizó una encuesta telefónica para identificar a individuos que vieron los comerciales. A estas personas se les pidió su opinión sobre cuál era el principal mensaje de los anuncios. Se obtuvieron los siguientes resultados.

	Comercial A	**Comercial B**
Número de personas que vio el comercial	150	200
Número de personas que recordaba el mensaje	63	60

a) Use $\alpha = 0.05$ y pruebe la hipótesis de que entre los dos comerciales no hay diferencia en las proporciones poblacionales de personas que recordaron el mensaje.

b) Calcule un intervalo de 95% de confianza para la diferencia entre las proporciones de personas que recordaron el mensaje en las dos poblaciones.

36. Durante el SuperBowl de 2003, un comercial de la cerveza Miller Lite Beer, conocido como "The Miller Lite Girls", fue uno de los tres anuncios más eficaces televisados en el evento (*USA Today,* 29 de diciembre de 2003). Un estudio para determinar la eficacia de los comerciales, conducido por *USA Today's* Ad Track, empleó muestras separados por grupos de edades para ver el efecto de la publicidad en el SuperBowl sobre los distintos sectores demográficos. A continuación se presentan los resultados muestrales respecto del anuncio de cerveza.

Grupo de edad	**Tamaño de la muestra**	**Le gustó mucho el comercial**
Menos de 30 años	100	49
De 30 a 49 años	150	54

a) Formule una prueba de hipótesis para determinar si las proporciones poblacionales de los dos grupos de edad difieren.

b) ¿Cuál es la estimación puntual de la diferencia entre las dos proporciones poblacionales?
c) Realice la prueba de hipótesis y proporcione el valor-*p*. Con $\alpha = 0.05$, ¿cuál es su conclusión?
d) Analice la forma en que el comercial llama la atención del grupo de menor y de mayor edad. ¿Le parecerá a la empresa cervecera que los resultados de la encuesta de *USA Today's* Ad Track le son favorables? Explique.

37. En 2003, *The New York Times*/CBS News tomó una muestra de 523 personas adultas que planeaban ir de vacaciones en los próximos seis meses, y encontró que 141 pensaban trasladarse en avión (*New York Times* News Service, 2 de marzo de 2003). En un sondeo similar que realizó en mayo de 1993, de los 477 adultos que planeaban ir de vacaciones en los próximos seis meses, 81 pensaban trasladarse en avión.
a) Establezca las hipótesis para determinar si en este periodo de 10 años hubo un cambio significativo en la proporción de personas que pensaban trasladarse en avión en sus vacaciones.
b) ¿Cuál es la proporción muestral de las personas que pensaban viajar en avión en el estudio de 2003? ¿Y en el de 1993?
c) Con $\alpha = 0.01$ pruebe si hay diferencia significativa. ¿A qué conclusión llega?
d) Analice las razones que puedan servir como explicación para esta conclusión.

Resumen

En este capítulo se estudian los procedimientos para desarrollar estimaciones por intervalo y pruebas de hipótesis cuando se tienen dos poblaciones. Primero se explica cómo hacer inferencias acerca de la diferencia entre dos medias poblacionales cuando se seleccionan muestras aleatorias simples independientes. Se considera primero el caso donde las desviaciones estándar poblacionales, σ_1 y σ_2, se suponen conocidas. La distribución normal estándar z se utiliza para desarrollar la estimación por intervalo y es útil como estadístico de prueba para las pruebas de hipótesis. Después se estudia el caso en que las desviaciones estándar poblacionales no se conocen y se estiman mediante las desviaciones estándar muestrales s_1 y s_2. En esta circunstancia, se usa la distribución t para obtener una estimación por intervalo que sirve como estadístico de prueba en la prueba de hipótesis.

A continuación se estudian las inferencias acerca de las diferencias entre dos medias poblacionales con el diseño de muestras pareadas. En este diseño, cada elemento proporciona un par de datos, uno de cada población. La diferencia entre los pares de valores de datos se usa para el análisis estadístico. El diseño de muestras pareadas suele preferirse al diseño de muestras independientes debido a que con ellas se suele mejorar la precisión de la estimación.

Por último, se estudian las estimaciones por intervalo y las pruebas de hipótesis para la diferencia entre dos proporciones poblacionales. Los procedimientos estadísticos para analizar esta diferencia son similares a los procedimientos estadísticos para analizar la diferencia entre dos medias poblacionales.

Glosario

Estimador combinado de *p* Estimador de una proporción poblacional que se obtiene al calcular un promedio ponderado de los estimadores puntuales extraídos de dos muestras independientes.

Muestras aleatorias simples independientes Muestras seleccionadas de dos poblaciones, de manera que los elementos que constituyen una muestra se tomen independientemente de los elementos que constituyen la otra muestra.

Muestras pareadas Muestras en las que cada valor de dato de una muestra está pareado con el correspondiente valor de dato de otra muestra.

Fórmulas clave

Estimador puntual de la diferencia entre dos medias poblacionales

$$\bar{x}_1 - \bar{x}_2 \tag{10.1}$$

Error estándar de $\bar{x}_1 - \bar{x}_2$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{10.2}$$

Estimación por intervalo de la diferencia entre dos medias poblacionales: σ_1 y σ_2 conocidas

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{10.4}$$

Estadístico de prueba para pruebas de hipótesis acerca de $\mu_1 - \mu_2$: σ_1 y σ_2 conocidas

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{10.5}$$

Estimación por intervalo para la diferencia entre dos medias poblacionales: σ_1 y σ_2 desconocidas

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{10.6}$$

Grados de libertad: distribución *t* con dos muestras aleatorias independientes

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} \tag{10.7}$$

Estadístico de prueba para pruebas de hipótesis acerca de $\mu_1 - \mu_2$: σ_1 y σ_2 desconocidas

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{10.8}$$

Estadístico de prueba para pruebas de hipótesis con muestras pareadas

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} \tag{10.9}$$

Estimador puntual de la diferencia entre dos proporciones poblacionales

$$\bar{p}_1 - \bar{p}_2 \tag{10.10}$$

Error estándar de $\bar{p}_1 - \bar{p}_2$

$$\sigma_{\bar{p}_1-\bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \tag{10.11}$$

Estimación por intervalo de la diferencia entre dos proporciones poblacionales

$$\bar{p}_1 - \bar{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}} \tag{10.13}$$

Error estándar de $\bar{p}_1 - \bar{p}_2$ cuando $p_1 = p_2 = p$

$$\sigma_{\bar{p}_1-\bar{p}_2} = \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{10.14}$$

Estimador combinado de p cuando $p_1 = p_2 = p$

$$\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} \tag{10.15}$$

Estadístico de prueba para pruebas de hipótesis acerca de $p_1 - p_2$

$$z = \frac{(\bar{p}_1 - \bar{p}_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{10.16}$$

Ejercicios complementarios

38. Safegate Foods Inc. está rediseñando las cajas de cobro en sus supermercados en todo el país y prueba dos diseños. Ambos sistemas se instalaron en dos supermercados y se midió el tiempo que tardaban los clientes en pasar por la caja. Los resultados se presentan resumidos en la siguiente tabla.

Sistema A	Sistema B
$n_1 = 120$	$n_2 = 100$
$\bar{x}_1 = 4.1$ minutos	$\bar{x}_2 = 3.4$ minutos
$\sigma_1 = 2.2$ minutos	$\sigma_2 = 1.5$ minutos

Con 0.05 como nivel de significancia, realice una prueba de hipótesis para determinar si hay diferencia entre las medias poblacionales del tiempo que tardan los clientes en pasar por la caja con estos dos sistemas. ¿Cuál se preferirá?

WEB archivo
HomePrices

39. El valor de las casas tiende a incrementarse en el tiempo en condiciones normales, pero la recesión de 2008 y 2009 ocasionó la caída en Estados Unidos de los precios de venta inmobiliarios (*BusinessWeek*, 9 de marzo de 2009). Se quiere saber si los datos apoyan esta conclusión. El archivo HomePrices contiene datos de 30 ventas de casas en 2006 y 40 en 2009.

a) Proporcione una estimación puntual para la diferencia entre las medias poblacionales de los precios en los dos años.
b) Desarrolle una estimación por intervalo de 99% de confianza para la diferencia entre los precios de reventa de casas en 2006 y 2009.
c) ¿Consideraría justificado concluir que los precios de reventa han disminuido de 2006 a 2009? ¿Por qué?

40. Los fondos de inversión (Mutual Funds) se clasifican en fondos *con comisión* (Load) y *sin comisión* (No Load). En los primeros se requiere que el inversor pague una cantidad base inicial o un porcentaje de la cantidad invertida en el fondo. En los fondos sin comisión no se requiere este pago inicial. Algunos asesores financieros aseguran que vale la pena pagar la comisión de los fondos con comisión, porque ofrecen tasas medias de rendimiento mayores que los otros. Se seleccionaron muestras de 30 fondos de inversión con comisión y de 30 sin comisión. Se recabaron los datos sobre su rendimiento anual en un periodo de cinco años. Estos datos se presentan en el conjunto de datos Mutual. La información de los cinco primeros fondos con comisión y de los cinco primeros fondos sin comisión se presentan en la tabla siguiente.

WEB archivo
Mutual

Mutual Funds–Load	Return	Mutual Funds–No Load	Return
American National Growth	15.51	Amana Income Fund	13.24
Arch Small Cap Equity	14.57	Berger One Hundred	12.13
Bartlett Cap Basic	17.73	Columbia International Stock	12.17
Calvert World International	10.31	Dodge & Cox Balanced	16.06
Colonial Fund A	16.23	Evergreen Fund	17.61

a) Formule H_0 y H_a de manera que el rechazo de H_0 lleve a la conclusión de que en este periodo de cinco años los fondos de inversión con comisión ofrecieron un mayor rendimiento medio anual.
b) Use los 60 fondos de inversión de la base de datos Mutual para realizar la prueba de hipótesis. ¿Cuál es el valor-*p*? Con $\alpha = 0.05$, ¿cuál es su conclusión?

41. La Asociación Estadounidense de Constructores de Casas presenta datos sobre los costos de las remodelaciones más frecuentes que se realizan en casas habitación. A continuación se presentan datos muestrales, en miles de dólares, de los dos tipos de proyectos de remodelación más frecuentes.

Cocina	Recámara principal	Cocina	Recámara principal
25.2	18.0	23.0	17.8
17.4	22.9	19.7	24.6
22.8	26.4	16.9	21.0
21.9	24.8	21.8	
19.7	26.9	23.6	

a) Desarrolle una estimación puntual de la diferencia entre las medias poblacionales de los costos de los dos tipos de remodelación.
b) Proporcione un intervalo de 90% de confianza para la diferencia entre estas dos medias poblacionales.

42. A principios de 2009 la economía experimentó una recesión. ¿Cómo afectó al mercado accionario? Se presentan datos acerca del precio por acción registrado el 1 de enero (January 1) y el 30 de abril (April 30) de una muestra de 15 empresas (*The Wall Street Journal*, 1 de mayo de 2009).

WEB archivo
PriceChange

Company	January 1 ($)	April 30 ($)
Applied Materials	10.13	12.21
Bank of New York	28.33	25.48
Chevron	73.97	66.10
Cisco Systems	16.30	19.32
Coca-Cola	45.27	43.05
Comcast	16.88	15.46
Ford Motors	2.29	5.98
General Electric	16.20	12.65
Johnson & Johnson	59.83	52.36
JP Morgan Chase	31.53	33.00
Microsoft	19.44	20.26
Oracle	17.73	19.34
Pfizer	17.71	13.36
Philip Morris	43.51	36.18
Procter & Gamble	61.82	49.44

a) ¿Cuál es el cambio en el precio medio por acción en el periodo de cuatro meses?
b) Proporcione una estimación por intervalo de 90% de confianza del cambio en el precio medio por acción. Interprete los resultados.
c) ¿De cuánto fue el cambio porcentual en el precio medio por acción en el periodo de cuatro meses?
d) Si este mismo cambio porcentual hubiera ocurrido en los siguientes cuatro meses y de nuevo en los cuatro meses posteriores, ¿cuál hubiera sido el precio medio por acción al final de 2009?

43. Jupiter Media realizó una encuesta para determinar en qué emplean su tiempo libre las personas. Ver la televisión es la actividad más popular para pasar el tiempo libre, tanto de los hombres como de las mujeres (*The Wall Street Journal,* 26 de enero de 2004). La proporción de ambos grupo que prefieren ver televisión para pasar su tiempo libre se estima a partir de los siguientes datos muestrales.

Género	Tamaño de la muestra	Ver televisión
Hombres	800	248
Mujeres	600	156

a) Establezca las hipótesis pertinentes para probar la diferencia entre la proporción poblacional de hombres y la de mujeres que prefieren ver televisión para pasar su tiempo libre.
b) ¿Cuál es la proporción muestral de hombres que destinan su tiempo libre a ver televisión? ¿Cuál es la proporción muestral de mujeres?
c) Lleve a cabo la prueba de hipótesis y calcule el valor-*p*. ¿Cuál es la conclusión con 0.05 como nivel de significancia?
d) ¿Cuál es el margen de error y la estimación por intervalo de 95% de confianza para la diferencia entre las proporciones poblacionales?

44. Una empresa grande de seguros de automóvil toma muestras de hombres asegurados, casados y solteros, y determina cuántos reclamaron el seguro en los tres años anteriores.

Asegurados solteros	Asegurados casados
$n_1 = 400$	$n_2 = 900$
Cantidad que utilizó el seguro = 76	Cantidad que utilizó el seguro = 90

a) Use $\alpha = 0.05$. Haga una prueba para determinar si la razón de reclamaciones es diferente entre asegurados solteros y casados.
b) Proporcione un intervalo de 95% de confianza para la diferencia entre las proporciones de las dos poblaciones.

45. Se realizaron pruebas médicas para probar la resistencia a fármacos contra la tuberculosis. En Nueva Jersey, de 142 casos, 9 mostraron resistencia a los medicamentos. En Texas, de 268 casos, 5 fueron resistentes a los fármacos. ¿Estos datos indican que existe una diferencia estadísticamente significativa entre la proporción de casos resistentes en estos dos estados? Utilice 0.02 como nivel de significancia. ¿Cuál es el valor-*p* y cuál es la conclusión a que se llega?

WEB archivo
Occupancy

46. En marzo de 2008 se esperaba que las tasas de ocupación aumentaran en Myrtle Beach, Carolina del Sur (*The Sun News*, 29 de febrero de 2008). Los datos en el archivo Occupancy permitirán responder a los hallazgos presentados por el periódico. La información presenta las unidades alquiladas y las no alquiladas de una muestra aleatoria de propiedades para vacacionar durante la primera semana de marzo de 2007 y marzo de 2008.
a) Estime la proporción de unidades alquiladas durante la primera semana de marzo de 2007 y la primera semana de marzo de 2008.
b) Proporcione un intervalo de 95% de confianza para la diferencia entre las proporciones.
c) Con base en sus hallazgos, ¿las tasas de alquiler de marzo de 2008 parecen haber aumentado con respecto a las del año anterior?

47. En la semana que terminaba el 15 de enero de 2009, el sentimiento alcista de los inversionistas individuales fue 27.6% (*AAII Journal*, febrero de 2009). Se reportó que el sentimiento alcista era de 48.7% una semana antes y de 39.7% un mes antes. La medición de este indicador se basa en una encuesta que realiza la Asociación Estadounidense de Inversionistas Individuales (AAII, por sus siglas en inglés). Suponga que cada medición del sentimiento alcista se basa en un tamaño muestral de 240.
a) Proporcione un intervalo de 95% de confianza para la diferencia entre las mediciones del sentimiento alcista en las dos semanas mencionadas.
b) Formule hipótesis de manera que el rechazo de la hipótesis nula permita concluir que el sentimiento alcista más reciente mencionado es más débil que el de un mes atrás.
c) Realice una prueba de hipótesis del inciso *b*) con $\alpha = 0.01$. ¿Cuál es su conclusión?

Caso a resolver Par, Inc.

Par, Inc. es un importante fabricante de equipo de golf. El gerente de Par piensa que la participación de la empresa en el mercado aumentará con la introducción de una pelota de golf de alta duración y resistente a los cortes. Con este objetivo, el grupo de investigación de Par ha estado probando un nuevo recubrimiento para las pelotas, diseñado para dar resistencia a los cortes y proporcionar un producto con mayor durabilidad. Las pruebas realizadas con el recubrimiento han sido promisorias.

Uno de los investigadores expresó su preocupación por el efecto del nuevo recubrimiento en la distancia de recorrido de la pelota. Par desea que el nuevo modelo, resistente a los cortes, ofrezca una distancia de recorrido comparable a la de las pelotas de golf actuales. Para comparar la distancia de recorrido de los dos tipos de pelotas, 40 de cada modelo fueron sometidas a pruebas de distancia, las cuales se realizaron con una máquina lanzadora de pelotas con objeto de que la diferencia entre las distancias medias de vuelo entre los dos modelos pudiera atribuirse a sus diferentes particularidades. Los resultados de las pruebas, con las distancias expresadas a la yarda más cercana, se presentan en la tabla de la página siguiente. Estos datos se encuentran en el sitio web del libro. (Current refiere las pelotas actuales y New las nuevas).

WEB archivo
Golf

Model		Model		Model		Model	
Current	New	Current	New	Current	New	Current	New
264	277	270	272	263	274	281	283
261	269	287	259	264	266	274	250
267	263	289	264	284	262	273	253
272	266	280	280	263	271	263	260
258	262	272	274	260	260	275	270
283	251	275	281	283	281	267	263
258	262	265	276	255	250	279	261
266	289	260	269	272	263	274	255
259	286	278	268	266	278	276	263
270	264	275	262	268	264	262	279

Informe gerencial

1. Formule y justifique una prueba de hipótesis que le sirva a Par para comparar las distancias de recorrido de la pelota actual y de la nueva pelota.
2. Analice los datos para formular la conclusión de la prueba de hipótesis. ¿Cuál es el valor-p de la prueba? ¿Qué le recomendaría a Par, Inc.?
3. Proporcione un resumen de estadística descriptiva con los datos de cada modelo.
4. ¿Cuál es el intervalo de 95% de confianza para la media poblacional de la distancia de recorrido de cada modelo, y cuál para la diferencia entre las medias de las dos poblaciones?
5. ¿Ve usted que haya necesidad de tomar muestras más grandes y de efectuar más pruebas con las pelotas de golf? Analícelo.

Apéndice 10.1

Inferencias acerca de dos poblaciones usando Minitab

Aquí se describe el uso de Minitab para calcular estimaciones por intervalo y realizar pruebas de hipótesis acerca de la diferencia entre dos medias poblacionales y entre dos proporciones poblacionales. Con Minitab se pueden calcular estimaciones por intervalo y efectuar pruebas de hipótesis dentro de un mismo módulo. Es decir, tiene un mismo procedimiento para los dos tipos de inferencias. En los ejemplos siguientes se mostrará cómo realizar los cálculos para una estimación por intervalo y para una prueba de hipótesis con las mismas dos muestras. Minitab no cuenta con una rutina para inferencias acerca de la diferencia entre dos medias poblacionales cuando las desviaciones estándar poblacionales σ_1 y σ_2 son conocidas.

Diferencia entre dos medias poblacionales: σ_1 y σ_1 desconocidas

WEB archivo
CheckAcct

Se utilizarán los datos sobre los saldos en las cuentas de cheques del ejemplo presentado en la sección 10.2. Los saldos en las cuentas de cheques de la sucursal Cherry Grove se encuentran en la columna C1 y los correspondientes a la sucursal Beechmont, en la columna C2. En este ejemplo se usará el procedimiento de Minitab 2-Sample t con el fin de obtener un intervalo de 95% de confianza para estimar la diferencia entre las medias poblacionales de los saldos de las cuentas de cheques para las dos sucursales. En el resultado de este procedimiento, Minitab proporciona también el valor-p para la prueba de hipótesis H_0: $\mu_1 - \mu_2 = 0$ frente a H_a: $\mu_1 - \mu_2 \neq 0$. Los pasos necesarios para realizar este procedimiento se indican a continuación.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Basic Statistics**.
Paso 3. Seleccione **2-Sample t**.

Paso 4. Cuando el cuadro de diálogo 2-Sample t (Test and Confidence Interval) aparezca:
Seleccione **Samples in different columns**.
Ingrese Cl en el cuadro **First**.
Ingrese C2 en el cuadro **Second**.
Seleccione **Options**.

Paso 5. Cuando el cuadro de diálogo 2-Sample t–Options aparezca:
Ingrese 95 en el cuadro **Confidence level**.
Introduzca 0 en el cuadro **Test difference**.
Ingrese not equal en el cuadro **Alternative**.
Haga clic en **OK**.

Paso 6. Haga clic en **OK**.

La estimación por intervalo de 95% de confianza va de \$37 a \$193, como se describió en la sección 10.2. El valor-p = 0.005 indica que la hipótesis nula de que las medias poblacionales son iguales puede rechazarse para el nivel de significancia $\alpha = 0.01$. El paso 5 puede modificarse para proporcionar otras aplicaciones con diferentes niveles de confianza, distintos valores hipotéticos y diversas formas de hipótesis.

Diferencia entre dos medias poblacionales con muestras pareadas

WEB archivo
Matched

Para ilustrar el procedimiento de muestras pareadas se usarán los datos de la tabla 10.2 sobre los tiempos de producción necesarios para realizar una tarea. Los tiempos con el método 1 se ingresan en la columna Cl y los tiempos con el método 2 en la columna C2. Los pasos a seguir usando Minitab para una prueba de muestras pareadas son los siguientes.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Basic Statistics**.
Paso 3. Seleccione **Paired t**.
Paso 4. Cuando el cuadro de diálogo Paired t (Test and Confidence Interval) aparezca:
Seleccione **Samples in columns**.
Ingrese Cl en el cuadro **First sample**.
Ingrese C2 en el cuadro **Second sample**.
Seleccione **Options**.

Paso 5. Cuando el cuadro de diálogo Paired t–Options aparezca:
Ingrese 95 en el cuadro **Confidence level**.
Introduzca 0 en el cuadro **Test mean**.
Ingrese not equal en el cuadro **Alternative**.
Haga clic en **OK**.

Paso 6. Haga clic en **OK**.

La estimación por intervalo de 95% de confianza va de −0.05 a 0.65, como se describió en la sección 10.3. El valor-p = 0.08 indica que la hipótesis nula de que no hay diferencia en los tiempos para realizar la tarea no puede rechazarse para el nivel de significancia $\alpha = 0.05$. El paso 5 puede modificarse para diferentes niveles de confianza, distintos valores hipotéticos y diversas formas de la prueba de hipótesis.

Diferencia entre dos proporciones poblacionales

WEB archivo
TaxPrep

Se emplearán los datos presentados en la sección 10.4 sobre los errores en las declaraciones de impuestos. Los resultados muestrales de 250 declaraciones fiscales elaboradas en la oficina 1 se encuentran en la columna C_1 – T y los resultados muestrales de 300 declaraciones elaboradas en la oficina 2 están en la columna C_2 – T. Sí indica que se identificó un error en la declaración de impuestos y No que no se encontró ningún error. Con el procedimiento que se describe a continuación se obtiene una estimación por intervalo de 95% de confianza para la diferencia entre las dos proporciones poblacionales, además de los resultados de la prueba de hipótesis de H_0: $p_1 - p_2 = 0$ y H_a: $p_1 - p_2 \neq 0$.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Basic Statistics**.

Paso 3. Elija **2 Proportions**.
Paso 4. Cuando el cuadro de diálogo 2 Proportions (Test and Confidence Interval) aparezca:
Seleccione **Samples in different columns**.
Ingrese Cl en el cuadro **First**.
Ingrese C2 en el cuadro **Second**.
Seleccione **Options**.
Paso 5. Cuando el cuadro de diálogo 2 Proportions-Options aparezca:
Ingrese 90 en el cuadro **Confidence level**.
Introduzca 0 en el cuadro **Test difference**.
Ingrese not equal en el cuadro **Alternative**.
Seleccione **Use pooled estimate of p for test**.
Haga clic en **OK**.
Paso 6. Haga clic en **OK**.

El intervalo de 90% de confianza va de 0.005 a 0.095, como se vio en la sección 10.4. El valor-p = 0.065 indica que la hipótesis nula de que no hay diferencia entre la tasa de errores se rechaza para $\alpha = 0.10$. El paso 5 puede modificarse para proporcionar diferentes niveles de confianza, distintos valores hipotéticos y diversas formas de las hipótesis.

En el ejemplo de las declaraciones de impuestos los datos son cualitativos. Yes y No sirven para indicar si hay o no un error. En los módulos para proporciones, Minitab calcula la proporción de la respuesta que tiene el segundo lugar en orden alfabético. Por tanto, en este ejemplo de las declaraciones de impuestos, Minitab calculará la proporción de respuestas Yes, que es la que se busca.

Si con el orden alfabético Minitab no calcula la proporción de la respuesta de interés, es posible modificarlo. Para ello, seleccione una celda en la columna de los datos, vaya a la barra del menú de Minitab y seleccione Editor > Column > Value Order. Esta secuencia proporcionará la opción de ingresar un orden especificado por el usuario. Sólo asegúrese de que la respuesta de interés esté listada en segundo lugar en el cuadro define-an-order. La rutina 2 Proportion de Minitab facilitará el intervalo de confianza y los resultados de la prueba de hipótesis para la proporción poblacional de interés.

Por último, la rutina 2 Proportion emplea un procedimiento de cálculo distinto al descrito en el libro. Por tanto, quizá los resultados suministrados por Minitab sean un tanto distintos y proporcionen una estimación por intervalo y el valor-p ligeramente diferentes. Sin embargo, los resultados de ambos métodos serán muy parecidos y se espera que conduzcan a la misma interpretación y conclusiones.

Apéndice 10.2 Inferencias acerca de dos poblaciones usando Excel

Se describirá el uso de Excel para realizar pruebas de hipótesis acerca de la diferencia entre dos medias poblacionales.* Se empieza con las inferencias acerca de las diferencias entre las medias de dos poblaciones cuando las desviaciones estándar poblacionales, σ_1 y σ_2, son conocidas.

Diferencia entre dos medias poblacionales: σ_1 y σ_2 conocidas

WEB archivo
ExamScores

Se usará el ejemplo de la sección 10.1 sobre las puntuaciones obtenidas en el examen para los dos centros de enseñanza. La etiqueta Center A se encuentra en la celda Al y la etiqueta Center B en la celda Bl. Las calificaciones de los exámenes del centro A se ubican en las celdas A2:A31 y las de los exámenes del centro B en las celdas B2:B41. Se supone que se conocen las desviaciones estándar poblacionales y que son $\sigma_1 = 10$ y $\sigma_2 = 10$. La rutina de Excel solicitará que se ingresen las varianzas $\sigma_1^2 = 100$ y $\sigma_2^2 = 100$. Para realizar una prueba de hi-

* Las herramientas de análisis de datos de Excel facilitan procedimientos para pruebas de hipótesis acerca de la diferencia entre dos medias poblacionales. Excel no cuenta con una rutina de estimación por intervalos para la diferencia entre dos medias poblacionales ni para inferencias acerca de la diferencia entre dos proporciones poblacionales.

pótesis acerca de la diferencia entre dos medias poblacionales se siguen los pasos que se indican a continuación.

Paso 1. Haga clic en la ficha **Data** en la cinta.
Paso 2. En el grupo **Analysis**, haga clic en **Data Analysis**.
Paso 3. Cuando el cuadro de diálogo Data Analysis aparezca:
Elija **z-Test: Two Sample for Means**.
Haga clic en **OK**.
Paso 4. Cuando el cuadro de diálogo z-Test: Two Sample for Means aparezca:
Ingrese Al:A31 en el cuadro **Variable 1 Range**.
Ingrese B1:B41 en el cuadro **Variable 2 Range**.
Introduzca 0 en el cuadro **Hypothesized Mean Difference**.
Ingrese 100 en el cuadro **Variable 1 Variance (known)**.
Introduzca 100 en el cuadro **Variable 2 Variance (known)**.
Seleccione **Labels**.
Ingrese 0.05 en el cuadro **Alpha**.
Seleccione **Output Range** e ingrese Cl en el cuadro.
Haga clic en **OK**.

El valor-*p* se denota P(Z<=z) two-tail. Su valor de 0.0977 no permite rechazar la hipótesis nula con $\alpha = 0.05$.

Diferencia entre dos medias poblacionales: σ_1 y σ_2 desconocidas

WEB archivo
SoftwareTest

Se usarán los datos que aparecen en la tabla 10.1 sobre el estudio de la prueba del software. Los datos ya se han ingresado en la hoja de trabajo de Excel con las etiquetas Current (Actual) en la celda Al y New (Nueva) en la celda Bl. Los tiempos requeridos con la tecnología actual se encuentran en las celdas A2:A13 y los tiempos requeridos con el nuevo software en las celdas B2:B13. Para realizar una prueba de hipótesis acerca de la diferencia entre dos medias poblacionales cuando no se conocen σ_1 y σ_2 se toman los siguientes pasos.

Paso 1. Haga clic en la ficha **Data** en la cinta.
Paso 2. En el grupo **Analysis**, dé clic en **Data Analysis**.
Paso 3. Cuando aparezca el cuadro de diálogo Data Analysis:
Elija **t-Test: Two Sample Assuming Unequal Variances**.
Haga clic en **OK**.
Paso 4. Cuando aparezca el cuadro de diálogo t-Test: Two Sample Assuming Unequal Variances:
Ingrese Al:A13 en el cuadro **Variable 1 Range**.
Introduzca B1:B13 en el cuadro **Variable 2 Range**.
Ingrese 0 en el cuadro **Hypothesized Mean Difference**.
Seleccione **Labels**.
Ingrese 0.05 en el cuadro **Alpha**.
Seleccione **Output Range** e ingrese Cl en el cuadro.
Haga clic en **OK**.

El valor-*p* apropiado, de una cola, se denota P(T<=t) one-tail. Su valor de 0.017 permite rechazar la hipótesis nula con $\alpha = 0.05$.

Diferencia entre dos medias poblacionales con muestras pareadas

WEB archivo
Matched

Para ilustrar este método se utilizan los pares de datos para la realización de una tarea que aparecen en la tabla 10.2. Estos datos se han ingresado en la hoja de cálculo con las etiquetas

Method 1 (método 1) en la celda Al y Method 2 (método 2) en la celda B2. Los tiempos requeridos para la realización de la tarea con el método 1 se localizan en las celdas A2:A7 y para la realización de la tarea con el método 2 se encuentran en las celdas B2:B7. En el procedimiento de Excel se emplean los pasos previamente descritos para la prueba *t*, excepto que en el paso 3 se debe elegir **t-Test: Paired Two Sample for Means** en la herramienta de análisis de datos. El rango para la variable 1 es A1:A7 y para la variable 2 es B1:B7.

El valor-*p* de dos colas se denota P(T<= t) two-tail. Su valor de 0.08 no permite rechazar la hipótesis nula con $\alpha = 0.05$.

Apéndice 10.3 Inferencias acerca de dos poblaciones usando StatTools

En este apéndice se muestra el uso de StatTools para calcular estimaciones por intervalos y realizar pruebas de hipótesis acerca de la diferencia entre dos medias poblacionales cuando se desconocen σ_1 y σ_2.

Estimación por intervalos de μ_1 y μ_2

WEB archivo
CheckAcct

Se empleará la información sobre los saldos en las cuentas de cheques del ejemplo presentado en la sección 10.2. Se inicia con el uso del Data Set Manager a efecto de crear un conjunto de datos de StatTools para los datos utilizados en el procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes se utilizan para calcular una estimación por intervalo de 95% de confianza para la diferencia entre las dos medias poblacionales.

Paso 1. Haga clic en la ficha **StatTools** en la cinta.
Paso 2. En el grupo **Analysis**, haga clic en **Statistical Inference**.
Paso 3. Seleccione la opción **Confidence Interval**.
Paso 4. Elija Mean/Std. Deviation
Paso 5. Cuando el cuadro de diálogo StatTools–Confidence Interval for Mean/Std. Deviation aparezca:
Para **Analysis Type**, elija **Two-Sample Analysis**.
En la sección **Variables**:
Seleccione **Cherry Grove**.
Elija **Beechmont**.
En la sección **Confidence Intervals to Calculate**:
Seleccione la opción **For the Difference of Means**.
Elija 95% para **Confidence Level**.
Haga clic en **OK**.

Debido a que el tamaño muestral de Cherry Grove ($n_1 = 28$) difiere del tamaño muestral de Beechmont ($n_2 = 22$), StatTools informará sobre esta diferencia antes de hacer clic en OK en el paso 4. Aparecerá un cuadro de diálogo con la leyenda: "The variable Beechmont contains missing data, which this analysis will ignore." Haga clic en OK. De inmediato aparecerá el cuadro de diálogo Choose Variable Ordering indicando que se compararán las diferencias entre los conjuntos de datos de Cherry Grove y de Beechmont. Haga clic en OK y aparecerá el resultado de la estimación por intervalo de StatTools.

Pruebas de hipótesis acerca de μ_1 y μ_2

WEB archivo
SoftwareTest

Se usarán los datos del estudio de evaluación de software presentados en la tabla 10.1. Se inicia con el uso del Data Set Manager para crear un conjunto de datos de StatTools sobre los datos utilizados en el procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes se utilizan para la prueba de hipótesis de H_0: $\mu_1 - \mu_2 \leq 0$ contra H_a: $\mu_1 - \mu_2 > 0$.

Paso 1. Haga clic en la ficha **StatTools** en la cinta.
Paso 2. En el grupo **Analysis**, haga clic en **Statistical Inference**.
Paso 3. Seleccione la opción **Hypothesis Test**.

Paso 4. Elija Mean/Std. Deviation.
Paso 5. Cuando el cuadro de diálogo StatTools–Hypothesis Test for Mean/Std. Deviation aparezca:
Para **Analysis Type**, elija **Two–Sample Analysis**.
En la sección **Variables**:
Seleccione **Current**.
Elija **New**.
En la sección **Hypothesis Test to Perform**:
Seleccione **Difference of Means**.
Ingrese 0 en el cuadro **Null Hypothesis Value**.
Seleccione **Greater Than Null Value (One–Tailed Test)** en el cuadro **Alternative Hypothesis**.
Haga clic en **OK**.
Cuando aparezca el cuadro de diálogo Choose Variable Ordering, haga clic en **OK**.

Aparecerá el resultado de la prueba de hipótesis.

Inferencias acerca de la diferencia entre dos medias poblacionales: muestras pareadas

WEB archivo
Matched

Se puede utilizar StatTools para calcular estimaciones por intervalo y realizar pruebas de hipótesis para la diferencia entre dos medias poblacionales en el caso de muestras pareadas. Para ilustrar estos procedimientos se emplean los pares de datos acerca de la realización de una tarea descritos en la tabla 10.2.

Se inicia con el uso del Data Set Manager para crear un conjunto de datos de StatTools sobre los datos utilizados en el procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes permiten calcular una estimación por intervalo de 95% de confianza para la diferencia entre las medias poblacionales de los tiempos necesarios para realizar una tarea.

Paso 1. Haga clic en la ficha **StatTools** en la cinta.
Paso 2. En el grupo **Analysis**, haga clic en **Statistical Inference**.
Paso 3. Seleccione la opción **Confidence Interval**.
Paso 4. Elija Mean/Std. Deviation.
Paso 5. Cuando el cuadro de diálogo StatTools–Confidence Interval for Mean/Std. Deviation aparezca:
Para **Analysis Type**, elija **Paired-Sample Analysis**.
En la sección **Variables**:
Elija **Method 1**.
Seleccione **Method 2**.
En la sección **Confidence Intervals to Calculate**:
Seleccione la opción **For the Difference of Means**.
Elija 95% para **Confidence Level**.
Si está seleccionado, retire la selección del cuadro **For the Standard Deviation**.
Haga clic en **OK**.
Cuando el cuadro de diálogo Choose Variable Ordering aparezca, haga clic en **OK**.

Aparecerá el intervalo de confianza.

Realizar pruebas de hipótesis en el caso de muestras pareadas es muy similar a realizarlas para la diferencia entre los dos medias mostradas previamente. Luego de seleccionar la opción Hypothesis Test en el paso 3, elija Paired-Sample Analysis en el paso 4.

CAPÍTULO 11

Inferencias acerca de varianzas poblacionales

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: U.S. GOVERNMENT ACCOUNTABILITY OFFICE

11.1 INFERENCIAS ACERCA DE UNA VARIANZA POBLACIONAL
Estimación por intervalo
Pruebas de hipótesis

11.2 INFERENCIAS ACERCA DE DOS VARIANZAS POBLACIONALES

ESTADÍSTICA *en* LA PRÁCTICA

U.S. GOVERNMENT ACCOUNTABILITY*

WASHINGTON, D.C.

La U.S. Government Accountability Office (GAO) es una organización de auditoría independiente, no política, perteneciente al área legislativa del gobierno federal. Los auditores de la GAO determinan la eficacia tanto de los programas federales existentes como de los que se proponen. Para realizar su labor, deben ser competentes en la revisión de documentos, investigación legislativa y técnicas de análisis estadístico.

En un caso, los auditores de la GAO estudiaron un programa del Departamento del Interior que tenía por objeto limpiar los ríos y lagos del país. Como parte del proyecto, se otorgaron subvenciones a las ciudades pequeñas de Estados Unidos. El Congreso pidió a la GAO que determinara la eficiencia con la que operaba este programa. Con tal objetivo, el organismo revisó documentos y visitó varias plantas de tratamiento de residuos.

Uno de los objetivos de la auditoría consistía en verificar en las plantas que las aguas residuales (desechos tratados) cumplieran determinadas normas. En las auditorías se revisaban, entre otros aspectos, datos muestrales sobre contenido de oxígeno, pH y cantidad de sólidos en suspensión en las aguas residuales. Un requerimiento del programa exigía que en cada planta se realizaran diversas pruebas diarias y que los datos obtenidos se enviaran al departamento de ingeniería del estado. Los datos de la investigación servían para determinar si las características de las aguas residuales se encontraban dentro de límites aceptables.

Así, por ejemplo, se analizaron cuidadosamente los valores promedio de pH y la varianza en los valores del nivel de pH de las aguas residuales. La prueba de hipótesis acerca de la varianza del pH en la población de aguas residuales fue la siguiente.

$$H_0: \sigma^2 = \sigma_0^2$$
$$H_a: \sigma^2 \neq \sigma_0^2$$

En esta prueba, σ_0^2 corresponde a la varianza poblacional esperada en los valores de pH de una planta funcionando adecuadamente. Para una de las centrales la hipótesis nula fue rechazada. Análisis más cuidadosos indicaron que en esa planta la varianza del pH era significativamente menor de lo normal.



Las aguas residuales de esta planta deben estar dentro de un determinado intervalo estadístico de valores de pH.
© John B. Boykin/CORBIS

Los auditores visitaron el sitio para revisar el equipo de medición y analizar los resultados estadísticos con el director de la planta. Encontraron que el equipo para la medición del pH no se usaba debido a que el operador no conocía su funcionamiento. Un ingeniero había informado al operador de los niveles de pH aceptables y éste sólo anotaba valores similares sin realizar ninguna medición. La varianza inusualmente baja de los datos de esta planta hizo que H_0 fuera rechazada. La GAO pensó que era probable que otras plantas tuvieran problemas similares y recomendó un programa de capacitación para los operadores con objeto de mejorar la recolección de datos para el programa de control de la contaminación.

En este capítulo se estudiará cómo hacer inferencias estadísticas acerca de las varianzas de una o dos poblaciones. También se presentarán dos nuevas distribuciones: la distribución ji-cuadrada y la distribución F. Ambas son útiles para obtener estimación por intervalos y realizar pruebas de hipótesis para la varianza poblacional.

* Los autores agradecen a Art Foreman y Dale Ledman de la U.S. Government Accountability Office por habe proporcionado este material para *Estadística en la práctica.*

En los cuatro capítulos anteriores se examinaron métodos de inferencia estadística para medias y proporciones poblacionales. En este capítulo se extiende este estudio a las varianzas poblacionales. Un ejemplo en el que la varianza brinda información importante para tomar una decisión es el caso de un proceso en el que se llenan recipientes con un detergente líquido. El mecanismo de llenado se ajusta para el proceso de manera que el peso de llenado medio sea 16 onzas por envase. Aunque la media de 16 onzas es relevante, la varianza en los pesos de llenado también es crítica.

En muchas aplicaciones de fabricación, controlar la varianza del proceso es de suma importancia para conservar la calidad.

Es decir, aun cuando el mecanismo de llenado tenga un ajuste adecuado para una media de 16 onzas, no es de esperar que todos los envases contengan exactamente esta cantidad. Para calcular la varianza muestral de la cantidad de onzas en cada envase, se toma una muestra de recipientes llenos. Este valor servirá como estimación de la varianza para la población de envases que están siendo llenados en el proceso de producción. Si la varianza muestral es moderada, el proceso de producción continúa. Pero si la varianza muestral es grande, puede estar ocurriendo exceso o insuficiencia de llenado aunque la media de 16 onzas sea la correcta. En este caso habrá que reajustar el mecanismo de producción con objeto de reducir la varianza de llenado de los recipientes.

En la primera sección se analizarán inferencias acerca de la varianza de una sola población. Después se estudiarán procedimientos para inferencias acerca de varianzas de dos poblaciones.

11.1 Inferencias acerca de una varianza poblacional

La varianza muestral

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} \tag{11.1}$$

es el estimador puntual de la varianza poblacional σ^2. Cuando se hacen inferencias acerca de la varianza poblacional mediante la varianza muestral, es útil la distribución de muestreo de la cantidad $(n - 1)s^2/\sigma^2$. Esta distribución se describe como sigue.

La distribución ji-cuadrada se basa en el muestreo de una población normal.

DISTRIBUCIÓN DE MUESTREO DE $(n - 1)s^2/\sigma^2$

Siempre que de una población normal se tome una muestra aleatoria simple de tamaño n, la distribución de muestreo de

$$\frac{(n - 1)s^2}{\sigma^2} \tag{11.2}$$

tendrá una distribución ji-cuadrada con $n - 1$ grados de libertad.

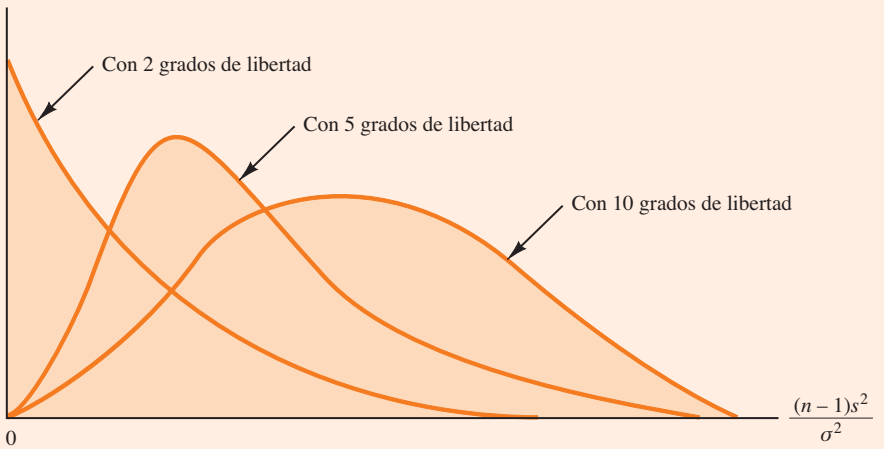
En la figura 11.1 se presentan algunas de las posibles formas que puede tomar la distribución de $(n - 1)s^2/\sigma^2$.

Como se sabe, siempre que se tome una muestra aleatoria simple de tamaño n de una población normal, la distribución muestral de $(n - 1)s^2/\sigma^2$ será una distribución ji-cuadrada, la cual sirve para obtener una estimación por intervalo y realizar pruebas de hipótesis acerca de la varianza poblacional.

Estimación por intervalo

Con el fin de exponer cómo usar la distribución ji-cuadrada para obtener una estimación de la varianza poblacional σ^2 mediante un intervalo de confianza, suponga que desea estimar la varianza poblacional del proceso industrial de llenado que se menciona al inicio de este capítulo. En una muestra de 20 envases se encuentra que la varianza muestral de las cantidades de llenado es $s^2 = 0.0025$. Sin embargo, sabe que no se puede esperar que la varianza de 20 envases corresponda al valor exacto de la varianza de toda la población de recipientes que se llenan en este proceso de producción. Así, deseará obtener una estimación por intervalo para la varianza poblacional.

FIGURA 11.1 Ejemplos de la distribución de muestreo de $(n-1)s^2/\sigma^2$ (distribución ji-cuadrada)



La expresión χ^2_α denota el valor de la distribución ji-cuadrada que propor-ciona un área o probabilidad de α a la *derecha* del valor χ^2_α. Por ejemplo, en la figura 11.2 aparece la distribución ji-cuadrada con 19 grados de libertad, en la que $\chi^2_{0.025} = 32.852$ indica que 2.5% de los valores ji-cuadrada se encuentran a la derecha de 32.852, y $\chi^2_{0.975} = 8.907$ significa que 97.5% de los valores ji-cuadrada están a la derecha de 8.907. Existen tablas que proporcionan las áreas o probabilidades de este tipo de distribución. Consulte la tabla 11.1 y verifique que los valores de ji-cuadrada con 19 grados de libertad (fila 19 de la tabla) son correctos. En la tabla 3 del apéndice B se encuentra una versión más completa con estos valores.

En la gráfica de la figura 11.2 se ve que 0.95 o 95% de los valores ji-cuadrada se encuentran entre $\chi^2_{0.975}$ y $\chi^2_{0.025}$. Es decir, hay 0.95 de probabilidad de obtener un valor χ^2 tal que.

$$\chi^2_{0.975} \leq \chi^2 \leq \chi^2_{0.025}$$

FIGURA 11.2 Distribución ji-cuadrada con 19 grados de libertad

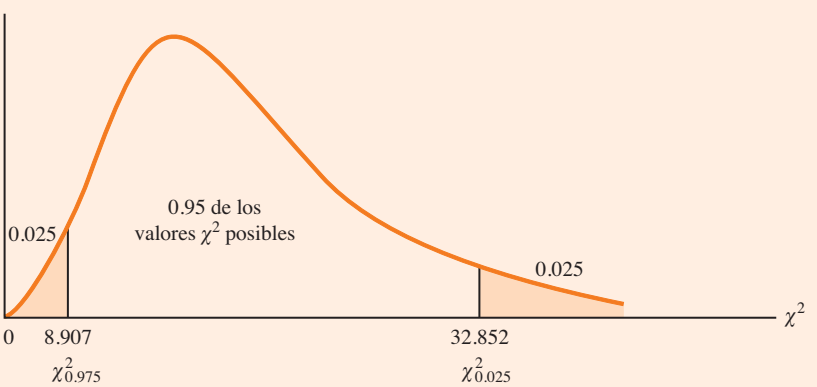
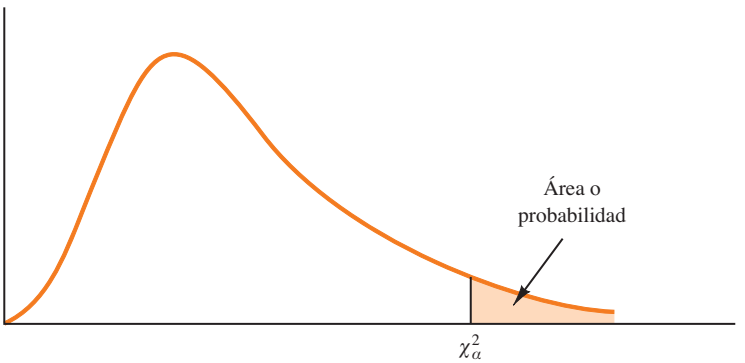


TABLA 11.1 Valores seleccionados de la tabla de distribución ji-cuadrada*



Grados de libertad	Área en la cola superior							
	0.99	**0.975**	**0.95**	**0.90**	**0.10**	**0.05**	**0.025**	**0.01**
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.878	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
100	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

**Nota*. En la tabla 3 del apéndice B se encuentra una versión más completa.

En la expresión (11.2) se estableció que $(n-1)s^2/\sigma^2$ sigue una distribución ji-cuadrada; por tanto, se puede sustituir $(n-1)s^2/\sigma^2$ por χ^2 y escribir

$$\chi^2_{0.975} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{0.025} \tag{11.3}$$

En efecto, la expresión (11.3) proporciona una estimación por intervalo en la que 0.95 o 95% de todos los valores que puede tomar $(n-1)s^2/\sigma^2$ están en el intervalo que va de $\chi^2_{0.975}$ a $\chi^2_{0.025}$. Ahora es necesario realizar algunas manipulaciones algebraicas con la expresión (11.3) para obtener una estimación por intervalo de la varianza poblacional σ^2. Al modificar la desigualdad de la izquierda tenemos

$$\chi^2_{0.975} \leq \frac{(n-1)s^2}{\sigma^2}$$

Por tanto,

$$\sigma^2\chi^2_{0.975} \leq (n-1)s^2$$

o

$$\sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{0.975}} \tag{11.4}$$

Con manipulaciones algebraicas similares a la desigualdad de la derecha de la expresión (11.3) tenemos:

$$\frac{(n-1)s^2}{\chi^2_{0.025}} \leq \sigma^2 \tag{11.5}$$

Los resultados de las expresiones (11.4) y (11.5) se combinan para obtener

$$\frac{(n-1)s^2}{\chi^2_{0.025}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{0.975}} \tag{11.6}$$

Como la expresión (11.3) es verdadera para 95% de los valores $(n-1)s^2/\sigma^2$, la expresión (11.6) proporciona una estimación por intervalo de 95% de confianza para la varianza poblacional σ^2.

Ahora, de regreso al problema de proporcionar una estimación por intervalo para la varianza poblacional de las cantidades de llenado, recuerde que en la muestra de 20 envases la varianza muestral fue $s^2 = 0.0025$. Dado que el tamaño de la muestra es 20, tenemos 19 grados de libertad. Como se aprecia en la figura 11.2, ya se determinó que $\chi^2_{0.975} = 8.907$, y $\chi^2_{0.025} = 32.852$. Con estos valores en la expresión (11.6) se obtiene la siguiente estimación por intervalo para la varianza poblacional:

$$\frac{(19)(0.0025)}{32.852} \leq \sigma^2 \leq \frac{(19)(0.0025)}{8.907}$$

Se puede obtener un intervalo de confianza para la desviación estándar poblacional calculando la raíz cuadrada del límite inferior y del límite superior del intervalo de confianza para la varianza poblacional.

o

$$0.0014 \leq \sigma^2 \leq 0.0053$$

Al calcular la raíz cuadrada de estos valores se obtiene el siguiente intervalo de 95% de confianza para la desviación estándar poblacional.

$$0.0380 \leq \sigma \leq 0.0730$$

De esta manera se ha ilustrado el proceso del uso de la distribución ji-cuadrada para obtener una estimación por intervalo de la varianza poblacional y la desviación estándar poblacional. Observe que, como se usaron $\chi^2_{0.975}$ y $\chi^2_{0.025}$, el coeficiente de confianza de la estimación por intervalo es 0.95. Extendiendo la expresión (11.6) al caso general con cualquier coeficiente de confianza, tenemos la siguiente estimación por intervalo para la varianza poblacional.

ESTIMACIÓN POR INTERVALO PARA UNA VARIANZA POBLACIONAL

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}} \qquad \textbf{(11.7)}$$

donde los valores χ^2 están basados en una distribución ji-cuadrada con $n-1$ grados de libertad, y $1-\alpha$ es el coeficiente de confianza.

Pruebas de hipótesis

Con σ_0^2 para denotar el valor hipotético de la varianza poblacional, las tres formas de una prueba de hipótesis para esta varianza son las siguientes.

$$\begin{array}{lll} H_0\colon \sigma^2 \geq \sigma_0^2 & H_0\colon \sigma^2 \leq \sigma_0^2 & H_0\colon \sigma^2 = \sigma_0^2 \\ H_a\colon \sigma^2 < \sigma_0^2 & H_a\colon \sigma^2 > \sigma_0^2 & H_a\colon \sigma^2 \neq \sigma_0^2 \end{array}$$

Estas tres formas son semejantes a las utilizadas para realizar las pruebas de hipótesis de una cola y de dos colas para las medias poblacionales y las proporciones poblacionales estudiadas en los capítulos 9 y 10.

En una prueba de hipótesis para la varianza poblacional se emplean el valor hipotético de la varianza poblacional σ_0^2 y la varianza muestral s^2 para calcular el valor del estadístico de prueba χ^2. Si la población tiene una distribución normal, el estadístico de prueba es el siguiente.

ESTADÍSTICO DE PRUEBA PARA PRUEBAS DE HIPÓTESIS ACERCA DE LA VARIANZA POBLACIONAL

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \qquad \textbf{(11.8)}$$

donde χ^2 tiene una distribución ji-cuadrada con $n-1$ grados de libertad.

Una vez calculado el valor del estadístico de prueba χ^2 para determinar si se rechaza la hipótesis nula, se emplea el método del valor-p o el método del valor crítico.

Considere ahora el siguiente ejemplo. St. Louis Metro Bus Company de Estados Unidos desea promover una imagen de confiabilidad haciendo que sus conductores sean puntuales en los horarios de llegada a las estaciones. La empresa desea que haya poca varianza en dichos tiempos. En términos de la varianza de los tiempos de arribo a las paradas, la empresa desea que sea de 4 minutos o menos. Se formula la siguiente prueba de hipótesis para que la empresa pueda determinar si la varianza poblacional en los tiempos de llegada a las estaciones es excesiva.

$$\begin{array}{l} H_0\colon \sigma^2 \leq 4 \\ H_a\colon \sigma^2 > 4 \end{array}$$

Suponer, tentativamente, que H_0 sea verdadera, es admitir que la varianza poblacional en los tiempos de llegada se encuentra dentro de los lineamientos establecidos por la empresa. La H_0 se rechaza si las evidencias muestrales indican que la varianza poblacional excede estos lineamientos. En tal caso habrá que tomar medidas para reducirla. Esta prueba de hipótesis se realiza usando $\alpha = 0.05$ como nivel de significancia.

WEB archivo
BusTimes

Asuma que, en una muestra aleatoria de 24 llegadas a cierta parada en una intersección en el centro de la ciudad, la varianza muestral encontrada es $s^2 = 4.9$. Si la distribución poblacional de los tiempos de llegada es aproximadamente normal, el valor del estadístico de prueba es el siguiente.

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(24-1)(4.9)}{4} = 28.18$$

En la figura 11.3 se presenta la distribución ji-cuadrada con $n - 1 = 24 - 1 = 23$ grados de libertad. Como ésta es una prueba de cola superior, el área bajo la curva a la derecha del valor del estadístico de prueba $\chi^2 = 28.18$ es el valor-p de la prueba.

Como ocurre con las tablas de la distribución t, las tablas de la distribución ji-cuadrada no son suficientemente detalladas para permitir determinar con exactitud el valor-p. Sin embargo, sí permiten obtener el intervalo en el que se encuentra el valor-p. Por ejemplo, usando la tabla 11.1 se encuentra la información siguiente para la distribución ji-cuadrada con 23 grados de libertad (gl).

Área en la cola superior	**0.10**	**0.05**	**0.025**	**0.01**
Valor χ^2 (23 gl)	32.007	35.172	38.076	41.638

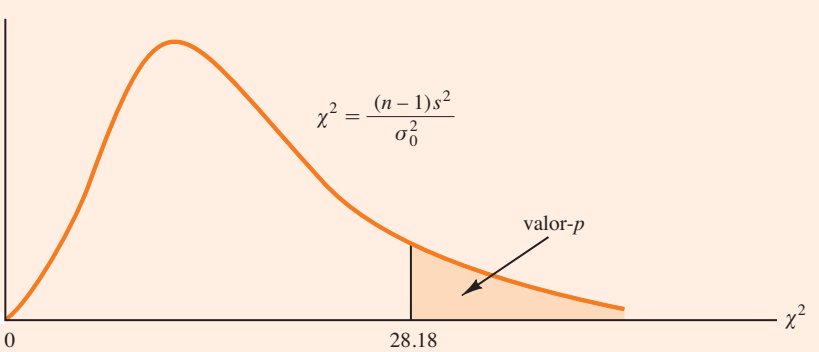
$\chi^2 = 28.18$

Dado que $\chi^2 = 28.18$ es menor que 32.007, el área en la cola superior (el valor-p) es mayor que 0.10. Como el valor-p es $> \alpha = 0.05$, no se puede rechazar la hipótesis nula. La muestra no lleva a la conclusión de que la varianza poblacional en los tiempos de llegada a las paradas sea excesiva.

Considerando la dificultad para determinar con exactitud el valor-p con las tablas de la distribución ji-cuadrada, es útil emplear un software como Minitab o Excel. En el apéndice F al final del libro se describe cómo calcular los valores-p. En el apéndice se especifica que el valor-p exacto que corresponde a $\chi^2 = 28.18$ es 0.2091.

Como ocurre con los otros procedimientos para pruebas de hipótesis, aquí también es posible recurrir al método del valor crítico para obtener la conclusión de la prueba de hipótesis. Con $\alpha = 0.05$, $\chi^2_{0.05}$ proporciona el valor crítico en la cola superior para esta prueba. Considerando

FIGURA 11.3 Distribución ji-cuadrada para el ejemplo de St. Louis Metro Bus



la tabla 11.1 y 23 grados de libertad, $\chi^2_{0.05} = 35.172$. De esta manera, la regla de rechazo para los tiempos de llegada a las paradas es la siguiente.

$$\text{Rechazar } H_0 \text{ si } \chi^2 \geq 35.172$$

Como el valor del estadístico de prueba es $\chi^2 = 28.18$, no se puede rechazar la hipótesis nula.

En la práctica, las pruebas de hipótesis para la varianza poblacional que se encuentran con más frecuencia son pruebas de cola superior como la aquí presentada. En situaciones que implican tiempos de arribo o de producción, pesos de llenado, dimensiones de piezas y mediciones parecidas, son deseables las varianzas pequeñas, en tanto que las grandes son inaceptables. Al establecer la varianza poblacional máxima permitida, es posible probar la hipótesis nula de que la varianza poblacional es menor o igual que el valor máximo permitido, contra la hipótesis alternativa de que es mayor que el valor máximo permitido. Con esta estructura de prueba deberán aplicarse acciones correctivas siempre que la hipótesis nula sea rechazada, lo que indica la presencia de una varianza poblacional excesiva.

Como ocurre con la media y la proporción poblacionales, también se realizan otras formas de pruebas de hipótesis. A continuación se presenta una prueba de dos colas para la varianza poblacional considerando una situación que suele presentarse en las oficinas de licencias para conducir vehículos motorizados en Estados Unidos. Históricamente, la varianza en las puntuaciones de los exámenes presentados por las personas que solicitan una licencia para conducir ha sido $\sigma^2 = 100$. Ahora se ha elaborado una versión con preguntas nuevas. Los administradores o gerentes desean que la varianza en las puntuaciones del examen se mantenga en los niveles históricos. Para evaluar la varianza en las puntuaciones del nuevo examen se propone la siguiente prueba de hipótesis de dos colas.

$$H_0: \sigma^2 = 100$$
$$H_a: \sigma^2 \neq 100$$

El rechazo de H_0 indicará que la varianza ha cambiado y que será necesario revisar algunas preguntas del nuevo examen para que la varianza en sus puntuaciones sea parecida a la de las puntuaciones del examen anterior. La nueva versión será aplicada a los integrantes de una muestra de 30 solicitantes de licencia de conducir. En esta prueba de hipótesis se usará como nivel de significancia $\alpha = 0.05$.

En este caso, la varianza muestral de las puntuaciones de 30 exámenes fue $s^2 = 162$. El valor del estadístico de prueba ji-cuadrada es el siguiente.

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(30-1)(162)}{100} = 46.98$$

Ahora queda calcular el valor-p. Mediante la tabla 11.1 y $n - 1 = 30 - 1 = 29$ grados de libertad, se encuentra lo siguiente.

Área en la cola superior	**0.10**	**0.05**	**0.025**	**0.01**
Valor χ^2 (29 *gl*)	39.087	42.557	45.722	49.588

$\chi^2 = 46.98$

De manera que el valor del estadístico de prueba $\chi^2 = 46.98$ corresponde a un área entre 0.025 y 0.01 en la cola superior de la distribución ji-cuadrada. Al duplicar este valor, tenemos que el

TABLA 11.2 Resumen de las pruebas de hipótesis para la varianza poblacional

	Prueba de cola inferior	Prueba de cola superior	Prueba de dos colas
Hipótesis	$H_0: \sigma^2 \geq \sigma_0^2$ $H_a: \sigma^2 < \sigma_0^2$	$H_0: \sigma^2 \leq \sigma_0^2$ $H_a: \sigma^2 > \sigma_0^2$	$H_0: \sigma^2 = \sigma_0^2$ $H_0: \sigma^2 \neq \sigma_0^2$
Estadístico de prueba	$\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$
Regla de rechazo: método del valor-*p*	Rechazar H_0 si el valor-$p \leq \alpha$	Rechazar H_0 si el valor-$p \leq \alpha$	Rechazar H_0 si el valor-$p \leq \alpha$
Regla de rechazo: método del valor crítico	Rechazar H_0 si $\chi^2 \leq \chi^2_{(1-\alpha)}$	Rechazar H_0 si $\chi^2 \geq \chi^2_{\alpha}$	Rechazar H_0 si $\chi^2 \leq \chi^2_{(1-\alpha/2)}$ o si $\chi^2 \geq \chi^2_{\alpha/2}$

valor-p de dos colas está entre 0.05 y 0.02. Con Excel o Minitab se encuentra el valor-p exacto $= 0.374$. Como el valor-$p \leq \alpha = 0.05$, H_0 es rechazada y se concluye que en el nuevo examen las puntuaciones presentan una varianza poblacional distinta a la varianza histórica de $\sigma^2 = 100$. Un resumen de las pruebas de hipótesis para la varianza poblacional se presenta en la tabla 11.2.

Ejercicios

Métodos

1. Utilizando la tabla 11.1 o la tabla 3 del apéndice B, encuentre los valores siguientes de la distribución ji-cuadrada.
   a) $\chi^2_{0.05}$, con $gl = 5$.
   b) $\chi^2_{0.25}$, con $gl = 15$.
   c) $\chi^2_{0.975}$, con $gl = 20$.
   d) $\chi^2_{0.01}$, con $gl = 10$.
   e) $\chi^2_{0.95}$, con $gl = 18$.

AUTOevaluación

2. En una muestra de 20 elementos, la desviación estándar muestral es 5.
   a) Calcule una estimación por intervalo de 90% de confianza para la varianza poblacional.
   b) Defina una estimación por intervalo de 95% de confianza para la varianza poblacional.
   c) Calcule una estimación por intervalo de 95% de confianza para la desviación estándar poblacional.

3. En una muestra de 16 elementos la desviación estándar muestral es 9.5. Pruebe la hipótesis siguiente usando $\alpha = 0.05$. ¿A qué conclusión llega? Utilice tanto el método del valor-p como el del valor crítico.

$$H_0: \sigma^2 \leq 50$$
$$H_a: \sigma^2 > 50$$

Aplicaciones

4. La varianza en los pesos de los medicamentos es un aspecto crítico en la industria farmacéutica. Considere un medicamento cuyo peso está dado en gramos y una muestra de 18 unidades del producto. La varianza muestral es $s^2 = 0.36$.
   a) Proporcione un intervalo de 90% de confianza para estimar la varianza poblacional de los pesos del medicamento.
   b) Proporcione un intervalo de 90% de confianza para estimar la desviación estándar poblacional.

5. A continuación se presentan los precios de renta diaria de un automóvil en ocho ciudades.

Ciudad	Renta diaria de un automóvil ($)
Atlanta	47
Chicago	50
Dallas	53
Nueva Orleáns	45
Phoenix	40
Pittsburgh	43
San Francisco	39
Seattle	37

a) Calcule la varianza y la desviación estándar para estos datos.
b) ¿Cuál es la estimación por intervalo de 95% de confianza para la varianza poblacional de los precios de renta de un automóvil por día?
c) ¿Cuál es la estimación por intervalo de 95% de confianza para la desviación estándar poblacional?

6. La Fidelity Growth & Income recibió fondos de inversión de tres estrellas, o neutrales, clasificados por Morningstar. A continuación se presentan los rendimientos porcentuales trimestrales en el periodo de cinco años que va de 2001 a 2005 (*Morningstar Funds 500*, 2006).

WEB archivo
Return

	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
2001	−10.91	5.80	−9.64	6.45
2002	0.83	−10.48	−14.03	5.58
2003	−2.27	10.43	0.85	9.33
2004	1.34	1.11	−0.77	8.03
2005	−2.46	0.89	2.55	1.78

a) Calcule la media, la varianza y la desviación estándar de estos rendimientos trimestrales.
b) Los analistas financieros suelen usar la desviación estándar como una medida del riesgo para las acciones y fondos de inversión. Proporcione un intervalo de 95% de confianza para la desviación estándar poblacional del rendimiento trimestral de los fondos de inversión de Fidelity Growth & Income.

7. Para analizar el riesgo o la volatilidad asociados con invertir en las acciones comunes de Chevron Corporation, se toma una muestra del rendimiento porcentual total mensual. A continuación se presentan los rendimientos de los 12 meses de 2005 (*Compustat,* 24 de febrero de 2006). El rendimiento total es el precio más cualquier dividendo pagado.

Mes	Rendimiento (%)	Mes	Rendimiento (%)
Enero	3.60	Julio	3.74
Febrero	14.86	Agosto	6.62
Marzo	−6.07	Septiembre	5.42
Abril	−10.82	Octubre	−11.83
Mayo	4.29	Noviembre	1.21
Junio	3.98	Diciembre	−0.94

a) Calcule la varianza muestral y la desviación estándar muestral como medidas de la volatilidad del rendimiento mensual total de Chevron.
b) Proporcione un intervalo de 95% de confianza para la varianza poblacional.
c) Construya un intervalo de 95% de confianza para la desviación estándar poblacional.

8. El 4 de marzo de 2009 fue uno de los pocos buenos días para el mercado de valores a principios de 2009. El índice Dow Jones alcanzó 149.82 puntos (*The Wall Street Journal*, 5 de marzo de 2009).La tabla siguiente presenta los cambios en los precios (Price Change) de las acciones de una muestra de 12 empresas (Company) que cotizaron ese día.

WEB archivo
PriceChange

Company	Price Change ($)	Company	Price Change ($)
Aflac	0.81	John.&John.	1.46
Bank of America	−0.05	Loews Cp	0.92
Cablevision	0.41	Nokia	0.21
Diageo	1.32	SmpraEngy	0.97
Flour Cp	2.37	Sunoco	0.52
Goodrich	0.3	Tyson Food	0.12

a) Calcule la varianza muestral del cambio de precio diario.
b) Determine la desviación estándar muestral de la variación de precio.
c) Proporcione una estimación por intervalo de 95% de confianza para la varianza poblacional y la desviación estándar poblacional.

AUTO evaluación

9. Una pieza para automóviles debe fabricarse con medidas de tolerancia muy estrechas para que sea aceptada por el cliente. Las especificaciones de producción indican que la varianza máxima en la longitud de la pieza debe ser 0.0004. Suponga que en 30 piezas la varianza muestral encontrada es $s^2 = 0.0005$. Use $\alpha = 0.05$ para probar si se está violando la especificación para la varianza poblacional.

10. La desviación estándar promedio del rendimiento anual de los grandes fondos de inversión de acciones de capital es 18.2% (*The Top Mutual Funds*, AAII, 2004). La desviación estándar muestral basada en una muestra de 36 fondos de inversión Vanguard PRIMECAP es 22.2%. Realice una prueba de hipótesis para determinar si la desviación estándar de los fondos Vanguard es mayor que la desviación estándar promedio de los grandes fondos de inversión de capital. Con un nivel de significancia de 0.05, ¿cuál es su conclusión?

11. A fines de 2008 la varianza en los rendimientos semestrales de los bonos gubernamentales en el extranjero fue $\sigma^2 = 0.70$. Un grupo de inversionistas se reunió entonces para analizar las tendencias futuras en los rendimientos de los bonos en el extranjero. Algunos esperaban una variabilidad en aumento de los rendimientos mientras que otros adoptaron el punto de vista contrario. La tabla siguiente muestra los rendimientos (Yield) semestrales en 12 países (Country) del 6 de marzo de 2009 (*Barron's*, 9 de marzo de 2009).

WEB archivo
Yields

Country	Yield (%)	Country	Yield (%)
Australia	3.98	Italy	4.51
Belgium	3.78	Japan	1.32
Canada	2.95	Netherlands	3.53
Denmark	3.55	Spain	3.90
France	3.44	Sweden	2.48
Germany	3.08	U.K.	3.76

a) Calcule la media, la varianza y la desviación estándar de estos rendimientos trimestrales.
b) Establezca la hipótesis para probar si los datos muestrales indican que la varianza en el rendimiento de bonos ha cambiado desde fines de 2008.
c) Utilice $\alpha = 0.05$ para realizar la prueba de la hipótesis establecida en el inciso *b*). ¿Cuál es su conclusión?

12. En un estudio de *Fortune* se encontró que la varianza en la cantidad de vehículos que poseen o rentan los suscriptores de la revista *Fortune* es 0.94. Suponga que en una muestra de 12 suscriptores de otra revista se encuentran los datos siguientes sobre la cantidad de vehículos que poseen o rentan dichos clientes: 2, 1, 2, 0, 3, 2, 2, 1, 2, 1, 0 y 1.
a) Calcule la varianza muestral del número de vehículos que poseen o rentan estos 12 suscriptores.
b) Pruebe la hipótesis H_0: $\sigma^2 = 0.94$ para determinar si la varianza del número de vehículos que poseen o rentan los suscriptores de la otra revista difiere de la propia de *Fortune*, que es $\sigma^2 = 0.94$. Con un nivel de significancia de 0.05, ¿cuál es su conclusión?

11.2 Inferencias acerca de dos varianzas poblacionales

En algunas aplicaciones estadísticas interesa comparar las varianzas en la calidad de un producto obtenida mediante dos métodos de producción diferentes, las varianzas de tiempos de ensamblado empleando dos métodos distintos o las varianzas de las temperaturas que se obtienen con dos dispositivos diversos de calentamiento. Para comparar las varianzas de dos poblaciones se emplean datos obtenidos de dos muestras aleatorias independientes: una de la población 1 y otra de la población 2. Para hacer las inferencias acerca de las dos varianzas poblacionales σ_1^2 y σ_2^2 se usan las dos varianzas muestrales s_1^2 y s_2^2. Cuando las varianzas de dos poblaciones normales son iguales ($\sigma_1^2 = \sigma_2^2$), la distribución de muestreo de la proporción entre las dos varianzas muestrales s_1^2/s_2^2 es la siguiente.

DISTRIBUCIÓN DE MUESTREO DE s_1^2/s_2^2 CUANDO $\sigma_1^2 = \sigma_2^2$

Cuando se toman muestras aleatorias simples independientes de tamaños n_1 y n_2 de dos poblaciones normales con varianzas iguales, la distribución de muestreo de

$$\frac{s_1^2}{s_2^2} \tag{11.9}$$

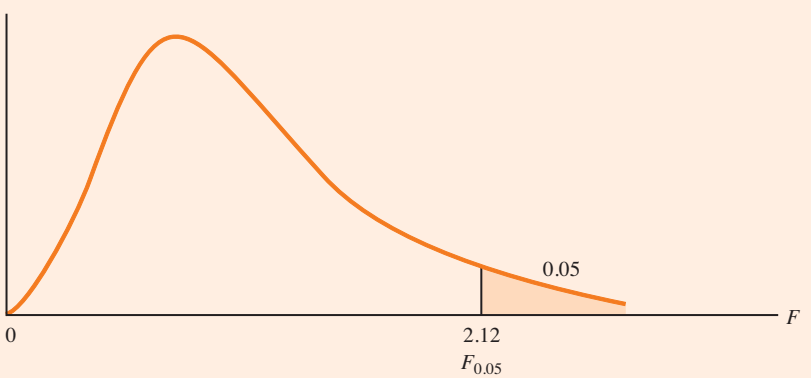
es una distribución F con $n_1 - 1$ grados de libertad en el numerador, y $n_2 - 1$ grados de libertad en el denominador; s_1^2 es la varianza muestral de la muestra aleatoria de n_1 elementos tomados de la población 1, y s_2^2 es la varianza muestral de la muestra aleatoria de n_2 elementos tomados de la población 2.

La distribución F se basa en muestras de dos poblaciones normales.

La figura 11.4 es una gráfica de la distribución F con 20 grados de libertad tanto en el numerador como en el denominador. Como se ve en esta gráfica, la distribución F no es simétrica y los valores F nunca pueden ser negativos. La forma de cada distribución F depende de los grados de libertad tanto del numerador como del denominador.

Para denotar el valor F correspondiente a un área o probabilidad de α en la cola superior de la distribución se utiliza la notación F_α. Por ejemplo, como aparece en la figura 11.4, $F_{0.05}$ corresponde a un área de 0.05 en la cola superior de la distribución F con 20 grados de libertad en el numerador y 20 grados de libertad en el denominador. El valor específico de $F_{0.05}$ se

FIGURA 11.4 Distribución F con 20 grados de libertad en el numerador y 20 grados de libertad en el denominador



encuentra en la tabla de la distribución F, parte de la cual se presenta en la tabla 11.3. Usando 20 grados de libertad tanto en el numerador como en el denominador y la fila correspondiente a un área de 0.05 en la cola superior de la distribución, se encuentra $F_{0.05} = 2.12$. Observe que la tabla sirve para determinar valores de F correspondientes a áreas de 0.10, 0.05, 0.025 y 0.01 en la cola superior. En la tabla 4 del apéndice B se encuentra una versión más completa de la distribución F.

Ahora se verá cómo usar la distribución F para realizar una prueba de hipótesis para las varianzas de dos poblaciones. Se empieza con una prueba sobre la igualdad de las dos varianzas poblacionales. Las hipótesis son las siguientes.

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_a: \sigma_1^2 \neq \sigma_2^2$$

De manera tentativa se supone que las varianzas poblacionales son iguales. Si se rechaza H_0, se concluirá que las varianzas poblacionales no son iguales.

Para realizar esta prueba de hipótesis se requieren dos muestras aleatorias independientes, una de cada población. Se calculan las dos varianzas muestrales. A la población en la que se encuentre la *mayor* varianza muestral se le considera población 1. De manera que el tamaño de muestra n_1 y la varianza muestral s_1^2 corresponden a la población 1, y el tamaño de muestra n_2 y la varianza muestral s_2^2 corresponden a la población 2. Con base en el supuesto de que ambas poblaciones tienen una distribución normal, la relación entre las varianzas muestrales proporciona el siguiente estadístico de prueba F.

ESTADÍSTICO DE PRUEBA PARA PRUEBAS DE HIPÓTESIS ACERCA DE VARIANZAS POBLACIONALES CON $\sigma_1^2 = \sigma_2^2$

$$F = \frac{s_1^2}{s_2^2} \qquad \textbf{(11.10)}$$

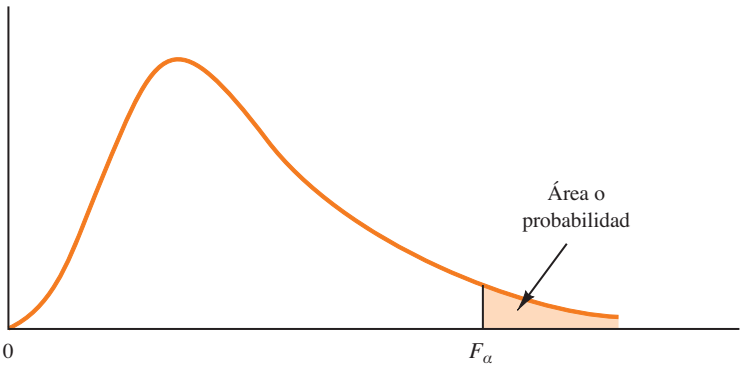
Al denotar como población 1 a la que posee mayor varianza muestral, el estadístico de prueba tiene una distribución F con $n_1 - 1$ grados de libertad en el numerador y $n_2 - 1$ grados de libertad en el denominador.

Como el estadístico de prueba F se construye con la varianza muestral más grande s_1^2 en el numerador, el valor de este estadístico se encontrará siempre en la cola superior de la distribución F. Por tanto, las tablas de la distribución F, como la 11.3 y la 4 del apéndice B, únicamente necesitan proporcionar áreas o probabilidades en la cola superior. Si no se construyera de este modo el estadístico de prueba, serían necesarias áreas o probabilidades en la cola inferior. En tal caso se requerirían más cálculos o tablas más extensas para la distribución F. A continuación se presenta un ejemplo de una prueba de hipótesis para la igualdad de dos varianzas poblacionales.

Dullus County School está por renovar el contrato del servicio de autobús para el año próximo y debe decidirse entre dos empresas que proporcionan el servicio: Milbank Company y Gulf Park Company. Como medida de la calidad del servicio de traslado se emplea la varianza de los tiempos en que el autobús llega a recoger/dejar a las personas. Baja varianza indica el servicio más constante y de mayor calidad. Si las varianzas de los tiempos de llegada asociados con ambos servicios son iguales, Dullus School optará por la empresa que ofrezca mejores condiciones financieras. Pero si los datos muestrales de los tiempos de llegada de ambas compañías indican una diferencia significativa en las varianzas, los administradores preferirán a la empresa con la menor varianza o el mejor servicio. Las hipótesis son las siguientes:

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_a: \sigma_1^2 \neq \sigma_2^2$$

TABLA 11.3 Valores seleccionados de la tabla de distribución *F**



Grados de libertad en el denominador	Área en la cola superior	Grados de libertad en el numerador				
		10	15	20	25	30
10	0.10	2.32	2.24	2.20	2.17	2.16
	0.05	2.98	2.85	2.77	2.73	2.70
	0.025	3.72	3.52	3.42	3.35	3.31
	0.01	4.85	4.56	4.41	4.31	4.25
15	0.10	2.06	1.97	1.92	1.89	1.87
	0.05	2.54	2.40	2.33	2.28	2.25
	0.025	3.06	2.86	2.76	2.69	2.64
	0.01	3.80	3.52	3.37	3.28	3.21
20	0.10	1.94	1.84	1.79	1.76	1.74
	0.05	2.35	2.20	2.12	2.07	2.04
	0.025	2.77	2.57	2.46	2.40	2.35
	0.01	3.37	3.09	2.94	2.84	2.78
25	0.10	1.87	1.77	1.72	1.68	1.66
	0.05	2.24	2.09	2.01	1.96	1.92
	0.025	2.61	2.41	2.30	2.23	2.18
	0.01	3.13	2.85	2.70	2.60	2.54
30	0.10	1.82	1.72	1.67	1.63	1.61
	0.05	2.16	2.01	1.93	1.88	1.84
	0.025	2.51	2.31	2.20	2.12	2.07
	0.01	2.98	2.70	2.55	2.45	2.39

* *Nota*. La tabla 4 del apéndice B ofrece una versión más completa.

Si se rechaza H_0, se concluirá que los servicios no tienen la misma calidad. Para realizar esta prueba de hipótesis se usa $\alpha = 0.10$ como nivel de significancia.

WEB archivo
SchoolBus

En una muestra de 26 tiempos de llegada de Milbank, la varianza muestral es 48, y en una muestra de 16 tiempos de llegada de Gulf Park la varianza muestral es 20. Como la varianza en la muestra de Milbank es la mayor, se etiqueta como población 1. Usando la ecuación (11.10) se encuentra el valor del estadístico de prueba.

$$F = \frac{s_1^2}{s_2^2} = \frac{48}{20} = 2.40$$

La distribución F correspondiente tiene $n_1 - 1 = 26 - 1 = 25$ grados de libertad en el numerador y $n_2 - 1 = 16 - 1 = 15$ grados de libertad en el denominador.

Igual que en las otras pruebas de hipótesis, para llegar a una conclusión se puede emplear el método del valor-p o el método del valor crítico. En la tabla 11.3 se encuentran las siguientes áreas en la cola superior correspondientes a los valores F de una distribución F con 25 grados de libertad en el numerador y los correspondientes 15 grados en el denominador.

Área en la cola superior	**0.10**	**0.05**	**0.025**	**0.01**
Valor F ($gl_1 = 25$; $gl_2 = 15$)	1.89	2.28	2.69	3.28

$F = 2.40$

Como $F = 2.40$ está entre 2.28 y 2.69, el área en la cola superior de la distribución se ubica entre 0.05 y 0.025. Dado que se trata de una prueba de dos colas, se duplica el área de la cola superior y se obtiene un valor-p entre 0.10 y 0.05. Como se eligió $\alpha = 0.10$ como nivel de significancia, el valor-$p < \alpha = 0.10$. Por tanto, se rechaza la hipótesis nula. Esto lleva a la conclusión de que los dos servicios de autobús difieren en términos de la varianza de los tiempos en que llegan a recoger/dejar a las personas. Se le recomienda a la escuela el servicio de menor varianza o el mejor, que es el ofrecido por la empresa Gulf Park.

Usando Excel o Minitab se encuentra que el estadístico de prueba $F = 2.40$ corresponde a un valor-$p = 0.0811$. Como $0.0811 < \alpha = 0.10$, se rechaza la hipótesis nula de que las dos varianzas poblacionales son iguales.

Para usar el método del valor crítico en una prueba de hipótesis de dos colas con $\alpha = 0.10$, se toman los valores críticos correspondientes a un área de $\alpha/2 = 0.10/2 = 0.05$ en cada cola de la distribución. Como el valor del estadístico de prueba calculado con la ecuación (11.10) está siempre en la cola superior, basta determinar el valor crítico en esta cola. En la tabla 11.3 se encuentra que $F_{0.05} = 2.28$. Así, aun cuando se trata de una prueba de dos colas, la regla de rechazo es la siguiente.

$$\text{Rechazar } H_0 \text{ si } F \geq 2.28$$

Como el estadístico de prueba es $F = 2.40$ es mayor que 2.28, se rechaza H_0 y se concluye que los dos servicios difieren en términos de la varianza en los tiempos en que llegan a recoger/dejar a las personas.

También se pueden realizar pruebas de una cola para dos varianzas poblacionales. En estos casos se usa la distribución F para determinar si una varianza poblacional es significativamente mayor que la otra. Una prueba de hipótesis de una cola para dos varianzas poblacionales se formula siempre como una prueba de *cola superior*:

Una prueba de hipótesis de una cola para dos varianzas poblacionales siempre se formula como una prueba de cola superior. Esto elimina la necesidad de tener valores F de cola inferior.

$$H_0: \sigma_1^2 \leq \sigma_2^2$$
$$H_a: \sigma_1^2 > \sigma_2^2$$

Con esta forma de una prueba de hipótesis, el valor-p y el valor crítico siempre se encuentran en la cola superior de la distribución F. De esta manera, sólo se necesitan los valores F de la cola superior, lo cual simplifica tanto los cálculos como la tabla de la distribución F.

A continuación se ilustra con una encuesta sobre opinión pública el uso de la distribución F para realizar una prueba de una cola acerca de las varianzas de dos poblaciones. Para estudiar las actitudes frente a los asuntos políticos actuales se elige una muestra de 31 hombres y otra de 41 mujeres. Al investigador que realiza el estudio le interesa saber si los datos muestrales indican que entre las mujeres hay mayor variación en las actitudes respecto de los asuntos políticos

TABLA 11.4 Resumen de las pruebas de hipótesis acerca de dos varianzas poblacionales

	Prueba de cola superior	Prueba de dos colas
Hipótesis	$H_0: \sigma_1^2 \leq \sigma_2^2$ $H_a: \sigma_1^2 > \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2$ $H_a: \sigma_1^2 \neq \sigma_2^2$ *Nota*. La población 1 tiene la varianza muestral más grande
Estadístico de prueba	$F = \dfrac{s_1^2}{s_2^2}$	$F = \dfrac{s_1^2}{s_2^2}$
Regla de rechazo: método del valor-*p*	Rechazar H_0 si el valor-$p \leq \alpha$	Rechazar H_0 si el valor-$p \leq \alpha$
Regla de rechazo: método del valor crítico	Rechazar H_0 si $F \geq F_\alpha$	Rechazar H_0 si $F \geq F_{\alpha/2}$

que entre los hombres. En la forma de la prueba de hipótesis de una cola analizada previamente, las mujeres representarán la población 1 y los hombres la población 2. La prueba de hipótesis se plantea como sigue.

$$H_0: \sigma^2_{\text{mujer}} \leq \sigma^2_{\text{hombre}}$$
$$H_a: \sigma^2_{\text{mujer}} > \sigma^2_{\text{hombre}}$$

Rechazar H_0 dará al investigador el respaldo estadístico necesario para concluir que las mujeres muestran mayor variación en las actitudes respecto de los asuntos políticos.

Con la varianza muestral de las mujeres en el numerador y la de los hombres en el denominador, la distribución F tendrá $n_1 - 1 = 41 - 1 = 40$ grados de libertad en el numerador, y $n_2 - 1 = 31 - 1 = 30$ grados de libertad en el denominador. En esta prueba de hipótesis se usa $\alpha = 0.05$ como nivel de significancia. Como resultado, se encuentra una varianza muestral $s_1^2 = 120$ para las mujeres y $s_2^2 = 80$ para los hombres. El estadístico de prueba es el siguiente.

$$F = \frac{s_1^2}{s_2^2} = \frac{120}{80} = 1.50$$

En la tabla 4 del apéndice B, la distribución F con 40 grados de libertad en el numerador y 30 en el denominador muestra $F_{0.10} = 1.57$. Como el estadístico de prueba, $F = 1.50$, es menor que 1.57, el área en la cola superior debe ser mayor que 0.10. Por ende, el valor-p es mayor que 0.10. Con Excel o Minitab se encuentra que el valor-$p = 0.1256$. Como el valor-$p > \alpha = 0.05$, no se puede rechazar H_0. Por tanto, los resultados muestrales no favorecen la conclusión de que entre las mujeres haya mayor variación en la actitud frente a los asuntos políticos que entre los hombres. En la tabla 11.4 se presenta un resumen de las pruebas de hipótesis para dos varianzas poblacionales.

NOTAS Y COMENTARIOS

Las investigaciones confirman el hecho de que para usar la distribución F es importante suponer que las poblaciones tienen una distribución normal. La distribución F no se puede usar a menos que sea razonable suponer que ambas poblaciones tienen una distribución por lo menos aproximadamente normal.

Ejercicios

Métodos

13. En la tabla 4 del apéndice B halle los valores siguientes de la distribución F.
    a) $F_{0.05}$ con 5 y 10 grados de libertad.
    b) $F_{0.025}$ con 20 y 15 grados de libertad.
    c) $F_{0.01}$ con 8 y 12 grados de libertad.
    d) $F_{0.10}$ con 10 y 20 grados de libertad.

14. En una muestra de 16 elementos de la población 1 la varianza muestral es $s_1^2 = 5.8$, y en una muestra de 21 elementos de la población 2 la varianza muestral es $s_2^2 = 2.4$. Pruebe las hipótesis siguientes usando 0.05 como nivel de significancia.

$$H_0: \sigma_1^2 \leq \sigma_2^2$$
$$H_a: \sigma_1^2 > \sigma_2^2$$

    a) ¿A qué conclusión se llega usando el método del valor-p?
    b) Repita la prueba usando el método del valor crítico.

AUTOevaluación

15. Considere la prueba de hipótesis siguiente.

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_a: \sigma_1^2 \neq \sigma_2^2$$

    a) ¿A qué conclusión se llega si $n_1 = 21$, $s_1^2 = 8.2$, $n_2 = 26$ y $s_2^2 = 4.0$? Use $\alpha = 0.05$ y el método del valor-p.
    b) Repita la prueba utilizando el método del valor crítico.

Aplicaciones

16. Por lo general, los inversionistas utilizan la desviación estándar del porcentaje de rendimiento mensual de un fondo de inversión como medida del riesgo del fondo; en tales casos, un fondo con una desviación estándar grande se considera con más riesgo que otro con una desviación estándar más baja. Las desviaciones estándar de los fondos American Century Equity Growth y Fidelity Growth Discovery fueron reportadas recientemente con 15% y 18.9%, respectivamente (*The Top Mutual Funds*, AAII, 2009). Suponga que cada una de las desviaciones estándar se basa en una muestra de 60 meses de rendimientos. ¿Los resultados de la muestra sustentan la conclusión de que el fondo Fidelity Growth Discovery tiene una varianza poblacional mayor que el de American Century Equity? ¿Cuál fondo es el de mayor riesgo?

AUTOevaluación

17. La mayoría de los conductores sabe que el gasto anual medio en reparaciones de un automóvil depende de la antigüedad del vehículo. Un investigador desea saber si la varianza de los gastos anuales que se aplican en reparación también aumenta con la antigüedad del vehículo. En una muestra de 26 automóviles de 4 años de antigüedad, la desviación estándar muestral para los gastos anuales en reparación fue de $170, y en una muestra de 25 automóviles de 2 años de antigüedad fue de $100.
    a) Establezca las versiones nula y alternativa de la hipótesis de investigación de que la varianza en los gastos anuales por reparación es mayor entre más viejos son los automóviles.
    b) Empleando 0.01 como nivel de significancia, ¿cuál es su conclusión? ¿Cuál es el valor-p? Analice lo razonable de sus hallazgos.

18. *Barron's* obtuvo datos de los 1 000 principales asesores financieros (*Barron's*, 9 de febrero de 2009). Merrill Lynch tiene 239 personas en la lista y Morgan Stanley, 121. Una muestra de 16 asesores de Merrill Lynch y 10 de Morgan Stanley evidenció que manejan muchas cuentas muy grandes con una gran varianza en el total de fondos administrados. La desviación estándar de la cantidad administrada por los asesores de Merrill Lynch fue $s_1 = \$587$ millones, y la del monto manejado por los asesores de Morgan Stanley fue $s_2 = \$489$ millones. Realice

una prueba de hipótesis con $\alpha = 0.10$ para determinar si existe una diferencia significativa entre las varianzas poblacionales de los montos administradas por las dos empresas. ¿Cuál es su conclusión acerca de la variabilidad en la cantidad de fondos manejada por los asesores de las dos firmas?

19. La varianza en un proceso de producción es un indicador importante de la calidad del proceso. Las varianzas grandes representan una oportunidad para mejorarlo, buscando maneras de reducir la varianza del proceso. Realice una prueba estadística para determinar si existe una diferencia significativa entre las varianzas de los pesos de las bolsas procesadas con dos máquinas diferentes (Machine 1 y Machine 2). Use 0.05 como nivel de significancia. ¿Cuál es su conclusión? ¿Alguna de las dos máquinas representa una oportunidad para mejorar la calidad?

WEB archivo

Bags

Machine 1	2.95	3.45	3.50	3.75	3.48	3.26	3.33	3.20
	3.16	3.20	3.22	3.38	3.90	3.36	3.25	3.28
	3.20	3.22	2.98	3.45	3.70	3.34	3.18	3.35
	3.12							
Machine 2	3.22	3.30	3.34	3.28	3.29	3.25	3.30	3.27
	3.38	3.34	3.35	3.19	3.35	3.05	3.36	3.28
	3.30	3.28	3.30	3.20	3.16	3.33		

20. Con base en los datos obtenidos en un estudio, se sabe que en las firmas de contadores públicos la varianza de los sueldos anuales de los empleados con mayor antigüedad es aproximadamente 2.1, y la de los sueldos anuales de los gerentes es alrededor de 11.1. Estos datos están calculados en miles de dólares. Si se obtuvieron de muestras de 25 empleados con mayor antigüedad y 26 gerentes, pruebe la hipótesis de que las varianzas poblacionales de los sueldos son iguales. Con 0.05 como nivel de significancia, ¿cuál es su conclusión?

21. Fidelity Magellan es un fondo de inversión de capital grande y Fidelity Small Cap Stock lo es de capital pequeño (*Morningstar Funds 500*, 2006). La desviación estándar de ambos fondos se calculó empleando muestras aleatorias de tamaño 26. La desviación estándar muestral de Fidelity Magellan resultó de 8.89% y la de Fidelity Small Cap Stock de 13.03%. Los analistas financieros suelen usar la desviación estándar como una medida del riesgo. Realice una prueba de hipótesis para determinar si los fondos de capital pequeño son más riesgosos que los de capital grande. Utilice $\alpha = 0.05$ como nivel de significancia.

22. Una hipótesis de investigación sostiene que la varianza de las distancias de frenado de los automóviles sobre pavimento húmedo es mayor que la varianza de las distancias de frenado sobre pavimento seco. En un estudio de 16 automóviles desplazándose a una misma velocidad, se les hizo frenar sobre pavimento húmedo y después sobre pavimento seco. En el primer caso la desviación estándar de las distancias de frenado fue de 32 pies. Sobre pavimento seco fue de 16 pies.
    a) Con 0.05 como nivel de significancia, ¿los datos muestrales justifican la conclusión de que en las distancias de frenado sobre pavimento húmedo la varianza es mayor que sobre pavimento seco? ¿Cuál es el valor-*p*?
    b) ¿Qué significan las conclusiones estadísticas de este estudio en términos de las recomendaciones para la seguridad al manejar?

Resumen

En este capítulo se presentaron los procedimientos estadísticos que se usan en las inferencias acerca de las varianzas poblacionales. Se introdujeron dos distribuciones de probabilidad nuevas: la distribución ji-cuadrada y la distribución *F*. La primera se usa en la estimación por intervalos y en pruebas de hipótesis para la varianza de una población normal.

Se ilustró el uso de la distribución *F* en pruebas de hipótesis para las varianzas de dos poblaciones normales. En particular, se indicó que si tenemos muestras aleatorias simples in-

dependientes de tamaños n_1 y n_2, tomadas de dos poblaciones normales con varianzas iguales $\sigma_1^2 = \sigma_2^2$, la distribución muestral de la razón entre las dos varianzas muestrales s_1^2/s_2^2 tiene una distribución F con $n_1 - 1$ grados de libertad en el numerador, y $n_2 - 1$ grados de libertad en el denominador.

Fórmulas clave

Estimación por intervalo para una varianza poblacional

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}} \tag{11.7}$$

Estadístico de prueba para pruebas de hipótesis acerca de la varianza poblacional

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \tag{11.8}$$

Estadístico de prueba para pruebas de hipótesis acerca de varianzas poblacionales con $\sigma_1^2 = \sigma_2^2$

$$F = \frac{s_1^2}{s_2^2} \tag{11.10}$$

Ejercicios complementarios

23. Por cuestiones de personal, los administradores del hotel Gibson-Marimont desean conocer la variabilidad en la cantidad de habitaciones ocupadas por día en una determinada temporada del año. En una muestra de 20 días de operación la media muestral es 290 habitaciones ocupadas por día y la desviación estándar muestral es 30 habitaciones.
    a) ¿Cuál es la estimación puntual de la varianza poblacional?
    b) Proporcione una estimación por intervalo de 90% de confianza para la varianza poblacional.
    c) Desarrolle una estimación por intervalo de 90% de confianza para la desviación estándar poblacional.

24. Las ofertas públicas iniciales (OPI) de acciones suelen estar subvaluadas. La desviación estándar mide la dispersión o variación del indicador subvaluación-sobrevaluación. En una muestra de 13 OPI canadienses, negociadas después en la Bolsa de Valores de Toronto, esta desviación estándar fue de 14.95. Proporcione una estimación por intervalo de 95% de confianza para la desviación estándar poblacional del indicador subvaluación-sobrevaluación.

25. A continuación se presentan los costos estimados de manutención por día (Daily Living Cost) de un ejecutivo que viaja a varias ciudades (City) importantes. Las estimaciones comprenden una habitación individual en un hotel de cuatro estrellas, bebidas, desayuno, taxis y costos incidentales.

WEB archivo
Travel

City	Daily Living Cost ($)	City	Daily Living Cost ($)
Bangkok	242.87	Mexico City	212.00
Bogota	260.93	Milan	284.08
Cairo	194.19	Mumbai	139.16
Dublin	260.76	Paris	436.72
Frankfurt	355.36	Rio de Janeiro	240.87
Hong Kong	346.32	Seoul	310.41
Johannesburg	165.37	Tel Aviv	223.73
Lima	250.08	Toronto	181.25
London	326.76	Warsaw	238.20
Madrid	283.56	Washington, D.C.	250.61

a) Calcule la media muestral.
b) Determine la desviación estándar muestral.
c) Calcule un intervalo de 95% de confianza para la desviación estándar poblacional.

26. La variabilidad es crucial en la fabricación de cojinetes de balines. Una varianza grande en su tamaño ocasiona que no trabajen bien y que se desgasten rápidamente. Las normas de producción exigen una varianza máxima de 0.0001 en la medida de los cojinetes dada en pulgadas. En una muestra de 15 unidades, la desviación estándar muestral fue de 0.014 pulgadas.
a) Use $\alpha = 0.10$ para determinar si la muestra indica que se ha excedido la varianza máxima exigida.
b) Proporcione una estimación por intervalo de 90% de confianza para la varianza poblacional de los cojinetes.

27. La varianza en el llenado de cajas de cereal debe ser 0.02 o menos. En una muestra de 41 cajas la desviación estándar muestral es de 0.16 onzas. Use $\alpha = 0.05$ para determinar si la varianza en el llenado de los envases de cereal está excediendo la especificación.

28. City Trucking, Inc. asegura tiempos uniformes de entregas a sus clientes rutinarios. En una muestra de 22 entregas, la varianza muestral fue de 1.5. Realice una prueba de hipótesis para determinar si se puede rechazar H_0: $\sigma^2 \leq 1$. Use $\alpha = 0.10$.

29. En una muestra de 9 días de los últimos seis meses se encontró que un dentista había tratado a los siguientes números de pacientes: 22, 25, 20, 18, 15, 22, 24, 19 y 26. Si el número de sujetos atendidos por día tiene una distribución normal, ¿un análisis de estos datos muestrales permitiría rechazar la hipótesis de que la varianza de la cantidad de pacientes atendidos por día es 10? Use un nivel de significancia de 0.10. ¿Cuál es su conclusión?

30. La desviación estándar muestral del número de pasajeros que toman determinado vuelo de una línea aérea es 8. Una estimación por intervalo de 95% de confianza para la desviación estándar poblacional va de 5.86 a 12.62 pasajeros.
a) ¿El tamaño de la muestra usado en este análisis estadístico fue 10 o 15?
b) Suponga que la desviación estándar muestral $s = 8$ se obtuvo de una muestra de 25 vuelos. ¿Qué cambio esperaría en el intervalo de confianza para la desviación estándar poblacional? Calcule una estimación por intervalo de 95% de confianza para σ con un tamaño de muestra de 25.

31. ¿Existe alguna diferencia en la variabilidad de las puntuaciones de golf de las jugadoras del LPGA Tour (el *tour* de mujeres golfistas profesionales) y los jugadores del PGA Tour (el *tour* de hombres golfistas profesionales)? Una muestra de 20 puntuaciones de torneo de los eventos de la LPGA arrojó una desviación estándar de 2.4623 golpes, y una muestra de 30 puntuaciones de los torneos de la PGA dio una desviación estándar de 2.2118 (*Golfweek*, 7 de febrero y 7 de marzo de 2009). Formule una prueba de hipótesis para iguales varianzas poblacionales con objeto de determinar si existe una diferencia estadísticamente significativa en la variabilidad de las puntuaciones de golf de los hombres y las mujeres golfistas profesionales. Utilice $\alpha = 0.10$. ¿Cuál es su conclusión?

32. En los promedios de calificaciones de 352 estudiantes que terminaron un curso de contabilidad financiera la desviación estándar es de 0.940. En los promedios de calificaciones de 73 estudiantes que no aprobaron el mismo curso la desviación estándar es de 0.797. ¿Estos datos indican alguna diferencia entre las varianzas de los promedios de las calificaciones de quienes aprobaron el curso y de los que no lo aprobaron? Use 0.05 como nivel de significancia. *Nota.* $F_{0.025}$ con 351 y 72 grados de libertad es 1.466.

33. El área de contabilidad analiza la varianza de los costos unitarios semanales en los informes de dos departamentos de producción. En una muestra de 16 reportes de costos de cada uno de los departamentos, las varianzas de los costos fueron 2.3 y 5.4, respectivamente. ¿La muestra es suficiente para concluir que los dos departamentos difieren en términos de la varianza de los costos unitarios? Use $\alpha = 0.10$.

34. Al probar dos nuevos métodos de fabricación se reporta el tiempo requerido por cada uno de ellos. Use $\alpha = 0.10$ para probar la igualdad de las dos varianza poblacionales.

	Método A	Método B
Tamaño muestral	$n_1 = 31$	$n_2 = 25$
Variación muestral	$s_1^2 = 25$	$s_2^2 = 12$

Caso a resolver

Programa de capacitación de la Fuerza Aérea

En un curso introductorio de electrónica para la fuerza aérea se emplea un sistema personalizado en el que cada estudiante ve una clase grabada en una videocinta y después se le proporciona un texto de enseñanza programada. Los estudiantes trabajan con el libro en forma independiente hasta que terminan y aprueban un examen. Aquí preocupan los diferentes ritmos a los cuales los estudiantes completan esta etapa de su capacitación. Algunos asimilan relativamente pronto el libro de enseñanza programada, mientras que otros trabajan mucho más con él y necesitan más tiempo para completar el curso. Entonces los primeros deben esperar hasta que los alumnos más lentos estén listos y todo el grupo pueda pasar a otra etapa de la capacitación.

Se ha propuesto un sistema alterno con enseñanza asistida por computadora. Este método consiste en que todos los estudiantes vean la misma clase grabada y después a cada uno se le asigne una terminal de computadora para continuar con la capacitación. La computadora guía al estudiante, quien trabaja en forma independiente en esta parte del curso.

Para comparar estos dos métodos, el propuesto y el actual, a los integrantes de un nuevo grupo de 122 estudiantes se les asigna en forma aleatoria uno de los métodos de capacitación. Un grupo de 61 sujetos usa el método del libro programado y otro grupo de 61 sujetos, el método de enseñanza asistida por computadora. Se registra el tiempo, en horas, que necesita cada estudiante. La información que se presenta a continuación se encuentra en la bse de datos Training.

Course Completion Times (hours) for Current Training Method
[Tiempo para completar el curso (en horas) con el método de enseñanza actual]

76	76	77	74	76	74	74	77	72	78	73
78	75	80	79	72	69	79	72	70	70	81
76	78	72	82	72	73	71	70	77	78	73
79	82	65	77	79	73	76	81	69	75	75
77	79	76	78	76	76	73	77	84	74	74
69	79	66	70	74	72					

WEB archivo
Training

Course Completion Times (hours) for Proposed Computer-Assisted Method
[Tiempo para completar el curso (en horas) con el método asistido por computadora]

74	75	77	78	74	80	73	73	78	76	76
74	77	69	76	75	72	75	72	76	72	77
73	77	69	77	75	76	74	77	75	78	72
77	78	78	76	75	76	76	75	76	80	77
76	75	73	77	77	77	79	75	75	72	82
76	76	74	72	78	71					

Informe gerencial

1. Use la estadística descriptiva adecuada para resumir las horas que se necesitaron con cada método. ¿Qué semejanzas y diferencias observa entre estos datos muestrales?

2. Utilice los métodos del capítulo 10 para comentar las diferencias entre las medias poblacionales de los dos métodos. Analice sus hallazgos.
3. Calcule la desviación estándar y la varianza de los datos obtenidos con cada método. Realice una prueba de hipótesis para la igualdad de las varianzas poblacionales en los datos obtenidos con los dos enfoques. Explique sus hallazgos.
4. ¿Qué conclusión obtiene acerca de las diferencias entre los dos métodos? ¿Qué recomienda? Explique.
5. ¿Sugiere otros datos o pruebas que sean de utilidad antes de decidir qué programa de capacitación usar?

Apéndice 11.1 Varianzas poblacionales con Minitab

Aquí se describe cómo usar Minitab para realizar una prueba de hipótesis que involucre dos varianzas poblacionales.

WEB archivo SchoolBus

Se usarán los datos de la sección 11.2 sobre el estudio para la elección del servicio de autobús escolar de Dullus County School. Los tiempos correspondientes a la empresa Milbank se encuentran en la columna Cl y los tiempos correspondientes a Gulf Park, en la columna C2. Para realizar la prueba de hipótesis H_0: $\sigma_1^2 = \sigma_2^2$ y H_a: $\sigma_1^2 \neq \sigma_2^2$, se sigue el procedimiento de Minitab que se describe a continuación.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Basic Statistics**.
Paso 3. Elija **2-Variances**.
Paso 4. Cuando el cuadro de diálogo 2-Variances aparezca:
Seleccione **Samples in different columns**.
Ingrese Cl en el cuadro **First**.
Ingrese C2 en el cuadro **Second**.
Haga clic en **OK**.

En la sección titulada F-Test (normal distribution) se desplegará la información sobre la prueba; el estadístico de prueba $F = 2.40$ y el valor-$p = 0.81$. Con este procedimiento de Minitab se realiza una prueba de dos colas para la igualdad de las varianzas poblacionales. Por tanto, cuando se usa esta rutina para una prueba de una cola, debe recordar que el área en ésta es la mitad del área del valor-p para dos colas; entonces será relativamente fácil calcular el valor-p para la prueba de una cola.

Apéndice 11.2 Varianzas poblacionales con Excel

Aquí se describe cómo usar Excel para realizar una prueba de hipótesis con dos varianzas poblacionales.

WEB archivo SchoolBus

Se usarán los datos de la sección 11.2 sobre el estudio para la elección del servicio de autobús escolar de Dullus County School. En la hoja de cálculo de Excel aparece en la celda Al la etiqueta Milbank y la etiqueta Gulf Park en la celda Bl. Los datos muestrales del tiempo de Milbank se encuentran en las celdas A2:A27, y los datos muestrales del tiempo de Gulf Park, en las celdas B2:B17. Los pasos para realizar la prueba de hipótesis H_0: $\sigma_1^2 = \sigma_2^2$ y H_a: $\sigma_1^2 \neq \sigma_2^2$ se presentan a continuación.

Paso 1. Haga clic en la ficha **Data** en la cinta.
Paso 2. En el grupo **Analysis** haga clic en **Data Analysis**.
Paso 3. Cuando el cuadro de diálogo Data Analysis aparezca:
Elija **F-Test Two-Sample for Variances**.
Haga clic en **OK**.
Paso 4. Cuando el cuadro de diálogo F-Test Two Sample for Variances aparezca:
Ingrese Al:A27 en el cuadro **Variable 1 Range**.
Ingrese B1:B17 en el cuadro **Variable 2 Range**.

Seleccione **Labels**.
Ingrese 0.05 en cuadro **Alpha**.
(*Nota.* En este procedimiento, Excel usa alfa como área en la cola superior.)
Seleccione **Output Range** e ingrese Cl en el cuadro.
Haga clic en **OK**.

El resultado P(F<=f) one-tail = 0.0405 es el área en una cola correspondiente al estadístico de prueba $F = 2.40$. Por tanto, el valor-p para dos colas es $2(0.0405) = 0.081$. Si se trata de una prueba de hipótesis de una cola, el área en una cola que aparece en la celda titulada P(F<=f) one tail proporciona la información necesaria para determinar el valor-p de la prueba.

Apéndice 11.3 Desviación estándar poblacional simple con StatTools

WEB archivo
BusTimes

En este apéndice se muestra el uso de StatTools para realizar pruebas de hipótesis acerca de una desviación estándar poblacional. StatTools realiza estas pruebas sobre la desviación estándar poblacional, no directamente sobre la varianza poblacional. Como ilustración se usará el ejemplo analizado en la sección 11.1 acerca de los horarios de llegada del autobús a una estación.

Se inicia con el uso del Data Set Manager a efecto de crear un conjunto de datos de StatTools para los datos del archivo BusTimes utilizando el procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes pueden utilizarse para probar la hipótesis H_0: $\sigma \leq 2$ frente a H_a: $\sigma > 2$.

Paso 1. Haga clic en la ficha **StatTools** en la cinta.
Paso 2. En el grupo **Analyses**, haga clic en **Statistical Inference**.
Paso 3. Elija la opción **Hypothesis Test**.
Paso 4. Seleccione **Mean/Std. Deviation**.
Paso 5. Cuando el cuadro de diálogo StatTools-Hypothesis Test for Mean/Std. Deviation aparezca:
En **Analysis Type**, elija **One-Sample Analysis**.
En la sección **Variables**, seleccione **Times**.
En la sección **Hypothesis Tests to Perform**:
Retire la marca de selección de la caja **Mean**.
Elija la opción **Standard Deviation**.
Ingrese 2 en la caja **Null Hypothesis Value**.
Seleccione **Greater Than Null Value (One-Tailed Test)** en la caja **Alternative Hypothesis**.
Haga clic en **OK**.

Aparecerán los resultados de la prueba de hipótesis, en los que se incluyen el valor-p y el valor del estadístico de prueba χ^2.

CAPÍTULO 12

Pruebas de bondad de ajuste e independencia

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: UNITED WAY

12.1 PRUEBA DE BONDAD DE AJUSTE: UNA POBLACIÓN MULTINOMIAL

12.2 PRUEBA DE INDEPENDENCIA

12.3 PRUEBA DE BONDAD DE AJUSTE: DISTRIBUCIONES DE POISSON Y NORMAL
Distribución de Poisson
Distribución normal

ESTADÍSTICA *en* LA PRÁCTICA

UNITED WAY*

ROCHESTER, NUEVA YORK

United Way of Greater Rochester es una organización sin fines de lucro que se dedica a mejorar la calidad de vida de las personas en los siete condados en los que proporciona servicios para satisfacer las necesidades humanas de cuidado más importantes de las comunidades.

La campaña anual de United Way/Cruz Roja para recolección de fondos que se realiza en cada primavera, patrocina cientos de programas ofrecidos por más de 200 proveedores de servicios, los cuales atienden una amplia variedad de necesidades humanas (físicas, mentales y sociales) y a personas de cualquier edad, origen y situación económica.

Debido a la gran cantidad de voluntarios involucrados, United Way of Great Rochester mantiene sus costos de operación en sólo 8 centavos por dólar recaudado.

La organización decidió ordenar un estudio para conocer la percepción de la comunidad sobre la caridad. Se realizaron sesiones de grupo (*focus-group*) con profesionales, personal de servicio y trabajadores generales para recabar información preliminar sobre sus percepciones. La información obtenida se utilizó para elaborar cuestionarios del estudio. Éstos fueron probados, modificados y distribuidos a 440 personas, y se obtuvieron 323 cuestionarios contestados.

A partir de los datos recolectados se prepararon diversos estadísticos descriptivos, como distribuciones de frecuencias y tabulaciones cruzadas. Una parte importante del análisis incluyó el manejo de tablas de contingencia y de pruebas ji-cuadrada de independencia. Uno de los usos de esas pruebas estadísticas consistió en determinar si las percepciones sobre los gastos administrativos eran independientes de la ocupación.

Las hipótesis para la prueba de independencia fueron las siguientes.

H_0: la percepción sobre los gastos administrativos de United Way es independiente de la ocupación del entrevistado.



Los programas de United Way atienden necesidades de niños y adultos. © Ed Bock/CORBIS

H_a: la percepción sobre los gastos administrativos de United Way no es independiente de la ocupación del entrevistado.

Dos de las preguntas del estudio suministraron los datos para la prueba estadística. Con una se obtenía información sobre las percepciones acerca de los recursos que se destinaban a gastos administrativos (hasta 10, 11-20 y 21% o más). Con la otra se indagaba sobre la ocupación del entrevistado.

La prueba ji-cuadrada con 0.05 como nivel de significancia llevó a rechazar la hipótesis nula de independencia y, de esta manera, a la conclusión de que las percepciones sobre los gastos administrativos variaban según la ocupación. Estos gastos eran en realidad menores de 9%, pero 35% de los entrevistados tenía la idea de que representaban 21% o más. Así que muchos tenían una percepción inadecuada sobre los costos administrativos. De este conjunto, los empleados de línea de producción y de oficina, los vendedores y los técnicos profesionales tenían percepciones más equivocadas que otros grupos.

El estudio sobre la percepción de la comunidad sirvió para que United Way of Greater Rochester hiciera ajustes a sus programas y actividades de recaudación de fondos. En este capítulo usted aprenderá cómo se realiza una prueba estadística de independencia como la descrita aquí.

* Los autores agradecen al doctor Philip R. Tyler, consultor de Marketing de United Way, por proporcionar este material para *Estadística en la práctica*.

En el capítulo 11 se vio cómo se utiliza la distribución ji-cuadrada en estimaciones y pruebas de hipótesis para la varianza poblacional. En este capítulo se presentan otras dos pruebas de hipótesis, ambas basadas en el uso de la distribución ji-cuadrada. Como otros procedimientos de pruebas de hipótesis, en éstas se comparan los resultados muestrales con los resultados esperados cuando la hipótesis nula es verdadera. La conclusión de la prueba de hipótesis se basa en qué tan "cerca" se encuentran los resultados muestrales de los esperados.

En la siguiente sección se presenta la prueba de bondad de ajuste para una población multinomial. Más adelante se aborda la prueba de independencia usando tablas de contingencia y después las pruebas de bondad de ajuste para distribuciones normales y de Poisson.

12.1 Prueba de bondad de ajuste: una población multinomial

En esta sección se estudia el caso en que cada elemento de una población corresponde a una y sólo a una de varias clases o categorías. A una población como esta se le denomina **población multinomial**, la cual se puede entender como una extensión de la distribución binomial al caso en el que hay tres o más categorías de resultados. En cada ensayo de un experimento multinomial, uno y sólo uno de los resultados ocurre. Se supone que cada ensayo es independiente y que en todos ellos las probabilidades para los resultados permanecen constantes.

Los supuestos en un experimento multinomial son los mismos que en un experimento binomial, salvo que en el primero hay tres o más resultados por ensayo.

Como ejemplo, considere el estudio sobre participación de mercado que realizó la firma Scott Marketing Research. A lo largo de los años, las participaciones en el mercado se han estabilizado en 30% para la empresa A, 50% para la B y 20% para la C. Hace poco, la empresa C desarrolló un producto nuevo y mejorado para sustituir uno de los que tenía en circulación, y pidió a la firma Scott Marketing Research que determinara si el nuevo producto modificaría su participación de mercado.

En este caso la población de interés es multinomial y cada cliente se clasifica como comprador de la empresa A, de la B o de la C. De manera que tenemos una población multinomial con tres resultados. Para las proporciones se usa la siguiente notación.

p_A = participación de mercado de la empresa A
p_B = participación de mercado de la empresa B
p_C = participación de mercado de la empresa C

Scott Marketing Research realizará un estudio muestral y calculará la proporción poblacional que prefiere el producto de cada empresa. Después aplicará una prueba de hipótesis para ver si el nuevo producto modifica la participación de mercado. Suponga que el producto no altera dicha participación; entonces, las hipótesis nula y alternativa serán las siguientes.

H_0: $p_A = 0.30$; $p_B = 0.50$, y $p_C = 0.20$
H_a: las proporciones poblacionales no son $p_A = 0.30$; $p_B = 0.50$, y $p_C = 0.20$

Si los resultados muestrales llevan al rechazo de H_0, Scott Marketing Research tendrá evidencias de que la introducción del nuevo producto afecta la participación de mercado.

Considere que para este estudio la firma de investigación ha empleado un panel de 200 consumidores. A cada individuo se le pide que indique su preferencia entre el producto de la empresa A, el de la empresa B o el nuevo producto de la empresa C. Las 200 respuestas obtenidas se presentan a continuación en forma resumida.

El panel de 200 consumidores en el que a cada consumidor se le pide que elija una de tres alternativas, es equivalente a un experimento multinomial consistente en 200 ensayos.

	Frecuencia observada	
Producto de la empresa A	**Producto de la empresa B**	**Producto de la empresa C**
48	98	54

Ahora se realiza la **prueba de bondad de ajuste** para determinar si la muestra de las 200 preferencias de los clientes coincide con la hipótesis nula. La prueba de bondad de ajuste se

basa en la comparación de los resultados muestrales *observados* con los resultados *esperados* bajo el supuesto de que la hipótesis nula es verdadera. Por tanto, el paso siguiente es calcular las preferencias esperadas en los 200 clientes asumiendo que $p_A = 0.30$, $p_B = 0.50$ y $p_C = 0.20$. Al hacerlo, se tendrán los resultados esperados.

Frecuencia esperada

Producto de la empresa A	Producto de la empresa B	Nuevo producto de la empresa C
200(0.30) = 60	200(0.50) = 100	200(0.20) = 40

Como se observa, la frecuencia esperada de cada categoría se encuentra al multiplicar el tamaño de la muestra, 200, por la proporción hipotética de esa categoría.

En la prueba de bondad de ajuste lo que interesa son las diferencias entre frecuencias observadas y esperadas. Grandes diferencias entre estas frecuencias harán dudar sobre el supuesto de que las proporciones o participación de mercado hipotética son correctas.

El siguiente estadístico de prueba ayuda a responder la pregunta de si las diferencias entre las frecuencias observadas y esperadas son "grandes" o "pequeñas".

ESTADÍSTICO DE PRUEBA PARA LA BONDAD DE AJUSTE

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \qquad (12.1)$$

donde:

f_i = frecuencia observada en la categoría i
e_i = frecuencia esperada en la categoría i
k = número de categorías

Nota. El estadístico de prueba tiene una distribución ji-cuadrada con $k - 1$ grados de libertad, siempre que en todas las categorías las frecuencias esperadas sean 5 *o más*.

Ahora, de regreso al ejemplo de Scott Marketing Research, los datos muestrales se emplearán para probar la hipótesis de que en la población multinomial las proporciones sigan siendo $p_A = 0.30$, $p_B = 0.50$ y $p_C = 0.20$. El nivel de significancia que se usará es $\alpha = 0.05$.Mediante las frecuencias observadas y esperadas se calcula el valor del estadístico de prueba.Como las frecuencias esperadas son todas 5 o más, se calcula el estadístico de prueba ji-cuadrada como se indica en la tabla 12.1, y se obtiene $\chi^2 = 7.34$.

La prueba de bondad de ajuste es siempre una prueba de una cola en la que el rechazo se presenta en la cola superior de la distribución ji-cuadrada.

La hipótesis nula es rechazada si las diferencias entre las frecuencias observadas y esperadas son *grandes*. Estas diferencias darán un valor grande del estadístico de prueba. Entonces, la prueba de bondad de ajuste será siempre una prueba de cola superior. El área en la cola superior se emplea en el estadístico de prueba y en el método del valor-p para determinar si se puede rechazar la hipótesis nula. Para $k - 1 = 3 - 1 = 2$ grados de libertad, la tabla de ji-cuadrada (tabla 3 del apéndice B) proporciona lo siguiente.

En la sección 11.1 se presentó una introducción a la distribución ji-cuadrada y al uso de la tabla correspondiente.

Área en la cola superior	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
Valor χ^2 (2 *gl*)	4.605	5.991	7.378	9.210	10.597

$\chi^2 = 7.34$

TABLA 12.1 Cálculo del estadístico de prueba ji-cuadrada para el estudio de participación de mercado realizado por Scott Marketing Research

Categoría	Proporción hipotética	Frecuencia observada (f_i)	Frecuencia esperada (e_i)	Diferencia ($f_i - e_i$)	Cuadrado de la diferencia $(f_i - e_i)^2$	Cuadrado de la diferencia dividido entre la frecuencia esperada $(f_i - e_i)^2/e_i$
Empresa A	0.30	48	60	−12	144	2.40
Empresa B	0.50	98	100	−2	4	0.04
Empresa C	0.20	54	40	14	196	4.90
Total		200				$\chi^2 = 7.34$

El estadístico de prueba $\chi^2 = 7.34$ lo encontramos entre 5.991 y 7.378. Por consiguiente, el área correspondiente en la cola superior o valor-p debe estar entre 0.05 y 0.025. Como el valor-$p \leq \alpha = 0.05$, H_0 es rechazada y se concluye que la introducción del nuevo producto de la empresa C sí modificará la estructura de participación de mercado actual. Se pueden utilizar los procedimientos de Excel y Minitab que se presentan en el apéndice F, al final del libro, para demostrar que $\chi^2 = 7.34$ da un valor-$p = 0.0255$.

En lugar del método del valor-p se puede utilizar el método del valor crítico, con el que se llega a la misma conclusión. Como $\alpha = 0.05$ y los grados de libertad son 2, el valor crítico para el estadístico de prueba es $\chi^2_{0.05} = 5.991$. La regla de rechazo de la cola superior se convierte en

$$\text{Rechazar } H_0 \text{ si } \chi^2 \geq 5.991$$

Como $7.34 > 5.991$, se rechaza H_0. Con los métodos del valor crítico o del valor-p se llega a la misma conclusión.

Aunque no se obtienen más conclusiones como resultado de la prueba, es posible comparar las frecuencias observadas y esperadas de manera informal para tener una idea de cómo ha cambiado la estructura de participación de mercado. Se observa que para la empresa C, la frecuencia observada, 54, es mayor que la frecuencia esperada, 40. Como esta última se basaba en la participación de mercado existente, el que la frecuencia observada sea mayor indica que el nuevo producto tendrá un efecto positivo sobre la participación de mercado de la empresa. Al comparar las frecuencias observadas y esperadas de las otras dos empresas, se observa que la C gana en participación de mercado afectando más a la A que a la B.

A continuación se presentan, en forma resumida, los pasos generales que se siguen en una prueba de bondad de ajuste para una distribución poblacional multinomial hipotética.

DISTRIBUCIÓN MULTINOMIAL DE PRUEBAS DE BONDAD DE AJUSTE: RESUMEN

1. Establecer las hipótesis nula y alternativa:

   H_0: la población tiene una distribución multinomial con la probabilidad específica de cada una de las k categorías

   H_a: la población no tiene una distribución multinomial con la probabilidad específica de cada una de las k categorías

2. Seleccionar una muestra aleatoria y anotar las frecuencias observadas f_i en cada categoría.
3. Suponer que la hipótesis nula es verdadera y determinar la frecuencia esperada e_i en cada categoría multiplicando la probabilidad de esa categoría por el tamaño de la muestra.

4. Calcular el valor del estadístico de prueba.

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

5. Regla de rechazo:

Método del valor-p: Rechazar H_0 si el valor-$p \leq \alpha$
Método del valor crítico: Rechazar H_0 si $\chi^2 \geq \chi^2_\alpha$

donde α es el nivel de significancia utilizado para la prueba y se tienen $k - 1$ grados de libertad.

Ejercicios

Métodos

AUTOevaluación

1. Pruebe las hipótesis siguientes usando la prueba de bondad de ajuste χ^2.

$$H_0: p_A = 0.40; p_B = 0.40, \text{ y } p_C = 0.20$$
$$H_a: \text{las proporciones poblacionales no son}$$
$$p_A = 0.40; p_B = 0.40, \text{ y } p_C = 0.20$$

En una muestra de 200 elementos, se tiene que 60 pertenecen a la categoría A, 120 a la B y 20 a la C. Use $\alpha = 0.01$ y pruebe si las proporciones son las establecidas en H_0.

a) Use el método del valor-*p*.
b) Repita la prueba usando el método del valor crítico.

2. Suponga que tiene una población multinomial con cuatro categorías: A, B, C y D. La hipótesis nula indica que la proporción de elementos es la misma en todas las categorías. La hipótesis nula es:

$$H_o: p_A = p_B = p_C = p_D = 0.25$$

Una muestra de tamaño 300 proporciona los resultados siguientes.

A: 85 B: 95 C: 50 D: 70

Use $\alpha = 0.05$ para determinar si se rechaza H_0. ¿Cuál es el valor-*P*?

Aplicaciones

AUTOevaluación

3. Durante las primeras 13 semanas de la temporada de televisión en Estados Unidos se registraron las proporciones siguientes de audiencia los sábados de 8:00 p.m. a 9:00 p.m.: ABC 29%, CBS 28%, NBC 25% e independientes 18%. Dos semanas después, en una muestra de 300 hogares se obtuvieron las audiencias siguientes en sábado por la noche: ABC 95 hogares, CBS 70, NBC 89, e independientes 46 hogares. Use $\alpha = 0.05$ para determinar si han variado las proporciones en la audiencia de televidentes.

4. M&M/MARS, fabricante de los chocolates M&M®, realizó un sondeo nacional en el que más de 10 millones de personas dieron su preferencia para un nuevo color. El resultado de este sondeo fue el remplazo del color café claro por uno azul. En el folleto "Colors", distribuido por

el área de Asuntos del Consumidor de M&M/Mars, la distribución de los colores de las lunetas (chocolates en forma de gragea) es la siguiente.

Café	Amarillo	Rojo	Naranja	Verde	Azul
30%	20%	20%	10%	10%	10%

En un estudio posterior se emplearon como muestras bolsas de 1 libra para determinar si los porcentajes reportados eran válidos. En una muestra de 506 lunetas se obtuvieron los siguientes resultados.

Café	Amarillo	Rojo	Naranja	Verde	Azul
177	135	79	41	36	38

Use $\alpha = 0.05$ para determinar si estos datos coinciden con los porcentajes reportados por la empresa.

5. ¿Dónde es más frecuente que las mujeres compren ropa informal? De los datos de la U.S. Shopper Database se obtuvieron los porcentajes siguientes acerca de las compras de ropa que realizan las mujeres en distintas tiendas (*The Wall Street Journal*, 28 de enero de 2004).

Tienda	Porcentaje	Tienda	Porcentaje
Wal-Mart	24	Kohl's	8
Tiendas departamentales tradicionales	11	Por correo	12
JC Penney	8	Otras	37

La categoría "Otras" incluye tiendas como Target, Kmart y Sears, así como numerosos establecimientos especializados. Ninguna de las tiendas de este grupo registra más de 5% de las compras femeninas. Un estudio reciente en el que se usó una muestra de 140 mujeres en Atlanta, Georgia, reveló las siguientes cifras: Wal-Mart 42, tiendas departamentales tradicionales 20, JC Penney 8, Kohl's 10, por correo 21, y otras 39. ¿Esta muestra indica que las compras femeninas en Atlanta difieren de las preferencias que indica la U.S. Shopper Database? ¿Cuál es el valor-p? Utilice $\alpha = 0.05$. ¿Cuál es su conclusión?

6. La American Bankers Association recoge datos sobre el uso de tarjetas de crédito o débito, cheques personales y efectivo para el pago de compras en tienda (*The Wall Street Journal*, 16 de diciembre de 2003). En 1999 los datos encontrados fueron los siguientes.

Compras en tienda	Porcentaje
Tarjeta de crédito	22
Tarjeta de débito	21
Cheque personal	18
Efectivo	39

En una muestra tomada en 2003 se encontró que de cada 220 compras en tienda, en 46 se usó tarjeta de crédito, en 67 tarjeta de débito, en 33 cheque personal y en 74 pago en efectivo.

a) Con $\alpha = 0.01$, ¿se puede concluir que en este periodo de cuatro años, de 1999 a 2003, se ha generado un cambio en la manera en que los clientes pagan sus compras en las tiendas? ¿Cuál es el valor-p?

b) A partir de los datos muestrales de 2003, calcule el porcentaje de uso de cada método de pago. ¿Cuál parece haber sido el principal o los principales cambios ocurridos en este periodo de cuatro años?

c) ¿Qué porcentaje de los pagos se efectuó con tarjeta (de crédito o de débito) en 2003?

7. En el Shareholder Scoreboard de *The Wall Street Journal* se sigue el desempeño de las 1 000 principales empresas de Estados Unidos (*The Wall Street Journal*, 10 de marzo de 2003). El comportamiento de cada firma se califica con base en los rendimientos anuales totales, que comprenden cambios en los precios de las acciones y la reinversión de dividendos. Las calificaciones se asignan clasificando las 1 000 empresas en cinco grupos, del A (20% mejor), B (siguiente 20%), hasta el E (20% inferior). Aquí se muestran las calificaciones obtenidas en un año por una muestra de 60 de las empresas más grandes. ¿El comportamiento de éstas difiere del desempeño de las 1 000 empresas del Shareholder Scoreboard? Utilice $\alpha = 0.05$.

A	B	C	D	E
5	8	15	20	12

8. ¿Qué tan bueno es el servicio proporcionado por las aerolíneas a sus clientes? Un estudio reveló las siguientes evaluaciones de los viajeros: 3% excelente, 28% bueno, 45% aceptable y 24% malo (*BusinessWeek*, 11 de septiembre de 2000). Un estudio de seguimiento del servicio de las compañías telefónicas, con una muestra de 400 adultos, reveló las siguientes evaluaciones: 24 excelente, 124 bueno, 172 aceptable y 80 malo. ¿La distribución de las evaluaciones de las compañías telefónicas difiere de la distribución de las evaluaciones de las líneas aéreas? Pruebe con $\alpha = 0.01$. ¿Cuál es su conclusión?

12.2 Prueba de independencia

Otra aplicación importante de la distribución ji-cuadrada implica el uso de datos muestrales para probar la independencia de dos variables. Para ilustrar la prueba de independencia se considerará la realizada por Alber's Brewery, de Tucson, Arizona. La empresa produce y distribuye tres tipos de cerveza: ligera, clara y oscura. Al analizar los segmentos de mercado de las tres bebidas, el grupo de investigación de mercados se preguntó si la inclinación de los consumidores por estos tipos de cerveza difería entre hombres y mujeres. En caso de que las preferencias fueran independientes del género del consumidor, se iniciaría una campaña publicitaria para todas las cervezas de Alber's. Pero si las preferencias por los distintos tipos de cerveza dependían del género del consumidor, la empresa ajustaría sus promociones a los diferentes mercados meta.

Se usó una prueba de independencia para determinar si la preferencia por un tipo de cerveza (ligera, clara u oscura) era independiente del género del consumidor (hombre o mujer). Las hipótesis fueron las siguientes.

H_0: la preferencia por un tipo de cerveza es independiente del género del consumidor
H_a: la preferencia por un tipo de cerveza no es independiente del género del consumidor

Para describir la situación a estudiar se usa la tabla 12.2. Después de identificar la población como todos los consumidores de cerveza, hombres y mujeres, se toma una muestra y a cada

TABLA 12.2 Tabla de contingencia de cerveza preferida y género del consumidor

		Cerveza preferida		
		Ligera	**Clara**	**Oscura**
Género	**Hombre**	celda (1,1)	celda (1,2)	celda (1,3)
	Mujer	celda (2,1)	celda (2,2)	celda (2,3)

TABLA 12.3 Resultados muestrales del tipo de cerveza que prefieren hombres y mujeres (frecuencias observadas)

		Cerveza preferida			
		Ligera	**Clara**	**Oscura**	**Total**
Género	**Hombre**	20	40	20	80
	Mujer	30	30	10	70
	Total	50	70	30	150

Para probar si dos variables son independientes, se toma una muestra y se prepara una tabulación cruzada para resumir los datos de las dos variables simultáneamente.

individuo se le pide que indique cuál de las tres cervezas de Alber's prefiere. Cada sujeto de la muestra se clasificará en una de las seis celdas de la tabla. Así, por ejemplo, se puede tener un individuo hombre que prefiera la cerveza clara [celda (1,2)], o una mujer que se incline más por la ligera [celda (2,1)], o una mujer que prefiera la cerveza oscura [celda (2,3)], y así sucesivamente. Dado que en la tabla se han enumerado todas las posibles combinaciones de cerveza preferida y género o, en otras palabras, todas las posibles contingencias, a la tabla 12.2 se le llama **tabla de contingencia**. Como en la prueba de independencia se usa el formato de este tipo de tabla, a esta prueba también se le suele llamar *prueba de tabla de contingencia*.

Suponga que toma una muestra aleatoria simple de 150 consumidores. Cada individuo de la muestra prueba los tres tipos de cerveza y después se le pide que indique cuál prefiere o cuál es su primera elección. En la tabulación cruzada de la tabla 12.3 se presenta el resumen de las respuestas recabadas en el estudio. Como se ve, los datos para la prueba de independencia se obtienen contando las cantidades o frecuencias correspondientes a cada celda o categoría. De las 150 personas de la muestra, 20 hombres prefirieron la cerveza ligera, 40 la clara, 20 la oscura, etcétera.

Los datos de la tabla 12.3 son las frecuencias observadas para cada una de las seis clases o categorías. Si se determinan las frecuencias esperadas bajo el supuesto de independencia entre cerveza preferida y género del consumidor, se puede emplear la distribución ji-cuadrada para establecer si existe diferencia significativa entre las frecuencias observadas y las esperadas.

Las frecuencias esperadas para las celdas de la tabla de contingencia se basan en la idea siguiente. Primero se supone que la hipótesis nula es verdadera; es decir, que la cerveza preferida es independiente del género del consumidor. Después se observa que en la muestra de 150 consumidores, 50 prefirieron la cerveza ligera, 70 la clara y 30 la oscura. En términos de proporciones se concluye que $^{50}/_{150} = {}^{1}/_{3}$ de los consumidores eligió la cerveza ligera; $^{70}/_{150} = {}^{7}/_{15}$ la cerveza clara, y $^{30}/_{150} = {}^{1}/_{5}$ la cerveza oscura. Si el supuesto de independencia es correcto, estas proporciones serán las que se observen tanto entre los hombres como entre las mujeres. Por consiguiente, bajo el supuesto de independencia, es de esperarse que en la muestra de 80 sujetos del sexo masculino, $({}^{1}/_{3})80 = 26.67$ prefieran la cerveza ligera, $({}^{7}/_{15})80 = 37.33$ elijan la cerveza clara y $({}^{1}/_{5})80 = 16$ prefieran la oscura. Al aplicar las proporciones correspondientes a los 70 consumidores del sexo femenino, se obtienen las frecuencias esperadas que aparecen en la tabla 12.4.

Sea e_{ij} la frecuencia esperada en la fila i, columna j de la tabla de contingencia. Mediante dicha notación, ahora se reconsidera el cálculo de la frecuencia esperada correspondiente a los

TABLA 12.4 Frecuencias esperadas si la preferencia por uno de los tipos de cerveza es independiente del género del consumidor

		Cerveza preferida			
		Ligera	**Clara**	**Oscura**	**Total**
Género	**Hombre**	26.67	37.33	16.00	80
	Mujer	23.33	32.67	14.00	70
	Total	50.00	70.00	30.00	150

hombres (fila $i = 1$) que prefieren la cerveza clara (columna $j = 2$), es decir, la frecuencia esperada e_{12}. Siguiendo el argumento anterior para el cálculo de esta frecuencia, vemos que

$$e_{12} = (7/15)80 = 37.33$$

Esta expresión se formula de una manera ligeramente diferente como

$$e_{12} = (7/15)80 = (70/150)80 = \frac{(80)(70)}{150} = 37.33$$

Observe que en esta expresión, 80 es el número total de hombres (total de la fila 1), 70 es la cantidad total de individuos que prefieren la cerveza clara (total de la columna 2) y 150 es el tamaño total de la muestra. Vemos entonces que

$$e_{12} = \frac{(\text{total de la fila 1})(\text{total de la columna 2})}{\text{tamaño de la muestra}}$$

La generalización de esta expresión lleva a la fórmula siguiente para obtener las frecuencias esperadas en una tabla de contingencia para una prueba de independencia.

FRECUENCIAS ESPERADAS PARA TABLAS DE CONTINGENCIA BAJO EL SUPUESTO DE INDEPENDENCIA

$$e_{ij} = \frac{(\text{total de la fila } i)(\text{total de la columna } j)}{\text{tamaño de la muestra}} \qquad \textbf{(12.2)}$$

Al aplicar esta fórmula para los consumidores hombres que prefieren la cerveza oscura, encontramos que la frecuencia esperada es $e_{13} = (80(30)/150 = 16.00$, como se ilustra en la tabla 12.4. Use la ecuación (12.2) para verificar las otras frecuencias esperadas que se presentan en esta tabla.

El procedimiento de prueba para comparar las frecuencias esperadas de la tabla 12.4 con las frecuencias observadas de la tabla 12.3 es semejante a los cálculos para la prueba de bondad de ajuste de la sección 12.1. En concreto, el valor χ^2 que se basa en las frecuencias observadas y esperadas se calcula como se indica a continuación.

ESTADÍSTICO DE PRUEBA PARA INDEPENDENCIA

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \qquad \textbf{(12.3)}$$

Donde:

f_{ij} = frecuencia observada en la categoría de la fila i y columna j de la tabla de contingencia

e_{ij} = frecuencia esperada en la categoría de la fila i y columna j de la tabla de contingencia, basada en el supuesto de independencia

Nota. Si una tabla de contingencia tiene n filas y m columnas, el estadístico de prueba tiene una distribución ji-cuadrada con $(n - 1)(m - 1)$ grados de libertad, siempre y cuando las frecuencias esperadas sean cinco o más en todas las categorías.

La doble sumatoria de la ecuación (12.3) indica que el cálculo debe efectuarse con todas las celdas que aparecen en la tabla de contingencia.

En las frecuencias esperadas registradas en la tabla 12.4 se ve que en cada categoría esta frecuencia es de cinco o más. Por tanto, se puede proceder a calcular el estadístico de prueba ji-cuadrada. En la tabla 12.5 se presentan los cálculos necesarios para obtener el estadístico de prueba ji-cuadrada que se utiliza para determinar si la preferencia por una cerveza es independiente del género del consumidor. Como se observa, el valor del estadístico de prueba es $\chi^2 = 6.12$.

El número de grados de libertad para la distribución ji-cuadrada adecuada se obtiene al multiplicar el número de filas menos 1 por el número de columnas menos 1. Como se tienen dos filas y tres columnas, los grados de libertad son $(2 - 1)(3 - 1) = 2$. Igual que con la prueba de bondad de ajuste, en la prueba de independencia H_0 es rechazada si las diferencias entre frecuencias observadas y esperadas dan un valor grande para el estadístico de prueba. De manera que la prueba de independencia es también una prueba de cola superior. La tabla de la distribución ji-cuadrada (tabla 3 del apéndice B), proporciona la información siguiente para 2 grados de libertad.

La prueba de independencia es siempre una prueba de una cola en la que la región de rechazo se encuentra en la cola superior de la distribución ji-cuadrada.

Área en la cola superior	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
Valor χ^2 (2 *gl*)	4.605	5.991	7.378	9.210	10.597

$\chi^2 = 6.12$

El estadístico de prueba $\chi^2 = 6.12$ se encuentra entre 5.991 y 7.378. Por tanto, el área correspondiente en la cola superior o valor-*p* está entre 0.05 y 0.025. Utilizando los procedimientos de Minitab o de Excel que se presentan en el apéndice F, se obtiene el valor-$p = 0.0469$. Como el valor-$p \leq \alpha = 0.05$, la hipótesis nula es rechazada y se concluye que la preferencia por una cerveza no es independiente del género del consumidor.

Para simplificar los cálculos que se requieren en una prueba de independencia, se usa software como Minitab o Excel. La información a suministrar en estos procedimientos es la tabla de contingencia de las frecuencias observadas, como se indican en la tabla 12.3. El software calcula automáticamente las frecuencias esperadas, el valor del estadístico de prueba χ^2 y el valor-*p*. En los apéndices 12.1 y 12.2 se presentan los procedimientos de Minitab y de Excel para esta prueba de independencia. En la figura 12.1 aparecen los resultados que proporciona Minitab para la prueba de Alber's Brewery.

Aunque no se pueden obtener conclusiones adicionales como resultado de la prueba, es posible realizar una comparación informal de las frecuencias observadas y esperadas para darse una idea de la dependencia entre cerveza preferida y género. Al observar las tablas 12.3 y 12.4, es notorio que en los consumidores de sexo masculino las frecuencias observadas son más altas que las esperadas en la preferencia por las cervezas clara y oscura, mientras que en las mujeres

TABLA 12.5 Cálculo del estadístico de prueba ji-cuadrada para determinar si la preferencia por un tipo de cerveza es independiente del género del consumidor

Género	**Cerveza preferida**	**Frecuencia observada (f_{ij})**	**Frecuencia esperada (e_{ij})**	**Diferencia ($f_{ij} - e_{ij}$)**	**Cuadrado de la diferencia $(f_{ij} - e_{ij})^2$**	**Cuadrado de la diferencia dividido entre la frecuencia esperada $(f_{ij} - e_{ij})^2/e_{ij}$**
Hombre	Ligera	20	26.67	−6.67	44.44	1.67
Hombre	Clara	40	37.33	2.67	7.11	0.19
Hombre	Oscura	20	16.00	4.00	16.00	1.00
Mujer	Ligera	30	23.33	6.67	44.44	1.90
Mujer	Clara	30	32.67	−2.67	7.11	0.22
Mujer	Oscura	10	14.00	−4.00	16.00	1.14
	Total	150				$\chi^2 = 6.12$

FIGURA 12.1 Resultados de Minitab para la prueba de independencia de Alber's Brewery

```
Expected counts are printed below observed counts

          Light    Regular     Dark    Total
    1        20         40       20       80
          26.67      37.33    16.00

    2        30         30       10       70
          23.33      32.67    14.00

Total        50         70       30      150

Chi-Sq = 6.122, DF = 2, P-Value = 0.047
```

la frecuencia observada en la elección de cerveza ligera es mayor que la frecuencia esperada. Dichas observaciones permiten comprender las diferentes preferencias de cerveza entre los hombres y las mujeres.

A continuación se resumen los pasos para una prueba de independencia de la tabla de contingencia.

PRUEBA DE INDEPENDENCIA: RESUMEN

1. Establecer las hipótesis nula y alternativa:

   H_0: la variable de las columnas es independiente de la variable de las filas

   H_a: la variable de las columnas no es independiente de la variable de las filas

2. Seleccionar una muestra aleatoria y anotar las frecuencias observadas en cada celda de la tabla de contingencia.
3. Utilizar la ecuación (12.2) para calcular la frecuencia esperada de cada celda.
4. Usar la ecuación (12.3) para determinar el valor del estadístico de prueba.
5. Regla de rechazo:

   Método del valor-p: Rechazar H_0 si el valor-$p \leq \alpha$

   Método del valor crítico: Rechazar H_0 si $\chi^2 \geq \chi^2_\alpha$

   donde α es el nivel de significancia, con n filas y m columnas que proporcionan $(n - 1)(m - 1)$ grados de libertad.

NOTAS Y COMENTARIOS

El estadístico de prueba para las pruebas ji-cuadrada de este capítulo requiere una frecuencia esperada de cinco para cada categoría. Cuando una categoría tiene menos de cinco, es conveniente combinar dos categorías adyacentes para tener una frecuencia esperada de cinco o más en cada categoría.

Ejercicios

Métodos

AUTOevaluación

9. La siguiente tabla de contingencia de 2×3 contiene las frecuencias observadas en una muestra de tamaño 200. Pruebe la independencia de las variables de fila y de columna utilizando la prueba χ^2 con $\alpha = 0.05$.

	Variable de las columnas		
Variable de las filas	A	B	C
P	20	44	50
Q	30	26	30

10. La siguiente tabla de contingencia 3 × 3 contiene las frecuencias observadas en una muestra de 240. Pruebe la independencia de la variable de las filas y la variable de las columnas utilizando la prueba χ^2 con $\alpha = 0.05$.

	Variable de las columnas		
Variable de las filas	A	B	C
P	20	30	20
Q	30	60	25
R	10	15	30

Aplicaciones

AUTOevaluación

11. Una de las preguntas del Subscriber Studio de *BusinessWeek* fue: "En sus viajes de negocios de los últimos 12 meses, ¿qué tipo de boleto de avión compró con mayor frecuencia?" Los datos obtenidos se presentan en la siguiente tabla de contingencia.

	Tipo de vuelo	
Tipo de boleto	Nacional	Internacional
Primera clase	29	22
Negocios/clase ejecutiva	95	121
Vuelo tradicional/clase económica	518	135

Use $\alpha = 0.05$ y pruebe la independencia entre tipo de vuelo y tipo de boleto. ¿Cuál es su conclusión?

12. Visa Card USA estudió la frecuencia con que los consumidores de diversos rangos de edad usan tarjetas plásticas (de crédito o de débito) para pagar sus compras (Associated Press, 16 de enero de 2006). A continuación se presentan los datos muestrales de 300 clientes divididos en cuatro grupos de edad.

	Grupo de edad			
Forma de pago	18–24	25–34	35–44	45 y más
Plástico	21	27	27	36
Efectivo o cheque	21	36	42	90

a) Pruebe la independencia entre el método de pago y el grupo de edad. ¿Cuál es el valor-*p*? Usando $\alpha = 0.05$, ¿cuál es su conclusión?

b) Si la forma de pago y el grupo de edad no son independientes, ¿qué observación puede formular acerca de la diferencia en el uso del plástico en los diversos grupos de edad?

c) ¿Qué consecuencias tiene este estudio para empresas como Visa, MasterCard y Discover?

13. Dados los incrementos porcentuales anuales de dos dígitos en los costos de los seguros médicos en Estados Unidos, cada día más trabajadores carecen de un seguro de esta naturaleza (*USA Today*, 23 de enero de 2004). Los siguientes datos muestrales proporcionan una comparación entre los trabajadores con y sin seguro médico en empresas pequeñas, medianas y grandes.

Para los propósitos de este estudio, las empresas pequeñas tienen menos de 100 empleados; las medianas de 100 a 999, y las grandes 1 000 o más. Los datos muestrales corresponden a 50 empleados de compañías pequeñas, 75 de medianas y 100 de grandes.

	Seguro médico		
Tamaño de la empresa	**Sí**	**No**	**Total**
Pequeño	36	14	50
Mediano	65	10	75
Grande	88	12	100

a) Realice una prueba de independencia para determinar si tener un seguro médico es independiente del tamaño de la empresa. Utilice $\alpha = 0.05$. ¿Cuál es el valor-*p* y cuál su conclusión?

b) El artículo de *USA Today* considera más probable que los empleados de empresas pequeñas carezcan de un seguro médico. Use porcentajes basados en la tabla anterior para apoyar dicha conclusión.

14. *Consumer Reports* mide la satisfacción de los propietarios de automóviles diversos con la siguiente pregunta de encuesta: "Considerando factores como precio, desempeño, confiabilidad, comodidad y disfrute, ¿compraría usted de nuevo este automóvil?" (sitio web de Consumer Reports, enero de 2009). La muestra de datos de 300 propietarios de cuatro populares sedanes medianos es la siguiente.

	Automóvil				
Volvería a comprar	**Chevrolet Impala**	**Ford Taurus**	**Honda Accord**	**Toyota Camry**	**Total**
Sí	49	44	60	46	199
No	37	27	18	19	101

a) Realice una prueba de independencia para determinar si la intención de recompra del propietario es independiente del automóvil. Utilice un nivel de significancia de 0.05. ¿Cuál es su conclusión?

b) *Consumer Reports* otorga una puntuación de satisfacción del propietario para cada automóvil, reportando el porcentaje de propietarios que comprarían el mismo tipo de unidad si pudieran volver a hacerlo. ¿Cuáles son las puntuaciones de satisfacción del automovilista para el Chevrolet Impala, el Ford Taurus, el Honda Accord y el Toyota Camry? Jerarquice los cuatro automóviles en términos de la satisfacción del propietario.

c) En el *Consumer Reports* de la clase sedán mediano se revisaron 23 diferentes automóviles. La puntuación general de satisfacción para todos los vehículos en esta clase fue 69. ¿Cómo se pueden comparar los automóviles estadounidenses (Impala y Taurus) con los japoneses (Accord y Camry) en términos de satisfacción del propietario? ¿Cuáles serán los efectos de tales hallazgos sobre la participación de mercado a futuro de esos vehículos?

15. FlightStats, Inc. recolecta datos sobre el número de vuelos programados y realizados en los principales aeropuertos de Estados Unidos. Sus datos indican que 56% de los vuelos programados en los aeropuertos de Newark, La Guardia y Kennedy se efectuaron durante una tormenta de nieve que duró tres días (*The Wall Street Journal*, 21 de febrero de 2006). Todas las aerolíneas afirman que operan siempre dentro de parámetros de seguridad establecidos: si las condiciones son muy malas, no vuelan. Los datos en la tabla superior de la siguiente página presentan una muestra de 400 vuelos programados durante tormentas de nieve.

	Aerolínea				
¿Voló?	American	Continental	Delta	United	Total
Sí	48	69	68	25	210
No	52	41	62	35	190

Use la prueba de independencia ji-cuadrada con un nivel de significancia de 0.05 para analizar estos datos. ¿Cuál es su conclusión? ¿Qué aerolínea elegiría para volar en condiciones de tormentas de nieve semejantes? Explique.

16. El interés por fuentes alternativas de energía se incrementa a medida que aumenta el precio del petróleo. En un estudio de *Financial Times*/Harris Poll se entrevistó a ciudadanos en seis países para evaluar sus actitudes hacia diversas formas alternas de energía (sitio web de Harris Interactive, 27 de febrero de 2008). Los datos de la siguiente tabla representan una parte de los hallazgos de la encuesta acerca de si las personas están a favor o en contra de la construcción de nuevas plantas de energía nuclear.

	País					
Respuesta	Gran Bretaña	Francia	Italia	España	Alemania	Estados Unidos
Muy a favor	141	161	298	133	128	204
Más a favor que en contra	348	366	309	222	272	326
En contra más que a favor	381	334	219	311	322	316
Muy en contra	217	215	219	443	389	174

a) ¿Qué tan grande fue la muestra en esta encuesta?
b) Realice una prueba de hipótesis para determinar si la actitud hacia la construcción de nuevas plantas de energía nuclear es independiente del país. ¿Cuál es su conclusión?
c) Utilizando el porcentaje de respuestas "muy a favor" y "más a favor que en contra", ¿qué país tiene la actitud más favorable hacia la construcción de nuevas plantas de energía nuclear? ¿Cuál tiene la actitud menos favorable?

17. La National Sleep Foundation utilizó una encuesta para determinar si las horas de sueño por noche son independientes de la edad (*Newsweek*, 19 de enero de 2004). Las siguientes son las horas de sueño entre semana en una muestra de personas de 49 años de edad o menos, y en otra muestra de personas de 50 años de edad o más.

	Horas de sueño				
Edad	Menos de 6	6 a 6.9	7 a 7.9	8 o más	Total
49 o menos	38	60	77	65	240
50 o más	36	57	75	92	260

a) Realice una prueba de independencia para determinar si las horas de sueño entre semana son independientes de la edad. Use a = 0.05. ¿Cuál es el valor-*p* y cuál es su conclusión?
b) ¿Cuál es su estimación del porcentaje de personas que duermen menos de 6 horas, de 6 a 6.9 horas, de 7 a 7.9, y 8 horas o más entre semana?

18. Muestras tomadas en tres ciudades, Anchorage, Atlanta y Minneapolis, se usaron para obtener información acerca del porcentaje de parejas casadas en las que los dos cónyuges trabajan (*USA Today*, 15 de enero de 2006). Analice los datos siguientes para determinar si el hecho de que los dos cónyuges trabajen es independiente del lugar donde viven. Use 0.05 como nivel

de significancia. ¿Cuál es su conclusión? Proporcione la estimación general del porcentaje de parejas casadas en las que ambos cónyuges trabajan.

	Ciudades		
Trabajan	**Anchorage**	**Atlanta**	**Minneapolis**
Ambos	57	70	63
Sólo uno	33	50	90

19. En un programa de televisión, los dos presentadores suelen dar la impresión de no estar en absoluto de acuerdo en cuanto a cuáles películas son mejores. En la evaluación de un filme pueden estar a favor ("pulgar hacia arriba"), en contra ("pulgar hacia abajo") o indiferentes. Se muestran las evaluaciones de 160 películas expresadas por los dos presentadores.

	Presentador B		
Presentador A	**En contra**	**Indiferente**	**A favor**
En contra	24	8	13
Indiferente	8	13	11
A favor	10	9	64

Para analizar estos datos, use la prueba ji-cuadrada de independencia con 0.01 como nivel de significancia. ¿Cuál es su conclusión?

12.3 Prueba de bondad de ajuste: distribuciones de Poisson y normal

En la sección 12.1 se introdujo la prueba de bondad de ajuste para poblaciones multinomiales. En general, esta prueba puede usarse con cualquier distribución de probabilidad hipotética. En esta sección se ilustra su uso para el caso en que tenemos la hipótesis de que la población tiene una distribución de Poisson o una distribución normal. Como verá, en la prueba de bondad de ajuste y en el uso de la distribución ji-cuadrada se sigue el mismo procedimiento general aplicado para la prueba de bondad de ajuste de la sección 12.1.

Distribución de Poisson

El uso de la prueba de bondad de ajuste se ilustra en el caso de una distribución poblacional que hipotéticamente tiene una distribución de Poisson. Considere, por ejemplo, las llegadas de los clientes al Dubek's Food Market en Tallahassee, Florida. Debido a recientes problemas de personal, los gerentes solicitan los servicios de una firma de consultoría para que les ayude en la programación de los empleados de caja. Después de revisar el avance de las filas en las cajas, la firma de consultoría sugerirá un procedimiento para la programación de los empleados. Este procedimiento se basa en un análisis matemático de las filas y sólo es aplicable si el número de clientes que llegan durante un determinado lapso sigue una distribución de Poisson. Por tanto, antes de poner en marcha el procedimiento de programación, habrá que recabar datos sobre las llegadas de los clientes y realizar una prueba estadística para ver si es razonable suponer que los arribos siguen una distribución de Poisson.

Las llegadas a la tienda se definen en términos de *cantidad de clientes* que entran en el establecimiento durante intervalos de 5 minutos. Por tanto, las hipótesis nula y alternativa que se indican enseguida son apropiadas para el estudio de Dubek's Food Market.

H_0: el número de clientes que entra en la tienda durante intervalos de 5 minutos tiene una distribución de probabilidad de Poisson

H_a: el número de clientes que entra en la tienda durante intervalos de 5 minutos no tiene una distribución de probabilidad de Poisson

Si una muestra de llegadas de clientes indica que no se puede rechazar H_0, Dubeck's procederá a poner en marcha el proceso de programación de la firma de consultoría. Pero si la muestra lleva a rechazar H_0, no se podrá suponer que los arribos siguen una distribución de Poisson y habrá que considerar otro procedimiento de programación.

Para probar el supuesto de que las llegadas de los clientes en las mañanas de los días entre semana siguen una distribución de Poisson, un empleado de la tienda toma una muestra aleatoria de 128 intervalos de 5 minutos en las mañanas de tres semanas consecutivas. Durante cada uno de los intervalos de 5 minutos que forman la muestra, el empleado registra el número de llegadas de clientes. Para resumir los datos, determina el número de intervalos de 5 minutos en los que no hubo ninguna llegada, el número de intervalos de 5 minutos en los que se registró una, el número de intervalos de 5 minutos en los que hubo dos, y así sucesivamente. Estos datos se presentan en la tabla 12.6.

TABLA 12.6

Frecuencias observadas en las llegadas de los clientes a Dubek's en una muestra de 128 intervalos de 5 minutos

Número de clientes que llegan	Frecuencia observada
0	2
1	8
2	10
3	12
4	18
5	22
6	22
7	16
8	12
9	6
Total	128

La tabla proporciona las frecuencias observadas en las 10 categorías. Ahora se usa la prueba de bondad de ajuste para determinar si la muestra de los 128 lapsos favorece la hipótesis relacionada con la distribución de Poisson. Para usar la prueba de bondad de ajuste se deben considerar las frecuencias esperadas para cada una de las 10 categorías, bajo el supuesto de que la distribución de las llegadas sigue dicha distribución. Es decir, si en realidad esto ocurre, es necesario calcular el número esperado de lapsos en los que llegarán cero clientes, un cliente, dos clientes, etcétera.

La función de probabilidad de Poisson, que ya se presentó en el capítulo 5, es

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} \qquad \textbf{(12.4)}$$

En esta función, μ representa la media o el número esperado de clientes que llegan en lapsos de 5 minutos, x representa la variable aleatoria del número de arribos en un lapso de 5 minutos y $f(x)$ es la probabilidad de que x clientes llegarán en un lapso de 5 minutos.

Antes de usar la ecuación (12.4) para calcular las probabilidades de Poisson, se necesita una estimación de μ, el número medio de llegadas de clientes en un lapso de 5 minutos. La media muestral de los datos de la tabla 12.6 proporciona esta estimación. Como se tienen 2 lapsos de 5 minutos en los que no llegó ningún cliente, 8 lapsos de 5 minutos en los que llegó un cliente, etc., el número total de clientes que llegan en los 128 lapsos de 5 minutos es $0(2) + 1(8) + 2(10) + \cdots + 9(6) = 640$. Este total de arribos en los 128 lapsos de la muestra dan una media de llegadas de $\mu = 640/128 = 5$ clientes por periodos de 5 minutos. Con este valor como media para la distribución de Poisson, una estimación de la función de probabilidad de Poisson en el caso de Dubek's Food Market es

$$f(x) = \frac{5^x e^{-5}}{x!} \qquad \textbf{(12.5)}$$

Esta función de probabilidad puede evaluarse para distintos valores de x y determinar así la probabilidad que corresponde a las diferentes categorías de llegadas. En la tabla 12.7 se presentan tales probabilidades, las cuales se encuentran también en la tabla 7 del apéndice B. Por ejemplo, la probabilidad de que lleguen 0 clientes en un lapso de cinco minutos es $f(0) = 0.0067$, la probabilidad del arribo de un cliente en un lapso de 5 minutos es $f(1) = 0.0337$, y así sucesivamente. Como se vio en la sección 12.1, la frecuencia esperada en cada una de las categorías se encuentra al multiplicar su probabilidad por el tamaño de la muestra. Por ejemplo, el número de lapsos de tiempo con 0 llegadas es $(0.0067)(128) = 0.86$; el número esperado de lapsos con 1 llegada es $(0.0337)(128) = 4.31$, y así sucesivamente.

Antes de hacer los cálculos de ji-cuadrada habituales para comparar las frecuencias observadas y esperadas, hay que notar que en la tabla 12.7 hay cuatro categorías que tienen una

TABLA 12.7 Frecuencias esperadas en las llegadas de clientes a Dubek's, suponiendo que sigan una distribución de Poisson con $\mu = 5$

Número de clientes que llegan (x)	Probabilidad de Poisson $f(x)$	Número esperado de lapsos de 5 minutos con x llegadas, $128\,f(x)$
0	0.0067	0.86
1	0.0337	4.31
2	0.0842	10.78
3	0.1404	17.97
4	0.1755	22.46
5	0.1755	22.46
6	0.1462	18.71
7	0.1044	13.36
8	0.0653	8.36
9	0.0363	4.65
10 o más	0.0318	4.07
	Total	128.00

Cuando en alguna categoría el número esperado es menor de cinco, no se satisfacen las condiciones para la prueba χ^2. Cuando esto ocurre, se pueden combinar categorías adyacentes para incrementar a cinco el número esperado.

frecuencia esperada menor que cinco. Esta condición incumple los requerimientos para el uso de la distribución ji-cuadrada. Sin embargo, las categorías con frecuencias esperadas menores de cinco no son una dificultad, ya que se pueden combinar categorías adyacentes para satisfacer la condición de que la frecuencia esperada sea "por lo menos de cinco". En particular, se combinan 0 y 1 en una sola categoría y también se combinan 9 y "10 o más" en otra categoría simple. De esta manera se satisface la regla de un mínimo de cinco como frecuencia esperada en cada categoría. En la tabla 12.8 se presentan las frecuencias observadas y las esperadas después de combinar categorías.

Como en la sección 12.1, la prueba de bondad de ajuste se centra en las diferencias entre frecuencias observadas y esperadas, $f_i - e_i$. Por tanto, para calcular el estadístico de prueba ji-cuadrada se usarán las frecuencias observadas y esperadas de la tabla 12.8.

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

TABLA 12.8 Frecuencias observadas y esperadas en las llegadas de clientes a Dubek's, después de combinar categorías

Número de clientes que llegan	Frecuencia observada (f_i)	Frecuencia esperada (e_i)
0 o 1	10	5.17
2	10	10.78
3	12	17.97
4	18	22.46
5	22	22.46
6	22	18.72
7	16	13.37
8	12	8.36
9 o más	6	8.72
Total	128	128.00

TABLA 12.9 Cálculo del estadístico de prueba ji-cuadrada para el estudio de Dubek's Food Market

Número de clientes que llegan (x)	Frecuencia observada (f_i)	Frecuencia esperada (e_i)	Diferencia ($f_i - e_i$)	Cuadrado de la diferencia $(f_i - e_i)^2$	Cuadrado de la diferencia dividido entre la frecuencia esperada $(f_i - e_i)^2/e_i$
0 o 1	10	5.17	4.83	23.28	4.50
2	10	10.78	−0.78	0.61	0.06
3	12	17.97	−5.97	35.62	1.98
4	18	22.46	−4.46	19.89	0.89
5	22	22.46	−0.46	0.21	0.01
6	22	18.72	3.28	10.78	0.58
7	16	13.37	2.63	6.92	0.52
8	12	8.36	3.64	13.28	1.59
9 o más	6	8.72	−2.72	7.38	0.85
Total	128	128.00			$\chi^2 = 10.96$

En la tabla 12.9 se muestran los cálculos necesarios para obtener el valor del estadístico de prueba ji-cuadrada. El valor del estadístico de prueba es $\chi^2 = 10.96$.

En general, en una prueba de bondad de ajuste la distribución ji-cuadrada tiene $k - p - 1$ grados de libertad, donde k es el número de categorías y p es el número de parámetros poblacionales estimados a partir de los datos muestrales. Para la prueba de bondad de ajuste de la distribución de Poisson, la tabla 12.9 indica que $k = 9$ categorías. Como los datos muestrales se usaron para estimar la media de la distribución de Poisson, $p = 1$, por ende tenemos $k - p - 1 = k - 2$ grados de libertad. Como $k = 9$, tenemos $9 - 2 = 7$ grados de libertad.

Suponga que en la prueba de la hipótesis nula de que la distribución de probabilidad de las llegadas de los clientes es una distribución de Poisson, se usa 0.05 como nivel de significancia. Para probar esta hipótesis, es necesario determinar el valor-p para el estadístico de prueba $\chi^2 = 10.96$ hallando el área en la cola superior de la distribución ji-cuadrada con 7 grados de libertad. En la tabla 3 del apéndice B se encuentra que $\chi^2 = 10.96$ corresponde a un área en la cola superior mayor que 0.10. Por consiguiente, sabemos que el valor-p es mayor que 0.10. Con los procedimientos de Minitab y de Excel que se describen en el apéndice F se obtiene que el valor-$p = 0.1404$. Como el valor-$p > \alpha = 0.05$, no se puede rechazar H_0. En consecuencia, no se puede descartar el supuesto de que las llegadas de los clientes, en las mañanas entre semana, sigan una distribución de probabilidad de Poisson. De esta manera, los gerentes de Dubek's pueden continuar con el procedimiento de programación para las mañanas de los días entre semana.

PRUEBA DE BONDAD DE AJUSTE PARA LA DISTRIBUCIÓN DE POISSON: RESUMEN

1. Establecer las hipótesis nula y alternativa.

H_0: la población tiene una distribución de Poisson
H_a: la población no tiene una distribución de Poisson

2. Tomar una muestra aleatoria y
   a) Registrar la frecuencia observada f_i para cada valor de la variable aleatoria de Poisson.
   b) Calcular el número medio μ de las ocurrencias.

3. Calcular, para cada valor de la variable aleatoria de Poisson, la frecuencia esperada e_i de ocurrencias. Multiplicar el tamaño de la muestra por la probabilidad de Poisson de ocurrencia para cada valor de la variable aleatoria de Poisson. Si para algún valor hay menos de cinco ocurrencias esperadas, combinar valores adyacentes y reducir el número de categorías tanto como sea necesario.
4. Determinar el valor del estadístico de prueba.

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

5. Regla de rechazo:

Método del valor-p: Rechazar H_0 si el valor-$p \leq \alpha$
Método del valor crítico: Rechazar H_0 si $\chi^2 \geq \chi^2_\alpha$

donde α es el nivel de significancia y los grados de libertad son $k - 2$.

Distribución normal

La prueba de bondad de ajuste para la distribución normal también se basa en el uso de la distribución ji-cuadrada. Se sigue un procedimiento similar al aplicado para la distribución de Poisson. Las frecuencias observadas en las diversas categorías de los datos muestrales se comparan con las frecuencias esperadas, en particular cuando se supone que la población tiene una distribución normal. Como esta distribución es continua, es necesario modificar la manera en que se definen las categorías y en que se calculan las frecuencias esperadas. La prueba de bondad de ajuste para una distribución normal se ilustrará con los datos de los exámenes presentados por las personas que solicitan empleo en Chemline, Inc. Estos datos se presentan en la tabla 12.10.

TABLA 12.10
Puntuaciones obtenidas en una muestra aleatoria de 50 solicitantes de empleo en la prueba de aptitudes de Chemline

71	66	61	65	54	93
60	86	70	70	73	73
55	63	56	62	76	54
82	79	76	68	53	58
85	80	56	61	61	64
65	62	90	69	76	79
77	54	64	74	65	65
61	56	63	80	56	71
79	84				

Cada año Chemline contrata a cerca de 400 nuevos empleados para sus cuatro plantas en Estados Unidos. El director de personal se pregunta si la población de puntuaciones de los exámenes de los solicitantes tendrá una distribución normal. Si es así, esta distribución podría servir para evaluar las puntuaciones; es decir, podrían identificarse fácilmente las que se ubican en el 20% superior, el 40% inferior, etc. Por tanto, se desea probar la hipótesis nula de que la población de las puntuaciones de estos exámenes tiene una distribución normal.

Para empezar, se obtienen estimaciones de la media y la desviación estándar de la distribución normal que se considerará en la hipótesis nula, considerando los datos de la tabla 12.10. La media muestral $\bar{x}$ y la desviación estándar muestral s se usan como estimadores puntuales de la media y la desviación estándar de la distribución normal. Los cálculos son los siguientes.

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{3\,421}{50} = 68.42$$

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{5\,310.0369}{49}} = 10.41$$

WEB archivo
Chemline

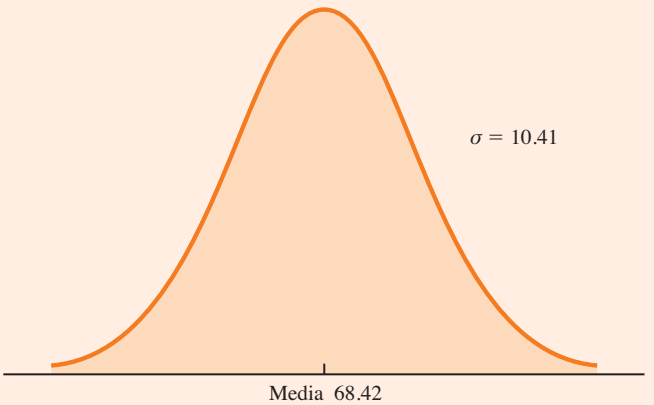
Con estos valores se establecen las siguientes hipótesis acerca de la distribución de las puntuaciones del examen de los aspirantes.

H_0: la población de las puntuaciones del examen tiene una distribución normal, con una media de 68.42 y una desviación estándar de 10.41.

H_a: la población de las puntuaciones del examen no tiene una distribución normal, con una media de 68.42 y una desviación estándar de 10.41.

En la figura 12.2 se ilustra esta distribución normal hipotética.

FIGURA 12.2 Distribución normal hipotética de las puntuaciones de los exámenes para los solicitantes de empleo en Chemline



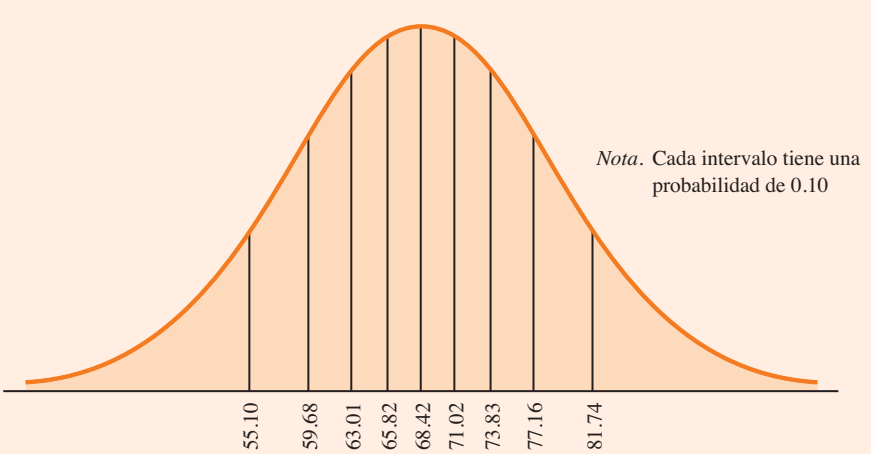
Ahora se verá cómo definir las categorías de una prueba de bondad de ajuste para una distribución normal. En el caso de la distribución de probabilidad discreta en la prueba para la distribución de Poisson fue fácil definir las categorías en términos del número de clientes que llegan, 0, 1, 2, etc. Sin embargo, para la distribución de probabilidad normal continua es necesario emplear un procedimiento diferente para definir las categorías, esto es, en términos de *intervalos* de puntuaciones de examen.

Recuerde la regla de que en cada intervalo o categoría la frecuencia esperada debe ser por lo menos de cinco. Las categorías para las puntuaciones de examen se definen de manera que la frecuencia esperada en cada una sea por lo menos de cinco. Como el tamaño de la muestra es 50, una manera de establecer las categorías es dividir la distribución normal en 10 intervalos con una misma probabilidad (vea la figura 12.3). Dado que el tamaño de la muestra es 50, se espera tener cinco resultados en cada intervalo o categoría, con lo que se satisface la regla de las frecuencias esperadas.

Con una distribución de probabilidad continua, se establecen intervalos de manera que en cada uno la frecuencia esperada sea de cinco o más.

Veamos más de cerca el procedimiento para calcular los límites de las categorías. Como se trata de una distribución de probabilidad normal, para determinar estos límites se emplean las tablas de probabilidad normal estándar. Primero se determina la puntuación de examen que

FIGURA 12.3 Distribución normal en el ejemplo de Chemline con 10 intervalos de probabilidad igual



separa el 10% inferior de las puntuaciones. En la tabla 1 del apéndice B se encuentra que el valor z correspondiente a esta puntuación de examen es -1.28. Por tanto, la puntuación $x = 68.42 - 1.28(10.41) = 55.10$ es el valor que separa el 10% inferior de las puntuaciones de examen. Para el 20% inferior tenemos $z = -0.84$ y, por tanto, $x = 68.42 - 0.84(10.41) = 59.68$. Al continuar de esta manera con la distribución normal se obtienen los valores siguientes para las puntuaciones de examen.

Porcentaje	z	Puntuación de examen
10%	−1.28	68.42 − 1.28(10.41) = 55.10
20%	−0.84	68.42 − 0.84(10.41) = 59.68
30%	−0.52	68.42 − 0.52(10.41) = 63.01
40%	−0.25	68.42 − 0.25(10.41) = 65.82
50%	0.00	68.42 + 0(10.41) = 68.42
60%	+0.25	68.42 + 0.25(10.41) = 71.02
70%	+0.52	68.42 + 0.52(10.41) = 73.83
80%	+0.84	68.42 + 0.84(10.41) = 77.16
90%	+1.28	68.42 + 1.28(10.41) = 81.74

En la gráfica 12.3 se observan estos puntos de separación o límites de los intervalos.

Una vez definidas las categorías o intervalos de las puntuaciones de examen y dado que la frecuencia esperada en cada categoría es de cinco, se usan los datos muestrales de la tabla 12.10 y se determinan las frecuencias observadas en estas categorías. Con esto se obtienen los resultados que aparecen en la tabla 12.11.

Una vez que tenemos los resultados de la tabla 12.11, el cálculo de la prueba de bondad de ajuste procede exactamente como antes. Es decir, se comparan los resultados observados y esperados calculando el valor de χ^2. En la tabla 12.12 se indican los procedimientos necesarios para obtener el estadístico de prueba ji-cuadrada. Como se ve, el valor del estadístico de prueba es $\chi^2 = 7.2$.

A efecto de determinar si este valor de 7.2 obtenido para χ^2 es suficientemente grande para rechazar H_0, se necesita consultar la tabla de la distribución ji-cuadrada. Al aplicar la regla para calcular el número de grados de libertad en la prueba de bondad de ajuste tenemos, $k - p - 1 = 10 - 2 - 1 = 7$ grados de libertad, ya que hay 10 categorías y $p = 2$ parámetros (media y desviación estándar) estimados mediante los datos muestrales.

Suponga que se prueba la hipótesis nula de que la distribución de las puntuaciones de examen es una distribución normal, utilizando 0.10 como nivel de significancia. Para probar

TABLA 12.11 Frecuencias observadas y esperadas para las puntuaciones de examen de los solicitantes de empleo en Chemline

Intervalo de puntuaciones de examen	Frecuencia observada (f_i)	Frecuencia esperada (e_i)
Menores que 55.10	5	5
55.10 a 59.68	5	5
59.68 a 63.01	9	5
63.01 a 65.82	6	5
65.82 a 68.42	2	5
68.42 a 71.02	5	5
71.02 a 73.83	2	5
73.83 a 77.16	5	5
77.16 a 81.74	5	5
81.74 y más	6	5
Total	50	50

TABLA 12.12 Cálculo del estadístico de prueba ji-cuadrada en el ejemplo de las puntuaciones de examen de los solicitantes de empleo en Chemline

Intervalos de puntuaciones de examen	Frecuencia observada (f_i)	Frecuencia esperada (e_i)	Diferencia ($f_i - e_i$)	Cuadrado de la diferencia $(f_i - e_i)^2$	Cuadrado de la diferencia dividido entre la frecuencia esperada $(f_i - e_i)^2/e_i$
Menos que 55.10	5	5	0	0	0.0
55.10 a 59.68	5	5	0	0	0.0
59.68 a 63.01	9	5	4	16	3.2
63.01 a 65.82	6	5	1	1	0.2
65.82 a 68.42	2	5	−3	9	1.8
68.42 a 71.02	5	5	0	0	0.0
71.02 a 73.83	2	5	−3	9	1.8
73.83 a 77.16	5	5	0	0	0.0
77.16 a 81.74	5	5	0	0	0.0
81.74 y más	6	5	1	1	0.2
Total	50	50			$\chi^2 = 7.2$

Dado que se estiman dos parámetros de la distribución normal, se pierden 2 grados de libertad para la prueba χ^2.

esta hipótesis se necesita calcular el valor-p del estadístico de prueba $\chi^2 = 7.2$ determinando el área correspondiente en la cola superior de la distribución ji-cuadrada con 7 grados de libertad. Al consultar la tabla 3 del apéndice B encontramos que el área en la cola superior correspondiente a $\chi^2 = 7.2$ es mayor que 0.10. Por consiguiente, sabemos que el valor-p es mayor que 0.10. Con los procedimientos de Minitab y Excel presentados en el apéndice F al final del libro, vemos que $\chi^2 = 7.2$ da un valor-$p = 0.4084$. Con el valor-$p > \alpha = 0.10$ no se puede rechazar la hipótesis nula de que la distribución de probabilidad de las puntuaciones de examen de los solicitantes de empleo en Chemline sea una distribución normal. Esta distribución se puede usar como ayuda en la interpretación de las puntuaciones de examen. A continuación se presenta un resumen de la prueba de bondad de ajuste para una distribución normal.

PRUEBA DE BONDAD DE AJUSTE PARA LA DISTRIBUCIÓN NORMAL: RESUMEN

1. Establecer las hipótesis nula y alternativa:

   H_0: la población tiene una distribución normal
   H_a: la población no tiene una distribución normal

2. Tomar una muestra aleatoria y
   a) Calcular la media muestral y la desviación estándar muestral.
   b) Definir intervalos de valores de manera que la frecuencia esperada en cada intervalo sea por lo menos de cinco. Usar intervalos de probabilidad igual es un buen enfoque.
   c) En cada uno de los intervalos definidos, anotar la frecuencia observada f_i en los valores de los datos.
3. Calcular el número esperado de ocurrencias e_i para cada uno de los intervalos de valores definidos en el paso 2*b*). Multiplicar el tamaño de la muestra por la probabilidad de que una variable aleatoria normal pertenezca al intervalo.
4. Determinar el valor del estadístico de prueba

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

5. Regla de rechazo:

Método del valor-p: Rechazar H_0 si el valor-$p \leq \alpha$
Método del valor crítico: Rechazar H_0 si $\chi^2 \geq \chi^2_\alpha$

donde α es el nivel de significancia y los grados de libertad son $k - 3$.

Ejercicios

Métodos

AUTOevaluación

20. A continuación se presenta el número de ocurrencias por periodo y su frecuencia observada. Use $\alpha = 0.05$ y la prueba de bondad de ajuste para determinar si estos datos se ajustan a una distribución de Poisson.

Número de ocurrencias	Frecuencia observada
0	39
1	30
2	30
3	18
4	3

AUTOevaluación

21. Se cree que los siguientes datos provienen de una distribución normal. Use la prueba de bondad de ajuste con $\alpha = 0.05$ para probar tal supuesto.

17	23	22	24	19	23	18	22	20	13	11	21	18	20	21
21	18	15	24	23	23	43	29	27	26	30	28	33	23	29

Aplicaciones

22. Al parecer, el número de accidentes automovilísticos por día en una determinada ciudad tiene una distribución de Poisson. A continuación se presentan los datos de una muestra de 80 días del año anterior. ¿Estos datos apoyan la creencia de que el número de accidentes por día tiene una distribución de Poisson? Use $\alpha = 0.05$.

Número de accidentes	Frecuencia observada (días)
0	34
1	25
2	11
3	7
4	3

23. Se cree que el número de llamadas telefónicas que llegan por minuto al conmutador de una empresa tiene una distribución de Poisson. Use $\alpha = 0.10$ y los datos de la página siguiente para probar este supuesto.

Número de llamadas telefónicas que llegan por minuto	Frecuencia observada
0	15
1	31
2	20
3	15
4	13
5	4
6	2
Total	100

24. Se tiene la percepción de que la demanda semanal de un producto tiene una distribución normal. Aplique una prueba de bondad de ajuste y los datos siguientes para probar este supuesto. Use $\alpha = 0.10$. La media muestral es 24.5 y la desviación estándar es 3.

18	20	22	27	22
25	22	27	25	24
26	23	20	24	26
27	25	19	21	25
26	25	31	29	25
25	28	26	28	24

25. Use $\alpha = 0.01$ y realice una prueba de bondad de ajuste para comprobar si la siguiente muestra fue tomada de una distribución normal.

55	86	94	58	55	95	55	52	69	95	90	65	87	50	56
55	57	98	58	79	92	62	59	88	65					

Una vez realizada la prueba de bondad de ajuste, elabore un histograma con todos estos datos. ¿Este gráfico respalda la conclusión a la que se llegó con la prueba de bondad de ajuste? (*Nota*. $\bar{x} = 71$ y $s = 17$.)

Resumen

En este capítulo se estudió la prueba de bondad de ajuste y la prueba de independencia, las cuales se basan en el uso de la distribución ji-cuadrada. El propósito de la prueba de bondad de ajuste es determinar si una distribución de probabilidad hipotética sirve como modelo para una determinada población de interés. Al efectuar los cálculos en una prueba de bondad de ajuste se comparan las frecuencias observadas en una muestra con las frecuencias esperadas, suponiendo que la distribución de probabilidad hipotética sea verdadera. La distribución ji-cuadrada se usa para determinar si las diferencias entre frecuencias observadas y esperadas son suficientemente grandes para rechazar la distribución de probabilidad hipotética. También se ilustra la prueba de bondad de ajuste para las distribuciones multinomial, de Poisson y normal.

Una prueba de independencia de dos variables es una extensión de la metodología empleada en la prueba de bondad de ajuste para una población multinomial. Una tabla de contingencia sirve para determinar las frecuencias observadas y esperadas. Luego se calcula un valor de

ji-cuadrada. Valores altos de ji-cuadrada, producto de diferencias grandes entre las frecuencias observadas y esperadas, lleva al rechazo de la hipótesis de independencia nula.

Glosario

Población multinomial Población en la que cada elemento corresponde a una y sólo a una de varias categorías. Una distribución multinomial es una extensión de la distribución binomial de dos a tres o más resultados.

Prueba de bondad de ajuste Prueba estadística que se realiza para determinar si una distribución de probabilidad hipotética es rechazada como distribución de una población.

Tabla de contingencia Tabla que se emplea para resumir las frecuencias observadas y esperadas de una prueba de independencia.

Fórmulas clave

Estadístico de prueba para la bondad de ajuste

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \tag{12.1}$$

Frecuencias esperadas para tablas de contingencia bajo el supuesto de independencia

$$e_{ij} = \frac{(\text{total de la fila } i)(\text{total de la columna } j)}{\text{tamaño de la muestra}} \tag{12.2}$$

Estadístico de prueba para independencia

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \tag{12.3}$$

Ejercicios complementarios

26. Para establecer cuotas de ventas, el gerente de marketing supone que en los cuatro territorios de ventas el potencial es el mismo. A continuación se presenta una muestra de 200 ventas. ¿Debe rechazarse el supuesto del gerente? Use $\alpha = 0.05$.

Territorios de ventas			
I	II	III	IV
60	45	59	36

27. Del total de los inversionistas en fondos de inversión, 7% considera que las acciones corporativas son "muy seguras", 58% las califica de "relativamente seguras", 24% de "no muy seguras", 4% de "no del todo seguras" y 7% de "no seguras". En un estudio de *BusinessWeek*/Harris se preguntó a 529 inversionistas de fondos de inversión cómo calificarían los bonos corporativos respecto de su seguridad. Las respuestas fueron las siguientes

Nivel de seguridad	Frecuencia
Muy seguros	48
Relativamente seguros	323
No muy seguros	79
No del todo seguros	16
No seguros	63
Total	529

¿La actitud de los inversionistas de fondos de inversión respecto de los bonos corporativos difiere de su actitud frente a las acciones corporativas? Apoye su conclusión con una prueba estadística. Use $\alpha = 0.01$.

28. Desde el año 2000, Toyota Camry, Honda Accord y Ford Taurus han sido los tres automóviles de pasajeros más vendidos en Estados Unidos. Los datos de ventas de 2003 indican que las participaciones en el mercado de estos tres modelos son las siguientes: Toyota Camry 37%, Honda Accord 34% y Ford Taurus 29% (*The World Almanac*, 2004). Suponga que en una muestra de 1 200 ventas de automóviles de pasajeros durante el primer trimestre de 2004 encontramos los datos siguientes.

Automóviles de pasajeros	Unidades vendidas
Toyota Camry	480
Honda Accord	390
Ford Taurus	330

¿Estos datos sirven para concluir que la participación de mercado de estos tres automóviles cambió en el primer trimestre de 2004? ¿Cuál es el valor-*p*? Use un nivel de significancia de 0.05. ¿Cuál es su conclusión?

29. Una autoridad regional de tránsito está preocupada por el número de pasajeros que viajan en una de las rutas de autobús. Al establecer la ruta se supuso que el número de pasajeros era el mismo para todos los días de la semana, de lunes a viernes. Con los datos siguientes y usando $\alpha = 0.05$, determine si el supuesto de la autoridad de tránsito es correcto.

Día	Número de pasajeros
Lunes	13
Martes	16
Miércoles	28
Jueves	17
Viernes	16

30. La encuesta anual de satisfacción laboral de *Computerworld* reveló que 28% de los gerentes de sistemas de información (SI) estaba muy satisfecho con su trabajo, 46% se sentía moderadamente satisfecho, 12% no estaba satisfecho ni insatisfecho, 10% estaba ligeramente insatisfecho y 4% estaba muy insatisfecho. Suponga que en una muestra de 500 programadores se encontraron los resultados que se listan a continuación.

Categoría	Número de entrevistados
Muy satisfechos	105
Moderadamente satisfechos	235
Ni satisfechos ni insatisfechos	55
Ligeramente insatisfechos	90
Muy insatisfechos	15

Use $\alpha = 0.05$ y realice una prueba para determinar si la satisfacción laboral entre los programadores de computadoras es diferente de la satisfacción laboral de los gerentes de SI.

31. De una muestra de piezas se obtiene la siguiente tabla de contingencia sobre la calidad, con base en el turno de producción.

Turno	Número de piezas	Número de defectos
Primero	368	32
Segundo	285	15
Tercero	176	24

Use $\alpha = 0.05$ para probar la hipótesis de que la calidad de las piezas es independiente del turno de producción. ¿Cuál es su conclusión?

32. El Suscriber Study de *The Wall Street Journal* dio a conocer datos sobre el tipo de empleo de sus suscriptores. Los siguientes datos muestrales corresponden a los clientes de las ediciones del este y del oeste.

	Región	
Tipo de empleo	Edición del este	Edición del oeste
Tiempo completo	1 105	574
Medio tiempo	31	15
Autoempleo/consultor	229	186
No empleado	485	344

Use $\alpha = 0.05$ para probar la hipótesis de que el tipo de empleo es independiente de la región. ¿Cuál es su conclusión?

33. Una institución de crédito facilitó los datos siguientes sobre préstamos aprobados por cuatro de sus agentes. Use $\alpha = 0.05$ y realice una prueba para determinar si la aprobación de las decisiones de préstamo es independiente del agente que recibe la solicitud respectiva.

	Decisión de aprobar el préstamo	
Agente de préstamo	Aprobada	Rechazada
Miller	24	16
McMahon	17	13
Games	35	15
Runk	11	9

34. En un estudio de Pew Research Center se preguntó a los entrevistados si preferían vivir en un lugar con un ritmo de vida más lento o en uno con un ritmo de vida más acelerado (*USA Today*, 13 de febrero de 2009). Tome en cuenta los siguientes datos que constituyen una muestra de las preferencias expresadas por 150 hombres y 150 mujeres.

	Ritmo de vida preferido		
Entrevistado	**Lento**	**Sin preferencia**	**Acelerado**
Hombre	102	9	39
Mujer	111	12	27

a) Combine las muestras de hombres y mujeres. ¿Cuál es el porcentaje general de entrevistados que prefieren vivir en un lugar con un ritmo de vida más lento? ¿Cuál es el porcentaje general de los que prefieren vivir en un lugar con un ritmo de vida más acelerado? ¿Cuál es su conclusión?

b) ¿Es el ritmo de vida preferido independiente del entrevistado? Use $\alpha = 0.05$. ¿Cuál es su conclusión y su recomendación?

35. Barna Research Group recolectó datos que ilustran la asistencia a la iglesia por grupos de edad (*USA Today*, 20 de noviembre de 2003). Utilice los datos muestrales para determinar si la asistencia a los templos es independiente de la edad. Use 0.05 como nivel de significancia. ¿Cuál es su conclusión? ¿Qué conclusión se puede formular acerca de la asistencia a la iglesia a medida que las personas envejecen?

	Asistencia a la iglesia		
Edad	**Sí**	**No**	**Total**
20–29	31	69	100
30–39	63	87	150
40–49	94	106	200
50–59	72	78	150

36. Los siguientes datos sobre el número de llamadas de emergencia solicitando una ambulancia se tomaron en una zona rural y en una zona urbana de Virginia.

		Día de la semana							
		Domingo	**Lunes**	**Martes**	**Miércoles**	**Jueves**	**Viernes**	**Sábado**	**Total**
Zona	**Urbana**	61	48	50	55	63	73	43	393
	Rural	7	9	16	13	9	14	10	78
	Total	68	57	66	68	72	87	53	471

Realice una prueba de independencia usando $\alpha = 0.05$. ¿Cuál es su conclusión?

37. La siguiente es una muestra aleatoria de calificaciones de los exámenes finales en un curso universitario.

55	85	72	99	48	71	88	70	59	98	80	74	93	85	74
82	90	71	83	60	95	77	84	73	63	72	95	79	51	85
76	81	78	65	75	87	86	70	80	64					

Use $\alpha = 0.05$ y realice una prueba para determinar si se debe rechazar que una distribución normal sea representativa de la distribución poblacional de las calificaciones.

38. Se reportó el índice de ocupación de las oficinas en cuatro zonas metropolitanas de California. ¿Los datos siguientes indican que la cantidad de oficinas libres es independiente de la zona metropolitana? Use 0.05 como nivel de significancia. ¿Cuál es su conclusión?

Estatus de ocupación	Los Ángeles	San Diego	San Francisco	San José
Ocupado	160	116	192	174
Vacante	40	34	33	26

39. Un vendedor realiza cuatro llamadas por día. En una muestra de 100 días, las frecuencias de los volúmenes de ventas son los siguientes.

Número de ventas	Frecuencia observada (días)
0	30
1	32
2	25
3	10
4	3
Total	100

Los registros indican que 30% de las llamadas llevan a una venta. Si las llamadas son independientes, el número de ventas por día deberá seguir una distribución binomial. La función de probabilidad binomial presentada en el capítulo 5 es

$$f(x) = \frac{n!}{x!(n - x)!} p^x(1 - p)^{n-x}$$

En este ejercicio, suponga que la población tiene una distribución binomial con $n = 4$, $p = 0.30$ y $x = 0, 1, 2, 3$ y 4.

a) Mediante la función de probabilidad binomial, calcule las frecuencias esperadas para $x = 0, 1, 2, 3$ y 4. Si es necesario, combine categorías para satisfacer el requerimiento de que la frecuencia esperada en cada categoría debe ser de cinco o más.

b) Use la prueba de bondad de ajuste para determinar si el supuesto de una distribución binomial debe ser rechazado. Use $\alpha = 0.05$. Como no fue necesario estimar ninguno de los parámetros de la distribución binomial a partir de los datos muestrales, los grados de libertad son $k - 1$, donde k es el número de categorías.

Caso a resolver

Una agenda bipartidista para el cambio

En un estudio realizado por Zogby International para *Democrat and Chronicle*, se entrevistó a más de 700 habitantes del estado de Nueva York para determinar si su gobierno estatal funcionaba bien. Los asuntos sobre los que se interrogaba a los entrevistados incluían reducciones en la remuneración de los legisladores, restricciones al cabildeo, límites de mandato para los legisladores y si los ciudadanos podían incluir sus temas en las consultas ciudadanas (*Democrat and Chronicle*, 7 de diciembre de 1997). Los resultados evidenciaron un amplio apoyo a varias reformas en los niveles político y demográfico.

Suponga que en un estudio subsecuente se entrevista a 100 individuos que viven en la región oeste de Nueva York. De cada entrevistado se registra su afiliación partidista (demócrata, independiente o republicano), así como sus respuestas a las siguientes tres preguntas.

1. ¿Se debe reducir el sueldo a los legisladores por cada día que se retrasan en el presupuesto estatal?
   Sí_____ No_____
2. ¿Debe haber más restricciones al cabildeo?
   Sí_____ No_____
3. ¿Debe haber límites para que el mandato de los legisladores sea de un número determinado de años?
   Sí_____ No_____

WEB archivo
NYReform

Las respuestas fueron codificadas usando 1 para Sí y 2 para No. La base completa de datos está disponible en el archivo titulado *NYReform*.

Informe gerencial

1. Use la estadística descriptiva para resumir los datos de este estudio. ¿Cuáles son, respecto de cada pregunta, las conclusiones preliminares acerca de la independencia entre la respuesta (Sí, No) y la afiliación política para cada una de las tres interrogantes en la encuesta?
2. Para la pregunta 1, pruebe la independencia entre la respuesta (Sí, No) y la afiliación partidista. Use $\alpha = 0.05$.
3. Con respecto a la pregunta 2, pruebe la independencia entre la respuesta (Sí, No) y la afiliación política. Use $\alpha = 0.05$.
4. Con respecto a la pregunta 3, pruebe la independencia entre la respuesta (Sí, No) y la afiliación política. Use $\alpha = 0.05$.
5. ¿Le parece que hay un amplio apoyo para los cambios en todos los aspectos políticos? Explique.

Apéndice 12.1 Pruebas de bondad de ajuste e independencia con Minitab

Prueba de bondad de ajuste

Este procedimiento de Minitab puede utilizarse en la prueba de bondad de ajuste de la población multinomial de la sección 12.1. El usuario debe obtener la frecuencia observada y la proporción hipotética para cada una de las k categorías. Las frecuencias observadas se ingresan en la columna C1 y las proporciones hipotéticas en la columna C2. Utilizando el ejemplo de Scott Marketing Research presentado en la sección 12.1, la columna C1 se etiqueta como Observed y la columna C2 como Proportion. Se introducen las frecuencias observadas 48, 98 y 54 en la columna C1 y las proporciones hipotéticas 0.30, 0.50 y 0.20 en la columna C2. Los pasos para la prueba de bondad de ajuste usando Minitab son los siguientes.

Paso l. Seleccione el menú **Stat**.
Paso 2. Escoja **Tables**.
Paso 3. Elija **Chi-Square Goodness of Fit Test (One Variable)**.
Paso 4. Cuando el cuadro de diálogo Chi-Square Goodness of Fit Test aparezca:
Seleccione **Observed counts**.
Introduzca C1 en el cuadro **Observed counts**.
Seleccione **Specific proportions**.
Introduzca C2 en el cuadro **Specific proportions**.
Haga clic en **OK**.

Prueba de independencia

Se empieza con una nueva hoja de cálculo de Minitab y se ingresan los datos de las frecuencias observadas en las columnas 1, 2 y 3, respectivamente, del ejemplo de Albert's Brewery de la sección 12.2. Esto es, se ingresan las frecuencias observadas que corresponden a las preferencias por la cerveza ligera (20 y 30) en C1, a las preferencias por la cerveza clara (40 y 30) en C2 y a las preferencias por la cerveza oscura (20 y 10) en C3. Los pasos para la prueba de independencia usando Minitab son los siguientes.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Escoja **Tables**.
Paso 3. Elija **Chi-Square Test (Two-Way Table in Worksheet)**.
Paso 4. Cuando el cuadro de diálogo Chi-Square Test aparezca:
Ingrese C1-C3 en el cuadro **Columns containing the table**.
Haga clic en **OK**.

Apéndice 12.2 Pruebas de bondad de ajuste e independencia con Excel

Prueba de bondad de ajuste

WEB archivo
FitTest

Este procedimiento puede ser utilizado en una prueba de bondad de ajuste para la distribución multinomial de la sección 12.1 y las distribuciones de Poisson y normal de la sección 12.3. El usuario tendrá que obtener las frecuencias observadas, calcular las frecuencias esperadas e ingresar ambas frecuencias en una hoja de cálculo de Excel.

Las frecuencias observadas y esperadas del ejemplo de Scott Market Research de la sección 12.1 se ingresan en las columnas A y B, como se ilustra en la figura 12.4. El estadístico de prueba $\chi^2 = 7.34$ se calcula en la columna D. Como hay $k = 3$ categorías, el usuario ingresa los grados de libertad $k - 1 = 3 - 1 = 2$ en la celda D11. La función CHIDIST proporciona el valor-p en la celda D13. La hoja de cálculo presenta en segundo plano las fórmulas de la celda.

Prueba de independencia

WEB archivo
Independence

En el procedimiento de Excel para pruebas de independencia se requiere que el usuario obtenga las frecuencias observadas y las ingrese en una hoja de cálculo. En el ejemplo de Alber's Brewery presentado en la sección 12.2 se proporcionan las frecuencias observadas, las cuales se ingresan en las celdas B7 a D8, como se indica en la hoja de cálculo de la figura 12.5. Las fórmulas que aparecen en las celdas de la hoja de cálculo en segundo plano indican el procedimiento empleado para calcular las frecuencias esperadas. En la celda E22 se ingresan los grados de libertad que, como se tienen dos filas y tres columnas, serán $(2 - 1)(3 - 1) = 2$. La función CHITEST proporciona el valor-p en la celda E24.

FIGURA 12.4 Hoja de cálculo de Excel para la prueba de bondad de ajuste en el ejemplo de Scott Marketing Research

	A	B	C	D	E
1	**Goodness of Fit Test**				
2					
3	Observed	Expected			
4	Frequency	Frequency		Calculations	
5	48	60		=(A5-B5)^2/B5	
6	98	100		=(A6-B6)^2/B6	
7	54	40		=(A7-B7)^2/B7	
8					
9		Test Statistic		=SUM(D5:D7)	
10					
11		Degrees of Freedom		2	
12					
13			p-Value	=CHIDIST(D9,D11)	
14					

	A	B	C	D	E
1	**Goodness of Fit Test**				
2					
3	Observed	Expected			
4	Frequency	Frequency		Calculations	
5	48	60		2.40	
6	98	100		0.04	
7	54	40		4.90	
8					
9		Test Statistic		7.34	
10					
11		Degrees of Freedom		2	
12					
13			p-Value	0.0255	
14					

FIGURA 12.5 Hoja de cálculo de Excel para la prueba de independencia de Alber's Brewery

	A	B	C	D	E	F
1	**Test of Independence**					
2						
3	**Observed Frequencies**					
4						
5		**Beer Preference**				
6	**Gender**	Light	Regular	Dark	Total	
7	Male	20	40	20	=SUM(B7:D7)	
8	Female	30	30	10	=SUM(B8:D8)	
9	Total	=SUM(B7:B8)	=SUM(C7:C8)	=SUM(D7:D8)	=SUM(E7:E8)	
10						
11						
12	**Expected Frequencies**					
13						
14		**Beer Preference**				
15	**Gender**	Light	Regular	Dark	Total	
16	Male	=E7*B$9/$E$9	=E7*C$9/$E$9	=E7*D$9/$E$9	=SUM(B16:D16)	
17	Female	=E8*B$9/$E$9	=E8*C$9/$E$9	=E8*D$9/$E$9	=SUM(B17:D17)	
18	Total	=SUM(B16:B17)	=SUM(C16:C17)	=SUM(D16:D17)	=SUM(E16:E17)	
19						
20				**Test Statistic**	=CHIINV(E24,E22)	
21						
22			**Degrees of Freedom**		2	
23						
24				**p-value**	=CHITEST(B7:D8,B16:D17)	
25						

	A	B	C	D	E	F
1	**Test of Independence**					
2						
3	**Observed Frequencies**					
4						
5		**Beer Preference**				
6	**Gender**	Light	Regular	Dark	Total	
7	Male	20	40	20	80	
8	Female	30	30	10	70	
9	Total	50	70	30	150	
10						
11						
12	**Expected Frequencies**					
13						
14		**Beer Preference**				
15	**Gender**	Light	Regular	Dark	Total	
16	Male	26.67	37.33	16	80	
17	Female	23.33	32.67	14	70	
18	Total	50	70	30	150	
19						
20			**Test Statistic**		6.12	
21						
22		**Degrees of Freedom**			2	
23						
24				**p-value**	0.0468	
25						

CAPÍTULO 13

Diseño de experimentos y análisis de varianza

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: BURKE MARKETING SERVICES, INC.

13.1 INTRODUCCIÓN AL DISEÑO DE EXPERIMENTOS Y AL ANÁLISIS DE VARIANZA
- Recolección de datos
- Supuestos para el análisis de varianza
- Análisis de varianza: una perspectiva conceptual

13.2 ANÁLISIS DE VARIANZA Y EL DISEÑO COMPLETAMENTE ALEATORIZADO
- Estimación de la varianza poblacional entre tratamientos
- Estimación de la varianza poblacional dentro de los tratamientos
- Comparación de las estimaciones de las varianzas: la prueba *F*
- Tabla de ANOVA
- Resultados de computadora para el análisis de varianza
- Prueba para la igualdad de *k* medias poblacionales: un estudio observacional

13.3 PROCEDIMIENTOS DE COMPARACIÓN MÚLTIPLE
- LSD de Fisher
- Tasas de error tipo I

13.4 DISEÑO DE BLOQUES ALEATORIZADO
- Prueba de estrés para controladores de tráfico aéreo
- Procedimiento ANOVA
- Cálculos y conclusiones

13.5 EXPERIMENTO FACTORIAL
- Procedimiento ANOVA
- Cálculos y conclusiones

ESTADÍSTICA *en* LA PRÁCTICA

BURKE MARKETING SERVICES, INC.*

CINCINNATI, OHIO

Burke Marketing Services, Inc. es una de las firmas de investigación de mercados con más experiencia en la industria. Cada día desarrolla más propuestas de proyectos que cualquier otra empresa del ramo en el mundo. Apoyada con tecnología de punta, Burke ofrece una amplia variedad de posibilidades de investigación, con lo que soluciona casi cualquier problema de marketing.

En un estudio, cierta empresa solicitó los servicios de Burke para evaluar nuevas versiones potenciales de un cereal para niños. Por razones de confidencialidad, aquí se nombrará a la empresa como Anon Company. Los cuatro factores clave que los desarrolladores de productos de Anon consideraban que podrían mejorar el sabor del cereal eran los siguientes.

1. Razón entre trigo y maíz en las hojuelas de cereal.
2. Tipo de edulcorante: azúcar, miel o artificial.
3. Presencia o ausencia de trozos con sabor a fruta.
4. Tiempo de cocción, largo o corto.

Burke diseñó un experimento para determinar los efectos de estos cuatro factores en el sabor del cereal. Por ejemplo, se elaboró un cereal de prueba con una razón específica de trigo y maíz, azúcar como edulcorante, trozos de sabor y un tiempo corto de cocción; se preparó otro cereal de prueba con diferente proporción de trigo y maíz, dejando igual los otros tres factores, y así de manera sucesiva. Después, grupos de niños probaron el sabor de los productos y dieron su opinión acerca de cada uno.



Burke aplica pruebas de degustación para obtener valiosa información estadística acerca de qué quieren los clientes de un producto. ©JLP/Sylvia Torres/CORBIS

El método estadístico para estudiar los datos obtenidos en las pruebas de degustación fue el análisis de varianza. Los resultados de los análisis mostraron lo siguiente.

- La composición de las hojuelas y el tipo de edulcorante influyeron de manera importante en la evaluación del sabor.
- Los trozos con sabor en realidad tuvieron un efecto negativo en el sabor del cereal.
- El tiempo de cocción no tuvo ninguna influencia en el sabor.

Esta información ayudó a Anon a identificar los factores que llevarían al cereal con mejor sabor.

El diseño de experimentos que Burke empleó y el subsecuente análisis de varianza le ayudaron emitir una recomendación para el diseño del producto. En este capítulo se verá cómo se realizan estos procedimientos.

* Los autores agradecen al doctor Ronald Tatham, de Burke Marketing Services, por haber proporcionado este material para *Estadística en la práctica*.

En el capítulo 1 se estableció que los estudios estadísticos se pueden clasificar como experimentales u observacionales. En un estudio estadístico experimental se realiza un experimento para generar los datos. Éste comienza con la identificación de la variable de interés. A continuación se identifican y controlan una o más variables que se considera que están relacionadas con la variable de interés, y se recogen datos de cómo influyen en ella.

En un estudio observacional los datos se obtienen por lo general mediante encuestas por muestreo y no por un experimento controlado. Aunque también se emplean los principios para un buen estudio, en estas investigaciones no es posible el control riguroso que se tiene en un trabajo experimental. Por ejemplo, en un estudio de la relación entre fumar y cáncer de pulmón, el investigador no puede asignarle a los sujetos el hábito de fumar. En realidad está restringido a la simple observación de los efectos de fumar en las personas que ya tienen este hábito, y los efectos de no fumar en las personas que no lo tienen.

Sir Ronald Alymer Fisher (1890-1962) inventó la rama de la estadística conocida como diseño de experimentos. *Además de sus aportaciones a la estadística, fue un científico sobresaliente en el campo de la genética.*

En este capítulo se presentan tres tipos de diseños de experimentos: diseño completamente aleatorizado, diseño de bloques aleatorizado y experimento factorial. Para cada uno se muestra cómo usar un procedimiento estadístico conocido como *análisis de varianza* (*ANOVA*, por su acrónimo en inglés) para analizar los datos disponibles. El ANOVA también se puede usar para analizar los datos obtenidos en un estudio observacional. Por ejemplo, se verá que en un diseño de experimentos completamente aleatorizado este procedimiento también sirve para probar la igualdad de tres o más medias poblacionales cuando se obtienen los datos mediante un estudio observacional. En los capítulos siguientes se verá que el ANOVA tiene gran importancia en el análisis de los resultados de estudios de regresión que incluyen tanto datos experimentales como observacionales.

En la primera sección se mencionan los principios básicos de un estudio experimental y se describe la forma de emplearlos en un diseño completamente aleatorizado. En la segunda sección se muestra cómo usar el ANOVA para analizar los datos de un diseño de experimento completamente aleatorizado. En la última sección se estudian procedimientos de comparación múltiple y otros dos diseños de experimentos ampliamente utilizados: el de bloques aleatorizado y el experimento factorial.

13.1 Introducción al diseño de experimentos y al análisis de varianza

Las relaciones de causa y efecto pueden ser difíciles de establecer en estudios observacionales, pero son más fáciles de establecer en estudios experimentales.

Como ejemplo de un estudio estadístico experimental, veamos el problema que enfrenta Chemitech, Inc. Esta empresa desarrolló un nuevo sistema de filtración para los suministros de aguas municipales. Los componentes del sistema se comprarán a varios proveedores y Chemitech los ensamblará en su planta de Columbia, Carolina del Sur. El grupo de ingeniería industrial es el encargado de determinar el mejor método de ensamble para el nuevo sistema de filtración. Después de considerar varios métodos posibles, el grupo redujo el número de alternativas a tres: el método A, el método B y el método C. Éstos difieren en el orden de los pasos que establecen para armar el sistema. Los directivos de Chemitech quieren determinar con qué método de ensamble se puede producir el mayor número de sistemas de filtración por semana.

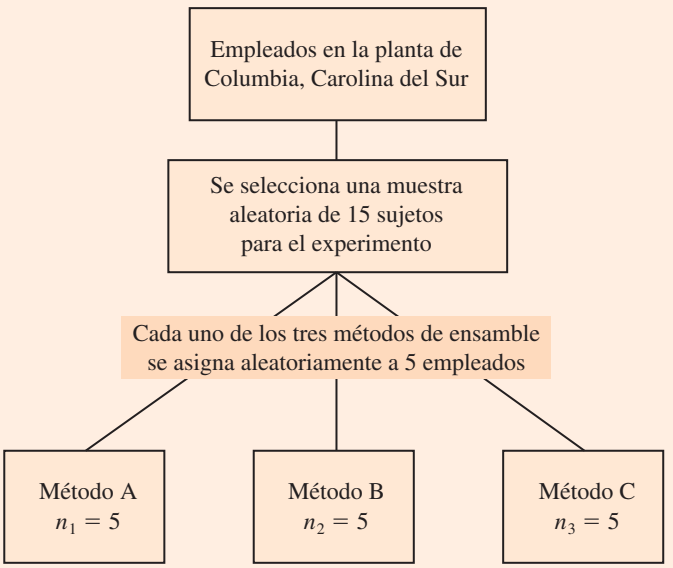
En el experimento de Chemitech, el método de ensamble es la variable independiente o **factor**. Como a este factor le corresponden tres tipos de ensamble, se dice que hay tres tratamientos asociados con este experimento; cada **tratamiento** corresponde a uno de los tres métodos. El problema de Chemitech es ejemplo de un **experimento de un solo factor**, porque involucra sólo un factor cualitativo (el método de ensamble). Los experimentos más complejos pueden consistir de múltiples factores; algunos son cualitativos y otros cuantitativos.

Los tres métodos de ensamble o tratamientos constituyen las tres poblaciones de interés para el experimento de Chemitech. Una población está conformada por todos los trabajadores que emplean el método de ensamble A, otra por la de aquellos que utilizan el método B y otra más por los que emplean el método C. Observe que en cada población la variable dependiente o **variable de respuesta** es el número de sistemas de filtración ensamblado por semana, y el principal objetivo estadístico del experimento es determinar si el número medio de unidades producidas por semana es el mismo para las tres poblaciones (métodos).

La aleatorización es el procedimiento de asignar al azar los tratamientos a las unidades experimentales. Antes del trabajo de Sir R. A. Fisher, los tratamientos se asignaban sobre bases sistemáticas o subjetivas.

Suponga que se selecciona una muestra aleatoria de tres sujetos de entre todos los trabajadores de ensamble en las instalaciones de manufactura de Chemitech. En la terminología del diseño de experimentos, los tres trabajadores seleccionados aleatoriamente son las **unidades experimentales**. El diseño de experimentos que se usará para el problema se llama **diseño completamente aleatorizado**. En este tipo de diseño se requiere que cada uno de los métodos de ensamble o tratamientos sea asignado de manera aleatoria a cada una de las unidades experimentales o trabajadores. Por ejemplo, el método A se asigna aleatoriamente al segundo trabajador, el método B al primero y el método C al tercero. El concepto de *aleatorización,* como se ilustró en este ejemplo, es un principio importante en todos los diseños de experimentos.

FIGURA 13.1 Diseño completamente aleatorizado para evaluar el método experimental de ensamble de Chemitech



Observe que el resultado de este experimento será sólo una medición o número de unidades ensambladas por cada tratamiento. Para obtener datos adicionales de cada método de ensamble se debe repetir o *replicar* el proceso experimental básico. Suponga, por ejemplo, que en lugar de tomar al azar sólo a tres trabajadores, se seleccionan 15, y luego se les asigna en forma aleatoria cada uno de los tratamientos a cinco de ellos. Como cada uno de los métodos de ensamble es asignado a cinco trabajadores, se dice que se obtienen cinco réplicas. El proceso de *replicación* es otro principio importante en el diseño de experimentos. En la figura 13.1 se presenta el diseño completamente aleatorizado para el experimento de Chemitech.

Recolección de datos

Una vez realizado el diseño del experimento, se procede a recolectar y analizar los datos. En el caso de Chemitech, se le explica a los trabajadores cómo emplear el método de ensamble que les ha sido asignado y empezarán a armar los sistemas de filtración con ese método. En la tabla 13.1 se presenta el número de unidades ensambladas por cada empleado en una semana. También se proporciona la media muestral, la varianza muestral y la desviación estándar muestral obtenidas con cada proceso de ensamble. Así, la media muestral del número de unidades producidas con el método A es 62; con el método B es 66, y usando el método C es 52. Con base en estos datos, parece que B proporciona las tasas más altas de producción que cualquiera de los otros métodos.

El punto a considerar es si cualquiera de las tres medias muestrales observadas difiere lo suficiente como para concluir que las medias poblacionales correspondientes a estos tres métodos de ensamble son diferentes. Para escribir esto en términos estadísticos, se introduce la notación siguiente.

μ_1 = número medio de unidades producidas por semana con el método A
μ_2 = número medio de unidades producidas por semana con el método B
μ_3 = número medio de unidades producidas por semana con el método C

TABLA 13.1 Número de unidades producidas por 15 trabajadores

WEB archivo
Chemitech

	Method		
	A	B	C
	58	58	48
	64	69	57
	55	71	59
	66	64	47
	67	68	49
Sample mean (media muestral)	62	66	52
Sample variance (varianza muestral)	27.5	26.5	31.0
Sample standard deviation (desviación estándar muestral)	5.244	5.148	5.568

Aunque nunca se podrá saber cuáles son los verdaderos valores de μ_1, μ_2 y μ_3, se utilizan las medias muestrales para probar las hipótesis siguientes.

$$H_0: \mu_1 = \mu_2 = \mu_3$$
$$H_a: \text{no todas las medias poblacionales son iguales}$$

Si H_0 es rechazada, no podemos concluir que todas las medias poblacionales son diferentes. Rechazar H_0, significa que por lo menos dos medias poblacionales tienen valores diferentes.

Como se demostrará más adelante, el análisis de varianza (ANOVA) es el procedimiento estadístico que se emplea para determinar si las diferencias observadas entre las tres medias muestrales son lo suficientemente grandes para rechazar H_0.

Supuestos para el análisis de varianza

Los supuestos requeridos para usar el análisis de varianza son tres.

Si los tamaños de las muestras son iguales, el análisis de varianza no es sensible a divergencias del supuesto de que las poblaciones están normalmente distribuidas.

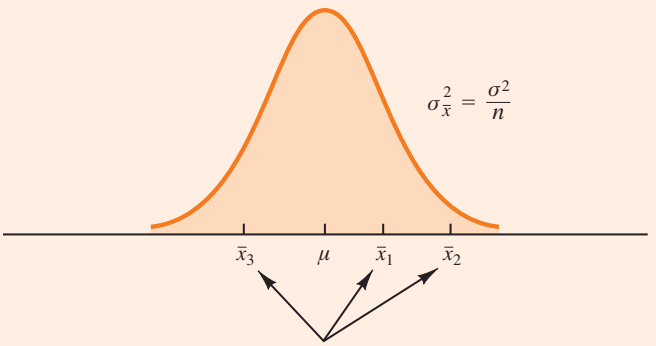
1. **En cada población, la variable de respuesta está normalmente distribuida.** *Implicación*. En el experimento de Chemitech, el número de unidades producidas por semana (variable de respuesta) debe estar normalmente distribuido para cada método de ensamble.
2. **La varianza de la variable de respuesta, denotada como σ^2, es la misma en todas las poblaciones.** *Implicación*. En el experimento de Chemitech, la varianza en el número de unidades producido por semana debe ser el mismo para cada método de ensamble.
3. **Las observaciones deben ser independientes.** *Implicación*. En el experimento de Chemitech la cantidad de unidades producida por semana por un empleado debe ser independiente del número de unidades producidas por semana por cualquier otro empleado.

Análisis de varianza: una perspectiva conceptual

Si las medias de las tres poblaciones son iguales, se esperaría que las tres medias muestrales fueran muy parecidas. De hecho, entre más parecidas sean éstas, mayor será la evidencia para concluir que las medias poblacionales son iguales. De otra forma, entre mayor sea la diferencia entre las medias muestrales, mayor será la evidencia para concluir que las medias poblacionales no son iguales. Esto es, si la variabilidad entre las medias muestrales es "pequeña", esto favorece H_0; si la variabilidad entre las medias muestrales es "grande", esto favorece H_a.

Si la hipótesis nula, H_0: $\mu_1 = \mu_2 = \mu_3$, es verdadera, se usa la variabilidad entre las medias muestrales para estimar σ^2. Primero, observe que si se satisfacen los supuestos para el

FIGURA 13.2 Distribución muestral de $\bar{x}$ si H_0 es verdadera



análisis de varianza, cada muestra provendrá de la misma distribución normal con media μ y varianza σ^2. Recuerde que en el capítulo 7 se vio que la distribución muestral de la media muestral $\bar{x}$ de una muestra aleatoria simple de tamaño n tomada de una población normal tendrá una distribución normal con media μ y varianza σ^2/n. En la figura 13.2 se ilustra una distribución muestral de este tipo.

Por consiguiente, si la hipótesis nula es verdadera, se considera cada una de las tres medias muestrales, $\bar{x}_1 = 62$, $\bar{x}_2 = 66$ y $\bar{x}_3 = 52$ de la tabla 13.1, como valores obtenidos aleatoriamente de la distribución muestral que aparece en la figura 13.2. En este caso, la media y la varianza de los tres valores $\bar{x}$ se pueden usar para estimar la media y la varianza de la distribución muestral. Cuando los tamaños de las muestras son iguales, como en el caso de Chemitech, la mejor estimación de la media de la distribución muestral de $\bar{x}$ es la media o el promedio de las medias muestrales. Por tanto, en el experimento de Chemitech, una estimación de la media de la distribución muestral de $\bar{x}$ es $(62 + 66 + 52)/3 = 60$, a la cual se le conoce como *media muestral general.* A su vez, una estimación de la varianza de la distribución muestral de $\bar{x}$, $\sigma_{\bar{x}}^2$, se obtiene de la varianza de las tres medias muestrales.

$$s_{\bar{x}}^2 = \frac{(62 - 60)^2 + (66 - 60)^2 + (52 - 60)^2}{3 - 1} = \frac{104}{2} = 52$$

Como $\sigma_{\bar{x}}^2 = \sigma^2/n$, al resolver para σ^2 obtenemos

$$\sigma^2 = n\sigma_{\bar{x}}^2$$

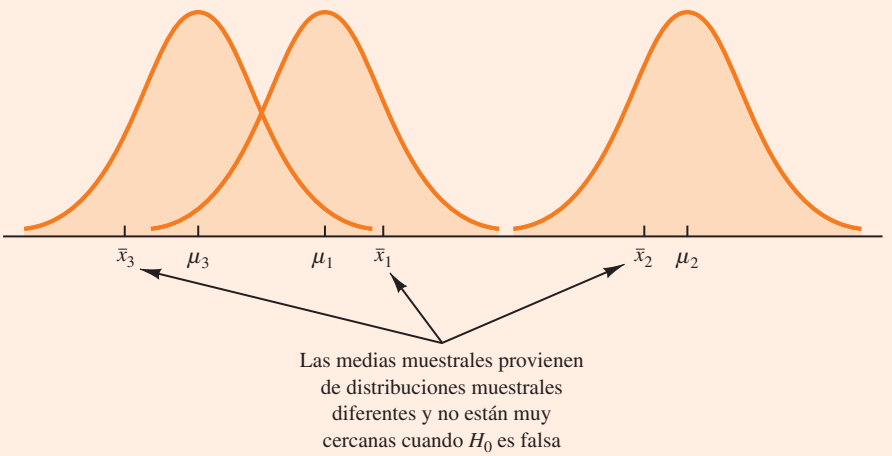
Por tanto,

$$\text{Estimación de } \sigma^2 = n \text{ (estimación de } \sigma_{\bar{x}}^2) = ns_{\bar{x}}^2 = 5(52) = 260$$

Al resultado, $ns_{\bar{x}}^2 = 260$, se le conoce como estimación de σ^2 *entre tratamientos*.

La estimación σ^2 entre tratamientos se basa en el supuesto de que la hipótesis nula es verdadera. En este caso cada una de las muestras proviene de la misma población y sólo hay una

FIGURA 13.3 Distribuciones muestrales de $\bar{x}$ si H_0 es falsa



distribución muestral de $\bar{x}$. Para ilustrar qué ocurre cuando H_0 es falsa, suponga que las medias poblacionales son todas diferentes. Observe que como las tres muestras provienen de poblaciones normales con medias diferentes, darán tres distribuciones muestrales distintas. En la figura 13.3 se advierte que en este caso las medias muestrales no están tan cerca unas de otras como cuando H_0 es verdadera. Entonces $s_{\bar{x}}^2$ será mayor, haciendo que la estimación entre tratamientos de σ^2 también lo sea. En general, cuando las medias poblacionales no son iguales, la estimación entre tratamientos sobreestimará la varianza poblacional σ^2.

La variación dentro de cada una de las muestras también tiene efecto sobre la conclusión a la que se arriba con el análisis de varianza. Cuando se selecciona una muestra aleatoria simple de cada población, cada una de las varianzas muestrales proporciona una estimación insesgada de σ^2. Por tanto, se combinan o juntan las estimaciones individuales de σ^2 en una general. A la estimación de σ^2 obtenida de esta manera se le conoce como estimación *conjunta* o *dentro de los tratamientos* de σ^2. Debido a que cada varianza muestral proporciona una estimación de σ^2 que se basa sólo en la variación dentro de cada muestra, a la estimación de σ^2 dentro de los tratamientos no le afecta que las medias poblacionales sean iguales. Cuando los tamaños de las muestras son iguales, la estimación dentro de los tratamientos de σ^2 se obtiene al calcular el promedio de las varianzas muestrales. En el experimento de Chemitech obtenemos

$$\text{Estimación de } \sigma^2 \text{ dentro de los tratamientos} = \frac{27.5 + 26.5 + 31.0}{3} = \frac{85}{3} = 28.33$$

En el experimento de Chemitech, la estimación de σ^2 *entre* los tratamientos (260) es mucho mayor que *dentro* de los tratamientos (28.33). De hecho, el cociente entre estas dos estimaciones es 260/28.33 = 9.18. Pero debe recordarse que el método entre tratamientos sólo proporciona una buena estimación de σ^2 si la hipótesis nula es verdadera; si es falsa, este método sobreestima σ^2. El método dentro de los tratamientos proporciona una buena estimación de σ^2 en cualquiera de los casos. Por tanto, si la hipótesis nula es verdadera, las dos estimaciones serán semejantes y su cociente será cercano a 1. Si la hipótesis es falsa, la estimación entre tratamientos será mayor que la estimación dentro de los tratamientos y su cociente será grande. En la sección siguiente se muestra qué tan grande debe ser este cociente para que H_0 sea rechazada.

En resumen, la lógica detrás del ANOVA se basa en obtener dos estimaciones independientes de la varianza poblacional común de σ^2. Una estimación de σ^2 se funda en la variabilidad entre las medias muestrales mismas y la otra en la variabilidad entre los datos dentro de cada muestra. Al comparar estas dos estimaciones de σ^2, podrá determinarse si las medias poblacionales son iguales.

NOTAS Y COMENTARIOS

1. En el diseño de experimentos, la aleatorización es análoga al muestreo probabilístico en un estudio observacional.
2. En muchos estudios médicos los sesgos potenciales se eliminan con el uso de un diseño de experimento doble ciego en el cual ni el médico que aplica el tratamiento ni el paciente saben qué tratamiento se está administrando. Este tipo de diseño también es útil en muchos otros tipos de experimentos.
3. En esta sección se presentó una perspectiva conceptual de cómo puede utilizarse el análisis de varianza para probar la igualdad de k medias poblacionales en un diseño experimental completamente aleatorizado. Veremos que este mismo procedimiento también se usa para probar la igualdad de k medias poblacionales en un estudio observacional o no experimental.
4. En las secciones 10.1 y 10.2 se presentaron métodos estadísticos para probar las hipótesis de que las medias de dos poblaciones son iguales. El ANOVA también puede utilizarse para probar estas mismas hipótesis. Sin embargo, en la práctica el análisis de varianza no es usualmente utilizado, excepto cuando se tienen tres o más medias poblacionales.

13.2 Análisis de varianza y el diseño completamente aleatorizado

En esta sección se muestra el uso del análisis de varianza para probar la igualdad de k medias poblacionales en un diseño completamente aleatorizado. La forma general de esta prueba de hipótesis es

$$H_0: \mu_1 = \mu_2 \cdots = \mu_k$$
$$H_a: \text{no todas las medias poblacionales son iguales}$$

donde

$$\mu_j = \text{media de la } j\text{-ésima población}$$

Se asume que de cada una de las k poblaciones o tratamientos se toma una muestra aleatoria simple de tamaño n_j. Para los datos muestrales resultantes, sean

$$x_{ij} = \text{valor de la observación } i \text{ del tratamiento } j$$
$$n_j = \text{número de observaciones en el tratamiento } j$$
$$\bar{x}_j = \text{media muestral del tratamiento } j$$
$$s_j^2 = \text{varianza muestral del tratamiento } j$$
$$s_j = \text{desviación estándar muestral del tratamiento } j$$

Las fórmulas para la media muestral y la varianza muestral del tratamiento j son las siguientes:

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j} \tag{13.1}$$

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n_j - 1} \tag{13.2}$$

La media muestral general, que se denota $\bar{\bar{x}}$, es la suma de todas las observaciones divididas entre la cantidad total de las observaciones. Es decir,

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{n_T} \tag{13.3}$$

donde

$$n_T = n_1 + n_2 + \cdots + n_k \tag{13.4}$$

Si el tamaño de cada muestra es de n, $n_T = kn$, en este caso la ecuación (13.3) se reduce a

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{kn} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}/n}{k} = \frac{\sum_{j=1}^{k} \bar{x}_j}{k} \tag{13.5}$$

En otras palabras, si todas las muestras son del mismo tamaño, la media muestral general es el promedio de las k medias muestrales.

En el experimento de Chemitech, como todas las muestras constaban de $n = 5$ observaciones, la media muestral general se calcula utilizando la fórmula (13.5). Con base en los datos de la tabla 13.1 obtenemos el siguiente resultado.

$$\bar{\bar{x}} = \frac{62 + 66 + 52}{3} = 60$$

Si la hipótesis nula es verdadera ($\mu_1 = \mu_2 = \mu_3 = \mu$), la media muestral general de 60 es la mejor estimación de la media poblacional μ.

Estimación de la varianza poblacional entre tratamientos

En la sección anterior se presentó el concepto de estimación de σ^2 entre tratamientos y se mostró cómo calcularla cuando todas las muestras son del mismo tamaño. A esta estimación de σ^2 se le llama *cuadrado medio debido a los tratamientos* y se denota como CMTR. La fórmula general para calcularlo es

$$\text{CMTR} = \frac{\sum_{j=1}^{k} n_j (\bar{x}_j - \bar{\bar{x}})^2}{k - 1} \tag{13.6}$$

Al numerador de la ecuación (13.6) se le llama *suma de cuadrados debido a los tratamientos* y se denota SCTR. El denominador, $k - 1$, representa los grados de libertad asociados con la SCTR. Por tanto, el cuadrado medio debido a los tratamientos se calcula con la fórmula siguiente.

CUADRADO MEDIO DEBIDO A LOS TRATAMIENTOS

$$\text{CMTR} = \frac{\text{SCTR}}{k - 1} \tag{13.7}$$

donde

$$\text{SCTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 \tag{13.8}$$

Si H_0 es verdadera, el CMTR proporciona una estimación insesgada de σ^2. No obstante, si las medias de las k poblaciones no son iguales, el CMTR no es un estimador insesgado de σ^2; en este caso, de hecho, sobreestima σ^2.

Para los datos de Chemitech de la tabla 13.1 obtenemos los siguientes resultados.

$$\text{SCTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 = 5(62 - 60)^2 + 5(66 - 60)^2 + 5(52 - 60)^2 = 520$$

$$\text{CMTR} = \frac{\text{SCTR}}{k - 1} = \frac{520}{2} = 260$$

Estimación de la varianza poblacional dentro de los tratamientos

Ya se presentó el concepto de estimación de σ^2 dentro de los tratamientos y cómo calcularla cuando todas las muestras son del mismo tamaño. A esta estimación de σ^2 se le llama *cuadrado medio debido al error* y se denota como CME. La fórmula general para calcularlo es

$$\text{CME} = \frac{\sum_{j=1}^{k}(n_j - 1)s_j^2}{n_T - k} \tag{13.9}$$

Al numerador de la ecuación (13.9) se le llama *suma de cuadrados debido al error*, y se denota como SCE. El denominador del CME son los grados de libertad correspondientes a la SCE. Por tanto, la fórmula para el CME también se expresa como sigue.

CUADRADO MEDIO DEBIDO AL ERROR

$$\text{CME} = \frac{\text{SCE}}{n_T - k} \tag{13.10}$$

donde

$$\text{SCE} = \sum_{j=1}^{k}(n_j - 1)s_j^2 \tag{13.11}$$

Observe que el CME está basado en la variación dentro de cada tratamiento; el que la hipótesis nula sea o no verdadera no tiene ninguna influencia. Por tanto, el CME proporciona siempre una estimación insesgada de σ^2.

Con base en los datos de la tabla 13.1 para el caso de Chemitech, obtenemos los resultados siguientes.

$$\text{SCE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = (5 - 1)27.5 + (5 - 1)26.5 + (5 - 1)31 = 340$$

$$\text{CME} = \frac{\text{SCE}}{n_T - k} = \frac{340}{15 - 3} = \frac{340}{12} = 28.33$$

Comparación de las estimaciones de las varianzas: la prueba *F*

En la sección 11.2 se presentó una introducción a la distribución F y al uso de las tablas de la distribución F.

Si la hipótesis nula es verdadera, el CMTR y el CME proporcionan dos estimaciones insesgadas e independientes de σ^2. Con base en lo estudiado en el capítulo 11 sabemos que cuando se tienen poblaciones normales la distribución muestral del cociente de dos estimaciones independientes de σ^2 sigue una distribución *F*. Por tanto, si la hipótesis nula es verdadera y se satisfacen los supuestos del ANOVA, la distribución muestral del CMTR/CME es una distribución *F* con $k - 1$ grados de libertad en el numerador y $n_T - k$ grados de libertad en el denominador. En otras palabras, si la hipótesis nula es verdadera, el valor del CMTR/CME parecerá que es un valor tomado de esta distribución *F*.

No obstante, si la hipótesis nula es falsa, el valor del CMTR/CME será muy grande debido a que el CMTR sobreestima σ^2. Por tanto, si el valor de CMTR/CME resulta ser demasiado grande para haber sido tomado de la distribución *F* con $k - 1$ grados de libertad en el numerador y $n_T - k$ grados de libertad en el denominador, H_0 será rechazada. Como la decisión de descartar H_0 está basada en el valor del CMTR/CME, el estadístico de prueba que se usa para probar la igualdad de *k* poblaciones es el siguiente.

ESTADÍSTICO DE PRUEBA PARA LA IGUALDAD DE *k* MEDIAS POBLACIONALES

$$F = \frac{\text{CMTR}}{\text{CME}} \qquad \textbf{(13.12)}$$

Este estadístico de prueba sigue una distribución *F* con $k - 1$ grados de libertad en el numerador y $n_T - k$ grados de libertad en el denominador.

Ahora bien, en el experimento de Chemitech se usará $\alpha = 0.05$ como nivel de significancia para realizar la prueba de hipótesis. El valor del estadístico de prueba es

$$F = \frac{\text{CMTR}}{\text{CME}} = \frac{260}{28.33} = 9.18$$

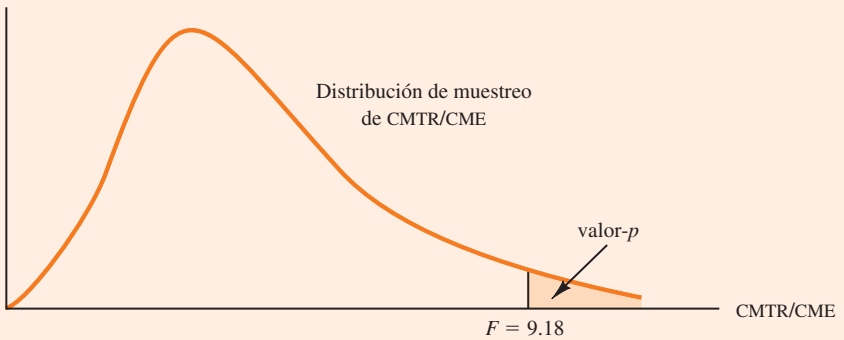
Los grados de libertad en el numerador son $k - 1 = 3 - 1 = 2$, y los grados de libertad para el denominador son $n_T - k = 15 - 3 = 12$. Como la hipótesis nula sólo será rechazada si obtenemos un valor grande para el estadístico de prueba, el valor-*p* será el área en la cola superior de la distribución *F* a la derecha del estadístico de prueba $F = 9.18$. En la figura 13.4 se presenta la distribución muestral de $F = \text{CMTR/CME}$, el valor del estadístico de prueba y el área en la cola superior que es el valor-*p* de esta prueba de hipótesis.

En la tabla 4 del apéndice B se encuentran las áreas siguientes en la cola superior de la distribución *F* con 2 grados de libertad en el numerador y 12 grados de libertad en el denominador.

Área en la cola superior	**0.10**	**0.05**	**0.025**	**0.01**
Valor *F* ($gl_1 = 2; gl_2 = 12$)	2.81	3.89	5.10	6.93

$F = 9.18$

FIGURA 13.4 Cálculo del valor-*p* a partir de la distribución de muestreo de CMTR/CME



En el apéndice F se explica cómo calcular el valor-p con Minitab o Excel.

Como $F = 9.18$ es mayor que 6.93, el área en la cola superior correspondiente a $F = 9.18$ es menor que 0.01. Por tanto, el valor-*p* es menor que 0.01. Para obtener el valor-*p* exacto, que es 0.004, se puede usar Minitab o Excel. Como el valor-$p \leq \alpha = 0.05$, H_0 es rechazada. La prueba proporciona evidencias suficientes para concluir que las medias de las tres poblaciones no son iguales. En otras palabras, el análisis de varianza favorece la conclusión de que las medias poblacionales del número de unidades producidas por semana con cada uno de los tres métodos de ensamble no son iguales.

Como en otros procedimientos de pruebas de hipótesis, aquí también puede emplearse el método del valor crítico. Como $\alpha = 0.05$, el valor crítico de F es aquel que deja un área de 0.05 en la cola superior de la distribución F con 2 y 12 grados de libertad. En las tablas de la distribución F se encuentra $F_{0.05} = 3.89$. Por tanto, la regla de rechazo en el caso del experimento de Chemitech es

$$\text{Rechazar } H_0 \text{ si } F \geq 3.89$$

Con $F = 9.18$, H_0 es rechazada, y concluimos que las medias de las tres poblaciones no son iguales. A continuación se presenta un resumen del procedimiento general para probar la igualdad de k medias poblacionales.

PRUEBA DE LA IGUALDAD DE k MEDIAS POBLACIONALES

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

H_a: no todas las medias poblacionales son iguales

ESTADÍSTICO DE PRUEBA

$$F = \frac{\text{CMTR}}{\text{CME}}$$

REGLA DE RECHAZO

Método del valor-*p*: Rechazar H_0 si el valor-$p \leq \alpha$

Método del valor crítico: Rechazar H_0 si $F \geq F_\alpha$

donde el valor de F_α está basado en una distribución F con $k - 1$ grados de libertad en el numerador y $n_T - k$ grados de libertad en el denominador.

Tabla de ANOVA

Los cálculos anteriores se pueden presentar de manera adecuada en un instrumento conocido como *tabla de análisis de varianza* o **tabla de ANOVA**. En la tabla 13.2 se observa la forma general de una tabla ANOVA para un diseño completamente aleatorizado; la tabla 13.3 corresponde a la tabla ANOVA del experimento de Chemitech. La suma de los cuadrados asociados con la fuente de variación que se indica como "Total" se conoce como *suma total de cuadrados* (STC). Observe que los resultados del experimento de Chemitech indican que STC = SCTR + SCE, y que los grados de libertad que corresponden a este resultado es la suma de los grados de libertad correspondiente a la suma de cuadrados debido a los tratamientos más la suma de cuadrados debido al error.

Cabe hacer notar que la STC dividida entre los grados de libertad $n_T - 1$ no es otra cosa que la varianza muestral general que se obtendría si se considerara la muestra de las 15 observaciones como un solo conjunto de datos. Si se toma todo el conjunto de datos como una sola muestra, la fórmula para calcular la suma total de cuadrados, STC, es

El análisis de varianza puede entenderse como un procedimiento estadístico de particionamiento de la suma total de los cuadrados en componentes separados.

$$\text{STC} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{\bar{x}})^2 \qquad \textbf{(13.13)}$$

Se puede demostrar que estos resultados observados para el análisis de la tabla de varianza en el caso del experimento de Chemitech también son aplicables a otros problemas. Es decir,

$$\text{STC} = \text{SCTR} + \text{SCE} \qquad \textbf{(13.14)}$$

En otras palabras, la STC se particiona en dos sumas de cuadrados: la suma de cuadrados debido a los tratamientos y la suma de cuadrados debido al error. Observe, además, que los grados de libertad que corresponden a la STC, $n_T - 1$, se pueden partir en grados de libertad correspondientes a SCTR, $k - 1$, y en grados de libertad correspondientes a SCE, $n_T - k$. El análisis de varianza se puede ver como el proceso de **partición** de la suma total de cuadrados y los grados de libertad en sus fuentes correspondientes: tratamientos y error. Al dividir las sumas de cuadrados entre los correspondientes grados de libertad, se obtienen las estimaciones de la varianza, el valor de F y el valor-p empleados en la prueba de hipótesis de igualdad entre las medias poblacionales.

TABLA 13.2 Tabla ANOVA para un diseño completamente aleatorizado

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	valor-p
Tratamientos	SCTR	$k - 1$	$\text{CMTR} = \dfrac{\text{SCTR}}{k-1}$	$\dfrac{\text{CMTR}}{\text{CME}}$	
Error	SCE	$n_T - k$	$\text{CME} = \dfrac{\text{SCE}}{n_T - k}$		
Total	STC	$n_T - 1$			

TABLA 13.3 Tabla de análisis de varianza para el experimento de Chemitech

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	valor-p
Tratamientos	520	2	260.00	9.18	0.004
Error	340	12	28.33		
Total	860	14			

FIGURA 13.5 Salida de Minitab para el análisis de varianza del experimento de Chemitech

```
Source     DF        SS        MS        F        P
Factor      2     520.0     260.0     9.18    0.004
Error      12     340.0      28.3
Total      14     860.0

S = 5.323      R-Sq = 60.47%      R-Sq(adj) = 53.88%

                                  Individual 95% CIs For Mean Based on
                                  Pooled StDev
 Level      N      Mean    StDev  ---+---------+---------+---------+------
 A          5    62.000    5.244               (-------*-------)
 B          5    66.000    4.148                       (------*-------)
 C          5    52.000    5.568  (------*-------)
                                  ---+---------+---------+---------+------
Pooled StDev = 5.323              49.0      56.0      63.0      70.0
```

Resultados de computadora para el análisis de varianza

Cuando se tienen muestras grandes o una cantidad grande de poblaciones, los cálculos del análisis de varianza se realizan con más facilidad mediante software para estadística. En los apéndices 13.1 a 13.3 se indican los pasos necesarios para realizar los cálculos del análisis de varianza con Minitab, Excel y StarTools. En la figura 13.5, aplicado al experimento de Chemitech, se presenta la pantalla de resultados de Minitab. En la primera parte de la pantalla se observa el formato ya conocido de la tabla ANOVA. Si se compara la figura 13.5 con la tabla 13.3, vemos que la información disponible es la misma, aunque algunos encabezados son ligeramente diferentes. El encabezado Source se usa en la columna correspondiente a la fuente de variación; Factor corresponde a la fila de tratamientos, y las columnas de las sumas de cuadrados y los grados de libertad están intercambiados.

Observe que, enseguida de la tabla ANOVA, la pantalla de la computadora proporciona los respectivos tamaños de las muestras, las medias muestrales y las desviaciones estándar. Además, Minitab presenta una figura con la estimación por intervalos de 95% de confianza para cada una de las medias poblacionales. Para obtener la estimación de estos intervalos, Minitab emplea el CME como estimación de σ^2. Por tanto, la raíz cuadrada del CME proporciona la mejor estimación de la desviación estándar poblacional σ. En la salida de la computadora esta estimación de σ es Pooled StDev, y su valor es 5.323. Para ilustrar cómo se calcula la estimación por intervalos se hará aquí la estimación por intervalo de 95% de confianza para la media poblacional del método A.

Con base en lo aprendido en el estudio de intervalos de confianza en el capítulo 8, sabemos que la forma general de una estimación por intervalo para una media poblacional es

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \tag{13.15}$$

donde s es la estimación de la desviación estándar poblacional σ. Como la mejor estimación de σ es la proporcionada por la Pooled StDev, se usa 5.323 en la expresión (13.15) como valor de s. Los grados de libertad para el valor de t son 12, los grados de libertad asociados con la suma de los cuadrados del error. Por tanto, como $t_{0.025} = 2.179$, obtenemos

$$62 \pm 2.179 \frac{5.323}{\sqrt{5}} = 62 \pm 5.19$$

Así, el intervalo de 95% de confianza para el método A va de $62 - 5.19 = 56.81$ a $62 + 5.19 = 67.19$. Como en el experimento de Chemitech los tamaños muestrales son iguales, también los intervalos de confianza para los métodos B y C se obtienen al sumar y restar 5.19 de la respectiva media muestral. En la salida de Minitab se aprecia que los anchos de los intervalos de confianza son los mismos.

Prueba para la igualdad de k medias poblacionales: un estudio observacional

Se ha revisado el uso del análisis de varianza para probar la igualdad de k medias poblacionales cuando se emplea un diseño experimental completamente aleatorizado. Es importante notar que el ANOVA también se puede utilizar para probar la igualdad de tres o más medias poblacionales usando datos de un estudio observacional. Para dar un ejemplo, se considerará el caso de National Computer Products, Inc. (NCP).

NCP fabrica impresoras y aparatos de fax en sus tres plantas situadas en Atlanta, Dallas y Seattle. Con el fin de medir los conocimientos de los empleados de estas tres plantas acerca de la administración de la calidad, se toma una muestra aleatoria de seis empleados de cada planta y se les aplica un examen acerca de su conocimiento sobre la calidad. En la tabla 13.4 se presentan las puntuaciones obtenidas en los exámenes por los 18 sujetos. En esta tabla se indican también la media, la varianza y la desviación estándar muestrales de cada grupo. Los gerentes de la empresa quieren usar estos datos para probar la hipótesis de que la media de las puntuaciones de los exámenes es la misma en las tres plantas.

Como población 1 se define a los empleados de la planta en Atlanta, como población 2 a los de la planta en Dallas y como población 3 a los de Seattle. Sean

μ_1 = media de las puntuaciones en los exámenes de la población 1
μ_2 = media de las puntuaciones en los exámenes de la población 2
μ_3 = media de las puntuaciones en los exámenes de la población 3

Aunque los verdaderos valores de μ_1, μ_2 y μ_3 nunca puedan conocerse, se usarán los resultados muestrales para probar las hipótesis siguientes.

H_0: $\mu_1 = \mu_2 = \mu_3$
H_a: no todas las medias poblacionales son iguales

Observe que la prueba de hipótesis para el estudio observacional de NCP es exactamente igual a la que se manejó para el experimento de Chemitech. También para analizar los datos del estudio

TABLA 13.4 Puntuaciones en los exámenes de 18 empleados

WEB archivo
NCP

	Planta 1 Atlanta	Planta 2 Dallas	Planta 3 Seattle
	85	71	59
	75	75	64
	82	73	62
	76	74	69
	71	69	75
	85	82	67
Media muestral	79	74	66
Varianza muestral	34	20	32
Desviación estándar muestral	5.83	4.47	5.66

El ejercicio 8 en este capítulo le pedirá analizar los datos de NCP utilizando el procedimiento de análisis de varianza.

observacional de NCP se emplea la misma metodología de análisis de varianza usada para el experimento de Chemitech.

Aun cuando en ambos casos se utiliza la misma metodología del ANOVA, vale la pena observar la diferencia entre el estudio estadístico observacional de NCP y la investigación estadística experimental de Chemitech. Las personas que realizaron el estudio de NCP no tuvieron control sobre la asignación de las plantas a cada uno de los empleados. Las plantas ya funcionaban y cada uno de los sujetos trabajaba en una de las tres. Lo único que se pudo hacer en este caso fue tomar una muestra aleatoria de seis empleados de cada una de las plantas y aplicarles el examen de conocimiento sobre la calidad. Para clasificarlo como un trabajo experimental, NPC tendría que haber tomado al azar 18 empleados y después, de manera aleatoria, asignar las plantas a cada uno.

NOTAS Y COMENTARIOS

1. La media muestral general también se calcula como media ponderada de las k medias muestrales.

$$\bar{\bar{x}} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{n_T}$$

En los problemas en que se proporcionan las medias muestrales, para calcular la media general es más sencillo utilizar esta fórmula que la expresión (13.3).

2. Si todas las muestras constan de n observaciones, la ecuación (13.6) puede escribe como

$$\text{CMTR} = \frac{n\sum_{j=1}^{k}(\bar{x}_j - \bar{\bar{x}})^2}{k-1} = n\left[\frac{\sum_{j=1}^{k}(\bar{x}_j - \bar{\bar{x}})^2}{k-1}\right] = ns_{\bar{x}}^2$$

Observe que este resultado es el mismo que el presentado en la sección 13.1 cuando se estudió el concepto de estimación de σ^2 entre tratamientos. La ecuación (13.6) es sólo una generalización de este resultado para el caso de tamaños muestrales distintos.

3. Si cada muestra tiene n observaciones, $n_T = kn$; por tanto, $n_T - k = k(n-1)$, y la ecuación (13.9) se puede reescribir como

$$\text{CME} = \frac{\sum_{j=1}^{k}(n-1)s_j^2}{k(n-1)} = \frac{(n-1)\sum_{j=1}^{k}s_j^2}{k(n-1)} = \frac{\sum_{j=1}^{k}s_j^2}{k}$$

En otras palabras, si los tamaños muestrales son iguales, el CME es simplemente el promedio de las k varianzas muestrales. Observe que éste es el mismo resultado que se usó en la sección 13.1 cuando se presentó el concepto de estimación de σ^2 dentro de los tratamientos.

Ejercicios

Métodos

AUTOevaluación

1. Los datos siguientes se obtuvieron de un diseño completamente aleatorizado.

	Tratamiento		
	A	B	C
	162	142	126
	142	156	122
	165	124	138
	145	142	140
	148	136	150
	174	152	128
Media muestral	156	142	134
Varianza muestral	164.4	131.2	110.4

a) Calcule la suma de cuadrados entre tratamientos.
b) Calcule el cuadrado medio entre tratamientos

c) Determine la suma de cuadrados debido al error.
d) Calcule el cuadrado medio debido al error.
e) Establezca la tabla ANOVA para este problema.
f) Con $\alpha = 0.05$, pruebe si las medias de los tres tratamientos son iguales.

2. En un diseño completamente aleatorizado, para cada uno de los cinco niveles del factor se usaron siete unidades experimentales. Complete la tabla ANOVA siguiente.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	*F*	valor-*p*
Tratamientos	300				
Error					
Total	460				

3. Remítase al ejercicio 2.
a) ¿Cuáles son las hipótesis implicadas en este problema?
b) Utilice el nivel de significancia $\alpha = 0.05$, ¿la hipótesis nula del inciso *a*) puede ser rechazada? Explique porqué.

4. En un experimento diseñado para probar los niveles de resultado de tres tratamientos diferentes se obtuvieron los resultados siguientes: STC = 400, SCTR = 150, $n_T = 19$. Establezca la tabla ANOVA y pruebe si hay alguna diferencia significativa entre las medias de los resultados de los tres tratamientos. Use $\alpha = 0.05$.

5. En un diseño completamente aleatorizado se usaron 12 unidades experimentales para el primer tratamiento, 15 para el segundo y 20 para el tercero. Complete el análisis de varianza siguiente. Emplee 0.05 como nivel de significancia, ¿hay diferencia significativa entre los tratamientos?

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	*F*	valor-*p*
Tratamientos	1 200				
Error					
Total	1 800				

6. Realice los cálculos del análisis de varianza para el siguiente diseño completamente aleatorizado. Con $\alpha = 0.05$, ¿la diferencia entre las medias de tratamiento (treatment) es significativa?

WEB archivo
Exer6

	Treatment		
	A	B	C
	136	107	92
	120	114	82
	113	125	85
	107	104	101
	131	107	89
	114	109	117
	129	97	110
	102	114	120
		104	98
		89	106
$\bar{x}_j$	119	107	100
s_j^2	146.86	96.44	173.78

Aplicaciones

7. Un ingeniero propone tres métodos distintos para ensamblar un producto. Para determinar el número de unidades ensambladas correctamente con cada método, se selecciona al azar a 30 empleados y se asignan de manera aleatoria a los tres enfoques propuestos, de manera que cada método sea empleado por 10 trabajadores. Se anota el número de unidades producidas correctamente y a estos datos se les aplica el análisis de varianza. Los resultados son los siguientes: STC = 10 800; SCTR = 4 560.
   a) Establezca la tabla ANOVA de este problema.
   b) Use $\alpha = 0.05$ para determinar si existen diferencias significativas entre las medias de los tres métodos de ensamble.
8. Vaya a la tabla 13.4 que presenta los datos de NCP. Establezca la tabla ANOVA y pruebe si existe diferencia significativa entre las medias de las puntuaciones de examen en las tres plantas. Use $\alpha = 0.05$.
9. Para estudiar el efecto de la temperatura en el rendimiento de un proceso químico, se produjeron cinco lotes con cada uno de tres niveles de temperatura. Los resultados se presentan a continuación. Establezca la tabla para el análisis de varianza. Use $\alpha = 0.05$ para probar si la temperatura afecta el rendimiento medio del proceso.

Temperatura		
50 °C	**60 °C**	**70 °C**
34	30	23
24	31	28
36	34	28
39	23	30
32	27	31

10. En una auditoría, los auditores tienen que emitir opiniones acerca de diversos aspectos con base en sus propias experiencias directas (Direct), indirectas (Indirect) o la combinación (Combination) de ambas. En un estudio se pidió a los auditores que dieran su opinión acerca de la frecuencia con que se presentan errores en una auditoría. Luego se compararon estas opiniones con los resultados reales. Suponga que los resultados que se presentan a continuación se obtuvieron de un estudio similar; los valores bajos indican opiniones más acertadas.

WEB archivo
AudJudg

Direct	Indirect	Combination
17.0	16.6	25.2
18.5	22.2	24.0
15.8	20.5	21.5
18.2	18.3	26.8
20.2	24.2	27.5
16.0	19.8	25.8
13.3	21.2	24.2

Use $\alpha = 0.05$ para determinar si el tipo de experiencia en que se basa la opinión afecta su calidad. ¿Cuál es su conclusión?

11. En la publicidad de cuatro pinturas (Paint 1, 2, 3 y 4) se dice que tienen el mismo tiempo de secado. Para verificarlo, se prueban cinco muestras de cada una de las pinturas. Se registra el tiempo en minutos necesario para que el secado sea suficiente para la aplicación de una segunda mano. Los datos obtenidos se listan a continuación.

WEB archivo
Paint

Paint 1	Paint 2	Paint 3	Paint 4
128	144	133	150
137	133	143	142
135	142	137	135
124	146	136	140
141	130	131	153

Con $\alpha = 0.05$ como nivel de significancia, realice una prueba para determinar si la media de los tiempos de secado es la misma en cada tipo de pintura.

12. La Encuesta de satisfacción de clientes de restaurantes de *Consumer Reports* se basa en más de 148 599 visitas a diferentes cadenas de restaurantes de servicio completo (sitio web de Consumer Reports). Una de las variables en el estudio es el precio de los alimentos, la cantidad promedio que paga una persona por la comida y la bebida, menos la propina. Suponga que un reportero del *Sun Coast Times* cree que sería de interés para sus lectores realizar un estudio similar en los restaurantes ubicados en la zona del Grand Strand en Myrtle Beach, Carolina del Sur. El reportero seleccionó una muestra de ocho restaurantes de mariscos (Seafood) ocho italianos (Italian) y ocho de carnes (Steakhouse). Los datos a continuación muestran los precios de la comida en dólares de los 24 negocios muestreados. Utilice $\alpha = 0.05$ para probar si hay una diferencia significativa entre el precio medio de la comida en los tres tipos de restaurantes.

WEB archivo
GrandStrand

Italian	Seafood	Steakhouse
$12	$16	$24
13	18	19
15	17	23
17	26	25
18	23	21
20	15	22
17	19	27
24	18	31

13.3 Procedimientos de comparación múltiple

Cuando se emplea el análisis de varianza para probar si las medias de k poblaciones son iguales, rechazar la hipótesis nula sólo permite concluir que las medias poblacionales *no son iguales*. En algunos casos se necesita dar un paso más y determinar dónde están las diferencias. El propósito de esta sección es mostrar el uso de **procedimientos de comparación múltiple** para establecer comparaciones entre pares de medias poblacionales.

LSD de Fisher

Suponga que en un análisis de varianza se encuentran evidencias estadísticas para rechazar la hipótesis nula que plantea la igualdad de las medias poblacionales. En tal caso, para determinar dónde están las diferencias se puede emplear el procedimiento de la diferencia mínima significativa (LSD, por sus siglas en inglés) de Fisher. Con el fin de ilustrar el uso del procedimiento de la LSD de Fisher para comparar pares de medias poblacionales, remítase al experimento de Chemitech presentado en la sección 13.1. A partir del análisis de varianza se concluyó que el número medio de unidades producidas por semana no era el mismo con los tres métodos de ensamble. En tal caso la siguiente pregunta es: se cree que hay diferencia entre los métodos pero, ¿dónde ocurren las diferencias? Es decir, las medias que difieren, ¿son las de las poblaciones 1 y 2? ¿O las de las poblaciones 1 y 3? ¿O las de las poblaciones 2 y 3?

En el capítulo 10 se presentó un procedimiento estadístico para probar la hipótesis de la igualdad de dos medias poblacionales. Con una ligera modificación en la manera de evaluar

la varianza poblacional, el procedimiento de la LSD de Fisher se basa en el estadístico de prueba t presentado para el caso de dos poblaciones. En la tabla siguiente se resume el procedimiento de la LSD de Fisher.

PROCEDIMIENTO DE LA LSD DE FISHER

$$H_0: \mu_i = \mu_j$$
$$H_a: \mu_i \neq \mu_j$$

ESTADÍSTICO DE PRUEBA

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\text{CME}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \qquad \textbf{(13.16)}$$

REGLA DE RECHAZO

Método del valor-p: Rechazar H_0 si el valor-$p \leq \alpha$

Método del valor crítico: Rechazar H_0 si $t \leq -t_{\alpha/2}$ o $t \geq t_{\alpha/2}$

donde el valor de $t_{\alpha/2}$ se basa en la distribución t con $n_T - k$ grados de libertad

A continuación se usará este procedimiento para determinar si existe alguna diferencia significativa entre la media de la población 1 (método A) y la media de la población 2 (método B) con $\alpha = 0.05$ como nivel de significancia. En la tabla 13.1 se indicó que las medias obtenidas con el método A es 62 y con el método B es 66. En la tabla 13.3 se observa que el valor del CME es 28.33; ésta es la estimación de σ^2 con 12 grados de libertad. Con los datos de Chemitech, el valor que se obtiene para el estadístico de prueba es

$$t = \frac{62 - 66}{\sqrt{28.33\left(\frac{1}{5} + \frac{1}{5}\right)}} = -1.19$$

Como se trata de una prueba de dos colas, el valor-p es el doble del área bajo la curva de la distribución t a la izquierda de $t = -1.19$. En la tabla 2 del apéndice B se encuentra la información siguiente para la distribución t con 12 grados de libertad.

Área en la cola superior	**0.20**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
Value *t* (12 *gl*)	0.873	1.356	1.782	2.179	2.681	3.055

$t = 1.19$

En el apéndice F se muestra cómo calcular los valores-p con Excel o Minitab.

La tabla de la distribución t sólo contiene valores positivos de t. Sin embargo, como la distribución t es simétrica, podemos determinar el área bajo la curva a la derecha de $t = 1.19$ y duplicarla para determinar el valor-p que corresponde a $t = -1.19$. En esta tabla vemos que $t = 1.19$ se encuentra entre 0.20 y 0.10. Al duplicar estas cantidades, tenemos que el valor-p debe estar entre 0.40 y 0.20. Se puede usar Excel o Minitab para ver que el valor-p exacto es 0.2571. Como este valor es mayor que $\alpha = 0.05$, la hipótesis nula no puede ser rechazada. Por tanto, no podemos concluir que la media poblacional del número de unidades producidas por semana con el método A sea diferente que la media poblacional del método B.

Muchas personas encuentran más fácil determinar qué tan grande tiene que ser la diferencia entre las medias muestrales para que H_0 sea rechazada. En este caso el estadístico de prueba es $\bar{x}_i - \bar{x}_j$, y la prueba se realiza siguiendo el procedimiento que se presenta a continuación.

PROCEDIMIENTO DE LA LSD DE FISHER BASADO EN EL ESTADÍSTICO DE PRUEBA $\bar{x}_i - \bar{x}_j$

$$H_0: \mu_i = \mu_j$$
$$H_a: \mu_i \neq \mu_j$$

ESTADÍSTICO DE PRUEBA

$$\bar{x}_i - \bar{x}_j$$

REGLA DE RECHAZO PARA EL NIVEL DE SIGNIFICANCIA α

$$\text{Rechazar } H_0 \text{ si } |\bar{x}_i - \bar{x}_j| \geq \text{LSD}$$

donde

$$\text{LSD} = t_{\alpha/2}\sqrt{\text{CME}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \qquad \textbf{(13.17)}$$

En el experimento de Chemitech, el valor de la LSD es

$$\text{LSD} = 2.179\sqrt{28.33\left(\frac{1}{5} + \frac{1}{5}\right)} = 7.34$$

Observe que si todos los tamaños muestrales son iguales, sólo se necesita calcular un valor de la LSD. En tales casos, basta comparar la magnitud de la diferencia entre dos medias muestrales con el valor de la LSD. Por ejemplo, la diferencia entre las medias muestrales de la población 1 (método A) y de la población 3 (método C) es $62 - 52 = 10$. Esta diferencia es mayor que la LSD $= 7.34$, lo que significa que se puede rechazar la hipótesis nula de que la media poblacional del número de unidades producidas por semana con el método A sea igual que la media poblacional del método C. De manera similar, entre las medias muestrales de las poblaciones 2 y 3 la diferencia es $66 - 52 = 14 > 7.34$, y se puede rechazar la hipótesis de que la media poblacional obtenida con el método B sea igual a la media poblacional del método C. Así, la conclusión es que tanto el método A como el B difieren del método C.

La LSD de Fisher también se usa para obtener una estimación mediante un intervalo de confianza de la diferencia entre las medias de dos poblaciones. El procedimiento general que se emplea es el siguiente.

ESTIMACIÓN POR INTERVALO DE CONFIANZA DE LA DIFERENCIA ENTRE DOS MEDIAS POBLACIONALES USANDO EL PROCEDIMIENTO DE LA LSD DE FISHER

$$\bar{x}_i - \bar{x}_j \pm \text{LSD} \qquad \textbf{(13.18)}$$

donde

$$\text{LSD} = t_{\alpha/2}\sqrt{\text{CME}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \qquad \textbf{(13.19)}$$

y $t_{\alpha/2}$ pertenece a la distribución t con $n_T - k$ grados de libertad.

Si el intervalo de confianza hallado con la expresión (13.18) incluye el valor cero, no se puede rechazar la hipótesis nula de que las dos medias poblacionales sean iguales. Pero si dicho intervalo no incluye al valor cero, podemos concluir que sí hay diferencia entre las medias poblacionales. En el caso del experimento de Chemitech, recuerde que la LSD $= 7.34$ (que corresponde a $t_{0.025} = 2.179$). Por tanto, una estimación de la diferencia entre las medias poblacionales 1 y 2 empleando un intervalo de 95% de confianza es $62 - 66 \pm 7.34 = -4 \pm 7.34 = -11.34$ a 3.34; como este intervalo incluye el cero, no se puede rechazar la hipótesis de que las dos medias sean iguales.

Tasas de error tipo I

El estudio del procedimiento de la LSD de Fisher se inició con la premisa de que el análisis de varianza proporcionaba evidencias estadísticas para rechazar la hipótesis nula de la igualdad entre medias poblacionales. Se mostró que en tales casos se puede emplear el procedimiento de la LSD de Fisher para determinar dónde están las diferencias. Técnicamente, a este procedimiento se le conoce como prueba *restringida o protegida* de la LSD debido a que sólo se usa si primero se ha encontrado un valor F significativo al aplicar el análisis de varianza. Para ver por qué es importante esta distinción en las pruebas de comparación múltiple es necesario explicar la diferencia entre tasa de error tipo I por *comparación* y tasa de error tipo I por *experimentación*.

En el experimento de Chemitech se usa el procedimiento de la LSD de Fisher para efectuar tres pares de comparaciones.

Prueba 1	**Prueba 2**	**Prueba 3**
$H_0: \mu_1 = \mu_2$	$H_0: \mu_1 = \mu_3$	$H_0: \mu_2 = \mu_3$
$H_a: \mu_1 \neq \mu_2$	$H_a: \mu_1 \neq \mu_3$	$H_a: \mu_2 \neq \mu_3$

En cada caso, el nivel de significancia empleado es $\alpha = 0.05$. Por tanto, en cada prueba, si la hipótesis nula es verdadera, la probabilidad de que se cometa un error tipo I es $\alpha = 0.05$; entonces, la probabilidad de no cometer un error tipo I es $1 - 0.05 = 0.95$. En el estudio de los procedimientos de comparación múltiple, a esta probabilidad de cometer un error tipo I ($\alpha = 0.05$) se le conoce como **tasa de error tipo I por comparación**, la cual indica el nivel de significancia que corresponde a una sola comparación por pares.

Considere ahora una cuestión ligeramente diferente. ¿Cuál es la probabilidad de que al hacer tres comparaciones por pares se cometa un error tipo I en por lo menos una de las tres pruebas? Para responder esta pregunta, observe que la probabilidad de que no se cometa un error tipo I en ninguna de las tres pruebas es $(0.95)(0.95)(0.95) = 0.8574$.[1] Por tanto, la probabilidad de cometer por lo menos un error tipo I es $1 - 0.8574 = 0.1426$. Entonces, cuando se usa el procedimiento de la LSD de Fisher para hacer los tres pares de comparaciones, la tasa de error tipo I correspondiente a este método no es 0.05, sino 0.1426, y se le conoce como **tasa de error tipo I por experimentación** o *general.* Para evitar confusiones, la tasa de error tipo I por experimentación se denota α_{EW}.

La tasa de error tipo I por experimentación es mayor en estudios con más poblaciones. Por ejemplo, en un problema con cinco poblaciones hay 10 pares de comparaciones. Si se prueban todas las comparaciones posibles por pares usando el procedimiento de la LSD de Fisher con una tasa de error por comparación de $\alpha = 0.05$, la tasa de error tipo I por experimentación será $1 - (1 - 0.05)^{10} = 0.40$. En tales casos se prefiere buscar otras alternativas que proporcionen un mejor control sobre la tasa de error por experimentación.

Una alternativa para controlar la tasa de error general por experimentación, conocida como *ajuste de Bonferroni*, consiste en usar en cada prueba tasas de error por comparación más pequeñas. Por ejemplo, si se quieren probar C comparaciones por pares y se desea que la pro-

[1] Se supone que las tres pruebas son independientes y, por tanto, la probabilidad conjunta de los tres eventos se obtiene con la simple multiplicación de las probabilidades individuales. De hecho, las tres pruebas no son independientes porque el CME se usa en cada prueba; en consecuencia, el error supuesto es mayor que el error mostrado.

babilidad máxima de cometer un error tipo I en todo el experimento sea α_{EW}, simplemente se usa una tasa de error por comparación igual a α_{EW}/C. En el experimento de Chemitech, si se desea emplear el procedimiento de la LSD de Fisher para probar los tres pares de comparaciones con una tasa de error máximo por experimentación de $\alpha_{EW} = 0.05$, se establece como tasa de error por comparación $\alpha = 0.05/3 = 0.017$. En un problema con cinco poblaciones y 10 comparaciones por pares, el ajuste de Bonferroni sugeriría una tasa de error por comparación de $0.05/10 = 0.005$. Recuerde que cuando se estudiaron las pruebas de hipótesis en el capítulo 9 se vio que para un tamaño de muestra dado, toda disminución en la probabilidad de cometer un error tipo I aumenta la probabilidad de cometer un error tipo II, el cual corresponde a aceptar la hipótesis de que las dos medias poblacionales son iguales cuando en realidad no lo son. Por tanto, suele haber renuencia a realizar pruebas individuales con una baja tasa de error tipo I por comparación debido a que aumenta el riesgo de cometer un error tipo II.

Como solución para tales situaciones se han elaborado otras alternativas, como el procedimiento de Turkey y la prueba de rango múltiple de Duncan. Sin embargo, en la comunidad estadística existe una gran controversia respecto de cuál es el "mejor" procedimiento. La verdad es que no hay uno que sea el mejor para todo tipo de problemas.

Ejercicios

Métodos

AUTOevaluación

13. Los datos siguientes se obtuvieron con un diseño completamente aleatorizado.

	Tratamiento A	Tratamiento B	Tratamiento C
	32	44	33
	30	43	36
	30	44	35
	26	46	36
	32	48	40
Media muestral	30	45	36
Varianza muestral	6.00	4.00	6.50

a) Con $\alpha = 0.05$ como nivel de significancia, ¿puede rechazar la hipótesis nula de que las medias de los tres tratamientos son iguales?

b) Use el procedimiento LSD de Fisher para probar si existe una diferencia significativa entre las medias de los tratamientos A y B, A y C, y B y C. Use $\alpha = 0.05$.

c) Utilice el procedimiento LSD de Fisher para obtener una estimación por intervalo de 95% de confianza para la diferencia entre las medias de los tratamientos A y B.

14. Los datos siguientes se obtuvieron con un diseño completamente aleatorizado. Para los cálculos correspondientes use $\alpha = 0.05$.

	Tratamiento 1	Tratamiento 2	Tratamiento 3
	63	82	69
	47	72	54
	54	88	61
	40	66	48
$\bar{x}_j$	51	77	58
s_j^2	96.67	97.34	81.99

a) Use el análisis de varianza para probar si hay una diferencia significativa entre las medias de los tres tratamientos.
b) Utilice el procedimiento LSD de Fisher para probar cuáles son las medias que difieren.

Aplicaciones

AUTOevaluación

15. Con el fin de probar si la media del tiempo necesario para mezclar un lote de un material es la misma si emplea las máquinas de tres fabricantes, Jacobs Chemical obtiene los datos siguientes sobre el tiempo (en minutos) requerido para mezclar el material.

Fabricantes		
1	**2**	**3**
20	28	20
26	26	19
24	31	23
22	27	22

a) Use estos datos para probar si las medias poblacionales de los tiempos necesarios para mezclar un lote de material usando las máquinas de estos tres fabricantes difieren. Use $\alpha = 0.05$.
b) Con $\alpha = 0.05$ como nivel de significancia, use el procedimiento LSD de Fisher para probar la igualdad entre las medias obtenidas con las máquinas del fabricante 1 y del fabricante 3. ¿Qué conclusión se obtiene después de realizar la prueba?

AUTOevaluación

16. Remítase al ejercicio 15. Use el procedimiento LSD de Fisher para obtener una estimación por intervalo de 95% de confianza para la diferencia entre las medias del fabricante 1 y del fabricante 2.

17. En un experimento diseñado para investigar la percepción de los valores éticos corporativos entre personas especializadas en marketing, se obtuvieron los datos siguientes (las puntuaciones más altas indican valores éticos más elevados).

Gerentes de marketing	Investigación de mercados	Publicidad
6	5	6
5	5	7
4	4	6
5	4	5
6	5	6
4	4	6

a) Use $\alpha = 0.05$ para probar si existe una diferencia significativa de percepción entre los tres grupos.
b) Con $\alpha = 0.05$ como nivel de significancia, podemos concluir que sí hay diferencias entre la percepción de los gerentes de marketing, los especialistas en investigación de mercados y los expertos en publicidad. Aplique los procedimientos estudiados en esta sección para determinar dónde están las diferencias. Use $\alpha = 0.05$.

18. Para probar si existe una diferencia significativa entre cuatro máquinas respecto del número de horas entre dos averías, se obtuvieron los datos siguientes.

Máquina 1	Máquina 2	Máquina 3	Máquina 4
6.4	8.7	11.1	9.9
7.8	7.4	10.3	12.8
5.3	9.4	9.7	12.1
7.4	10.1	10.3	10.8
8.4	9.2	9.2	11.3
7.3	9.8	8.8	11.5

a) Con $\alpha = 0.05$, como nivel de significancia, ¿cuál es la diferencia, si hay alguna, entre las medias poblacionales de los tiempos de las cuatro máquinas?

b) Use el procedimiento LSD de Fisher para probar la igualdad de las medias en las máquinas 2 y 4. Utilice 0.05 como nivel de significancia.

19. Remítase al ejercicio 18. Use el ajuste de Bonferroni para probar si hay diferencia significativa entre todos los pares de medias. Suponga que desea que el máximo de la tasa de error por experimentación sea 0.05.

20. La International League of Triple-A, una liga menor de beisbol, aglomera 14 equipos en tres divisiones: Norte, Sur y Oeste. A continuación se muestran los datos del promedio de asistencia (Attendance) a los juegos de los 14 equipos de la International League, incluyendo nombre del equipo (Team Name) y división (Division) (sitio web de The Biz of Baseball, enero de 2009). También se muestran los récords de los equipos: W indica el número de juegos ganados, L el número de juegos perdidos y PCT la proporción de juegos ganados.

WEB archivo

Triple-A

Team Name	Division	W	L	PCT	Attendance
Buffalo Bisons	Norte	66	77	0.462	8 812
Lehigh Valley IronPigs	Norte	55	89	0.382	8 479
Pawtucket Red Sox	Norte	85	58	0.594	9 097
Rochester Red Wings	Norte	74	70	0.514	6 913
Scranton-Wilkes Barre Yankees	Norte	88	56	0.611	7 147
Syracuse Chiefs	Norte	69	73	0.486	5 765
Charlotte Knights	Sur	63	78	0.447	4 526
Durham Bulls	Sur	74	70	0.514	6 995
Norfolk Tides	Sur	64	78	0.451	6 286
Richmond Braves	Sur	63	78	0.447	4 455
Columbus Clippers	Oeste	69	73	0.486	7 795
Indianapolis Indians	Oeste	68	76	0.472	8 538
Louisville Bats	Oeste	88	56	0.611	9 152
Toledo Mud Hens	Oeste	75	69	0.521	8 234

a) Utilice $\alpha = 0.05$ para probar cualquier diferencia en la media de asistencia para las tres divisiones.

b) Utilice el procedimiento de la LSD de Fisher para determinar dónde se presentan las diferencias. Use $\alpha = 0.05$

13.4 Diseño de bloques aleatorizado

Hasta ahora sólo se ha considerado el diseño de experimentos completamente aleatorizado. Como recordará, para probar la diferencia entre las medias de los tratamientos se calcula el valor de F mediante el cociente

$$F = \frac{\text{CMTR}}{\text{CME}} \tag{13.20}$$

Un diseño completamente aleatorizado es útil cuando las unidades experimentales son homogéneas. Si estas unidades son heterogéneas, suele emplearse la ***formación de bloques*** *para tener grupos homogéneos.*

Sin embargo, puede surgir un problema por diferencias debido a factores ajenos (no considerados en el experimento) que ocasionen que el término CME en este cociente se vuelva más grande. En estos casos, el valor de F en la ecuación (13.20) será más pequeño, haciendo que se concluya que no hay diferencia entre las medias de los tratamientos cuando en realidad sí la hay.

En esta sección se presenta un diseño de experimentos conocido como **diseño de bloques aleatorizado**, cuyo propósito es controlar algunas fuentes ajenas de variación eliminándolas del término CME. Este diseño tiende a proporcionar una mejor estimación de la varianza del error y conduce a pruebas de hipótesis más sólidas en términos de su capacidad para detectar

diferencias entre medias de tratamientos. Para ilustrar esto se retoma un estudio sobre el estrés que experimentan los controladores del tráfico aéreo.

Prueba de estrés para controladores de tráfico aéreo

Como resultado de un estudio para medir la fatiga y el estrés de los controladores de tráfico aéreo, se propusieron modificaciones y rediseños a su estación de trabajo. Después de evaluar diversos diseños, se seleccionaron tres alternativas consideradas con el mayor potencial para reducir el estrés en los controladores. La pregunta clave es: ¿en qué medida difieren estas tres alternativas en su efecto sobre el estrés de los sujetos de estudio? Para responder esta pregunta es necesario diseñar un experimento que proporcione mediciones del estrés de los controladores del tráfico aéreo bajo cada alternativa.

En los estudios experimentales relacionados con negocios suelen intervenir unidades experimentales muy heterogéneas; en consecuencia, los diseños de bloques aleatorizados se emplean con frecuencia.

Si se empleara un diseño completamente aleatorizado, una muestra al azar de controladores sería asignada a cada una de las alternativas de estaciones de trabajo. Sin embargo, se cree que los sujetos difieren de forma significativa en su habilidad para manejar situaciones estresantes. Lo que para un controlador implica una gran tensión, para otro puede ser sólo un estrés moderado e incluso pequeño. Por tanto, al considerar la fuente de variación dentro del grupo (CME), hay que reconocer que esta variación comprende tanto el error aleatorio como el error debido a las diferencias individuales de los sujetos. De hecho, los gerentes consideran que la variabilidad entre los controladores será la contribución principal al término CME.

En el diseño de experimentos, la formación de bloques es similar a la estratificación en el muestreo.

Una manera de hacer a un lado el efecto de las diferencias individuales es usar el diseño de bloques aleatorizado, en el cual se identifica la variabilidad debido a las diferencias individuales de los controladores y se elimina del término CME. En el diseño de bloques aleatorizado se emplea una sola muestra de controladores. Cada uno de ellos se prueba con cada una de las tres alternativas de puestos de trabajo. En la terminología del diseño de experimentos, el puesto de trabajo es el *factor de interés* y los controladores son los *bloques*. Los tres tratamientos o poblaciones asociados con el factor puesto de trabajo son las tres alternativas de puesto de trabajo. Para simplificar, a estas tres alternativas se les designará como sistema A, sistema B y sistema C.

El aspecto *aleatorizado* del diseño de bloques aleatorizado es el orden al azar en el que les son asignados los tratamientos (sistemas) a los controladores. Si cada sujeto probara los tres sistemas en el mismo orden, cualquier diferencia encontrada podría deberse al orden de la prueba más que a las verdaderas diferencias entre los sistemas.

Para obtener los datos necesarios, en el Centro de Control Cleveland en Oberlin, Ohio, se instalaron las tres alternativas de estación de trabajo. Se seleccionó a seis controladores en forma aleatoria y se le asignó a cada sujeto uno de los sistemas para que lo operara. Después de practicar una entrevista y un examen médico a cada uno de los participantes en el estudio, se obtuvieron las mediciones del estrés de cada controlador en cada uno de los sistemas. En la tabla 13.5 se presentan estos datos con las etiquetas Blocks (bloques), Controller (controlador), System (sistema) y Treatments (tratamientos).

En la tabla 13.6 aparece un resumen de los datos recabados sobre el estrés. En ella se presentan los totales de las columnas (tratamientos) y los totales de las filas (bloques), así como

TABLA 13.5 Diseño de bloques aleatorizado para la prueba de estrés en los controladores de tráfico aéreo

		Treatments		
		System A	**System B**	**System C**
Blocks	**Controller 1**	15	15	18
	Controller 2	14	14	14
	Controller 3	10	11	15
	Controller 4	13	12	17
	Controller 5	16	13	16
	Controller 6	13	13	13

WEB archivo
AirTraffic

TABLA 13.6 Resumen de los datos recolectados para la prueba de estrés en los controladores de tráfico aéreo.

		Tratamientos			Totales de fila o de bloque	Medias por bloque
		Sistema A	Sistema B	Sistema C		
Bloques	Controlador 1	15	15	18	48	$\bar{x}_{1\cdot} = 48/3 = 16.0$
	Controlador 2	14	14	14	42	$\bar{x}_{2\cdot} = 42/3 = 14.0$
	Controlador 3	10	11	15	36	$\bar{x}_{3\cdot} = 36/3 = 12.0$
	Controlador 4	13	12	17	42	$\bar{x}_{4\cdot} = 42/3 = 14.0$
	Controlador 5	16	13	16	45	$\bar{x}_{5\cdot} = 45/3 = 15.0$
	Controlador 6	13	13	13	39	$\bar{x}_{6\cdot} = 39/3 = 13.0$
Totales de columna o de tratamiento		81	78	93	252	$\bar{\bar{x}} = \frac{252}{18} = 14.0$
Medias por tratamiento		$\bar{x}_{\cdot 1} = \frac{81}{6} = 13.5$	$\bar{x}_{\cdot 2} = \frac{78}{6} = 13.0$	$\bar{x}_{\cdot 3} = \frac{93}{6} = 15.5$		

algunas medias muestrales necesarias que serán útiles para efectuar los cálculos de la suma de cuadrados del ANOVA. Dado que los valores bajos de estrés se consideran mejores, los datos muestrales parecen favorecer el sistema B, en el que la media de las mediciones del estrés es 13. Sin embargo, la pregunta persiste: ¿los resultados muestrales justifican la conclusión de que las medias poblacionales de los niveles de estrés con estos tres sistemas difieren? Es decir, ¿las diferencias son estadísticamente significativas? Para responder esta pregunta se emplea un análisis del cálculo de la varianza, similar al empleado en el diseño completamente aleatorizado.

Procedimiento ANOVA

El procedimiento ANOVA para el diseño de bloques aleatorizado requiere la partición de la suma total de los cuadrados (STC) en tres grupos: la suma de los cuadrados debido a los tratamientos (SCTR), la suma de los cuadrados debido a los bloques (SCBL) y la suma de los cuadrados debida al error (SCE). A continuación se proporciona la fórmula para este particionamiento.

$$\text{STC} = \text{SCTR} + \text{SCBL} + \text{SCE} \qquad \textbf{(13.21)}$$

Esta suma de la partición de cuadrados se presenta en la tabla ANOVA para el diseño de bloques aleatorizado como se muestra en la tabla 13.7. La notación empleada es

k = número de tratamientos
b = número de bloques
n_T = tamaño muestral total ($n_T = kb$)

Observe que en la tabla ANOVA también se indica la partición de los $n_T - 1$ grados de libertad totales de manera que $k - 1$ grados de libertad correspondan a los tratamientos, $b - 1$ a los bloques y $(k - 1)(b - 1)$ al término del error. En la columna cuadrado medio se proporcionan las sumas de los cuadrados divididas entre los grados de libertad, y F = CMTR/CME es el cociente F que se usa para probar si hay diferencias significativas entre las medias de los tratamientos. La contribución más importante del diseño de bloques aleatorizado radica en que, al emplear bloques, se eliminan del término CME las diferencias individuales de los controladores y se obtiene una prueba más sólida para las diferencias de estrés entre las tres alternativas de estaciones de trabajo.

TABLA 13.7 Tabla ANOVA para el diseño de bloques aleatorizado con k tratamientos y b bloques

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	valor-p
Tratamientos	SCTR	$k - 1$	$\text{CMTR} = \dfrac{\text{SCTR}}{k-1}$	$\dfrac{\text{CMTR}}{\text{CME}}$	
Bloques	SCBL	$b - 1$	$\text{CMBL} = \dfrac{\text{SCBL}}{b-1}$		
Error	SCE	$(k-1)(b-1)$	$\text{CME} = \dfrac{\text{SCE}}{(k-1)(b-1)}$		
Total	STC	$n_T - 1$			

Cálculos y conclusiones

Para calcular el estadístico F requerido para probar si existe diferencia entre las medias de los tratamientos en un diseño de bloques aleatorizado, se necesita calcular el CMTR y el CME. Para determinar estos dos cuadrados medios es preciso calcular primero la SCTR y la SCE; para esto también se calcula la SCBL y la STC. En forma más sencilla, estos procedimientos se realizan en cuatro pasos. Además de la notación k, b y n_T ya definida, se usará:

x_{ij} = valor de la observación correspondiente al tratamiento j en el bloque i
$\bar{x}_{\cdot j}$ = media muestral del tratamiento j-ésimo
$\bar{x}_{i\cdot}$ = media muestral para el bloque i-ésimo
$\bar{\bar{x}}$ = media muestral general

Paso 1. Calcular la suma total de cuadrados (STC).

$$\text{STC} = \sum_{i=1}^{b}\sum_{j=1}^{k}(x_{ij} - \bar{\bar{x}})^2 \qquad \textbf{(13.22)}$$

Paso 2. Estimar la suma de cuadrados debido a los tratamientos (SCTR).

$$\text{SCTR} = b\sum_{j=1}^{k}(\bar{x}_{\cdot j} - \bar{\bar{x}})^2 \qquad \textbf{(13.23)}$$

Paso 3. Calcular la suma de cuadrados debido a los bloques (SCBL).

$$\text{SCBL} = k\sum_{i=1}^{b}(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 \qquad \textbf{(13.24)}$$

Paso 4. Determinar la suma de cuadrados debido al error (SCE).

$$\text{SCE} = \text{STC} - \text{SCTR} - \text{SCBL} \qquad \textbf{(13.25)}$$

En el caso de los datos de la tabla 13.6 sobre los controladores del tráfico aéreo, con estos cálculos se obtienen las sumas de los cuadrados siguientes.

Paso 1. $\text{STC} = (15-14)^2 + (15-14)^2 + (18-14)^2 + \cdots + (13-14)^2 = 70$

Paso 2. $\text{SCTR} = 6[(13.5-14)^2 + (13.0-14)^2 + (15.5-14)^2] = 21$

Paso 3. $\text{SCBL} = 3[(16-14)^2 + (14-14)^2 + (12-14)^2 + (14-14)^2 + (15-14)^2 + (13-14)^2] = 30$

Paso 4. $\text{SCE} = 70 - 21 - 30 = 19$

TABLA 13.8 Tabla ANOVA para la prueba de estrés de los controladores de tráfico aéreo

Fuentes de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	*F*	valor-*p*
Tratamientos	21	2	10.5	10.5/1.9 = 5.53	0.024
Bloques	30	5	6.0		
Error	19	10	1.9		
Total	70	17			

Las sumas de cuadrados divididas entre sus grados de libertad proporcionan los correspondientes cuadrados medios que se presentan en la tabla 13.8.

Ahora, para realizar la prueba de hipótesis se usará $\alpha = 0.05$ como nivel de significancia. El valor del estadístico de prueba es

$$F = \frac{\text{CMTR}}{\text{CME}} = \frac{10.5}{1.9} = 5.53$$

Los grados de libertad en el numerador son $k - 1 = 3 - 1 = 2$, y en el denominador son $(k - 1)(b - 1) = (3 - 1)(6 - 1) = 10$. Como la prueba de hipótesis nula es rechazada sólo cuando los valores del estadístico de prueba son grandes, el valor-*p* es el área bajo la distribución *F* a la derecha de $F = 5.53$. En la tabla 4 del apéndice B se puede ver que para 2 y 10 grados de libertad, $F = 5.53$ se encuentra entre $F_{0.025} = 5.46$ y $F_{0.01} = 7.56$. Por tanto, el área en la cola superior, o valor-*p*, se ubica entre 0.01 y 0.025. Se puede usar también Excel o Minitab y encontrar que el valor-*p* exacto para $F = 5.53$ es 0.024. Como el valor-$p \leq \alpha = 0.05$, se rechaza la hipótesis nula H_0: $\mu_1 = \mu_2 = \mu_3$, y se concluye que las medias poblacionales de los niveles de estrés en las tres alternativas de estación de trabajo no son iguales.

Acerca de este diseño de bloques aleatorizado se pueden exponer algunos comentarios generales. El diseño de experimentos descrito en esta sección es un diseño de bloques *completo;* la palabra "completo" indica que cada bloque se somete a todos los *k* tratamientos. Es decir, todos los controladores (bloques) fueron probados con los tres sistemas (tratamientos). A los diseños de experimentos en los que a cada bloque se le aplican algunos, pero no todos los tratamientos, se les llama diseños de bloques *incompleto*. Su estudio queda fuera del alcance de este libro.

Como en la prueba sobre el estrés de los controladores de tráfico aéreo cada sujeto usó todos los sistemas, este método garantiza un diseño de bloques completo. En algunos casos la formación de los bloques se realiza con unidades experimentales "similares" en cada bloque. Por ejemplo, suponga que en una prueba preliminar realizada a los controladores se divide la población en grupos que van desde personas con mucho estrés hasta individuos con estrés sumamente bajo. Aquí también se puede tener la formación de bloques haciendo que en el estudio participen tres controladores de cada nivel de estrés. En este caso, cada bloque consistirá en tres sujetos de un mismo nivel de estrés. El aspecto aleatorizado del diseño de bloques será la designación aleatoria de los tres controladores de cada bloque a los tres sistemas.

Por último, observe que en la tabla ANOVA que se presenta en la tabla 13.7, se proporciona un valor *F* para probar los efectos de los tratamientos pero *no* de los bloques. La razón estriba en que el experimento se diseñó para probar un solo factor: el diseño de la estación de trabajo. La formación de bloques basada en las diferencias del estrés individuales se realizó para eliminar tal variación del término CME. El estudio no se diseñó para detectar las diferencias individuales de estrés.

Algunos analistas calculan $F = \text{CMBL/CME}$ y usan este estadístico para probar la significancia de los bloques. Después utilizan los resultados como guía para determinar si el mismo tipo de bloques puede ser útil en experimentos futuros. Sin embargo, si la diferencia en el estrés de las personas ha de ser un factor en el estudio, deberá emplearse un diseño de experimentos diferente. Una prueba de significancia sobre los bloques no debe hacerse como base para una conclusión acerca de un segundo factor.

NOTAS Y COMENTARIOS

En un diseño de bloques aleatorizado, los grados de libertad del error son menos que en un diseño completamente aleatorizado, debido a que en los b bloques se pierden $b - 1$ grados de libertad. Si n es pequeño, los efectos potenciales debido a los bloques pueden quedar ocultos por la pérdida de grados de libertad del error; con n grande, los efectos se minimizan.

Ejercicios

Métodos

AUTOevaluación

21. Considere los resultados experimentales del siguiente diseño de bloques aleatorizado. Realice los cálculos necesarios para establecer la tabla de análisis de varianza.

		Tratamientos		
		A	B	C
Bloques	1	10	9	8
	2	12	6	5
	3	18	15	14
	4	20	18	18
	5	8	7	8

Utilice $\alpha = 0.05$ para probar cualesquiera diferencias significativas.

22. Los siguientes datos se obtuvieron de un diseño de bloques aleatorizado con cinco tratamientos y tres bloques: STC = 430, SCTR = 310 y SCBL = 85. Establezca la tabla ANOVA y pruebe si existen cualesquiera diferencias significativas. Use $\alpha = 0.05$.

23. Se realizó un experimento con cuatro tratamientos y ocho bloques. Complete la siguiente tabla de análisis de varianza.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F
Tratamientos	900			
Bloques	400			
Error				
Total	1 800			

Use $\alpha = 0.05$ y pruebe si existen cualesquiera diferencias significativas.

Aplicaciones

24. Un vendedor de automóviles realiza una prueba para determinar si el tiempo en minutos que se necesita para afinar un motor pequeño depende de si se utiliza un analizador de motor computarizado o uno electrónico. Debido a que el tiempo de afinación varía entre automóviles compactos, medianos y grandes, en el experimento se utilizaron los tres tipos de vehículos como bloques. Los datos obtenidos se indican a continuación.

		Analizador	
		Computarizado	**Electrónico**
Automóvil	**Compacto**	50	42
	Mediano	55	44
	Grande	63	46

Use $\alpha = 0.05$ y pruebe si existen cualesquiera diferencias significativas.

25. Las vitaminas y otros suplementos para la salud se han encarecido durante los años recientes y, con frecuencia, los precios establecidos por los distintos minoristas varían en gran medida. Los datos a continuación listan los precios de 13 productos (Item) de cuatro minoristas en Rochester, Nueva York (*Democrat and Chronicle*, 13 de febrero de 2005).

WEB archivo
Vitamins

Item	**CVS**	**Kmart**	**Rite-Aid**	**Wegmans**
Caltrate +D (600 mg/60 tabletas)	8.49	5.99	7.99	5.99
Centrum (130 tabletas)	9.49	9.47	9.89	7.97
Aceite de hígado de bacalao (100 tabletas en gel)	2.66	2.59	1.99	2.69
Aceite de pescado (1,000 mg/60 tabletas)	6.19	4.99	4.99	5.99
Vitaminas para niños (60 tabletas)	7.69	5.99	5.99	6.29
Ácido fólico (400 mcg/250 tabletas)	2.19	2.49	3.74	2.69
One-a-Day Maximum (100 tabletas)	8.99	7.49	6.99	6.99
One-a-Day Scooby (50 tabletas)	7.49	5.99	6.49	5.47
Poly-Vi-Sol (gotas, 50 ml)	9.99	8.49	9.99	8.37
Vitamina B-12 (100 mcg/100 tabletas)	3.59	1.99	1.99	1.79
Vitamina C (500 mg/100 tabletas)	2.99	2.49	1.99	2.39
Vitamina E (200 UI/100 tabletas)	4.69	3.49	2.99	3.29
Zinc (50 mg/100 tabletas)	2.66	2.59	3.99	2.79

Use $\alpha = 0.05$ y pruebe si existe alguna diferencia significativa entre los precios medios de los cuatro minoristas.

26. El Examen de aptitud escolar (SAT, por sus siglas en inglés) contiene tres secciones: lectura crítica, matemáticas y redacción. Cada parte se califica en una escala de 800 puntos. La información de las puntuaciones del examen para la versión 2009 del SAT está disponible en el sitio web del College Board. Una muestra de las puntuaciones alcanzadas por seis estudiantes (Student) en el SAT se lista enseguida para lectura crítica (Critical Reading), matemáticas (Mathematics) y redacción (Writing).

WEB archivo
SATScores

Student	**Critical Reading**	**Mathematics**	**Writing**
1	526	534	530
2	594	590	586
3	465	464	445
4	561	566	553
5	436	478	430
6	430	458	420

a) Utilizando un nivel de significancia de 0.05, ¿los estudiantes se desempeñan de manera distinta en las tres partes del examen?

b) ¿Cuál sección parece darles más problemas? Explique.

27. El *Journal of the American Medical Association* publicó una investigación acerca de la demanda cardiaca por palear grandes cantidades de nieve. Diez hombres saludables se sometieron a pruebas de ejercicio empleando una caminadora y una bicicleta adaptada ergonómicamente para ejercitar los brazos. Después, estos mismos hombres limpiaron dos tramos de nieve mojada y pesada con una pala ligera para nieve y un lanzanieve eléctrico. Se midió el ritmo cardiaco, la presión sanguínea y el consumo de oxígeno de cada uno de los participantes en la prueba durante la remoción de nieve, y estos valores se compararon con los obtenidos durante

las pruebas con la caminadora (Treadmill) y la bicicleta adaptada (Arm-Crank Ergometer). En la tabla siguiente se presentan los valores de ritmo cardiaco expresados en pulsaciones por minuto, de cada uno de los 10 individuos (Subject). Se incluyen los valores de pala para nieve (Snow Shovel) y lanzanieve eléctrico (Snow Thrower).

WEB archivo
SnowShoveling

Subject	Treadmill	Arm-Crank Ergometer	Snow Shovel	Snow Thrower
1	177	205	180	98
2	151	177	164	120
3	184	166	167	111
4	161	152	173	122
5	192	142	179	151
6	193	172	205	158
7	164	191	156	117
8	207	170	160	123
9	177	181	175	127
10	174	154	191	109

Con un nivel de significancia de 0.05, pruebe si existen cualesquiera diferencias significativas.

13.5 Experimento factorial

Los diseños de experimentos estudiados hasta ahora permiten formular conclusiones estadísticas acerca de un solo factor. Sin embargo, en algunos experimentos tal vez se quieran formular conclusiones acerca de más de una variable o factor. Un **experimento factorial** es un diseño que permite obtener conclusiones simultáneas acerca de dos o más factores. El término *factorial* se utiliza porque las condiciones experimentales incluyen todas las posibles combinaciones de los factores. Por ejemplo, para a niveles de un factor A y b niveles de un factor B, el experimento incluirá una colección de datos en el tratamiento de las combinaciones ab. En esta sección mostraremos el análisis para un experimento factorial de dos factores. El enfoque básico puede ampliarse a más de dos factores.

Como ilustración de un experimento factorial de dos factores, veremos un estudio acerca del Examen de admisión de graduados en administración (GMAT, por sus siglas en inglés), una prueba estandarizada que utilizan las escuelas de negocios para evaluar una habilidad de los aspirantes a cubrir un programa de grado en ese campo. Las puntuaciones del GMAT están en el rango de 200 a 800; las de nivel más elevado significan una aptitud más alta.

Con la intención de mejorar el desempeño de los estudiantes en el GMAT, una de las principales universidades de Texas considera ofrecer los siguientes tres programas de preparación para ese examen.

1. Una sesión de repaso de tres horas, en la que se revisa el tipo de preguntas que suele encontrarse en el GMAT.
2. Un programa de un día en el que se ve el material más relevante del examen, junto con un examen muestra que se califica.
3. Un curso intensivo de 10 semanas en el que se identifican las debilidades de cada estudiante y se establecen programas individualizados de mejora.

Por tanto, un factor en este estudio es el programa de preparación, el cual tiene tres tratamientos: un repaso de tres horas, un programa de un día y un curso de 10 semanas. Antes de seleccionar la opción a adoptar, más estudios llevarán a determinar el efecto de cada uno de los programas sobre las puntuaciones obtenidas en este examen de admisión.

Por lo general, los aplicantes del GMAT son estudiantes de tres licenciaturas: negocios, ingeniería y artes y ciencias. En consecuencia, el segundo factor de interés en el experimento es si la licenciatura influye en la calificación del GMAT. Para este segundo factor hay también tres tratamientos: negocios, ingeniería y artes y ciencias. El diseño factorial de este experimento con tres tratamientos para el factor A, programa de preparación, y tres tratamientos para el

TABLA 13.9 Las nueve combinaciones de tratamiento en el experimento con dos factores del GMAT

		Factor B: licenciatura		
		Negocios	**Ingeniería**	**Artes y ciencias**
Factor A: programa de preparación	**Repaso de tres horas**	1	2	3
	Programa de un día	4	5	6
	Curso de 10 semanas	7	8	9

factor B, tipo de licenciatura, habrá un total de $3 \times 3 = 9$ combinaciones. En la tabla 13.9 se resumen estas combinaciones de tratamientos o condiciones experimentales.

Suponga que se toma una muestra de dos sujetos para cada una de las combinaciones de tratamientos de la tabla 13.9: dos estudiantes de negocios participarán en el repaso de tres horas, dos participarán en el programa de un día y otros dos en el curso de 10 semanas. Además, dos estudiantes de ingeniería y dos de artes y ciencias participarán en cada uno de los tres programas. En la terminología del diseño de experimentos, el tamaño muestral de dos para cada combinación de tratamientos indica que se tienen dos **replicaciones**. Se pueden usar también más replicaciones y tamaños muestrales mayores, pero elegimos minimizar los cálculos para este ejemplo.

En este diseño de experimentos se requiere que de *cada una* de las licenciaturas (negocios, ingeniería y artes y ciencias) se tomen aleatoriamente seis estudiantes que pretendan realizar este examen de admisión. Después, dos de cada licenciatura deben ser asignados de manera aleatoria a cada uno de los programas de preparación para el examen, con lo que en total participan 18 sujetos en el estudio.

Asumamos que los estudiantes seleccionados de manera aleatoria participaron en los programas de preparación y luego tomaron el GMAT. En la tabla 13.10 se presentan las calificaciones obtenidas en el programa de preparación (Preparation Program), que incluyó repaso de tres horas (Three-hour review), programa de un día (One-day program) y curso de 10 semanas (10-week course) para las licenciaturas (College) de negocios (Business), ingeniería (Engineering) y artes y ciencias (Arts and Sciences). Los cálculos para el análisis de varianza con los datos de la tabla 13.10 darán respuesta a las siguientes preguntas.

- **Efecto principal (factor A)**. ¿Los programas de preparación tienen efectos diferentes sobre la puntuación obtenida en el GMAT?
- **Efecto principal (factor B)**. ¿Las licenciaturas tienen efectos diferentes sobre la puntuación obtenida en el GMAT?
- **Efecto de interacción (factores A y B)**. ¿Es uno de los programas de preparación mejor para los estudiantes que provienen de una de las tres licenciaturas, mientras que para los de otras licenciaturas es mejor otro de los programas?

El término **interacción** se refiere a un nuevo efecto que es posible estudiar debido a que se emplea un experimento factorial. Si el efecto de interacción tiene algún impacto significativo

TABLA 13.10 Puntuaciones en el GMAT para el experimento de dos factores

		Factor B: College		
		Business	**Engineering**	**Arts and Sciences**
Factor A: Preparation Program	**Three-hour review**	500 580	540 460	480 400
	One-day program	460 540	560 620	420 480
	10-week course	560 600	600 580	480 410

WEB archivo
GMATStudy

TABLA 13.11 Tabla ANOVA para el experimento factorial de dos factores con r replicaciones

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	valor-p
Factor A	SCA	$a - 1$	$\text{CMA} = \dfrac{\text{SCA}}{a - 1}$	$\dfrac{\text{CMA}}{\text{CME}}$	
Factor B	SCB	$b - 1$	$\text{CMB} = \dfrac{\text{SCB}}{b - 1}$	$\dfrac{\text{CMB}}{\text{CME}}$	
Interacción	SCAB	$(a - 1)(b - 1)$	$\text{CMAB} = \dfrac{\text{SCAB}}{(a - 1)(b - 1)}$	$\dfrac{\text{CMAB}}{\text{CME}}$	
Error	SCE	$ab(r - 1)$	$\text{CME} = \dfrac{\text{SCE}}{ab(r - 1)}$		
Total	STC	$n_T - 1$			

sobre las puntuaciones del GMAT, se podrá concluir que el efecto del tipo de programa de preparación depende de la licenciatura.

Procedimiento ANOVA

El procedimiento ANOVA para el experimento factorial de dos factores requiere la partición de la suma total de cuadrados (STC) en cuatro grupos: suma de cuadrados del factor A (SCA), suma de cuadrados del factor B (SCB), suma de cuadrados de la interacción (SCAB) y suma de cuadrados debido al error (SCE). La fórmula para esta partición se da a continuación.

$$\text{STC} = \text{SCA} + \text{SCB} + \text{SCAB} + \text{SCE} \qquad \textbf{(13.26)}$$

En la tabla 13.11 se resumen las particiones de las sumas de cuadrados y de los grados de libertad. Se emplea la notación siguiente:

a = número de niveles del factor A

b = número de niveles del factor B

r = número de replicaciones

n_T = número total de observaciones realizadas en el experimento; $n_T = abr$

Cálculos y conclusiones

Para determinar los estadísticos F que se requieren en las pruebas de significancia del factor A, del factor B y de la interacción, es necesario calcular CMA, CMB, CMAB y CME. Para obtener estos cuatro cuadrados medios se debe calcular primero SCA, SCB, SCAB y SCE; con esto se calcula también STC. Para simplificar la presentación, los procedimientos se dividen en cinco pasos. Además de a, b, r y n_T definidos previamente, se emplea la siguiente notación.

x_{ijk} = observación correspondiente a la k-ésima réplica tomada del tratamiento i del factor A y del tratamiento j del factor B

$\bar{x}_{i\cdot}$ = media muestral de las observaciones en el tratamiento i (factor A)

$\bar{x}_{\cdot j}$ = media muestral de las observaciones en el tratamiento j (factor B)

$\bar{x}_{ij}$ = media muestral de las observaciones correspondientes a la combinación del tratamiento i (factor A) y el tratamiento j (factor B)

$\bar{\bar{x}}$ = media muestral general de todas las n_T observaciones

Paso 1. Calcular la suma total de cuadrados.

$$\text{STC} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(x_{ijk} - \bar{\bar{x}})^2 \tag{13.27}$$

Paso 2. Calcular la suma de cuadrados del factor A.

$$\text{SCA} = br\sum_{i=1}^{a}(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 \tag{13.28}$$

Paso 3. Calcular la suma de cuadrados del factor B.

$$\text{SCB} = ar\sum_{j=1}^{b}(\bar{x}_{\cdot j} - \bar{\bar{x}})^2 \tag{13.29}$$

Paso 4. Calcular la suma de cuadrados debido a la interacción.

$$\text{SCAB} = r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{x}_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})^2 \tag{13.30}$$

Paso 5. Calcular la suma de cuadrados debido al error.

$$\text{SCE} = \text{STC} - \text{SCA} - \text{SCB} - \text{SCAB} \tag{13.31}$$

En la tabla 13.12 se reportan los datos obtenidos en el experimento y las diversas sumas necesarias para los cálculos de las sumas de cuadrados. Mediante las ecuaciones (13.27) a (13.31) se calculan las siguientes sumas de cuadrados del experimento factorial de dos factores del GMAT.

Paso 1. $\text{STC} = (500 - 515)^2 + (580 - 515)^2 + (540 - 515)^2 + \cdots + (410 - 515)^2 = 82\,450$

Paso 2. $\text{SCA} = (3)(2)[(493.33 - 515)^2 + (513.33 - 515)^2 + (538.33 - 515)^2] = 6\,100$

Paso 3. $\text{SCB} = (3)(2)[(540 - 515)^2 + (560 - 515)^2 + (445 - 515)^2] = 45\,300$

Paso 4. $\text{SCAB} = 2[(540 - 493.33 - 540 + 515)^2 + (500 - 493.33 - 560 + 515)^2 + \cdots + (445 - 538.33 - 445 + 515)^2] = 11\,200$

Paso 5. $\text{SCE} = 82\,450 - 6\,100 - 45\,300 - 11\,200 = 19\,850$

Estas sumas divididas entre sus correspondientes grados de libertad proporcionan los valores de los cuadrados medios apropiados para estimar los dos efectos principales (programas de preparación y licenciatura) y el efecto de su interacción.

Debido a la gran cantidad de cálculos involucrada en cualquier experimento factorial desde uno modesto hasta uno de gran dimensión, usualmente la computadora juega un papel importante en la realización de los cálculos necesarios en el análisis de varianza mostrado antes y en la obtención de los valores-p que se emplean para tomar las decisiones en la prueba de hipótesis. En la figura 13.6 se presenta la pantalla de resultados de Minitab para el análisis de varianza del experimento factorial de dos factores del GMAT. Para realizar la prueba de hipótesis de dos factores en este estudio usaremos el resultado de Minitab y un nivel de significancia $\alpha = 0.05$. El valor-p utilizado para probar si hay diferencias significativas entre los tres programas de preparación (factor A) es 0.299. Como este valor-$p = 0.299$ es mayor que $\alpha = 0.05$, no existe diferencia significativa entre las medias de las puntuaciones obtenidas en el GMAT para los tres programas de preparación. Sin embargo, en relación con el efecto de la licenciatura, el valor-$p = 0.005$ es menor que $\alpha = 0.05$; por tanto, sí hay una diferencia significativa en las medias de las puntuaciones en el GMAT entre las tres licenciaturas. Por último, debido a que el

TABLA 13.12 Resumen de los datos del examen GMAT para el experimento de dos factores

	Totales de combinación de tratamiento	Factor B: licenciatura			Totales de fila	Medias del factor A
		Negocios	**Ingeniería**	**Artes y ciencias**		
Factor A: programa de preparación	**Repaso de tres horas**	500 580 1 080 $\bar{x}_{11} = \frac{1\,080}{2} = 540$	540 460 1 000 $\bar{x}_{12} = \frac{1\,000}{2} = 500$	480 400 880 $\bar{x}_{13} = \frac{880}{2} = 440$	2 960	$\bar{x}_{1.} = \frac{2\,960}{6} = 493.33$
	Programa de un día	460 540 1 000 $\bar{x}_{21} = \frac{1\,000}{2} = 500$	560 620 1 180 $\bar{x}_{22} = \frac{1\,180}{2} = 590$	420 480 900 $\bar{x}_{23} = \frac{900}{2} = 450$	3 080	$\bar{x}_{2.} = \frac{3\,080}{6} = 513.33$
	Curso de 10 semanas	560 600 1 160 $\bar{x}_{31} = \frac{1\,160}{2} = 580$	600 580 1 180 $\bar{x}_{32} = \frac{1\,180}{2} = 590$	480 410 890 $\bar{x}_{33} = \frac{890}{2} = 445$	3 230	$\bar{x}_{3.} = \frac{3\,230}{6} = 538.33$
	Totales de columna	3 240	3 360	2 670	9 270 ← Total general	
	Medias del factor B	$\bar{x}_{.1} = \frac{3\,240}{6} = 540$	$\bar{x}_{.2} = \frac{3\,360}{6} = 560$	$\bar{x}_{.3} = \frac{2\,670}{6} = 445$	$\bar{\bar{x}} = \frac{9\,270}{18} = 515$	

FIGURA 13.6 Pantalla de resultados de Minitab para el diseño de dos factores del examen GMAT

SOURCE	DF	SS	MS	F	P
Factor A	2	6100	3050	1.38	0.299
Factor B	2	45300	22650	10.27	0.005
Interaction	4	11200	2800	1.27	0.350
Error	9	19850	2206		
Total	17	82450			

valor-p de 0.350 correspondiente al efecto de la interacción es mayor que $\alpha = 0.05$, no hay un efecto significativo de interacción. Por tanto, en este estudio no se encuentran razones para pensar que los tres programas de preparación difieren en su capacidad para capacitar a estudiantes de las distintas licenciaturas para el GMAT.

Se encontró que la licenciatura sí es un factor significativo. Al revisar los cálculos de la tabla 13.12, vemos que las medias muestrales son: estudiantes de negocios $\bar{x}_{\cdot 1} = 540$, estudiantes de ingeniería $\bar{x}_{\cdot 2} = 560$ y estudiantes de artes y ciencias $\bar{x}_{\cdot 3} = 445$. Se pueden realizar pruebas para los distintos tratamientos; sin embargo, después de observar las tres medias muestrales es posible anticipar que no hay diferencia entre los alumnos con las licenciaturas de ingeniería y negocios. Pero los de artes y ciencias parecen estar menos preparados para este examen que los de las otras dos licenciaturas. Quizás esta observación haga que la universidad busque otras opciones para ayudar a este grupo a prepararse para el GMAT.

Ejercicios

Métodos

AUTO evaluación

28. En un experimento factorial con dos niveles para el factor A y tres niveles para el factor B se obtuvieron los datos siguientes.

		Factor B		
		Nivel 1	**Nivel 2**	**Nivel 3**
Factor A	**Nivel 1**	135 165	90 66	75 93
	Nivel 2	125 95	127 105	120 136

Realice una prueba para determinar si hay algunos efectos principales significativos y algún efecto de interacción. Use $\alpha = 0.05$.

29. De los cálculos de un experimento factorial con cuatro niveles para el factor A, tres niveles para el factor B y tres replicaciones se obtuvieron los datos siguientes: STC = 280, SCA = 26, SCB = 23 y SCAB = 175. Establezca la tabla ANOVA y pruebe si hay algunos efectos principales significativos y algún efecto de interacción. Use $\alpha = 0.05$.

Aplicaciones

30. Una empresa de ventas por catálogo realizó un experimento factorial para probar el efecto del tamaño de un anuncio de revista y su diseño sobre el número de solicitudes de catálogos recibido (datos en miles). Se pusieron a consideración tres diseños publicitarios y dos tamaños. Los datos obtenidos se presentan a continuación. Utilice el procedimiento ANOVA para un

diseño factorial a fin de probar si hay efectos significativos debido al tipo de diseño, al tamaño del anuncio o a la interacción. Use $\alpha = 0.05$.

		Tamaño del anuncio	
		Pequeño	**Grande**
Diseño	**A**	8 12	12 8
	B	22 14	26 30
	C	10 18	18 14

31. Un parque de diversión estudió algunos métodos para reducir el tiempo de espera (en minutos) al bajar y subir a los pasajeros a los juegos. Se propusieron dos métodos para realizar estas tareas. Para tomar en cuenta las diferencias potenciales debido al tipo de juego y a la interacción que puede haber entre tipo de juego y método de subir y bajar a los pasajeros, se diseñó un experimento factorial. Use los datos siguientes para pobrar cualquier efecto significativo debido al método de subir y bajar a los pasajeros, el tipo de juego y la interacción. Use $\alpha = 0.05$.

	Tipo de juego		
	Montaña rusa	**Rueda de la fortuna**	**Tobogán**
Método 1	41 43	52 44	50 46
Método 2	49 51	50 46	48 44

32. En un estudio diseñado para comparar vehículos híbridos (Hybrid) y convencionales (Conventional) con equipo similar, *Consumer Reports* probó varias clases de automóviles híbridos, automóviles a gasolina y vehículos utilitarios deportivos (SUV). Los datos siguientes muestran la clasificación en millas por galón que *Consumer Reports* obtuvo para dos automóviles compactos (Small Car) híbridos, dos automóviles medianos (Midsize Car) híbridos, dos SUV compactos (Small SUV) híbridos y dos SUV medianos (Midsize SUV) híbridos; también se muestra el rendimiento en millas por galón obtenidas de ocho modelos convencionales con equipo similar (*Consumer Reports*, octubre de 2008). Make/Model indica fabricante y modelo; Class (clase), Type (tipo) y MPG (millas por galón).

WEB archivo
HybridTest

Make/Model	Class	Type	MPG
Honda Civic	Small Car	Hybrid	37
Honda Civic	Small Car	Conventional	28
Toyota Prius	Small Car	Hybrid	44
Toyota Corolla	Small Car	Conventional	32
Chevrolet Malibu	Midsize Car	Hybrid	27
Chevrolet Malibu	Midsize Car	Conventional	23
Nissan Altima	Midsize Car	Hybrid	32
Nissan Altima	Midsize Car	Conventional	25
Ford Escape	Small SUV	Hybrid	27
Ford Escape	Small SUV	Conventional	21
Saturn Vue	Small SUV	Hybrid	28
Saturn Vue	Small SUV	Conventional	22
Lexus RX	Midsize SUV	Hybrid	23
Lexus RX	Midsize SUV	Conventional	19
Toyota Highlander	Midsize SUV	Hybrid	24
Toyota Highlander	Midsize SUV	Conventional	18

Realice pruebas para encontrar efectos significativos debido a la clase, tipo e interacción con un nivel de significancia $\alpha = 0.05$.

33. En un estudio publicado en *The Accounting Review* se examinaron los efectos separados y conjuntos de dos grados de presión de tiempo (bajo y moderado) y de tres niveles de conocimiento (inexperto, declarativo y de procedimiento) en la conducta de un grupo de sujetos al seleccionar palabras clave en una investigación de impuestos. A los sujetos se les presentaban casos de impuestos que contenían una serie de hechos, un asunto sobre impuestos y un índice con 1 336 palabras clave. Se les pedía que seleccionaran las palabras clave que creyeran que llevarían a una autoridad tributaria relevante a resolver el caso. Antes del experimento, un grupo de expertos fiscales determinó que en el texto había 19 palabras clave relevantes. Los sujetos en el grupo inexperto poseían poco o ningún conocimiento declarativo o de procedimiento; los sujetos en el grupo declarativo tenían un conocimiento declarativo significativo, pero poco o ninguno de procedimiento, y los sujetos en el grupo de procedimiento tenían considerables conocimientos declarativos y de procedimiento. El conocimiento declarativo abarca tanto las reglas impositivas aplicables como los términos técnicos empleados para describir esas reglas; y el de procedimiento es el conocimiento de las reglas que guían la búsqueda del investigador de impuestos para encontrar palabras clave. Los sujetos en el grupo de poca presión de tiempo contaron con 25 minutos para resolver el problema, cantidad de tiempo que debía ser "más que adecuada" para completar el caso; los sujetos en el grupo de presión de tiempo moderada tuvieron "sólo" 11 minutos para resolver el problema. Se seleccionaron 25 sujetos para cada una de las seis combinaciones de tratamientos y las medias muestrales de cada combinación de tratamientos se indican a continuación (las desviaciones estándar están entre paréntesis).

		Conocimiento		
		Inexperto	**Declarativo**	**De procedimiento**
Presión de tiempo	**Baja**	1.13 (1.12)	1.56 (1.33)	2.00 (1.54)
	Moderada	0.48 (0.80)	1.68 (1.36)	2.86 (1.80)

Use el procedimiento del ANOVA para probar si hay diferencias significativas debido a la presión de tiempo, al conocimiento o a la interacción. Use 0.05 como nivel de significancia. Asuma que la suma total de cuadrados en este experimento es 327.50.

Resumen

En este capítulo se estudió cómo usar el análisis de varianza para encontrar diferencias entre las medias de varias poblaciones o tratamientos. Se presentó el diseño completamente aleatorizado, el diseño de bloques aleatorizado y el experimento factorial de dos factores. Los diseños completamente aleatorizado y de bloques aleatorizado se usan para formular conclusiones acerca de las diferencias en las medias de un solo factor. El objetivo principal de la formación de bloques en el diseño de bloques aleatorizado es eliminar fuentes extrañas de variación del término del error. La formación de bloques proporciona una mejor estimación de la verdadera varianza del error y una mejor prueba para determinar si las medias de las poblaciones o tratamientos del factor difieren de forma significativa.

Se mostró que la base para las pruebas estadísticas empleadas en el análisis de varianza y en el diseño de experimentos es la obtención de dos estimaciones independientes de la varianza poblacional σ^2. En el caso de un solo factor, uno de los estimadores se basa en la variación entre los tratamientos; este estimador proporciona sólo un estimador insesgado de σ^2 si las medias $\mu_1, \mu_2, \ldots, \mu_k$ son iguales. El otro estimador de σ^2 se basa en la variación de las observaciones dentro de cada muestra y siempre proporciona un estimador insesgado de σ^2. Al calcular el cociente de estos dos estimadores (el estadístico F) se obtiene la regla de rechazo para determinar si la hipótesis nula, que establece que las medias poblacionales o de los tratamientos son iguales, es rechazada o no. En todos los diseños de experimentos aquí conside-

rados, la partición de las sumas de cuadrados y de los grados de libertad en sus diferentes fuentes permite calcular los valores necesarios para el análisis de varianza y las pruebas. Se mostró también cómo usar el procedimiento de la LSD de Fisher y el ajuste de Bonferroni para realizar comparaciones por pares y determinar cuáles medias son diferentes.

Glosario

Diseño completamente aleatorizado Diseño de experimentos en el que los tratamientos se asignan en forma aleatoria a las unidades experimentales.
Diseño de bloques aleatorizado Diseño de experimentos en el que se utiliza la formación de bloques.
Experimento de un solo factor Experimento en el que hay un solo factor con k poblaciones o tratamientos.
Experimento factorial Diseño de experimentos en el que se obtienen conclusiones simultáneas acerca de dos o más factores.
Factor Otro término empleado para la variable independiente de interés.
Formación de bloques Proceso que consiste en usar la misma o similares unidades experimentales para todos los tratamientos. El objetivo de la formación de bloques radica en eliminar una fuente de variación del término del error y con esto proporcionar una prueba más sólida para diferenciar las medias de las poblaciones o de los tratamientos.
Interacción Efecto que se produce cuando los niveles de un factor interaccionan con los niveles de otro factor e influyen en la variable de respuesta.
Partición Proceso de distribución de la suma total de cuadrados y de los grados de libertad en sus diversos componentes.
Procedimientos de comparación múltiple Procedimientos que se emplean para realizar comparaciones estadísticas entre pares de medias poblacionales.
Replicaciones Número de veces que se repite una condición experimental en un experimento.
Tabla ANOVA Tabla usada para resumir los cálculos y los resultados del análisis de varianza. Contiene columnas en las que se muestran las fuentes de variación, las sumas de cuadrados, los grados de libertad, los cuadrados medios y el(los) valor(es) de F.
Tasa de error tipo I por comparación Probabilidad de cometer un error tipo I en la comparación de un solo par.
Tasa de error tipo I por experimentación Probabilidad de cometer un error tipo I en al menos una de varias comparaciones por pares.
Tratamientos Los diferentes niveles de un factor.
Unidades experimentales Los objetos de interés en el experimento.
Variable de respuesta Otro término para denotar variable dependiente de interés.

Fórmulas clave

Diseño completamente aleatorizado

Media muestral del tratamiento j

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j} \tag{13.1}$$

Varianza muestral del tratamiento j

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n_j - 1} \tag{13.2}$$

Media muestral general

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{n_T} \tag{13.3}$$

$$n_T = n_1 + n_2 + \cdots + n_k \tag{13.4}$$

Cuadrado medio debido a los tratamientos

$$\text{CMTR} = \frac{\text{SCTR}}{k-1} \tag{13.7}$$

Suma de cuadrados debido a los tratamientos

$$\text{SCTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 \tag{13.8}$$

Cuadrado medio debido al error

$$\text{CME} = \frac{\text{SCE}}{n_T - k} \tag{13.10}$$

Suma de cuadrados debido al error

$$\text{SCE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 \tag{13.11}$$

Estadístico de prueba para la igualdad de *k* medias poblacionales

$$F = \frac{\text{CMTR}}{\text{CME}} \tag{13.12}$$

Suma total de cuadrados

$$\text{STC} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (x_{ij} - \bar{\bar{x}})^2 \tag{13.13}$$

Partición de la suma de cuadrados

$$\text{STC} = \text{SCTR} + \text{SCE} \tag{13.14}$$

Procedimientos de comparación múltiple

Estadístico de prueba para el procedimiento de la LSD de Fisher

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\text{CME}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \tag{13.16}$$

LSD de Fisher

$$\text{LSD} = t_{\alpha/2}\sqrt{\text{CME}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{13.17}$$

Diseño de bloques aleatorizado

Suma total de cuadrados

$$\text{STC} = \sum_{i=1}^{b}\sum_{j=1}^{k}(x_{ij} - \bar{\bar{x}})^2 \tag{13.22}$$

Suma de cuadrados debido a los tratamientos

$$\text{SCTR} = b\sum_{j=1}^{k}(\bar{x}_{\cdot j} - \bar{\bar{x}})^2 \tag{13.23}$$

Suma de cuadrados debido a los bloques

$$\text{SCBL} = k\sum_{i=1}^{b}(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 \tag{13.24}$$

Suma de cuadrados debido al error

$$\text{SCE} = \text{STC} - \text{SCTR} - \text{SCBL} \tag{13.25}$$

Experimentos factoriales

Suma total de cuadrados

$$\text{STC} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(x_{ijk} - \bar{\bar{x}})^2 \tag{13.27}$$

Suma de cuadrados del factor A

$$\text{SCA} = br\sum_{i=1}^{a}(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 \tag{13.28}$$

Suma de cuadrados del factor B

$$\text{SCB} = ar\sum_{j=1}^{b}(\bar{x}_{\cdot j} - \bar{\bar{x}})^2 \tag{13.29}$$

Suma de cuadrados debido a la interacción

$$\text{SCAB} = r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{x}_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})^2 \tag{13.30}$$

Suma de cuadrados debido al error

$$\text{SCE} = \text{STC} - \text{SCA} - \text{SCB} - \text{SCAB} \tag{13.31}$$

Ejercicios complementarios

34. En un diseño de experimentos completamente aleatorizado se probó la capacidad de absorción de agua de tres marcas de toallas de papel. Se usaron toallas de un mismo tamaño para probar cuatro secciones de toalla por marca. A continuación se proporcionan los datos de la capacidad de absorción. Con un nivel de significancia de 0.05, ¿parece haber alguna diferencia en la capacidad de absorción de estas marcas?

Marca		
x	*y*	*z*
91	99	83
100	96	88
88	94	89
89	99	76

35. En un estudio publicado en el *Journal of Small Business Management* se concluyó que los individuos que se autoemplean no experimentan tanta satisfacción laboral como los que no se autoemplean. En esta investigación, la satisfacción laboral se midió con 18 *items*, cada uno de los cuales se evaluaba con una escala de Likert con 1-5 opciones de respuesta, que iban de totalmente de acuerdo a totalmente en desacuerdo. En esta escala, una puntuación alta corresponde a mayor satisfacción laboral. La suma de las puntuaciones de los 18 *items*, que iban de 18-90, se usó para medir la satisfacción laboral. Suponga que se utiliza este método para medir la satisfacción laboral de abogados (Lawyer), terapeutas físicos (Physical Therapist), diseñadores (Cabinetmaker) y analistas de sistemas (Systems Analyst). A continuación se listan los resultados obtenidos en una muestra de 10 individuos de cada profesión.

WEB archivo
SatisJob

Lawyer	Physical Therapist	Cabinetmaker	Systems Analyst
44	55	54	44
42	78	65	73
74	80	79	71
42	86	69	60
53	60	79	64
50	59	64	66
45	62	59	41
48	52	78	55
64	55	84	76
38	50	60	62

Con $\alpha = 0.05$ como nivel de significancia, pruebe si hay diferencia en la satisfacción laboral entre las cuatro profesiones.

36. La revista *Money* publicó los rendimientos porcentuales y las proporciones de gastos para los fondos de capital y de bonos. Los datos siguientes representan las proporciones de gastos en 10 fondos de capital "midcap", 10 fondos de capital "small-cap", 10 fondos de capital Hybrid (híbrido) y 10 fondos de capital Specialty (especializado) (*Money,* marzo de 2003).

WEB archivo
Funds

Midcap	Small-cap	Hybrid	Specialty
1.2	2.0	2.0	1.6
1.1	1.2	2.7	2.7
1.0	1.7	1.8	2.6
1.2	1.8	1.5	2.5
1.3	1.5	2.5	1.9
1.8	2.3	1.0	1.5
1.4	1.9	0.9	1.6
1.4	1.3	1.9	2.7
1.0	1.2	1.4	2.2
1.4	1.3	0.3	0.7

Use $\alpha = 0.05$ para probar si hay diferencias significativas entre las proporciones de gastos medios de estos cuatro fondos de capital.

37. La Oficina del Censo de Estados Unidos (U.S. Census Bureau) realiza un cálculo trimestral de tasas de viviendas desocupadas y propiedad de vivienda por estado y área estadística metropolitana (MSA, por sus siglas en inglés). Cada MSA tiene al menos un área urbanizada de 50 000 o más habitantes. Los siguientes datos son las tasas (%) de casas desocupadas para renta por MSA en cuatro regiones geográficas de Estados Unidos (Midwest, oeste medio; Northeast, noreste; South, sur, y West, oeste) en el primer trimestre de 2008 (sitio web del U.S. Census Bureau, enero de 2009).

WEB archivo
RentalVacancy

Midwest	Northeast	South	West
16.2	2.7	16.6	7.9
10.1	11.5	8.5	6.6
8.6	6.6	12.1	6.9
12.3	7.9	9.8	5.6
10.0	5.3	9.3	4.3
16.9	10.7	9.1	15.2
16.9	8.6	5.6	5.7
5.4	5.5	9.4	4.0
18.1	12.7	11.6	12.3
11.9	8.3	15.6	3.6
11.0	6.7	18.3	11.0
9.6	14.2	13.4	12.1
7.6	1.7	6.5	8.7
12.9	3.6	11.4	5.0
12.2	11.5	13.1	4.7
13.6	16.3	4.4	3.3
		8.2	3.4
		24.0	5.5
		12.2	
		22.6	
		12.0	
		14.5	
		12.6	
		9.5	
		10.1	

Utilice $\alpha = 0.05$ para probar si la tasa media de viviendas desocupadas es la misma para cada región geográfica.

38. Se proponen tres nuevos métodos de ensamble para un producto nuevo. Para determinar con cuál de los métodos se producen más unidades por hora se elige un diseño de experimentos completamente aleatorizado y a 30 trabajadores seleccionados al azar se les asigna alguno de los métodos (Method) de fabricación. En la tabla siguiente se presenta el número de unidades producidas por cada uno de los trabajadores.

WEB archivo
Assembly

Method		
A	B	C
97	93	99
73	100	94
93	93	87
100	55	66
73	77	59
91	91	75
100	85	84
86	73	72
92	90	88
95	83	86

Utilice estos datos y realice una prueba para ver si el número medio de unidades producidas es la misma en los tres métodos de fabricación. Use $\alpha = 0.05$.

39. En un estudio realizado para investigar la actividad de los clientes en las grandes tiendas, a cada sujeto se le clasificó al inicio como poco activo (Nonbrowser), medianamente activo (Light Browser) y muy activo (Heavy Browser). Se obtuvo un valor para cada cliente que medía cuán a gusto se encontraba en la tienda. Los valores más altos indicaban que el cliente se encontraba más a gusto. Suponga que se obtuvieron los datos siguientes.

WEB archivo
Browsing

Nonbrowser	Light Browser	Heavy Browser
4	5	5
5	6	7
6	5	5
3	4	7
3	7	4
4	4	6
5	6	5
4	5	7

a) Use $\alpha = 0.05$ para probar si hay diferencia en el grado en que se sienten a gusto los tres tipos de clientes.

b) Use el procedimiento de la LSD de Fisher para comparar los grados en que se sienten a gusto los clientes poco activos y los medianamente activos. Use $\alpha = 0.05$ ¿Cuál es su conclusión?

40. Una empresa realiza una investigación para determinar el rendimiento característico, en millas por galón, de tres marcas de gasolina. Como cada gasolina produce rendimientos distintos en automóviles de marcas diferentes, se eligen cinco marcas de vehículos que se tratan como bloques en el experimento; es decir, el automóvil de cada marca se prueba con los tres tipos de gasolina. Los resultados del experimento (en millas por galón) se presentan a continuación.

		Marcas de gasolina		
		I	II	III
Automóviles	A	18	21	20
	B	24	26	27
	C	30	29	34
	D	22	25	24
	E	20	23	24

a) Con $\alpha = 0.05$, ¿se encuentra alguna diferencia significativa entre los rendimientos medios en millas por galón de las tres marcas de gasolina?

b) Analice los datos experimentales usando el ANOVA para diseños completamente aleatorizados. Compare sus hallazgos con los obtenidos en el inciso *a*). ¿Cuál es la ventaja de tratar de eliminar el efecto de bloque?

41. Wegman's Food Markets y Tops Friendly Markets son las cadenas más grandes de tiendas de abarrotes en el área de Rochester, Nueva York. Cuando Wal-Mart abrió uno de sus Supercenter en un suburbio de Rochester, los expertos predijeron que vendería más barato que sus competidores locales. El diario *Democrat and Chronicle* publicó los precios de los 15 artículos (Item) que se listan en la tabla que se presenta a continuación.

WEB archivo
MarketBasket

Item	Tops	Wal-Mart	Wegmans
Bananas (1 lb)	0.49	0.48	0.49
Sopa instantánea (10.75 oz)	0.60	0.54	0.77
Pechuga de pollo (3 lb)	10.47	8.61	8.07
Pasta dental (6.2 oz)	1.99	2.40	1.97
Huevos grandes (1 docena)	1.59	0.88	0.79
Salsa catsup (36 oz)	2.59	1.78	2.59
Gelatina (3 oz)	0.67	0.42	0.65
Crema de cacahuate (18 oz)	2.29	1.78	2.09
Leche (descremada, 1/2 gal)	1.34	1.24	1.34
Salchichas (1 lb)	3.29	1.50	3.39
Salsa para pasta (1 lb, 10 oz)	2.09	1.50	1.25
Galletas (1 lb)	3.29	2.00	3.39
Detergente (líquido, 100 oz)	6.79	5.24	5.99
Jugo de naranja (1/2 gal)	2.50	2.50	2.50
Caramelos (1 lb)	1.19	1.27	1.69

Con $\alpha = 0.05$ como nivel de significancia, pruebe si hay una diferencia significativa en el precio medio de estos 15 artículos entre las tres tiendas.

42. El U.S. Department of Housing and Urban Development publica datos que reflejan el mercado de rentas mensuales en las áreas metropolitanas. Los datos siguientes representan los precios de alquiler ($) mensuales aceptables en cinco zonas metropolitanas para departamentos de 1, 2 y 3 recámaras (*The New York Times Almanac,* 2006).

	Boston	Miami	San Diego	San José	Washington
1 recámara	1 077	775	975	1 107	1 045
2 recámaras	1 266	929	1 183	1 313	1 187
3 recámaras	1 513	1 204	1 725	1 889	1 537

Utilice 0.05 como nivel de significancia y pruebe si las medias de los precios de alquiler mensuales aceptables son iguales en estas cinco zonas metropolitanas.

43. Para identificar si hay alguna diferencia significativa en la rapidez de dos sistemas de software para traducir el inglés a otros idiomas, se diseñó un experimento factorial. Como el idioma al que se traduzca es también un factor importante, los dos sistemas se prueban traduciendo tres: español, francés y alemán. Utilice los datos siguientes del tiempo en horas necesario para efectuar esta tarea.

	Idioma		
	Español	**Francés**	**Alemán**
Sistema 1	8 12	10 14	12 16
Sistema 2	6 10	14 16	16 22

Realice una prueba para determinar si hay alguna diferencia significativa debido al idioma al que se traduce y si hay algún efecto de interacción. Use $\alpha = 0.05$.

44. En una fábrica se diseña un experimento factorial para determinar si hay diferencia entre el número de partes defectuosas producidas por dos máquinas y si el número de defectos depende también de si a estas máquinas se les suministra la materia prima manualmente o mediante un

sistema de alimentación automático. A continuación se presentan los datos del número de partes defectuosas producidas. Use $\alpha = 0.05$ para probar si hay algún efecto significativo debido a la máquina, al sistema de suministro de la materia prima y a la interacción.

	Suministro de la materia prima	
	Manual	**Automático**
Máquina 1	30 34	30 26
Máquina 2	20 22	24 28

Caso a resolver 1 Wentworth Medical Center

Como parte de un estudio a largo plazo de individuos de 65 años de edad o más, sociólogos y médicos del Wentworth Medical Center en Nueva York investigaron la relación entre ubicación geográfica y depresión. Se tomó una muestra de 60 personas, todas en buenas condiciones de salud; 20 eran residentes de Florida, 20 de Nueva York y 20 de Carolina del Norte. A cada individuo de la muestra se le aplicó una prueba estandarizada para medir la depresión. Los datos obtenidos se listan a continuación; las puntuaciones más altas indican niveles más acentuados de depresión. Esos datos se encuentran en el archivo Medical1.

En la segunda parte del estudio se consideró la relación entre ubicación geográfica y depresión en individuos de 65 años de edad o más con algún padecimiento crónico como artritis, hipertensión y/o cardiopatía. Se identificó una muestra de 60 individuos en estas condiciones. De nuevo, 20 eran residentes de Florida, 20 de Nueva York y 20 de Carolina del Norte. Los niveles de depresión registrados en este estudio se presentan a continuación. Estos datos se encuentran en el archivo Medical2.

WEB archivo
Medical1

WEB archivo
Medical2

Data from Medical1			Data from Medical2		
Florida	**New York**	**North Carolina**	**Florida**	**New York**	**North Carolina**
3	8	10	13	14	10
7	11	7	12	9	12
7	9	3	17	15	15
3	7	5	17	12	18
8	8	11	20	16	12
8	7	8	21	24	14
8	8	4	16	18	17
5	4	3	14	14	8
5	13	7	13	15	14
2	10	8	17	17	16
6	6	8	12	20	18
2	8	7	9	11	17
6	12	3	12	23	19
6	8	9	15	19	15
9	6	8	16	17	13
7	8	12	15	14	14
5	5	6	13	9	11
4	7	3	10	14	12
7	7	8	11	13	13
3	8	11	17	11	11

Informe gerencial

1. Use estadística descriptiva para resumir los datos de los dos estudios. ¿Cuáles son sus observaciones preliminares acerca de los valores de depresión?
2. Utilice el análisis de varianza para ambos conjuntos de datos. En cada caso proporcione las hipótesis a probar. ¿Cuáles son sus conclusiones?
3. Si es necesario use inferencias acerca de las medias de cada uno de los tratamientos. ¿Cuáles son sus conclusiones?

Caso a resolver 2 Compensación para profesionales de ventas

Suponga que un grupo local de profesionales de ventas de San Francisco realiza encuestas entre sus miembros para identificar si hay alguna relación entre los años de experiencia (Experience) y el sueldo (Salary) de los individuos empleados como vendedores en posición (Position) de internos (Inside) y externos (Outside). En esta encuesta se solicita a los sujetos de estudio que especifiquen uno de tres niveles de años de experiencia: baja/Low (1 a 10 años), media/Mediun (l1 a 20 años) o alta/High (21 o más años). A continuación se presenta una parte de los datos obtenidos. El conjunto completo, que contiene 120 observaciones (Observation), se encuentra en el archivo SalesSalary.

WEB archivo
SalesSalary

Observation	Salary $	Position	Experience
1	53 938	Inside	Medium
2	52 694	Inside	Medium
3	70 515	Outside	Low
4	52 031	Inside	Medium
5	62 283	Outside	Low
6	57 718	Inside	Low
7	79 081	Outside	High
8	48 621	Inside	Low
9	72 835	Outside	High
10	54 768	Inside	Medium
.	.	.	.
.	.	.	.
.	.	.	.
115	58 080	Inside	High
116	78 702	Outside	Medium
117	83 131	Outside	Medium
118	57 788	Inside	High
119	53 070	Inside	Medium
120	60 259	Outside	Low

Informe gerencial

1. Use la estadística descriptiva para resumir los datos.
2. Desarrolle una estimación con un intervalo de 95% de confianza del sueldo medio anual de todos los vendedores, sin importar los años de experiencia y el tipo de vendedor.
3. Calcule una estimación con un intervalo de 95% de confianza del sueldo medio anual de los vendedores internos.
4. Desarrolle una estimación con un intervalo de 95% de confianza del sueldo medio anual de los vendedores externos.
5. Utilice el análisis de varianza para determinar si hay diferencias significativas debido a la posición. Use 0.05 como nivel de significancia y, por ahora, ignore el efecto de los años de experiencia.

6. Use el análisis de varianza para determinar si hay diferencias significativas debido a los años de experiencia. Use 0.05 como nivel de significancia y, por ahora, ignore el efecto de la posición.
7. Con 0.05 como nivel de significancia, realice una prueba para determinar si hay diferencias significativas debido a la posición, a los años de experiencia o a la interacción.

Apéndice 13.1 Análisis de varianza con Minitab

Diseño completamente aleatorizado

En la sección 13.2 se estudió el uso del análisis de varianza para probar la igualdad de k medias poblacionales con los datos de un diseño completamente aleatorizado. Para ilustrar el uso de Minitab en este tipo de diseño de experimentos, se muestra cómo probar si son iguales las medias del número de unidades producidas semanalmente con cada uno de los métodos del experimento de Chemitech presentado en la sección 13.1. Los datos muestrales se han ingresado en las tres primeras columnas de la hoja de cálculo de Minitab. La columna 1 se etiquetó como A, la 2 se etiquetó como B y la 3 como C. Mediante los pasos siguientes se obtiene la pantalla de Minitab presentada en la figura 13.5.

WEB archivo
Chemitech

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **ANOVA**.
Paso 3. Escoja **One-way (Unstacked)**.
Paso 4. Cuando el cuadro de diálogo One-way Analysis of Variance aparezca:
Ingrese C1-C3 en el cuadro **Responses (in separate columns)**.
Haga clic en **OK**.

Diseño de bloques aleatorizado

En la sección 13.4 se trabajó el análisis de varianza para probar la igualdad de k medias poblacionales con los datos de un diseño de bloques aleatorizado. Para ilustrar el uso de Minitab en este tipo de diseño de experimentos, se muestra cómo probar si las medias de los niveles de estrés de los controladores de tráfico aéreo son las mismas en las tres estaciones de trabajo utilizando los datos de la tabla 13.5. Los bloques (controladores), los tratamientos (sistema) y las puntuaciones de los niveles de estrés mostrados en la tabla 13.5 se han ingresado en las columnas C1, C2 y C3, respectivamente, de una hoja de trabajo de Minitab. Con los siguientes pasos se obtiene la pantalla que corresponde a la tabla de ANOVA mostrada en la tabla 13.8.

WEB archivo
AirTraffic

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **ANOVA**.
Paso 3. Escoja **Two-way**.
Paso 4. Cuando el cuadro de diálogo Two-way Analysis of Variance aparezca:
Ingrese C3 en el cuadro **Response**.
Introduzca C2 en el cuadro **Row factor**.
Ingrese C1 en el cuadro **Column factor**.
Seleccione **Fit Additive Model**.
Haga clic en **OK**.

Los tratamientos se introducen en el cuadro Row factor y los bloques en el cuadro Column factor.

Experimento factorial

En la sección 13.5 se estudió el uso del análisis de varianza para probar la igualdad de k medias poblacionales con los datos de un experimento factorial. Para ilustrar el uso de Minitab en este tipo de diseño de experimentos, se muestra cómo analizar los datos del experimento de dos factores del GMAT presentado en esta sección. Las puntuaciones obtenidas en el GMAT e incluidas

WEB archivo
GMATStudy

en la tabla 13.11 se han ingresado en la columna 1 de una hoja de trabajo de Minitab; la columna 1 se ha etiquetado como Score, la 2 como Program y la 3 como College. Con los siguientes pasos se obtiene la pantalla de Minitab que corresponde a la tabla de ANOVA presentada en la fi-gura 13.6.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **ANOVA**.
Paso 3. Seleccione **Two-way**.
Paso 4. Cuando el cuadro de diálogo Two-way Analysis of Variance aparezca:
Ingrese C1 en el cuadro **Response**.
Introduzca C2 en el cuadro **Row factor**.
Ingrese C3 en el cuadro **Column factor**.
Haga clic en **OK**.

Apéndice 13.2 Análisis de varianza con Excel

Diseño completamente aleatorizado

En la sección 13.2 se abordó el análisis de varianza para probar la igualdad de k medias poblacionales con los datos de un diseño completamente aleatorizado. Para ilustrar cómo utilizar Excel para probar la igualdad de k medias poblacionales en este tipo de diseño de experimentos, se muestra cómo probar si las medias del número de unidades producidas semanalmente son iguales para cada uno de los métodos de ensamble del experimento de Chemitech presentado en la sección 13.1. Los datos muestrales se han ingresado en las filas 2 a 6 de las columnas A, B y C de la hoja de cálculo como se observa en la figura 13.7. Los pasos siguientes se utilizan para obtener el resultado que se indica en las celdas A8:G22; la parte de la ANOVA de este resultado corresponde a la tabla ANOVA presentada en la tabla 13.3.

WEB archivo
Chemitech

Paso 1. Haga clic en la ficha **Data** en la cinta.
Paso 2. En el grupo **Analysis**, haga clic en **Data Analysis**.
Paso 3. Elija **Anova: Single Factor** de la lista de Analysis Tools.
Haga clic en **OK**.
Paso 4. Cuando el cuadro de diálogo Anova: Single Factor aparezca:
Ingrese A1:C6 en el cuadro **Input Range**.
Seleccione **Columns**.
Elija **Labels in First Row**.
Seleccione **Output Range** e ingrese A8 en el cuadro.
Haga clic en **OK**.

Diseño de bloques aleatorizado

En la sección 13.4 se trabajó el análisis de varianza para probar la igualdad de k medias poblacionales con los datos de un diseño de bloques aleatorizado. Para ilustrar el uso de Excel en este tipo de diseño de experimentos, se muestra cómo probar si las medias de los niveles de estrés de los controladores de tráfico aéreo son las mismas en las tres estaciones de trabajo. Los niveles de estrés que se muestran en la tabla 13.5 se han ingresado en los filas 2 a 7 de las columnas B, C y D de la hoja de cálculo, como se ilustra en la figura 13.8. Las celdas de las filas 2 a 7 de la columna A contienen el número de cada controlador (1, 2, 3, 4, 5 y 6). Los pasos siguientes producen el resultado de Excel que se observa en las celdas A9:G30. La parte de la ANOVA de este resultado corresponde al resultado de Minitab presentado en la tabla 13.8.

WEB archivo
AirTraffic

Paso 1. Haga clic en la ficha **Data** en la cinta.
Paso 2. En el grupo **Analysis**, haga clic en **Data Analysis**.

FIGURA 13.7 Solución de Excel para el experimento de Chemitech

	A	B	C	D	E	F	G	H
1	**Method A**	**Method B**	**Method C**					
2	58	58	48					
3	64	69	57					
4	55	71	59					
5	66	64	47					
6	67	68	49					
7								
8	Anova: Single Factor							
9								
10	SUMMARY							
11	*Groups*	*Count*	*Sum*	*Average*	*Variance*			
12	Method A	5	310	62	27.5			
13	Method B	5	330	66	26.5			
14	Method C	5	260	52	31			
15								
16								
17	ANOVA							
18	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*	
19	Between Groups	520	2	260	9.1765	0.0038	3.8853	
20	Within Groups	340	12	28.3333				
21								
22	Total	860	14					
23								
24								

Paso 3. Elija **Anova: Two-Factor Without Replication** de la lista Analysis Tools.
Haga clic en **OK**.

Paso 4. Cuando el cuadro de diálogo Anova: Two-Factor Without Replication aparezca:
Ingrese A1:D7 en el cuadro **Input Range**.
Seleccione **Labels**.
Elija **Output Range** e ingrese A9 en el cuadro.
Haga clic en **OK**.

Experimento factorial

En la sección 13.5 se estudió el uso del análisis de varianza para probar la igualdad de k medias poblacionales con los datos de un experimento factorial. Para ilustrar el uso de Excel en este tipo de diseño de experimentos, se muestra cómo analizar los datos del experimento de dos factores para el GMAT presentado en esa sección. Las puntuaciones obtenidas en el GMAT, que se presentan en la tabla 13.10, se han ingresado en las filas 2 a 7 de las columnas B, C y D de la hoja de cálculo como se observa en la figura 13.9. Con los pasos siguientes se obtienen los resultados que se observan en las celdas A9:G44; la parte del ANOVA corresponde a la pantalla de Minitab de la figura 13.6.

WEB archivo
GMATStudy

Paso 1. Haga clic en la ficha **Data** en la cinta.

Paso 2. En el grupo **Analysis**, haga clic en **Data Analysis**.

Paso 3. Elija **Anova: Two-Factor Without Replication** de la lista Analysis Tools.
Haga clic en **OK**.

Paso 4. Cuando el cuadro de diálogo Anova: Two-Factor Without Replication aparezca:
Ingrese A1:D7 en el cuadro **Input Range**.
Ingrese 2 en el cuadro **Rows per sample**.

FIGURA 13.8 Solución de Excel para la prueba de estrés de los controladores de tráfico aéreo

	A	B	C	D	E	F	G	H
1	**Controller**	**System A**	**System B**	**System C**				
2	1	15	15	18				
3	2	14	14	14				
4	3	10	11	15				
5	4	13	12	17				
6	5	16	13	16				
7	6	13	13	13				
8								
9	Anova: Two-Factor Without Replication							
10								
11	*SUMMARY*	*Count*	*Sum*	*Average*	*Variance*			
12	1	3	48	16	3			
13	2	3	42	14	0			
14	3	3	36	12	7			
15	4	3	42	14	7			
16	5	3	45	15	3			
17	6	3	39	13	0			
18								
19	System A	6	81	13.5	4.3			
20	System B	6	78	13	2			
21	System C	6	93	15.5	3.5			
22								
23								
24	ANOVA							
25	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*	
26	Rows	30	5	6	3.16	0.0574	3.33	
27	Columns	21	2	10.5	5.53	0.0242	4.10	
28	Error	19	10	1.9				
29								
30	Total	70	17					
31								

Seleccione **Output Range** e ingrese A9 en el cuadro.
Haga clic en **OK**.

Apéndice 13.3 Análisis de un diseño completamente aleatorizado usando StatTools

En este apéndice se muestra el uso de StatTools para probar la igualdad de k medias poblacionales en un diseño completamente aleatorizado. Para ilustrarlo se utilizan los datos de Chemitech de la tabla 13.1. Se inicia con el uso del Data Set Manager a efecto de crear un conjunto de datos de StatTools para los datos utilizados en el procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes se utilizan para probar la igualdad de las tres medias poblacionales.

WEB archivo
Chemitech

Paso 1. Haga clic en la ficha **StatTools** en la cinta.
Paso 2. En el grupo **Analysis**, haga clic en **Statistical Inference**.
Paso 3. Elija la opción **One-way ANOVA**.

FIGURA 13.9 Solución de Excel para el experimento de dos factores del GMAT

	A	B	C	D	E	F	G	H
1		**Business**	**Engineering**	**Arts and Sciences**				
2	**3-hour review**	500	540	480				
3		580	460	400				
4	**1-day program**	460	560	420				
5		540	620	480				
6	**10-week course**	560	600	480				
7		600	580	410				
8								
9	Anova: Two-Factor With Replication							
10								
11	SUMMARY	Business	Engineering	Arts and Sciences	Total			
12	*3-hour review*							
13	Count	2	2	2	6			
14	Sum	1 080	1 000	880	2 960			
15	Average	540	500	440	493.33333			
16	Variance	3 200	3 200	3 200	3 946.6667			
17								
18	*1-day program*							
19	Count	2	2	2	6			
20	Sum	1 000	1 180	900	3 080			
21	Average	500	590	450	513.33333			
22	Variance	3 200	1 800	1 800	5 386.6667			
23								
24	*10-week course*							
25	Count	2	2	2	6			
26	Sum	1 160	1 180	890	3 230			
27	Average	580	590	445	538.33333			
28	Variance	800	200	2 450	5 936.6667			
29								
30	*Total*							
31	Count	6	6	6				
32	Sum	3 240	3 360	2 670				
33	Average	540	560	445				
34	Variance	2 720	3 200	1 510				
35								
36								
37	ANOVA							
38	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*	
39	Sample	6 100	2	3 050	1.38	0.2994	4.26	
40	Columns	45 300	2	22 650	10.27	0.0048	4.26	
41	Interaction	11 200	4	2 800	1.27	0.3503	3.63	
42	Within	19 850	9	2205.5556				
43								
44	Total	82 450	17					
45								

Paso 4. Cuando aparezca el cuadro de diálogo StatTools-One-way ANOVA:

En la sección **Variables:**

Haga clic en el botón **Format** y seleccione **Unstacked**.

Seleccione **Method A**.

Elija **Method B**.

Seleccione **Method C**.

Elija 95% en el cuadro **Confidence Level**.

Haga clic en **OK**.

Observe que en el paso 4 se seleccionó la opción Unstacked después de hacer clic en el botón Format. Esta opción significa que los datos de los tres tratamientos aparecen en columnas separadas de la hoja de cálculo. Sólo se podrían utilizar dos columnas en el formato apilado. Por ejemplo, los datos se organizarían de la siguiente forma.

	A	B	C
1	**Method A**	**Units Produced**	
2	Method A	58	
3	Method A	64	
4	Method A	55	
5	Method A	66	
6	Method A	67	
7	Method B	58	
8	Method B	69	
9	Method B	71	
10	Method B	64	
11	Method B	68	
12	Method C	48	
13	Method C	57	
14	Method C	59	
15	Method C	47	
16	Method C	49	
17			

Con frecuencia los datos son registrados en un formato apilado. Para apilar los datos, simplemente seleccione la opción Stacked después de hacer clic en el botón Format.

CAPÍTULO 14

Regresión lineal simple

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: ALLIANCE DATA SYSTEMS

14.1 MODELO DE REGRESIÓN LINEAL SIMPLE
- Modelo de regresión y ecuación de regresión
- Ecuación de regresión estimada

14.2 MÉTODO DE MÍNIMOS CUADRADOS

14.3 COEFICIENTE DE DETERMINACIÓN
- Coeficiente de correlación

14.4 SUPUESTOS DEL MODELO

14.5 PRUEBA DE SIGNIFICANCIA
- Estimación de σ^2
- Prueba t
- Intervalo de confianza para β_1
- Prueba F
- Algunas advertencias acerca de la interpretación de las pruebas de significancia

14.6 USO DE LA ECUACIÓN DE REGRESIÓN ESTIMADA PARA ESTIMACIÓN Y PREDICCIÓN
- Estimación puntual
- Estimación por intervalo
- Intervalo de confianza para el valor medio de y
- Intervalo de predicción para un solo valor de y

14.7 SOLUCIÓN POR COMPUTADORA

14.8 ANÁLISIS DE RESIDUALES: CONFIRMACIÓN DE LOS SUPUESTOS DEL MODELO
- Gráfica de residuales contra x
- Gráfica de residuales contra $\hat{y}$
- Residuales estandarizados
- Gráfica de probabilidad normal

14.9 ANÁLISIS DE RESIDUALES: OBSERVACIONES ATÍPICAS Y OBSERVACIONES INFLUYENTES
- Detección de observaciones atípicas
- Detección de observaciones influyentes

ESTADÍSTICA *en* LA PRÁCTICA

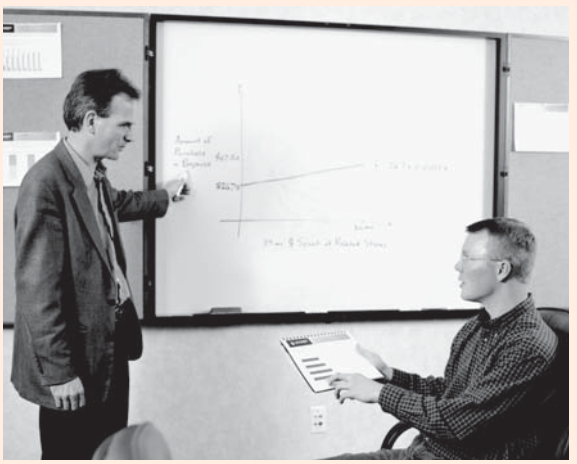
ALLIANCE DATA SYSTEMS*

DALLAS, TEXAS

Alliance Data Systems (ADS) ofrece procesamiento de transacciones y servicios de crédito y marketing en la creciente industria de la administración de relaciones con el cliente (CRM, por sus siglas en inglés). Los clientes de ADS están concentrados en cuatro industrias: ventas al detalle, derivados del petróleo/tiendas de conveniencia, servicios y transporte. En 1983, Alliance empezó ofreciendo servicios integrales de tramitación de crédito para los sectores minorista, de derivados del petróleo y de restaurantes de categoría media; actualmente emplea a más de 6 500 personas que proporcionan servicios a clientes en todo el mundo. Sólo en Estados Unidos, ADS opera más de 140 000 terminales de punto de venta y procesa más de 2 500 millones de transacciones al año. En este país es la segunda empresa en servicios de crédito de marca propia, pues maneja 49 programas con casi 72 millones de tarjetahabientes. En 2001 hizo una oferta pública inicial y ahora cotiza en la bolsa de Nueva York.

Como parte de sus servicios de marketing, ADS diseña campañas y promociones directas por correo. La empresa posee una base de datos con información sobre los hábitos de consumo de más de 100 millones de consumidores, lo que le permite dirigir sus acciones a quienes tienen la mayor probabilidad de beneficiarse de la publicidad por correo directo. El grupo de desarrollo analítico de ADS emplea el análisis de regresión para construir modelos que miden y predicen la receptividad del consumidor a las campañas de marketing directo. Algunos modelos de regresión predicen la probabilidad de compra de las personas que reciben la publicidad y otros pronostican la cantidad que gastarán cuando realicen una compra.

En una determinada campaña, una cadena de tiendas minoristas deseaba atraer a nuevos clientes. Para predecir el efecto de la campaña, los analistas de ADS tomaron de la base de datos una muestra de consumidores, les enviaron material promocional y después recabaron datos sobre su respuesta. Los datos se referían al monto de la compra realizada por los sujetos que respondieron a la campaña, así como a diversas variables específicas del consumidor que se consideraron útiles para predecir las ventas. La variable que más contribuyó a predecir el monto de compra fue la cantidad total de adquisiciones a crédito realizadas en tiendas semejantes en los últimos 39 meses. Los analistas de ADS obtuvieron una ecuación de regresión estimada en la que se relacionaba el monto de compra con la cantidad gastada en tiendas semejantes:



Los expertos de Alliance Data Systems analizan el uso de un modelo de regresión para predecir las ventas de una campaña de marketing directo. © Cortesía de Alliance Data Systems.

$$\hat{y} = 26.7 + 0.00205x$$

donde

$\hat{y}$ = monto de la compra
x = monto gastado en tiendas similares

Con esta ecuación pudieron predecir que una persona que hubiera gastado $10 000 en tiendas semejantes en los últimos 39 meses, gastaría $47.20 como respuesta a la publicidad por correo directo. En este capítulo se verá cómo obtener estas ecuaciones de regresión estimada.

El modelo final que obtuvieron los analistas de ADS también incluía otras variables que incrementaban el poder predictivo de la ecuación anterior, entre ellas la existencia o no de una tarjeta de crédito bancaria, el ingreso estimado y la cantidad promedio gastada en cada visita a la tienda seleccionada. En el capítulo siguiente se verá cómo incorporar estas variables adicionales a un modelo de regresión múltiple.

* Los autores agradecen a Philip Clemance, director de Desarrollo Analítico de Alliance Data Systems por haber proporcionado este material para *Estadística en la práctica.*

Con frecuencia las decisiones gerenciales se basan en la relación entre dos o más variables. Por ejemplo, al analizar la relación entre el gasto en publicidad y las ventas, un gerente de marketing puede tratar de predecir las ventas correspondientes a un determinado gasto en publicidad. En otro caso, una empresa de servicios públicos establece la relación entre la temperatura diaria y la demanda de electricidad para predecir la necesidad de fluido eléctrico considerando las temperaturas diarias que se esperan para el mes siguiente. Algunas veces los directivos se apoyan en la intuición para juzgar la relación entre dos variables. Sin embargo, cuando los datos están disponibles, puede emplearse un procedimiento estadístico llamado *análisis de regresión* para obtener una ecuación que indique cuál es la relación entre las variables.

Sir Francis Galton (1822-1911) fue el primero en emplear los métodos estadísticos para estudiar la relación entre dos variables. Galton estaba interesado en estudiar la relación entre las estaturas de padre e hijo. Karl Pearson (1857-1936), discípulo de Galton, analizó esta relación de estatura de padre e hijo en 1 078 parejas de individuos.

En la terminología que se emplea en la regresión, la variable a predecir se llama **variable dependiente**, y a la variable o variables que se usan para predecir su valor se les llama **variables independientes**. Por ejemplo, al analizar el efecto de los gastos en publicidad sobre las ventas, como lo que busca el gerente de marketing es predecir estas últimas, las ventas serán la variable dependiente.

En este capítulo se estudia el tipo más sencillo de análisis de regresión en el que interviene una variable independiente y una variable dependiente donde la relación entre estas variables se aproxima mediante una línea recta. A este tipo de análisis se le conoce como **regresión lineal simple**. Al análisis en el que intervienen dos o más variables independientes se le llama *análisis de regresión múltiple*; éste y los casos en los que la relación es curvilínea se estudian en los capítulos 15 y 16.

14.1 Modelo de regresión lineal simple

Armand's Pizza Parlors es una cadena de restaurantes de comida italiana que abarca un área de cinco estados. Las ubicaciones con mayor éxito se encuentran cerca de los campus universitarios. Los gerentes creen que las ventas trimestrales de estos restaurantes (denotadas por y) están directamente relacionadas con el tamaño de la población estudiantil (denotada por x); es decir, en los establecimientos que están cerca de algún campus con una población estudiantil grande se generan más ventas que en aquellos situados cerca de algún campus con una población estudiantil pequeña. Empleando el análisis de regresión, se puede desarrollar una ecuación que muestre cuál es la relación de la variable dependiente y con la variable independiente x.

Modelo de regresión y ecuación de regresión

En el ejemplo de Armand's Pizza Parlors, la población consta de todos los restaurantes de esta cadena. Para cada restaurante en la población hay un valor x (población estudiantil) y un valor correspondiente y (ventas trimestrales). La ecuación que describe cómo se relaciona y con x, y se da un término para el error, se llama **modelo de regresión**. El siguiente es el modelo que se emplea en la regresión lineal simple.

MODELO DE REGRESIÓN LINEAL SIMPLE

$$y = \beta_0 + \beta_1 x + \epsilon \qquad (14.1)$$

β_0 y β_1 se conocen como *parámetros del modelo*, y ϵ (la letra griega *épsilon*) es una variable aleatoria denominada *término del error*. Este último da cuenta de la variabilidad de y, que no puede ser explicada por la relación lineal entre x y y.

La población de los restaurantes Armand's puede verse también como un conjunto de subpoblaciones, una para cada uno de los valores de x. Por ejemplo, una subpoblación está formada por todos los restaurantes Armand's localizados cerca de los campus universitarios con 8 000 estudiantes; otra subpoblación consta de todos los restaurantes Armand's localizados cerca de los campus universitarios con 9 000 estudiantes, y así sucesivamente. Para cada subpoblación hay una distribución de valores y. Así, hay una distribución de valores y que corresponde a los restaurantes localizados cerca de los campus con 8 000 estudiantes y hay otra para los restaurantes ubicados cerca de los campus con 9 000 estudiantes, y así sucesivamente. Cada una de estas distribuciones tiene su propia media o valor esperado. A la ecuación que describe la relación entre el valor esperado de y, que se denota $E(y)$, y x se le llama **ecuación de regresión**. La siguiente es la ecuación de regresión para la regresión lineal simple.

ECUACIÓN DE REGRESIÓN LINEAL SIMPLE

$$E(y) = \beta_0 + \beta_1 x \qquad (14.2)$$

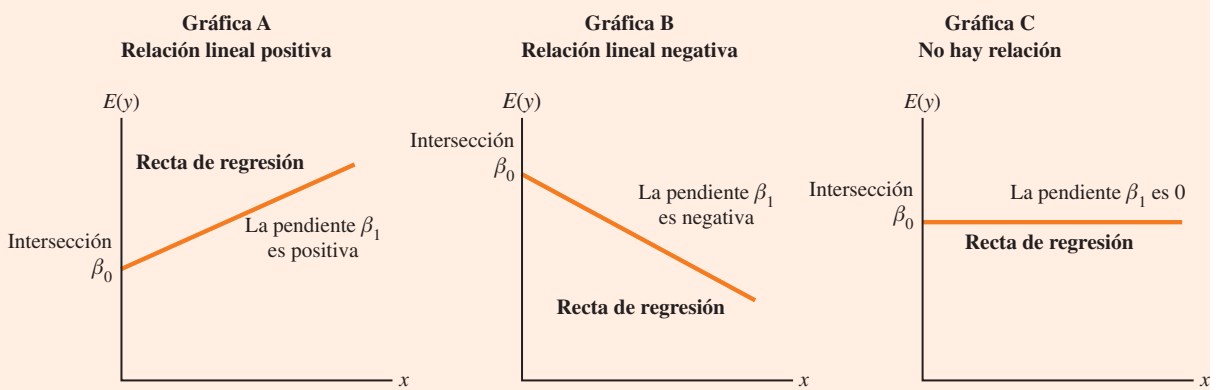
La gráfica de la ecuación de regresión lineal simple es una recta; β_0 es la intersección de la recta de regresión con el eje y, β_1 es la pendiente y $E(y)$ es la media o valor esperado de y para un valor dado de x.

En la figura 14.1 se presentan ejemplos de posibles rectas de regresión. La de la gráfica A indica que el valor medio de y está relacionado positivamente con x, con los valores mayores de $E(y)$ asociados a valores mayores de x. La recta de regresión de la gráfica B indica que el valor medio de y está relacionado negativamente con x, con valores menores de $E(y)$ que corresponden a valores mayores de x. La gráfica C muestra el caso en el que el valor medio de y no está relacionado con x; es decir, el valor medio de y es el mismo para todos los valores de x.

Ecuación de regresión estimada

Si se conocieran los valores de los parámetros poblacionales β_0 y β_1, se podría emplear la ecuación (14.2) para calcular el valor medio de y para un valor dado de x. Sin embargo, en la práctica no se conocen los valores de estos parámetros y es necesario estimarlos usando datos muestrales. Se calculan estadísticos muestrales (que se denotan como b_0 y b_1) como estimaciones de los parámetros poblacionales β_0 y β_1. Al sustituir b_0 y b_1 por β_0 y β_1 en la ecuación de

FIGURA 14.1 Ejemplos de líneas de regresión posibles en la regresión lineal simple



regresión, se obtiene la **ecuación de regresión estimada**. La ecuación de regresión estimada de una regresión lineal simple se da a continuación.

ECUACIÓN DE REGRESIÓN LINEAL SIMPLE ESTIMADA

$$\hat{y} = b_0 + b_1 x \qquad (14.3)$$

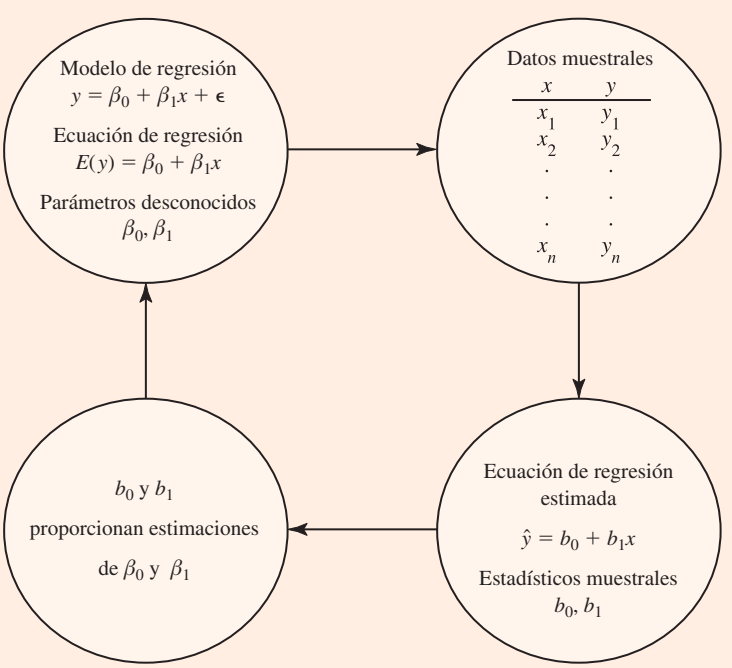
A la gráfica de la ecuación de regresión lineal simple estimada se le llama *recta de regresión estimada;* b_0 es la intersección con el eje y, y b_1 es la pendiente. En la sección siguiente se muestra el uso del método de mínimos cuadrados para calcular los valores de b_0 y b_1 en la ecuación de regresión estimada.

En general, $\hat{y}$ es el estimador puntual de $E(y)$, el valor medio de las y para un valor dado de x. Por tanto, para estimar la media o el valor esperado de las ventas trimestrales de todos los restaurantes situados cerca de los campus con 10 000 estudiantes, Armand's tendrá que sustituir x por 10 000 en la ecuación (14.3). Sin embargo, en algunos casos a la cadena le interesará predecir las ventas de un determinado restaurante. Por ejemplo, suponga que desea pronosticar las ventas trimestrales del que se encuentra cerca de Talbot College, una escuela con 10 000 estudiantes. Resulta que la mejor estimación de la y que corresponde a un determinado valor de x es también la proporcionada por $\hat{y}$. Por tanto, para predecir las ventas trimestrales del restaurante en cuestión, Armand's también sustituirá la x de la ecuación (14.3) por 10 000.

Como el valor de $\hat{y}$ proporciona tanto una estimación puntual de $E(y)$ para un valor dado de x, como una estimación puntual de un solo valor de y para un valor dado de x, a $\hat{y}$ se le llamará simplemente *valor estimado de y.* En la figura 14.2 se presenta en forma resumida el proceso de estimación en la regresión lineal simple.

La estimación de β_0 y β_1 es un proceso estadístico muy parecido a la estimación de μ estudiada en el capítulo 7. Tanto β_0 como β_1 son los parámetros de interés desconocidos, y b_0 y b_1 son los estadísticos muestrales que se usan para estimar los parámetros.

FIGURA 14.2 Proceso de estimación en la regresión lineal simple



NOTAS Y COMENTARIOS

1. El análisis de regresión no puede entenderse como un procedimiento para establecer una relación de causa y efecto entre las variables. Sólo indica cómo o en qué medida las variables están relacionadas una con otra. Cualesquiera conclusiones acerca de una relación causa y efecto deben basarse en los conocimientos de los especialistas en la aplicación de que se trate.
2. La ecuación de regresión en la regresión lineal simple es $E(y) = \beta_0 + \beta_1 x$. En libros más avanzados sobre análisis de regresión suele escribirse la ecuación como $E(y|x) = \beta_0 + \beta_1 x$, enfatizando así que lo que proporciona esta expresión es el valor medio de las y para un valor dado de x.

14.2 Método de mínimos cuadrados

En la regresión lineal simple cada observación consta de dos valores: uno para la variable independiente y otro para la variable dependiente.

El **método de mínimos cuadrados** es un procedimiento en el que se usan los datos muestrales para encontrar la ecuación de regresión estimada. Para ilustrar este método, suponga que se recolectan datos de una muestra de 10 restaurantes Armand's Pizza Parlors ubicados todos cerca de campus universitarios. Para la $i$ésima observación o restaurante en la muestra, x_i es el tamaño de la población de estudiantes (en miles) en el campus, y y_i son las ventas trimestrales (en miles de dólares). En la tabla 14.1 se presentan los valores de x_i y y_i en esta muestra de 10 restaurantes. Como se puede ver, el restaurante 1, con $x_1 = 2$ y $y_i = 58$, está cerca de un campus con 2 000 estudiantes y sus ventas trimestrales son de $58 000. El restaurante 2, con $x_2 = 6$ y $y_2 = 105$, está cerca de un campus con 6 000 estudiantes y sus ventas trimestrales son de $105 000. El valor mayor corresponde a las ventas del restaurante 10, que está cerca de un campus con 26 000 estudiantes y sus ventas trimestrales son de $202 000.

La figura 14.3 es el diagrama de dispersión de los datos de la tabla 14.1. La población de estudiantes (Student Population) se indica en el eje horizontal (en miles) y las ventas trimestrales (Quarterly Sales) en el eje vertical (en miles de $). Los **diagramas de dispersión** para el análisis de regresión se trazan colocando la variable indepediente x en el eje horizontal y la variable dependiente y en el eje vertical. Este diagrama permite observar gráficamente los datos y obtener conclusiones acerca de la relación entre las variables.

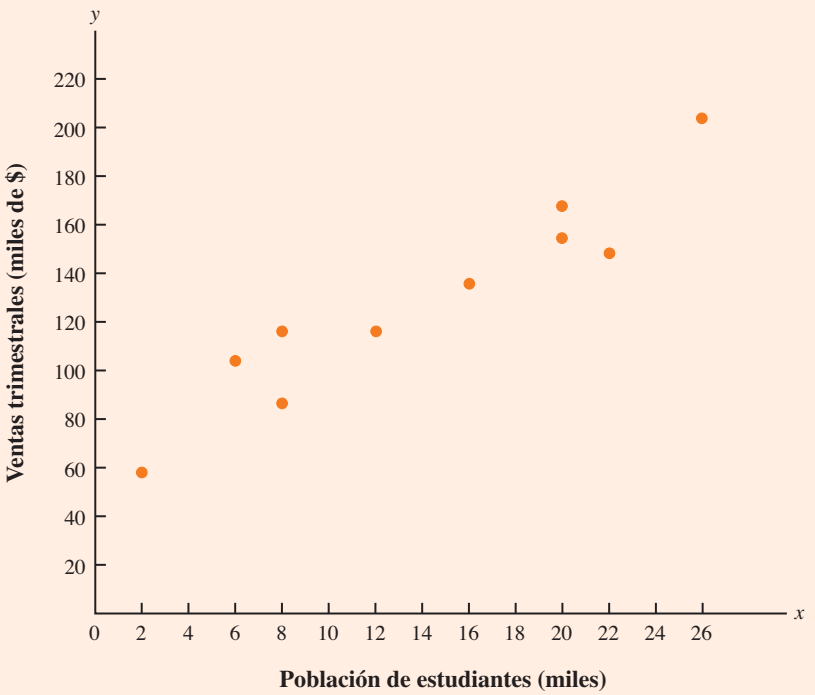
¿Qué conclusión preliminar se puede formular de la figura 14.3? Las ventas trimestrales parecen ser mayores cerca de los campus en los que la población de estudiantes es más grande. Además, en estos datos se observa que la relación entre el tamaño de la población y las ventas parece que puede aproximarse mediante una línea recta; en efecto, se observa que hay una re-

TABLA 14.1 Población de estudiantes y ventas trimestrales en 10 restaurantes Armand's Pizza Parlors

WEB archivo
Armand's

Restaurant i	Student Population (1 000s) x_i	Quarterly Sales ($1 000s) y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

FIGURA 14.3 Diagrama de dispersión de la población de estudiantes y ventas trimestrales en Armand's Pizza Parlors



lación lineal positiva entre x y y. Así, para representar la relación entre las ventas trimestrales y la población de estudiantes, se elige el modelo de regresión lineal simple. Decidido esto, la tarea siguiente es usar los datos muestrales de la tabla 14.1 para determinar los valores de b_0 y b_1 en la ecuación de regresión lineal simple estimada. Para el restaurante *i*ésimo, la ecuación de regresión simple estimada es

$$\hat{y}_i = b_0 + b_1 x_i \qquad \textbf{(14.4)}$$

donde

$\hat{y}_i$ = valor estimado de las ventas trimestrales (en miles de dólares) del *i*ésimo restaurante
b_0 = intersección de la recta de regresión estimada con el eje y
b_1 = pendiente de la recta de regresión estimada
x_i = tamaño de la población de estudiantes (en miles) del *i*ésimo restaurante

Como y_i denota ventas observadas (reales) para el restaurante i, y $\hat{y}_i$ representa el valor estimado de las ventas en la ecuación (14.4), para cada uno de los restaurantes de la muestra habrá un valor de ventas observadas y_i y un valor de ventas estimadas $\hat{y}_i$. Para que la recta de regresión estimada proporcione un buen ajuste a los datos, las diferencias entre los valores observados y estimados deben ser pequeñas.

En el método de mínimos cuadrados se usan los datos muestrales para obtener los valores de b_0 y b_1 que minimicen la *suma de los cuadrados de las desviaciones* (*diferencias*) entre los valores observados de la variable dependiente y_i y los valores estimados de la variable dependiente $\hat{y}_i$. El criterio que se emplea en el método de mínimos cuadrados se basa en la expresión (14.5).

Carl Friedrich Gauss (1777-1855) propuso el método de mínimos cuadrados.

CRITERIO DE MÍNIMOS CUADRADOS

$$\min \Sigma(y_i - \hat{y}_i)^2 \qquad \textbf{(14.5)}$$

donde:

y_i = valor observado de la variable dependiente en la observación *i*ésima
$\hat{y}_i$ = valor estimado de la variable dependiente en la observación *i*ésima

Se pueden usar cálculos diferenciales para demostrar (vea el apéndice 14.1) que los valores de b_0 y b_1 que minimiza la expresión (14.5) se pueden encontrar usando las ecuaciones (14.6) y (14.7).

Al determinar b_1 con ayuda de una calculadora, en los cálculos intermedios deben llevarse tantas cifras significativas como sea posible. Se recomiendan por lo menos cuatro dígitos significativos.

PENDIENTE E INTERSECCIÓN CON EL EJE *y* DE LA ECUACIÓN DE REGRESIÓN ESTIMADA[1]

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \qquad \textbf{(14.6)}$$

$$b_0 = \bar{y} - b_1\bar{x} \qquad \textbf{(14.7)}$$

donde:

x_i = valor de la variable independiente en la observación *i*ésima
y_i = valor de la variable dependiente en la observación *i*ésima
$\bar{x}$ = media de la variable independiente
$\bar{y}$ = media de la variable dependiente
n = número total de observaciones

En la tabla 14.2 se presentan algunos cálculos necesarios para desarrollar la ecuación de regresión estimada por mínimos cuadrados en el ejemplo de Armand's Piza Parlors. Como la muestra es de 10 restaurantes, tenemos $n = 10$ observaciones. Dado que en las ecuaciones (14.6) y (14.7) se necesitan $\bar{x}$ y $\bar{y}$, se empieza por calcularlas.

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{140}{10} = 14$$

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{1\,300}{10} = 130$$

Utilizando las ecuaciones (14.6) y (14.7) y la información de la tabla 14.2, se calcula la pendiente y la intersección de la ecuación de regresión estimada para Armand's Pizza Parlors. La pendiente (b_1) se calcula como se muestra enseguida.

[1] Otra fórmula para calcular b_1 es

$$b_1 = \frac{\Sigma x_i y_i - (\Sigma x_i \Sigma y_i)/n}{\Sigma x_i^2 - (\Sigma x_i)^2/n}$$

Esta forma de la ecuación (14.6) suele recomendarse cuando se emplea una calculadora para determinar b_1.

TABLA 14.2 Cálculos de la ecuación de regresión estimada por mínimos cuadrados para Armand's Pizza Parlors

Restaurante i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	−12	−72	864	144
2	6	105	−8	−25	200	64
3	8	88	−6	−42	252	36
4	8	118	−6	−12	72	36
5	12	117	−2	−13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totales	140	1300			2840	568
	Σx_i	Σy_i			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$
$$= \frac{2840}{568}$$
$$= 5$$

La intersección con el eje y (b_0) se calcula como sigue.

$$b_0 = \bar{y} - b_1\bar{x}$$
$$= 130 - 5(14)$$
$$= 60$$

Por tanto, la ecuación de regresión estimada es

$$\hat{y} = 60 + 5x$$

En la figura 14.4 se ilustra esta ecuación graficada sobre el diagrama de dispersión.

La pendiente de la ecuación de regresión estimada ($b_1 = 5$) es positiva, lo que implica que a medida que aumenta el tamaño de la población de estudiantes, las ventas se incrementan. Se concluye (con base en las ventas dadas en miles de $ y el tamaño de la población en miles) que un aumento de 1 000 en el tamaño de la población de estudiantes corresponde a un incremento de $5 000 en las ventas esperadas; es decir, se prevé que las ventas trimestrales se incrementen $5 por cada estudiante.

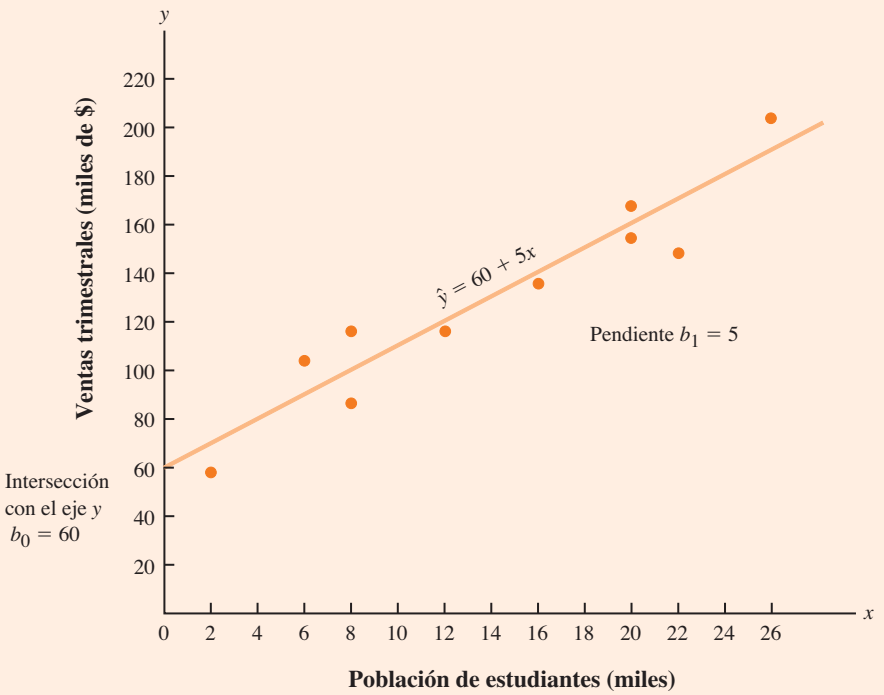
Debe tenerse mucho cuidado al usar la ecuación de regresión estimada para hacer predicciones fuera del rango de valores de la variable independiente, ya que fuera de ese rango no puede asegurarse que esta relación sea válida.

Si se considera que la ecuación de regresión estimada obtenida por el método de mínimos cuadrados describe adecuadamente la relación entre x y y, parecerá razonable usar esta ecuación a efecto de pronosticar el valor de y para un valor dado de x. Por ejemplo, si se quisieran predecir las ventas trimestrales de un restaurante ubicado cerca de un campo de 16 000 estudiantes, se calcularía, como sigue.

$$\hat{y} = 60 + 5(16) = 140$$

De manera que las ventas trimestrales pronosticadas para este restaurante serían de $140 000. En las secciones siguientes se verán los métodos para evaluar el uso correcto de la ecuación de regresión para hacer estimaciones y predicciones.

FIGURA 14.4 Gráfica de la ecuación de regresión estimada para Armand's Pizza Parlors: $\hat{y} = 60 + 5x$



NOTAS Y COMENTARIOS

El método de mínimos cuadrados proporciona una ecuación de regresión estimada que minimiza la suma de los cuadrados de las desviaciones entre los valores observados de la variable dependiente y_i y los valores estimados de la variable dependiente $\hat{y}_i$. El criterio de mínimos cuadrados permite elegir la ecuación de mejor ajuste. Si se empleara otro criterio, como minimizar la suma de las desviaciones absolutas entre y_i y $\hat{y}_i$, se obtendría una ecuación diferente. En la práctica el método de mínimos cuadrados es el más utilizado.

Ejercicios

Métodos

AUTOevaluación

1. Dadas las siguientes cinco observaciones de las variables x y y.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a) Trace el diagrama de dispersión correspondiente a estos datos.
b) ¿Qué indica el diagrama de dispersión del inciso *a*) respecto de la relación entre las dos variables?

c) Trate de aproximar la relación entre x y y trazando una línea recta que pase a través de los puntos de los datos.
d) Utilice las ecuaciones (14.6) y (14.7) para calcular b_0 y b_1, y desarrolle la ecuación de regresión estimada.
e) Use la ecuación de regresión estimada para predecir el valor de y cuando $x = 4$.

2. Dadas las siguientes cinco observaciones de las variables x y y.

x_i	3	12	6	20	14
y_i	55	40	55	10	15

a) Trace el diagrama de dispersión correspondiente a estos datos.
b) ¿Qué indica el diagrama de dispersión del inciso *a*) respecto de la relación entre las dos variables?
c) Trate de aproximar la relación entre x y y trazando una línea recta que pase a través de los puntos de los datos.
d) Utilice las ecuaciones (14.6) y (14.7) para calcular b_0 y b_1, y desarrolle la ecuación de regresión estimada.
e) Use la ecuación de regresión estimada para predecir el valor de y cuando $x = 10$.

3. Dadas las cinco observaciones siguientes obtenidas en un estudio de regresión sobre dos variables.

x_i	2	6	9	13	20
y_i	7	18	9	26	23

a) Trace el diagrama de dispersión correspondiente a estos datos.
b) Desarrolle la ecuación de regresión estimada correspondiente a estos datos.
c) Use la ecuación de regresión estimada para predecir el valor de y cuando $x = 6$.

Aplicaciones

AUTOevaluación

4. Los datos siguientes son estaturas (en pulgadas) y pesos (en libras) de nadadoras.

Estatura	68	64	62	65	66
Peso	132	108	102	115	128

a) Trace el diagrama de dispersión de estos datos usando la estatura como variable independiente.
b) ¿Qué indica el diagrama de dispersión del inciso *a*) respecto de la relación entre las dos variables?
c) Trate de aproximar la relación entre estatura y peso trazando una línea recta a través de los puntos de los datos.
d) Desarrolle la ecuación de regresión estimada calculando los valores de b_0 y b_1
e) Si la estatura de una nadadora es de 63 pulgadas, ¿cuál será su peso estimado?

5. Las ejercitadoras elípticas se están convirtiendo en una de las máquinas de ejercicio más populares. Su movimiento de bajo impacto es suave y estable, lo que las vuelve la elección preferida por las personas con problemas en rodillas y tobillos. Sin embargo, elegir la ejercitadora adecuada puede resultar un proceso difícil. El precio y la calidad son factores importantes en cualquier decisión de compra. ¿Están asociados los precios altos con las ejercitadoras elípticas de alta calidad? *Consumer Reports* realizó amplias pruebas para desarrollar una clasificación general basada en facilidad de uso, ergonomía, construcción y rango de ejercicio. A continua-

ción se muestran los datos de precio (Price) y calificación (Rating) de ocho ejercitadoras elípticas probadas, de las cuales se detallan marca y modelo (Brand and Model) (*Consumer Reports*, febrero de 2008).

WEB archivo
Ellipticals

Brand and Model	Price ($)	Rating
Precor 5.31	3 700	87
Keys Fitness CG2	2 500	84
Octane Fitness Q37e	2 800	82
LifeFitness X1 Basic	1 900	74
NordicTrack AudioStrider 990	1 000	73
Schwinn 430	800	69
Vision Fitness X6100	1 700	68
ProForm XP 520 Razor	600	55

a) Trace un diagrama de dispersión con estos datos empleando el precio como variable independiente.
b) Una tienda de equipo para ejercitarse que vende principalmente equipo caro puso un letrero sobre el área de exhibición que dice: "Calidad: usted obtiene lo que paga." Con base en su análisis de los datos, ¿considera usted que el letrero refleja de manera justa la relación precio-calidad de las ejercitadoras elípticas?
c) Utilice el método de mínimos cuadrados para obtener la ecuación de regresión estimada.
d) Utilice la ecuación de regresión estimada para predecir la clasificación de una ejercitadora elíptica con un precio de $1500.

6. El costo de un automóvil seminuevo depende de factores como marca y modelo, año, millas recorridas, condiciones y si se compra en una agencia o a un vendedor particular. Para investigar la relación entre millas (Miles) recorridas y precio (Price) de venta, se obtuvieron los datos de 10 operaciones de compra-venta entre particulares de un Honda Accord modelo 2000 (sitio web de PriceHub, octubre de 2008).

WEB archivo
HondaAccord

Miles (1 000s)	Price ($1 000s)
90	7.0
59	7.5
66	6.6
87	7.2
90	7.0
106	5.4
94	6.4
57	7.0
138	5.1
87	7.2

a) Trace un diagrama de dispersión tomando las millas como la variable independiente.
b) ¿Qué indica el diagrama de dispersión del inciso *a*) acerca de la relación entre las dos variables?
c) Use el método de mínimos cuadrados para desarrollar la ecuación de regresión estimada.
d) Proporcione una interpretación para la pendiente de la ecuación de regresión estimada.
e) Prediga el precio de venta de un Honda Accord 2000 con 100 000 millas.

7. Un gerente de ventas obtuvo los siguientes datos sobre ventas anuales (Annual Sales) y años de experiencia (Years of Experience) de 10 vendedores (Salesperson).

WEB archivo
Sales

Salesperson	Years of Experience	Annual Sales ($1 000s)
1	1	80
2	3	97
3	4	92
4	4	102
5	6	103
6	8	111
7	10	119
8	10	123
9	11	117
10	13	136

a) Trace un diagrama de dispersión con estos datos tomando como variable independiente los años de experiencia.
b) Obtenga una ecuación de regresión estimada que pueda utilizarse para predecir las ventas anuales proporcionando los años de experiencia.
c) Utilice la ecuación de regresión estimada para predecir las ventas anuales de un vendedor con 9 años de experiencia.

8. Bergans of Norway fabrica equipo para deportes a la intemperie desde 1908. Los siguientes datos muestran el rango de temperatura (Temperature Rating) en ºF y el precio (Price) en dólares de 11 modelos (Model) de *sleeping bags* fabricados por Bergans (*Backpacker* 2006 Gear Guide).

WEB archivo
SleepingBags

Model	Temperature Rating (°F)	Price ($)
Ranger 3-Seasons	12	319
Ranger Spring	24	289
Ranger Winter	3	389
Rondane 3-Seasons	13	239
Rondane Summer	38	149
Rondane Winter	4	289
Senja Ice	5	359
Senja Snow	15	259
Senja Zero	25	229
Super Light	45	129
Tight & Light	25	199

a) Trace un diagrama de dispersión con estos datos, en el que la variable independiente sea el rango de temperatura (°F).
b) ¿Qué indica el diagrama de dispersión del inciso *a*) respecto de la relación entre el rango de temperatura (°F) y precio?
c) Use el método de mínimos cuadrados para desarrollar la ecuación de regresión estimada.
d) Prediga cuál será el precio de un *sleeping bag* si el rango de temperatura (°F) es 20.

9. Con frecuencia, los viajeros de las aerolíneas empacan tanto como pueden en su maleta para evitar las tarifas por sobreequipaje. Encontrar una maleta rodante durable, con gran capacidad y fácil de mover puede ser difícil. La tabla siguiente muestra los resultados de pruebas realizadas por *Consumer Reports* en 10 maletas rodantes; las puntuaciones (Score) más altas indican mejores resultados en las pruebas en general (sitio web Consumer Reports, octubre de 2008). La tabla incluye marca (Brand) y precio (Price) de las maletas.

WEB archivo
Suitcases

Brand	Price ($)	Score
Briggs & Riley	325	72
Hartman	350	74
Heys	67	54
Kenneth Cole Reaction	120	54
Liz Claiborne	85	64
Samsonite	180	57
Titan	360	66
TravelPro	156	67
Tumi	595	87
Victorinox	400	77

a) Trace un diagrama de dispersión utilizando el precio como la variable independiente.
b) ¿Qué indica el diagrama de dispersión del inciso *a*) acerca de la relación entre las dos variables?
c) Use el método de mínimos cuadrados para desarrollar la ecuación de regresión estimada.
d) Proporcione una interpretación para la pendiente de la ecuación de regresión estimada.
e) La maleta de la marca Eagle Creek Hovercraft tiene un precio de $225. Usando la ecuación de regresión estimada desarrollada en el inciso *c*), prediga la puntuación para esta maleta.

10. Con base en la revisión anual de sueldos de *Advertising Age*, Mark Hurd, de 49 años, presidente (Chairman) y presidente ejecutivo (CEO) de Hewlett-Packard Co., recibió un sueldo anual de $817 000, un bono de más de $5 millones y otras compensaciones que superaron los $17 millones. Su compensación total fue ligeramente mejor que el pago total promedio de un CEO, $12.4 millones. La tabla siguiente muestra la edad (Age) y el sueldo anual (Salary) en miles de dólares de Mark Hurd y otros 14 ejecutivos (Executive) con su respectivo cargo (Title), quienes dirigen empresas que cotizan en la bolsa (*Advertising Age*, 5 de diciembre de 2006).

WEB archivo
ExecSalary

Executive	Title	Company	Age	Salary ($1 000s)
Charles Prince	Chmn/CEO	Citigroup	56	1 000
Harold McGraw III	Chmn/Pres/CEO	McGraw-Hill Cos.	57	1 172
James Dimon	Pres/CEO	JP Morgan Chase & Co.	50	1 000
K. Rupert Murdoch	Chmn/CEO	News Corp.	75	4 509
Kenneth D. Lewis	Chmn/Pres/CEO	Bank of America	58	1 500
Kenneth I. Chenault	Chmn/CEO	American Express Co.	54	1 092
Louis C. Camilleri	Chmn/CEO	Altria Group	51	1 663
Mark V. Hurd	Chmn/Pres/CEO	Hewlett-Packard Co.	49	817
Martin S. Sorrell	CEO	WPP Group	61	1 562
Robert L. Nardelli	Chmn/Pres/CEO	Home Depot	57	2 164
Samuel J. Palmisano	Chmn/Pres/CEO	IBM Corp.	55	1 680
David C. Novak	Chmn/Pres/CEO	Yum Brands	53	1 173
Henry R. Silverman	Chmn/CEO	Cendant Corp.	65	3 300
Robert C. Wright	Chmn/CEO	NBC Universal	62	2 500
Sumner Redstone	Exec Chmn/Founder	Viacom	82	5 807

a) Trace un diagrama de dispersión con estos datos utilizando la edad del ejecutivo como la variable independiente.
b) ¿Qué indica el diagrama de dispersión del inciso *a*) acerca de la relación entre las dos variables?
c) Use el método de mínimos cuadrados para obtener la ecuación de regresión estimada.
d) Suponga que Bill Gustin, de 72 años, es el presidente y CEO de una de las principales empresas de electrónica. Prediga su sueldo anual.

11. Los automóviles deportivos están diseñados para proporcionar mejor conducción, mayor aceleración y experiencia de manejo más agradable que un sedán típico. Sin embargo, incluso dentro de este exclusivo grupo de vehículos, tanto el desempeño como el precio pueden variar. *Consumer Reports* proporciona información de puntuaciones en pruebas de manejo (Road-Test Score) y precios (Price) de los siguientes 12 automóviles (Car) deportivos (sitio web de Consumer Reports, octubre de 2008). Los precios están en miles de dólares y las puntuaciones en pruebas de manejo se basan en una escala de 0 a 100, donde los valores más altos indican un mejor desempeño.

WEB archivo
SportyCars

Car	Price ($1 000s)	Road-Test Score
Chevrolet Cobalt SS	24.5	78
Dodge Caliber SRT4	24.9	56
Ford Mustang GT (V8)	29.0	73
Honda Civic Si	21.7	78
Mazda RX-8	31.3	86
Mini Cooper S	26.4	74
Mitsubishi Lancer Evolution GSR	38.1	83
Nissan Sentra SE-R Spec V	23.3	66
Suburu Impreza WRX	25.2	81
Suburu Impreza WRX Sti	37.6	89
Volkswagen GTI	24.0	83
Volkswagen R32	33.6	83

a) Trace un diagrama de dispersión tomando el precio como la variable independiente.
b) ¿Qué indica el diagrama de dispersión del inciso *a*) acerca de la relación entre las dos variables?
c) Use el método de mínimos cuadrados para desarrollar la ecuación de regresión estimada.
d) Proporcione una interpretación para la pendiente de la ecuación de regresión estimada.
e) Otro automóvil deportivo probado por *Consumer Reports* es el BMW 135i; el precio de este vehículo fue de $36 700. Prediga la puntuación en la prueba de manejo para el BMW 135i utilizando la ecuación de regresión estimada obtenida en el inciso *c*).

12. Una moto acuática es una embarcación con propulsión a chorro de agua, diseñada para ser conducida por una persona sentada, de pie o arrodillada sobre el casco. En la década de 1970, Kawasaki Motors Corp. U.S.A. introdujo la moto acuática JET SKI©, la primera que tuvo éxito comercial. En la actualidad, *jet ski* se usa en forma general como término genérico para motos acuáticas personales. En la tabla siguiente se dan pesos (Weight) redondeados a la decena de libras más cercana, y precios (Price) redondeados a los $50 más cercanos, de 10 motos acuáticas personales de tres plazas, con su respectivo fabricante y modelo (Make and Model) (sitio web de Jetski News, 2006).

WEB archivo
JetSki

Make and Model	Weight (lb)	Price ($)
Honda AquaTrax F-12	750	9500
Honda AquaTrax F-12X	790	10500
Honda AquaTrax F-12X GPScape	800	11200
Kawasaki STX-12F Jetski	740	8500
Yamaha FX Cruiser Waverunner	830	10000
Yamaha FX High Output Waverunner	770	10000
Yamaha FX Waverunner	830	9300
Yamaha VX110 Deluxe Waverunner	720	7700
Yamaha VX110 Sport Waverunner	720	7000
Yamaha XLT1200 Waverunner	780	8500

a) Trace el diagrama de dispersión correspondiente a estos datos empleando el peso como variable independiente.
b) ¿Qué indica el diagrama de dispersión del inciso *a*) respecto de la relación entre peso y precio?
c) Use el método de mínimos cuadrados para desarrollar la ecuación de regresión estimada.
d) Prediga el precio de una moto acuática de tres plazas cuyo peso sea de 750 libras.

e) La Honda Aqua Trax F-12 pesa 750 libras y su precio es de $9 500. ¿El precio pronosticado en el inciso *d*) para una moto acuática con peso de 750 libras no debería ser también de $9 500?

f) La Kawasaki SX-R 800 Jetski tiene capacidad sólo para una persona y pesa 350 libras. ¿Cree usted que la ecuación de regresión estimada obtenida en el inciso *c*) deba emplearse para pronosticar su precio?

13. Para el Internal Revenue Service (Servicio de Administración Tributaria de Estados Unidos), el carácter razonable de las deducciones declaradas por un contribuyente depende de su ingreso bruto ajustado. Deducciones grandes que comprenden donaciones de caridad o por atención médica son más apropiadas para contribuyentes que tengan un ingreso bruto ajustado grande. Si las deducciones de una persona son mayores que las deducciones declaradas promedio correspondientes a un determinado nivel de ingresos, aumentan las posibilidades de que se le realice una auditoría. Los datos (en miles de dólares) sobre ingreso bruto ajustado y el monto promedio o razonable de deducciones declaradas se listan a continuación.

Ingreso bruto ajustado (miles de $)	Monto razonable de las deducciones declaradas (miles de $)
22	9.6
27	9.6
32	10.1
48	11.1
65	13.5
85	17.7
120	25.5

a) Trace un diagrama de dispersión con estos datos empleando como variable independiente el ingreso bruto ajustado.

b) Use el método de mínimos cuadrados para desarrollar la ecuación de regresión estimada.

c) Calcule el monto razonable de deducciones declaradas de un contribuyente cuyo ingreso bruto ajustado es de $52 500. Si éste tiene deducciones declaradas por $20 400, ¿estará justificada una auditoria? Explique.

14. *PCWorld* valoró cuatro componentes de 10 computadoras laptop ultraportátiles: características, desempeño, diseño y precio. Cada elemento fue valorado utilizando una escala de puntos de 0 a 100. Luego se desarrolló un índice general, llamado *PCW World* Rating, para cada laptop. La siguiente tabla muestra el índice de características (Features Rating) y el índice *PCW World* para 10 modelos (Model) de computadoras (sitio web de PC World, 5 de febrero de 2009).

WEB archivo
Laptop

Model	Features Rating	*PCW World* Rating
Thinkpad X200	87	83
VGN-Z598U	85	82
U6V	80	81
Elitebook 2530P	75	78
X360	80	78
Thinkpad X300	76	78
Ideapad U110	81	77
Micro Express JFT2500	73	75
Toughbook W7	79	73
HP Voodoo Envy133	68	72

a) Trace el diagrama de dispersión correspondiente a estos datos empleando el índice de características como variable independiente.

b) ¿Qué indica el diagrama de dispersión del inciso *a*) respecto de la relación entre ambas variables?

c) Use el método de mínimos cuadrados para desarrollar la ecuación de regresión estimada.

d) Estime el Índice *PCW World* para una nueva computadora laptop que tiene un índice de características de 70.

14.3 Coeficiente de determinación

En el ejemplo de Armand's Pizza Parlors, para aproximar la relación lineal entre el tamaño de la población de estudiantes x y las ventas trimestrales y se desarrolló la ecuación de regresión estimada $\hat{y} = 60 + 5x$. Ahora la pregunta es: ¿qué tan bien se ajusta a los datos la ecuación de regresión estimada? En esta sección se muestra que el **coeficiente de determinación** proporciona una medida de la bondad de ajuste para la ecuación de regresión estimada.

A la diferencia que existe en la observación $i$ésima entre el valor observado de la variable dependiente y_i, y el valor estimado de la variable dependiente $\hat{y}_i$, se le llama **residual $i$ésimo**. Éste representa el error que existe al usar $\hat{y}_i$ para estimar y_i. Por tanto, para la observación $i$ésima el residual es $y_i - \hat{y}_i$. La suma de los cuadrados de estos residuales o errores es la cantidad que se minimiza empleando el método de los mínimos cuadrados. Esta cantidad, también conocida como *suma de cuadrados debido al error,* se denota como SCE.

SUMA DE CUADRADOS DEBIDO AL ERROR

$$\text{SCE} = \Sigma(y_i - \hat{y}_i)^2 \qquad \textbf{(14.8)}$$

El valor de la SCE es una medida del error al utilizar la ecuación de regresión estimada para calcular los valores de la variable dependiente de la muestra.

En la tabla 14.3 se indican los procedimientos que se requieren para calcular la suma de cuadrados debido al error en el caso de Armand's Pizza Parlors. Por ejemplo, los valores de las variables independiente y dependiente del restaurante 1 son $x_1 = 2$ y $y_1 = 58$. El valor estimado para sus ventas trimestrales obtenido con la ecuación de regresión estimada es $\hat{y}_1 = 60 + 5(2) = 70$. Por consiguiente, para el restaurante 1 el error al usar $\hat{y}_1$ para estimar y_1 es $y_1 - \hat{y}_1 = 58 - 70 = -12$. El error al cuadrado, $(-12)^2 = 144$, aparece en la última columna de la tabla 14.3. Después de calcular y elevar al cuadrado los residuales de cada uno de los restaurantes de la muestra, se suman y obtenemos que SCE = 1530. Por tanto, esta suma mide el error que existe al utilizar la ecuación de regresión estimada $\hat{y} = 60 + 5x$ para predecir las ventas.

Ahora suponga que se pide una estimación de las ventas trimestrales sin conocer el tamaño de la población de estudiantes. Sin tener conocimiento de ninguna otra variable relacionada, se emplearía la media muestral como una estimación de las ventas trimestrales de cualquiera de

TABLA 14.3 Cálculo de SCE en el ejemplo de Armand's Pizza Parlors

Restaurante i	x_i = población de estudiantes (miles)	y_i = ventas trimestrales (miles de $)	Ventas pronosticadas $\hat{y}_i = 60 + 5x_i$	Error $y_i - \hat{y}_i$	Error al cuadrado $(y_i - \hat{y}_i)^2$
1	2	58	70	−12	144
2	6	105	90	15	225
3	8	88	100	−12	144
4	8	118	100	18	324
5	12	117	120	−3	9
6	16	137	140	−3	9
7	20	157	160	−3	9
8	20	169	160	9	81
9	22	149	170	−21	441
10	26	202	190	12	144
					SCE = 1530

TABLA 14.4 Cálculo de la suma total de cuadrados en el ejemplo Armand's Pizza Parlors

Restaurante i	x_i = población de estudiantes (miles)	y_i = ventas trimestrales (miles de $)	Desviación $y_i - \bar{y}$	Desviación al cuadrado $(y_i - \bar{y})^2$
1	2	58	−72	5 184
2	6	105	−25	625
3	8	88	−42	1 764
4	8	118	−12	144
5	12	117	−13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1 521
9	22	149	19	361
10	26	202	72	5 184
				STC = 15 730

los restaurantes. En la tabla 14.2 se mostró que con base en los datos de las ventas, $\Sigma y_i = 1300$. Así, el valor medio de las ventas trimestrales en la muestra de los 10 restaurantes Armand's es $\bar{y} = \Sigma y_i/n = 1300/10 = 130$. En la tabla 14.4 se presenta la suma de las desviaciones al cuadrado que se obtiene cuando se usa la media muestral $\bar{y} = 130$ para estimar el valor de las ventas trimestrales de cada uno de los restaurantes. Para el restaurante *i*ésimo de la muestra, la diferencia $y_i - \bar{y}$ proporciona una medida del error que implica usar $\bar{y}$ para estimar las ventas. La correspondiente suma de cuadrados, llamada *suma total de cuadrados*, se denota STC.

SUMA TOTAL DE CUADRADOS

$$\text{STC} = \Sigma(y_i - \bar{y})^2 \qquad \textbf{(14.9)}$$

Como la STC = *15 730 y la* SCE = *1 530, la línea de regresión estimada se ajusta mucho mejor a los datos que la línea* $y = \bar{y}$.

La cifra debajo de la última columna de la tabla 14.4 es la suma total de cuadrados en el ejemplo de Armand's Pizza Parlors; ésta es STC = 15 730.

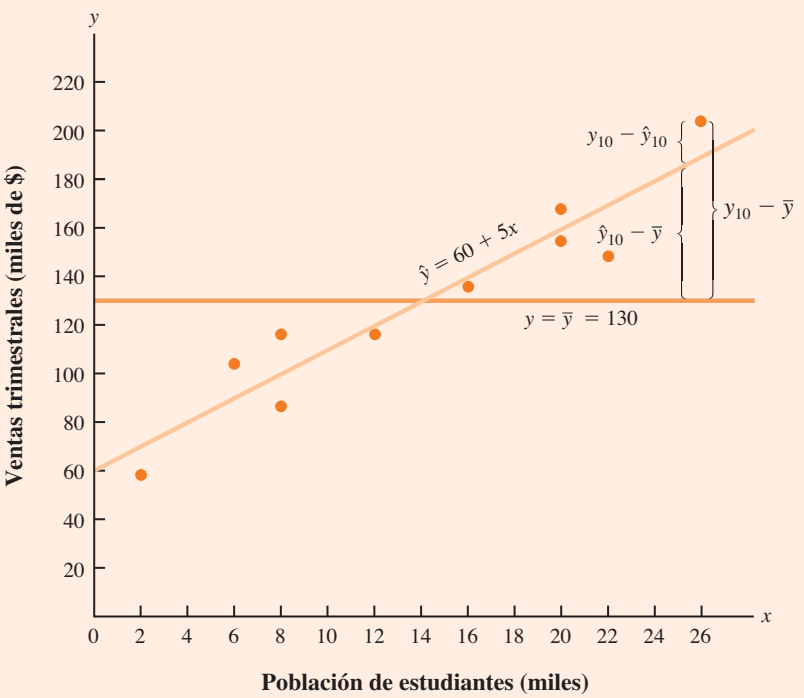
En la figura 14.5 se indica la línea de regresión estimada $\hat{y} = 60 + 5x$ y la línea correspondiente a $\bar{y} = 130$. Observe que los puntos se agrupan de manera más cerrada alrededor de la recta de regresión estimada que en torno de la línea $\bar{y} = 130$. Por ejemplo, vemos que para el 10o. restaurante de la muestra, el error es mucho más grande cuando se usa $\bar{y} = 130$ para estimar y_{10} que cuando se utiliza $\hat{y}_{10} = 60 + 5(26) = 190$. Se puede entender la STC como una medida de cuánto se agrupan las observaciones en torno a la línea $\bar{y}$, y la SCE como una medida de cuánto se agrupan las observaciones en torno de la recta $\hat{y}$.

Para medir cuánto se desvían de $\bar{y}$ los valores $\hat{y}$ en la recta de regresión, se calcula otra suma de cuadrados, la cual se llama *suma de cuadrados debido a la regresión* y se denota como SCR.

SUMA DE CUADRADOS DEBIDO A LA REGRESIÓN

$$\text{SCR} = \Sigma(\hat{y}_i - \bar{y})^2 \qquad \textbf{(14.10)}$$

FIGURA 14.5 Desviaciones respecto de la línea de regresión estimada y la línea $y = \bar{y}$ en el ejemplo de Armand's Pizza Parlors



Por lo antes dicho, se esperaría que hubiera alguna relación entre STC, SCR y SCE. En efecto, la relación entre estas tres sumas de cuadrados constituye uno de los resultados más importantes de la estadística.

La SCR *puede entenderse como la parte explicada de la* STC, *y la* SCE *puede entenderse como la parte no explicada de la* STC.

RELACIÓN ENTRE STC, SCR Y SCE

$$\text{STC} = \text{SCR} + \text{SCE} \qquad \textbf{(14.11)}$$

donde:

STC = suma total de cuadrados
SCR = suma de cuadrados debido a la regresión
SCE = suma de cuadrados debido al error

La ecuación (14.11) indica que la suma total de cuadrados puede ser dividida en dos componentes: la suma de cuadrados debido a la regresión y la suma de cuadrados debido al error. Por consiguiente, si se conocen los valores de dos de estas sumas, es fácil calcular la tercera suma de cuadrados. Por ejemplo, en el caso de Armand's Pizza Parlors se conocen SCE = 1 530 y STC = 15 730; por tanto, al despejar SCR en la ecuación (14.11), se encuentra que la suma de cuadrados debido a la regresión es

$$\text{SCR} = \text{STC} - \text{SCE} = 15\,730 - 1\,530 = 14\,200$$

Ahora se verá como se usan estas tres sumas de cuadrados, STC, SCR y SCE, para obtener una medida de la bondad de ajuste de la ecuación de regresión estimada. Esta ecuación se ajustaría perfectamente a los datos si cada uno de los valores de la variable dependiente y_i se encontrara sobre la recta de regresión. En este caso, para todas las observaciones se tendría que $y_i - \hat{y}_i$ sería igual a 0, con lo que SCE $= 0$. Como STC $=$ SCR $+$ SCE, vemos que para que haya un ajuste perfecto, SCR debe ser igual a STC, y el cociente (SCR/STC) debe ser igual a 1. Cuando los ajustes son malos, se tendrán valores altos para SCE. Si en la ecuación (14.11) despejamos ésta, tenemos que SCE $=$ STC $-$ SCR. Por consiguiente, el valor más grande de SCE (y por tanto el ajuste más pobre) se presenta cuando SCR $= 0$ y SCE $=$ STC.

El cociente SCR/STC, que puede tomar valores entre 0 y 1, se usa para evaluar la bondad de ajuste de la ecuación de regresión estimada. A este cociente se le llama *coeficiente de determinación* y se denota como r^2.

COEFICIENTE DE DETERMINACIÓN

$$r^2 = \frac{\text{SCR}}{\text{STC}} \tag{14.12}$$

En el ejemplo de Armand's Pizza Parlors, el valor del coeficiente de determinación es

$$r^2 = \frac{\text{SCR}}{\text{STC}} = \frac{14\,200}{15\,730} = 0.9027$$

Cuando se expresa el coeficiente de determinación en forma de porcentaje, r^2 se puede interpretar como el porcentaje de la suma total de cuadrados que se explica mediante el uso de la ecuación de regresión estimada. En el ejemplo de Armand's Pizza Parlors se concluye que 90.27% de la suma total de cuadrados se explica utilizando la ecuación de regresión estimada $\hat{y} = 60 + 5x$ para predecir las ventas trimestrales. En otras palabras, 90.27% de la variabilidad en las ventas se explica por la relación lineal que existe entre éstas y el tamaño de la población de estudiantes. Sería satisfactorio encontrar un buen ajuste para la ecuación de regresión estimada.

Coeficiente de correlación

En el capítulo 3 se presentó el **coeficiente de correlación** como una medida descriptiva de la intensidad de la relación lineal entre dos variables x y y. Los valores del coeficiente de correlación siempre estarán entre -1 y $+1$. Un valor de $+1$ indica que las dos variables x y y están perfectamente relacionadas en un sentido lineal positivo. Es decir, todos los puntos de los datos se encuentran en una línea recta que tiene pendiente positiva. Un valor de -1 indica que x y y están perfectamente relacionadas en un sentido lineal negativo con todos los puntos de los datos en una recta con pendiente negativa. Los valores del coeficiente de correlación cercanos a 0 indican que x y y no están relacionadas linealmente.

En la sección 3.5 se presentó la ecuación para calcular el coeficiente de correlación muestral. Cuando se ha realizado un análisis de regresión y calculado el coeficiente de determinación r^2, el coeficiente de correlación muestral se puede obtener como se indica a continuación.

COEFICIENTE DE CORRELACIÓN MUESTRAL

$$\begin{aligned} r_{xy} &= (\text{signo de } b_1)\sqrt{\text{coeficiente de determinación}} \\ &= (\text{signo de } b_1)\sqrt{r^2} \end{aligned} \tag{14.13}$$

donde:

b_1 = pendiente de la ecuación de regresión estimada $\hat{y} = b_0 + b_1 x$

El signo del coeficiente de correlación muestral es positivo si la ecuación de regresión estimada tiene pendiente positiva ($b_1 > 0$), y es negativo si la ecuación de regresión estimada tiene pendiente negativa ($b_1 < 0$).

En el ejemplo de Armand's Pizza Parlor, el valor del coeficiente de determinación correspondiente a la ecuación de regresión estimada $\hat{y} = 60 + 5x$ es 0.9027. Como la pendiente de esta ecuación es positiva, la ecuación (14.13) indica que el coeficiente de correlación muestral es $+\sqrt{0.9027} = +0.9501$. Con este coeficiente $r_{xy} = +0.9501$, concluimos que existe una fuerte relación lineal positiva entre x y y.

En el caso de una relación lineal entre dos variables, tanto el coeficiente de determinación como el coeficiente de correlación muestral proporcionan medidas de la fuerza de la relación. El primero provee una medida entre 0 y 1, mientras que el segundo proporciona una medida entre -1 y $+1$. Aunque el coeficiente de correlación muestral está restringido a la relación lineal entre dos variables, el coeficiente de determinación puede emplearse para relaciones no lineales y para otras en las que hay dos o más variables independientes. Por tanto, ofrece un rango de aplicación más amplio.

NOTAS Y COMENTARIOS

1. Al desarrollar la ecuación de regresión estimada mediante el método de mínimos cuadrados y calcular el coeficiente de determinación, no se hizo ningún supuesto probabilístico acerca del término del error ϵ; tampoco se realizó ninguna prueba estadística de significancia para la relación entre x y y. Los valores grandes de r^2 implican que la recta de mínimos cuadrados se ajusta mejor a los datos; es decir, las observaciones se encuentran más cerca de la recta de mínimos cuadrados. Sin embargo, usando sólo r^2 no se pueden formular conclusiones acerca de si la relación entre x y y es estadísticamente significativa. Tal conclusión debe basarse en consideraciones que implican el tamaño de la muestra y las propiedades de la distribución de muestreo adecuada de los estimadores de mínimos cuadrados.
2. Para fines prácticos, cuando se trata de datos que se manejan típicamente en las ciencias sociales, valores de r^2 tan pequeños como 0.25 suelen considerarse útiles. En datos de física o de ciencias de la vida suelen encontrarse valores r^2 de 0.60 o más, en algunos casos mayores de 0.90. En las aplicaciones de negocios, los valores de r^2 varían enormemente dependiendo de las características particulares de cada aplicación.

Ejercicios

Métodos

AUTOevaluación

15. Los siguientes datos provienen del ejercicio 1.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

La ecuación de regresión estimada para estos datos es $\hat{y} = 0.20 + 2.60x$.

a) Calcule las SCE, STC y SCR empleando las ecuaciones (14.8), (14.9) y (14.10).

b) Calcule el coeficiente de determinación r^2. Haga un comentario sobre la bondad del ajuste.

c) Determine el coeficiente de correlación muestral.

16. Los datos que se presentan a continuación son retomados del ejercicio 2.

x_i	3	12	6	20	14
y_i	55	40	55	10	15

La ecuación de regresión estimada para estos datos es $\hat{y} = 68 - 3x$.

a) Calcule las SCE, STC y SCR.

b) Calcule el coeficiente de determinación r^2. Exprese un comentario sobre la bondad del ajuste.

c) Determine el coeficiente de correlación muestral.

17. Los datos a continuación provienen los del ejercicio 3.

x_i	2	6	9	13	20
y_i	7	18	9	26	23

La ecuación de regresión estimada para estos datos es $\hat{y} = 7.6 + 0.9x$. ¿Qué porcentaje de la suma total de cuadrados puede explicarse mediante esta ecuación? ¿Cuál es el valor del coeficiente de correlación muestral?

Aplicaciones

AUTOevaluación

18. En los datos siguientes, y corresponde a los sueldos mensuales y x es el promedio obtenido por los estudiantes que terminaron su grado en administración con especialidad en sistemas de información. La ecuación de regresión estimada con estos datos es $\hat{y} = 1\,790.5 + 581.1x$.

Promedio	Sueldo mensual ($)
2.6	3 300
3.4	3 600
3.6	4 000
3.2	3 500
3.5	3 900
2.9	3 600

a) Calcule las SCE, STC y SCR.

b) Calcule el coeficiente de determinación r^2. Haga un comentario sobre la bondad del ajuste.

c) Determine el valor del coeficiente de correlación muestral.

19. En el ejercicio 7 un gerente de ventas obtuvo los siguientes datos sobre $x =$ ventas anuales (Annual Sales) y $y =$ años de experiencia (Years of Experience) de 10 vendedores (Salesperson). La ecuación de regresión estimada para estos datos es $\hat{y} = 80 + 4x$.

WEB archivo

Sales

Salesperson	Years of Experience	Annual Sales ($1 000s)
1	1	80
2	3	97
3	4	92
4	4	102
5	6	103
6	8	111
7	10	119
8	10	123
9	11	117
10	13	136

a) Calcule las SCE, STC y SCR.
b) Calcule el coeficiente de determinación r^2. Haga un comentario sobre la bondad del ajuste.
c) ¿Cuál es el valor del coeficiente de correlación muestral?

20. *Consumer Reports* publica extensas pruebas y evaluaciones de más de 100 televisores de alta definición. Para cada modelo se elaboró una puntuación general basada principalmente en la calidad de la imagen. Una puntuación más alta indica un mejor funcionamiento. En los datos siguientes se proporcionan el precio (Price) y la puntuación (Score) general de 10 televisores de plasma de 42 pulgadas con su respectiva marca (Brand) (*Consumer Reports*, marzo de 2006).

WEB archivo
PlasmaTV

Brand	Price ($)	Score
Dell	2800	62
Hisense	2800	53
Hitachi	2700	44
JVC	3500	50
LG	3300	54
Maxent	2000	39
Panasonic	4000	66
Phillips	3000	55
Proview	2500	34
Samsung	3000	39

a) Utilice estos datos para desarrollar una ecuación de regresión estimada que pueda emplearse para determinar la puntuación general de un televisor de plasma de 42 pulgadas dado el precio.
b) Calcule r^2. ¿Proporcionó un buen ajuste la ecuación de regresión estimada?
c) Estime la puntuación general de un televisor cuyo precio es de $3200.

21. Una aplicación importante del análisis de regresión en la contaduría es la estimación de costos. Partiendo de datos sobre volumen de producción y costos, y empleando el método de mínimos cuadrados para desarrollar una ecuación de regresión estimada que relacione ambos datos, un contador puede calcular los costos correspondientes a un determinado volumen de producción. Considere la siguiente muestra de datos sobre volumen de producción y costo total de una operación de manufactura.

Volumen de producción (unidades)	Costo total ($)
400	4000
450	5000
550	5400
600	5900
700	6400
750	7000

a) Utilice estos datos para desarrollar la ecuación de regresión estimada útil a efecto de pronosticar los costos totales dado un volumen de producción determinado.
b) ¿Cuál es el costo variable por unidad producida?
c) Calcule el coeficiente de determinación. ¿Qué porcentaje de la variación en los costos totales puede ser explicada por el volumen de producción?
d) Con base en el programa de producción de la empresa, el mes próximo se deberán producir 500 unidades. ¿Cuál es el costo total estimado para esta operación?

22. Remítase al ejercicio 5 donde se utilizaron los siguientes datos para investigar si, por lo general, los precios altos están o no asociados con las altas calificaciones de las ejercitadoras elípticas (*Consumer Reports*, febrero de 2008). La siguiente tabla presenta los datos de marca y modelo (Brand and Model), precio (Price) y calificación (Rating).

WEB archivo
Ellipticals

Brand and Model	Price ($)	Rating
Precor 5.31	3 700	87
Keys Fitness CG2	2 500	84
Octane Fitness Q37e	2 800	82
LifeFitness X1 Basic	1 900	74
NordicTrack AudioStrider 990	1 000	73
Schwinn 430	800	69
Vision Fitness X6100	1 700	68
ProForm XP 520 Razor	600	55

Con x = precio ($) y y = calificación, la ecuación de regresión estimada es $\hat{y} = 58.158 + 0.008449x$. Para esos datos, SCE = 173.88.

a) Calcule el coeficiente de determinación r^2.

b) ¿La ecuación de regresión estimada proporciona un buen ajuste? Explique su respuesta.

c) ¿Cuál es el valor del coeficiente de correlación muestral? ¿Cómo es la relación que refleja entre precio y calificación: fuerte o débil?

14.4 Supuestos del modelo

En un análisis de regresión se empieza por hacer un supuesto acerca del modelo apropiado para la relación entre las variables dependientes e independientes. En el caso de la regresión lineal simple, se supone que el modelo de regresión es

$$y = \beta_0 + \beta_1 x + \epsilon$$

Después, empleando el método de mínimos cuadrados, se obtienen los valores de b_0 y b_1, que son las estimaciones de los parámetros del modelo β_0 y β_1, respectivamente. Así se llega a la ecuación de regresión estimada

$$\hat{y} = b_0 + b_1 x$$

Como vimos, el valor del coeficiente de determinación (r^2) es una medida de la bondad de ajuste de la ecuación de regresión estimada. Sin embargo, aun cuando se obtenga un valor grande para r^2, la ecuación de regresión estimada no debe ser usada sino hasta que se realice un análisis para determinar si el modelo empleado es apropiado. Un paso importante para ver si el supuesto del modelo es adecuado, consiste en probar la significancia de la relación. Las pruebas de significancia en el análisis de regresión están basadas en los siguientes supuestos acerca del término del error ϵ.

SUPUESTOS ACERCA DEL TÉRMINO DEL ERROR ϵ EN EL MODELO DE REGRESIÓN

$$y = \beta_0 + \beta_1 x + \epsilon$$

1. El término del error ϵ es una variable aleatoria cuya media, o valor esperado, es cero; es decir, $E(\epsilon) = 0$.
Implicación. β_0 y β_1 son constantes, por tanto, $E(\beta_0) = \beta_0$ y $E(\beta_1) = \beta_1$; así, para un valor dado de x, el valor esperado de y es

$$E(y) = \beta_0 + \beta_1 x \qquad \textbf{(14.14)}$$

(*Continúa*)

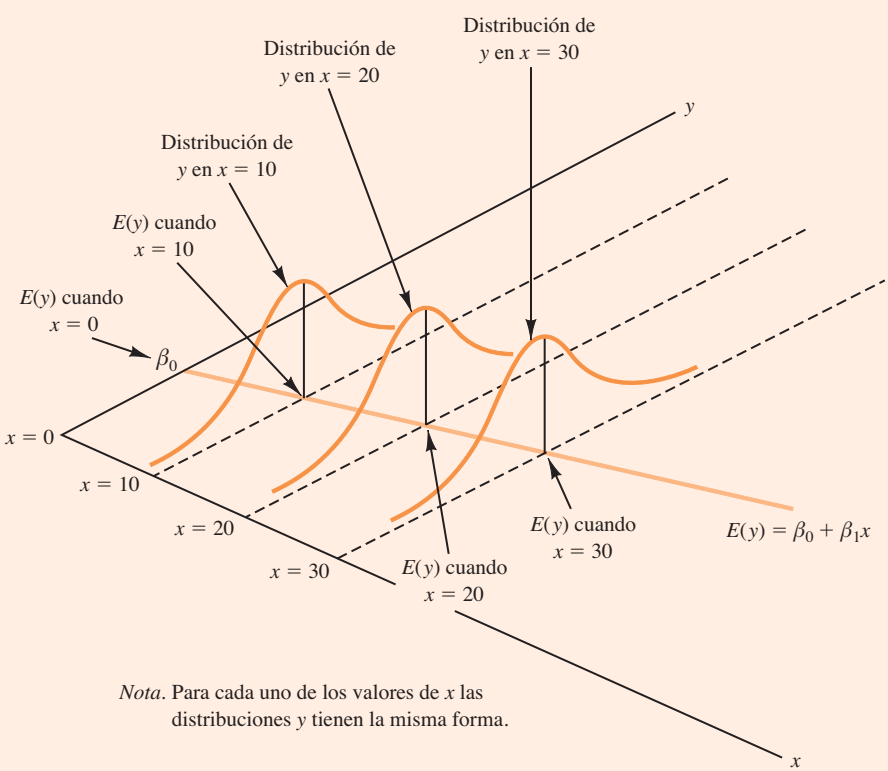
Como ya se indicó, a la ecuación (14.14) se le conoce como *ecuación de regresión*.

2. La varianza de ϵ, que se denota como σ^2, es la misma para todos los valores de x.
   Implicación. La varianza de y respecto de la recta de regresión es igual a σ^2 y es la misma para todos los valores de x.
3. Los valores de ϵ son independientes.
   Implicación. El valor de ϵ correspondiente a un determinado valor de x no está relacionado con el valor de ϵ para cualquier otro valor de x; por tanto, el valor de y correspondiente a un valor particular de x no está relacionado con el valor de y de ningún otro valor de x.
4. El término del error ϵ es una variable aleatoria distribuida normalmente.
   Implicación. Como y es una función lineal de ϵ, también y es una variable aleatoria distribuida normalmente.

En la figura 14.6 se ilustran los supuestos del modelo y sus implicaciones; observe que en esta interpretación gráfica el valor de $E(y)$ cambia con base en el valor específico de x que se considere. Sin embargo, sea cual fuere el valor de x, la distribución de probabilidad de ϵ y, por tanto, la distribución de probabilidad de y, son distribuciones normales que tienen todas la misma varianza. El valor específico del error ϵ en cualquier punto depende de si el valor real de y es mayor o menor que $E(y)$.

En este punto hay que tener presente que también se hace un supuesto o se tiene una hipótesis acerca de la forma de la relación entre x y y. Es decir, se supone que la base de la relación

FIGURA 14.6 Supuestos del modelo de regresión



Nota. Para cada uno de los valores de x las distribuciones y tienen la misma forma.

entre las variables es una recta representada por $\beta_0 + \beta_1 x$. No se debe perder de vista el hecho de que puede haber algún otro modelo, por ejemplo $y = \beta_0 + \beta_1 x^2 + \epsilon$, que resulte ser mejor para la relación subyacente.

14.5 Prueba de significancia

En una ecuación de regresión lineal simple, la media o valor esperado de y es una función lineal de x: $E(y) = \beta_0 + \beta_1 x$. Pero si el valor de β_1 es cero, $E(y) = \beta_0 + (0)x = \beta_0$. En este caso, el valor medio de y no depende del valor de x y, por tanto, podemos concluir que x y y no están relacionadas linealmente. De manera alterna, si el valor de β_1 es distinto de cero, se concluirá que las dos variables están relacionadas. Así, para probar si existe una relación de regresión significativa, se debe realizar una prueba de hipótesis a efecto de determinar si el valor de β_1 es distinto de cero. Hay dos pruebas que son las más usadas. En ambas se requiere una estimación de σ^2, la varianza de ϵ en el modelo de regresión.

Estimación de σ^2

Con base en el modelo de regresión y sus supuestos, podemos concluir que σ^2, la varianza de ϵ, representa también la varianza de los valores de y respecto de la recta de regresión. Recuerde que a las desviaciones de los valores de y respecto de la recta de regresión estimada se les conoce como *residuales*. Por tanto, SCE, la suma de los cuadrados de los residuales, es una medida de la variabilidad de las observaciones reales respecto de la línea de regresión estimada. El **error cuadrado medio** (ECM) proporciona una estimación de σ^2; esta estimación es de SCE dividida entre sus grados de libertad.

Como $\hat{y}_i = b_0 + b_1 x_i$, SCE se puede expresar como

$$\text{SCE} = \Sigma(y_i - \hat{y}_i)^2 = \Sigma(y_i - b_0 - b_1 x_i)^2$$

Cada suma de cuadrados está asociada con un número llamado *grados de libertad*. Los expertos en estadística han demostrado que la SCE tiene $n - 2$ grados de libertad, porque para calcularla es necesario estimar dos parámetros (β_0 y β_1). El error cuadrado medio se calcula al dividir SCE entre $n - 2$. El ECM proporciona un estimador insesgado de σ^2. Como el valor del ECM provee un estimado de σ^2, se emplea también la notación s^2.

ERROR CUADRADO MEDIO (ESTIMACIÓN DE σ^2)

$$s^2 = \text{ECM} = \frac{\text{SCE}}{n - 2} \qquad (14.15)$$

En la sección 14.3 se encontró que en el ejemplo de Armand's Pizza Parlors, SCE = 1 530; por tanto,

$$s^2 = \text{ECM} = \frac{1\,530}{8} = 191.25$$

proporciona un estimador insesgado de σ^2.

Para estimar σ se calcula la raíz cuadrada de s^2. Al valor que se obtiene, s, se le conoce como **error estándar de estimación**.

ERROR ESTÁNDAR DE ESTIMACIÓN

$$s = \sqrt{\text{ECM}} = \sqrt{\frac{\text{SCE}}{n - 2}} \qquad (14.16)$$

En el ejemplo de Armand´s Pizza Parlors, $s = \sqrt{\text{ECM}} = \sqrt{191.25} = 13.829$. El error estándar de estimación se emplea en el siguiente análisis acerca de las pruebas de significancia de la relación entre x y y.

Prueba *t*

El modelo de regresión lineal simple es $y = \beta_0 + \beta_1 x + \epsilon$. Si x y y están relacionadas linealmente, entonces $\beta_1 \neq 0$. El objetivo de la prueba t es determinar si se puede concluir que $\beta_1 \neq 0$. Para probar la hipótesis siguiente acerca del parámetro β_1 se emplearán los siguientes datos muestrales.

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

Si H_0 es rechazada, se concluirá que $\beta_1 \neq 0$ y que entre las dos variables existe una relación estadísticamente significativa. Si H_0 no es rechazada, habrá evidencia insuficiente para concluir que esta relación significativa existe. La base para esta prueba de hipótesis la proporcionan las propiedades de la distribución de muestreo de b_1, el estimador de β_1 obtenido mediante el método de mínimos cuadrados.

Primero, considere qué ocurriría si para el mismo estudio de regresión se usara una muestra aleatoria diferente. Suponga, por ejemplo, que Armand's Pizza Parlors usa los registros de ventas de una muestra diferente de 10 restaurantes. El análisis de regresión de esta otra muestra dará como resultado una ecuación de regresión parecida a la ecuación de regresión anterior $\hat{y} = 60 + 5x$. Sin embargo, no puede esperarse que se obtenga exactamente la misma ecuación (una ecuación en la que exactamente la intersección con el eje y sea 60 y la pendiente sea 5). Los estimadores b_0 y b_1, obtenidos por el método de mínimos cuadrados, son estadísticos muestrales que tienen su propia distribución de muestreo. A continuación se presentan las propiedades de la distribución de muestreo de b_1.

DISTRIBUCIÓN DE MUESTREO DE b_1

Valor esperado

$$E(b_1) = \beta_1$$

Desviación estándar

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\Sigma(x_i - \bar{x})^2}} \qquad \textbf{(14.17)}$$

Forma de distribución: normal

Observe que el valor esperado de b_1 es igual a β_1, por lo que b_1 es un estimador insesgado de β_1.

Como no se conoce el valor de σ, se obtiene una estimación de σ_{b_1}, que se denota s_{b_1}, estimando σ mediante s en la ecuación (14.17). De esta manera obtenemos el estimador siguiente de σ_{b_1}.

A la desviación estándar de b_1 se le conoce también como error estándar *de b_1. Así, s_{b_1} proporciona una estimación del error estándar de b_1.*

DESVIACIÓN ESTÁNDAR ESTIMADA DE b_1

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} \qquad \textbf{(14.18)}$$

En el ejemplo de Armand's Pizza Parlors, $s = 13.829$. Por tanto, dado que $\Sigma(x_i - \bar{x})^2 = 568$, como se aprecia en la tabla 14.2, tenemos

$$s_{b_1} = \frac{13.829}{\sqrt{568}} = 0.5803$$

que es la desviación estándar estimada de b_1.

La prueba t para determinar si la relación es significativa se basa en el hecho de que el estadístico de prueba

$$\frac{b_1 - \beta_1}{s_{b_1}}$$

sigue una distribución t con $n - 2$ grados de libertad. Si la hipótesis nula es verdadera, entonces $\beta_1 = 0$ y $t = b_1/s_{b_1}$.

Ahora se realizará esta prueba de significancia con los datos de Armand's Pizza Parlors empleando como nivel de significancia $\alpha = 0.01$. El estadístico de prueba es

$$t = \frac{b_1}{s_{b_1}} = \frac{5}{0.5803} = 8.62$$

En los apéndices 14.3 y 14.4 se describe el uso de Minitab y de Excel para calcular el valor-p.

En las tablas de la distribución t encontramos que para $n - 2 = 10 - 2 = 8$ grados de libertad, $t = 3.355$ proporciona un área de 0.005 en la cola superior. Por tanto, el área en la cola superior de la distribución t correspondiente al estadístico de prueba $t = 8.62$ debe ser menor de 0.005. Como ésta es una prueba de dos colas, este valor se duplica y concluimos que el valor-p asociado con $t = 8.62$ debe ser menor a $2(0.005) = 0.01$. Empleando Excel o Minitab se encuentra el valor-$p = 0.000$. Dado que el valor-p es menor que $\alpha = 0.01$, H_0 es rechazada y concluimos que β_1 no es igual a cero. Esto es suficiente evidencia para asegurar que existe una relación significativa entre la población de estudiantes y las ventas trimestrales. A continuación se presenta un resumen de la prueba t de significancia para la regresión lineal simple.

PRUEBA t DE SIGNIFICANCIA PARA LA REGRESIÓN LINEAL SIMPLE

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

ESTADÍSTICO DE PRUEBA t

$$t = \frac{b_1}{s_{b_1}} \qquad \textbf{(14.19)}$$

REGLA DE RECHAZO

Método del valor-p: Rechazar H_0 si el valor-$p \leq \alpha$

Método del valor crítico: Rechazar H_0 si $t \leq -t_{\alpha/2}$ o si $t \geq t_{\alpha/2}$

donde $t_{\alpha/2}$ se toma de la distribución t con $n - 2$ grados de libertad.

Intervalo de confianza para β_1

La fórmula de un intervalo de confianza para β_1 es la siguiente:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

El estimador puntual es b_1 y el margen de error $t_{\alpha/2}s_{b_1}$. El coeficiente de confianza para este intervalo es $1 - \alpha$, y $t_{\alpha/2}$ es el valor t que proporciona un área $\alpha/2$ en la cola superior de la distribución t con $n - 2$ grados de libertad. Suponga, por ejemplo, que en el caso de Armand's Pizza Parlors se desea obtener una estimación de β_1 mediante un intervalo de 99% de confianza. En la tabla 2 del apéndice B encontramos que el valor t correspondiente a $\alpha = 0.01$ y $n - 2 = 10 - 2 = 8$ grados de libertad es $t_{0.005} = 3.355$. Por tanto, la estimación mediante un intervalo de 99% de confianza de β_1 es

$$b_1 \pm t_{\alpha/2}s_{b_1} = 5 \pm 3.355(0.5803) = 5 \pm 1.95$$

o el intervalo que va de 3.05 a 6.95.

Al emplear la prueba t de significancia, las hipótesis probadas fueron

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

Utilizando $\alpha = 0.01$ como nivel de significancia, se puede usar el intervalo de 99% de confianza como alternativa para llegar a la conclusión de la prueba de hipótesis que se obtiene con los datos de Armand's. Como 0, que es el valor hipotético de β_1, no está comprendido en el intervalo de confianza (3.05 a 6.95), H_0 puede ser rechazada y concluimos que entre el tamaño de la población de estudiantes y las ventas trimestrales sí existe una relación estadísticamente significativa. En general, se puede usar un intervalo de confianza para probar cualquier hipótesis de dos colas acerca de β_1. Si el valor hipotético de β_1 está contenido en el intervalo de confianza, H_0 no es rechazada. De lo contrario, es rechazada.

Prueba *F*

Una prueba F basada en la distribución de probabilidad F también puede emplearse para probar la significancia en la regresión. Cuando sólo se tiene una variable independiente, la prueba F lleva a la misma conclusión que la prueba t; es decir, si esta t indica que $\beta_1 \neq 0$ y por tanto existe una relación significativa, la prueba F también indicará que existe esta relación. Pero cuando hay más de una variable independiente, sólo la prueba F puede usarse para probar que existe una relación significativa general.

La lógica detrás del uso de la prueba F para determinar si la relación de regresión es estadísticamente significativa se basa en el desarrollo de dos estimaciones independientes de σ^2. Se explicó cómo ECM proporciona una estimación de σ^2. Si la hipótesis nula H_0: $\beta_1 = 0$ es verdadera, la suma de cuadrados debido a la regresión, SCR, dividida entre sus grados de libertad proporciona otra estimación independiente de σ^2. A esta estimación se le llama *cuadrado medio debido a la regresión* o simplemente *cuadrado medio de la regresión*, y se denota como CMR. En general

$$\text{CMR} = \frac{\text{SCR}}{\text{grados de libertad de la regresión}}$$

En los modelos que se consideran en este libro, el número de grados de libertad de la regresión es siempre igual al número de variables independientes en el modelo:

$$\text{CMR} = \frac{\text{SCR}}{\text{número de variables independientes}} \qquad \textbf{(14.20)}$$

Como en este capítulo sólo se consideran modelos de regresión con una sola variable independiente, tenemos CMR = SCR/1 = SCR. Por tanto, en el ejemplo de Armand's Pizza Parlors, CMR = SCR = 14 200.

Si la hipótesis nula (H_0: $\beta_1 = 0$) es verdadera, CMR y ECM son dos estimaciones independientes de σ^2 y la distribución de muestreo de CMR/ECM sigue una distribución F en la que el

número de grados de libertad en el numerador es igual a 1 y el número de grados de libertad en el denominador es igual a $n - 2$. Por consiguiente, si $\beta_1 = 0$, el valor de CMR/ECM deberá ser un valor cercano a 1. Pero si la hipótesis nula es falsa, ($\beta_1 \neq 0$), CMR sobreestimará σ^2 y el valor de CMR/ECM se inflará; de esta manera valores grandes de CMR/ECM conducirán al rechazo de H_0 y a la conclusión de que la relación entre x y y es estadísticamente significativa.

A continuación se realizará la prueba F en el ejemplo de Armand's Pizza Parlors. El estadístico de prueba es

$$F = \frac{\text{CMR}}{\text{ECM}} = \frac{14\,200}{191.25} = 74.25$$

En la regresión lineal simple, la prueba F y la prueba t proporcionan resultados idénticos.

En la tabla de la distribución F (tabla 4 del apéndice B) se observa que con un grado de libertad en el numerador y $n - 2 = 10 - 2 = 8$ grados de libertad en el denominador, $F = 11.26$ proporciona un área de 0.01 en la cola superior. Por tanto, el área en la cola superior de la distribución F que corresponde al estadístico de prueba $F = 74.25$ debe de ser menor de 0.01. Por tanto, concluimos que el valor-p debe de ser menor a 0.01. Empleando Excel o Minitab se encuentra que el valor-$p = 0.000$. Como el valor-p es menor que $\alpha = 0.01$, H_0 es rechazada y concluimos que entre el tamaño de la población de estudiantes y las ventas trimestrales existe una relación significativa. A continuación se presenta un resumen de la prueba F de significancia para la regresión lineal simple

PRUEBA F DE SIGNIFICANCIA EN EL CASO DE LA REGRESIÓN LINEAL SIMPLE

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

ESTADÍSTICO DE PRUEBA

$$F = \frac{\text{CMR}}{\text{ECM}} \qquad \textbf{(14.21)}$$

REGLA DE RECHAZO

Método del valor-p: Rechazar H_0 si el valor-$p \leq \alpha$

Método del valor crítico: Rechazar H_0 si $F \geq F_\alpha$

donde F_α es un valor de distribución F con 1 grado de libertad en el numerador y $n - 2$ grados de libertad en el denominador.

Si H_0 es falsa, ECM proporciona una estimación insesgada de σ^2 y el CMR sobreestima σ^2. Si H_0 es verdadera, tanto ECM como CMR proporcionan una estimación insesgada de σ^2; en este caso el valor de CMR/ECM es cercano a 1.

En el capítulo 13 vimos que el análisis de varianza (ANOVA) y la **tabla de ANOVA** pueden utilizarse para proporcionar una visión resumida de los cálculos que se emplean en el análisis de varianza. Una tabla de ANOVA similar se emplea para resumir los cálculos de la prueba F de significancia para la regresión. En la tabla 14.5 se presenta la forma general de una tabla de ANOVA para la regresión lineal simple, y en la 14.6 la tabla de ANOVA con los cálculos para la prueba F del ejemplo de Armand's Pizza Parlors. Regresión, error y total son las etiquetas de las tres fuentes de variación, y SCR, SCE y STC son las sumas de cuadrados correspondientes que aparecen en la columna 2. En la columna 3 se indican los grados de libertad 1 para SCR, $n - 2$ para SCE y $n - 1$ para STC. Los valores de CMR y ECM aparecen en la columna 4, mientras que la 5 contiene el valor de F = CMR/ECM, y la 6 el valor-p que corresponde al valor de F de la columna 5. Casi todos los resultados proporcionados por computadora para el análisis de regresión incluyen un resumen de la tabla ANOVA de la prueba F de significancia.

TABLA 14.5 Forma general de la tabla de ANOVA para la regresión lineal simple

En toda tabla para el análisis de varianza, la suma total de cuadrados es la suma de la suma de cuadrados de la regresión más la suma de cuadrados del error; además, el total de los grados de libertad es la suma de los grados de libertad de la regresión más los grados de libertad del error.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	*F*	valor-*p*
Regresión	SCR	1	$\text{CMR} = \frac{\text{SCR}}{1}$	$F = \frac{\text{CMR}}{\text{ECM}}$	
Error	SCE	$n - 2$	$\text{ECM} = \frac{\text{SCE}}{n-2}$		
Total	STC	$n - 1$			

Algunas advertencias acerca de la interpretación de las pruebas de significancia

El análisis de regresión que se usa para identificar la relación entre las variables no puede emplearse como evidencia de una relación de causa y efecto.

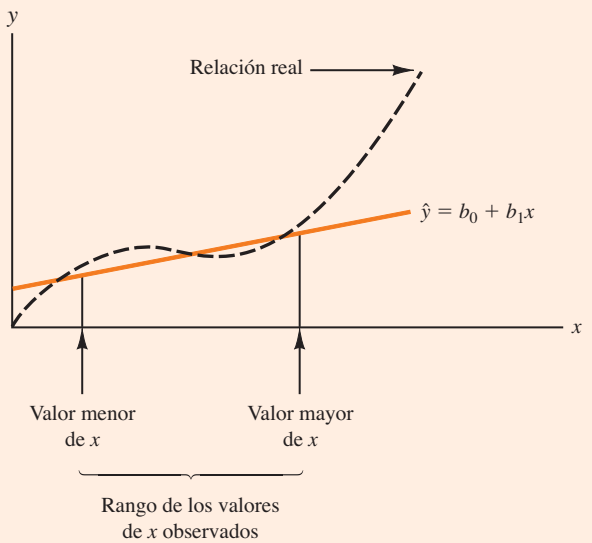
Cuando la hipótesis nula H_0: $\beta_1 = 0$ es rechazada, determinar que la relación que existe entre x y y es significativa no permite que se concluya que existe una relación de causa y efecto entre x y y. Sólo puede concluirse que existe esta relación cuando el analista pueda dar justificaciones teóricas de que, en efecto, la relación es causal. En el ejemplo de Armand's Pizza Parlors, concluimos que existe una relación significativa entre el tamaño de la población de estudiantes x y las ventas trimestrales y; aún más, la ecuación de regresión estimada $\hat{y} = 60 + 5x$ proporciona una estimación de la relación obtenida por el método de mínimos cuadrados. Sin embargo, por el solo hecho de que se haya encontrado que hay una relación estadísticamente significativa entre x y y, no podemos concluir que cambios en la población de estudiantes x *causen* cambios en las ventas trimestrales y. Lo apropiado de concluir que hay una relación de causa y efecto se deja a las justificaciones teóricas de soporte y al buen juicio de los analistas. Los gerentes de Armand's creían que el aumento en la población de estudiantes probablemente fuera una causa del aumento de las ventas trimestrales. Por tanto, el resultado de la prueba de significancia les permite concluir que hay una relación de causa y efecto.

Además, el hecho de que se pueda rechazar H_0: $\beta_1 = 0$ y demostrar que hay significancia estadística no permite concluir que la relación entre x y y sea lineal. Lo único que se puede establecer es que x y y están relacionadas y que la relación lineal explica una porción significativa de la variabilidad de y sobre el rango de los valores de x observados en la muestra. En la figura 14.7 se ilustra esta situación. La prueba de significancia lleva al rechazo de la hipótesis nula H_0: $\beta_1 = 0$ y a la conclusión de que x y y están significantemente relacionadas, pero en la figura se observa que la verdadera relación entre x y y no es lineal. Aunque la aproximación

TABLA 14.6 Tabla ANOVA para el ejemplo de Armand's Pizza Parlors

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	*F*	valor-*p*
Regresión	14 200	1	$\frac{14\,200}{1} = 14\,200$	$\frac{14\,200}{191.25} = 74.25$	0.000
Error	1 530	8	$\frac{1\,530}{8} = 191.25$		
Total	15 730	9			

FIGURA 14.7 Ejemplo de una aproximación lineal para una relación no lineal



lineal proporcionada por $\hat{y} = b_0 + b_1x$ es buena en el rango de los valores de x observados en la muestra, se vuelve deficiente para valores de x fuera de ese rango.

Dada una relación significativa, la ecuación de regresión estimada se puede usar con confianza para predicciones correspondientes a valores de x dentro del rango de los valores de x observados en la muestra. En el ejemplo de Armand's Pizza Parlors, este rango corresponde a los valores de x entre 2 y 26. A menos que haya otras razones que indiquen que el modelo es válido más allá de este rango, las predicciones fuera del rango de la variable independiente deben realizarse con cuidado. En el ejemplo de Armand's Pizza Parlors, como se ha encontrado que la relación de regresión es significativa al nivel de significancia de 0.01, se puede tener confianza de usarla para predecir las ventas de restaurantes en los que la población de estudiantes correspondiente esté en el intervalo de 2 000 a 26 000.

NOTAS Y COMENTARIOS

1. Los supuestos acerca del término del error (sección 14.4) permiten las pruebas de significancia estadística de esta sección. Las propiedades de la distribución de muestreo de b_1 y las subsiguientes pruebas t y F se derivan directamente de estos supuestos.
2. No se debe confundir la significancia estadística con la significancia práctica. Con tamaños de muestra muy grandes se pueden obtener resultados estadísticamente significativos para valores pequeños de b_1; en tales casos hay que tener cuidado al concluir que la relación tiene significancia práctica.
3. Una prueba de significancia para la relación lineal entre x y y también se puede realizar usando el coeficiente de correlación muestral r_{xy}. Con ρ_{xy} para denotar el coeficiente de correlación poblacional, las hipótesis son las siguientes.

$$H_0: \rho_{xy} = 0$$
$$H_a: \rho_{xy} \neq 0$$

Si H_0 es rechazada, podemos concluir que existe una relación significativa. En el apéndice 14.2 se proporcionan los detalles de esta prueba. Sin embargo, las pruebas t y F presentadas en esta sección aportan el mismo resultado que la prueba de significancia usando el coeficiente de correlación. Por consiguiente, si ya se ha realizado una prueba t o una prueba F, no es necesario realizar la de significancia usando el coeficiente de correlación.

Ejercicios

Métodos

AUTOevaluación

23. A continuación se presentan los datos del ejercicio 1.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a) Usando la ecuación (14.15) calcule el error cuadrado medio.
b) Calcule el error estándar de estimación con la ecuación (14.16).
c) Utilizando la ecuación (14.18), calcule la desviación estándar estimada de b_1.
d) Use la prueba *t* para probar las hipótesis siguientes ($\alpha = 0.05$).

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

e) Use la prueba *F* para probar las hipótesis del inciso *d*) empleando 0.05 como nivel de significancia. Presente los resultados en el formato de tabla del análisis de varianza.

24. A continuación se presentan los datos del ejercicio 2.

x_i	3	12	6	20	14
y_i	55	40	55	10	15

a) Usando la ecuación (14.15), calcule el error cuadrado medio.
b) Calcule el error estándar de estimación usando la ecuación (14.16).
c) Con la ecuación (14.18), calcule la desviación estándar estimada de b_1.
d) Use la prueba *t* para probar las hipótesis siguientes ($\alpha = 0.05$).

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

e) Aplique la prueba *F* para probar las hipótesis del inciso *d*) empleando 0.05 como nivel de significancia. Presente los resultados en el formato de tabla de análisis de varianza.

25. A continuación se presentan los datos del ejercicio 3.

x_i	2	6	9	13	20
y_i	7	18	9	26	23

a) ¿Cuál es el valor del error estándar de estimación?
b) Pruebe si existe una relación significativa utilizando la prueba *t*. Use $\alpha = 0.05$.
c) Emplee la prueba *F* para ver si existe una relación significativa. Use $\alpha = 0.05$. ¿Cuál es su conclusión?

Aplicaciones

AUTOevaluación

26. En el ejercicio 18, los datos sobre el promedio obtenido en la licenciatura y los sueldos mensuales fueron los siguientes.

Promedio	Sueldo mensual ($)	Promedio	Sueldo mensual ($)
2.6	3 300	3.2	3 500
3.4	3 600	3.5	3 900
3.6	4 000	2.9	3 600

a) ¿La prueba *t* indica que hay una relación significativa entre el promedio y el sueldo mensual? ¿Cuál es su conclusión? Use $\alpha = 0.05$.
b) Determine si la relación es significativa usando la prueba *F*. ¿Cuál es su conclusión? Utilice $\alpha = 0.05$.
c) Muestre la tabla de ANOVA.

27. La revista *Outside Magazine* probó 10 modelos de mochilas y botas para excursionismo. En la tabla siguiente se presentan los datos de soporte superior (Upper Support) y precio (Price) de cada fabricante y modelo (Manufacturer and Model). El soporte superior se midió con una escala del 1 al 5, en la que 1 significa aceptable y 5 denota excelente soporte superior (*Outside Magazine Buyer's Guide 2001*).

WEB archivo
Boots

Manufacturer and Model	Upper Support	Price ($)
Salomon Super Raid	2	120
Merrell Chameleon Prime	3	125
Teva Challenger	3	130
Vasque Fusion GTX	3	135
Boreal Maigmo	3	150
L.L. Bean GTX Super Guide	5	189
Lowa Kibo	5	190
Asolo AFX 520 GTX	4	195
Raichle Mt. Trail GTX	4	200
Scarpa Delta SL M3	5	220

a) Use estos datos para desarrollar la ecuación de regresión estimada a efecto de estimar el precio de las mochilas y las botas para excursionismo con base en el soporte superior.
b) Empleando un nivel de significancia de 0.05, determine si hay relación entre soporte superior y precio.
c) ¿Confiaría en usar la ecuación de regresión estimada desarrollada en el inciso *a*) para estimar el precio de las mochilas y las botas con base en la evaluación del soporte superior?
d) Estime el precio de una mochila que tiene 4 como evaluación del soporte superior.

WEB archivo
SleepingBags

28. En el ejercicio 8, con los datos x = rango de temperatura (°F) y y = precio ($) de 11 *sleeping bags* fabricados por Bergans of Norway se obtuvo la ecuación de regresión estimada $\hat{y} = 359.2668 - 5.2772x$. Empleando 0.05 como nivel de significancia, determine si el rango de temperatura y el precio están relacionados. Muestre la tabla de ANOVA. ¿Cuál es su conclusión?

29. Remítase al ejercicio 21, en el que se usaron los datos sobre volumen de producción y costos para desarrollar una ecuación de regresión estimada que relacionaba el volumen de producción y los costos de una determinada operación de producción. Use $\alpha = 0.05$ para probar si el volumen de producción está relacionado de manera significativa con los costos totales. Muestre la tabla de ANOVA. ¿Cuál es su conclusión?

30. Remítase al ejercicio 5 donde se utilizaron los siguientes datos para investigar si, por lo general, los altos precios (Price) están o no asociados con las altas calificaciones (Rating) de las ejercitadoras elípticas de acuerdo con la marca y modelo (Brand and Model) (*Consumer Reports*, febrero de 2008).

WEB archivo
Ellipticals

Brand and Model	Price ($)	Rating
Precor 5.31	3 700	87
Keys Fitness CG2	2 500	84
Octane Fitness Q37e	2 800	82
LifeFitness X1 Basic	1 900	74
NordicTrack AudioStrider 990	1 000	73
Schwinn 430	800	69
Vision Fitness X6100	1 700	68
ProForm XP 520 Razor	600	55

Con x = precio ($) y y = calificación, la ecuación de regresión estimada es $\hat{y} = 58.158 + 0.008449x$. Para esos datos, SCE = 173.88 y STC = 756. ¿La evidencia indica que hay una relación significativa entre precio y calificación?

31. En el ejercicio 20, con los datos de x = precio ($) y y = puntuación general de 10 televisores de plasma de 42 pulgadas probados por *Consumer Reports,* se proporcionó la ecuación de regresión estimada $\hat{y} = 12.0169 + 0.0127x$. Con estos datos se obtuvieron SCE = 540.04 y STC = 982.40. Use la prueba F para determinar si el precio de los televisores y la evaluación general están relacionados en un nivel de significancia de 0.05.

14.6 Uso de la ecuación de regresión estimada para estimación y predicción

Al usar el modelo de regresión lineal simple se hace un supuesto acerca de la relación entre x y y. Después se usa el método de mínimos cuadrados para obtener una ecuación de regresión lineal simple estimada. Si existe una relación significativa entre x y y, y el coeficiente de determinación indica que el ajuste es bueno, la ecuación de regresión estimada es útil para estimación y predicción.

Estimación puntual

En el ejemplo de Armand's Pizza Parlors, la ecuación de regresión estimada $\hat{y} = 60 + 5x$ proporciona una estimación de la relación entre el tamaño de la población de estudiantes x y las ventas trimestrales y. Con la ecuación de regresión estimada se puede obtener una estimación puntual del valor medio de y correspondiente a un determinado valor de x, o se puede predecir el valor individual de y que corresponde a un valor determinado de x. Por ejemplo, suponga que los gerentes de Armand's desean una estimación puntual de la media de las ventas trimestrales de todos los restaurantes que se encuentren cerca de campus universitarios con 10 000 estudiantes. Usando la ecuación de regresión estimada $\hat{y} = 60 + 5x$, con $x = 10$ (o 10 000 estudiantes) obtenemos $\hat{y} = 60 + 5(10) = 110$. Por tanto, una estimación puntual de la media de las ventas trimestrales de todos los restaurantes de este ejemplo con 10 000 estudiantes es $110 000.

Ahora suponga que los gerentes de Armand's desean predecir las ventas de un determinado restaurante ubicado cerca de Talbot College, una escuela con 10 000 alumnos. En este caso lo que interesa no es la media correspondiente a todos los restaurantes que están cerca de campus con 10 000 estudiantes, sino únicamente predecir las ventas trimestrales de uno en específico. En realidad, la estimación puntual de un solo valor de y es igual a la estimación puntual de la media de los valores de y. Así, la predicción de las ventas trimestrales de este restaurante en particular será $\hat{y} = 60 + 5(10) = 110$ o $110 000.

Estimación por intervalo

Tanto los intervalos de confianza como los de predicción indican la precisión de los resultados de la regresión. Los intervalos más estrechos proporcionan un mayor grado de precisión.

Las estimaciones puntuales no proporcionan información alguna acerca de la precisión de una estimación. Para eso es necesario obtener estimaciones por intervalo que son muy parecidas a las estudiadas en los capítulos 8, 10 y 11. El primer tipo de estimación por intervalo, el **intervalo de confianza** es una estimación del *valor medio de las y* que corresponden a un valor dado de x. El segundo tipo, el **intervalo de predicción**, se usa cuando se necesita una estimación por intervalo de un *solo valor de y* para un valor dado de x. La estimación puntual del valor medio de y es igual a la estimación puntual de un solo valor de y. Pero las estimaciones por intervalo que se obtienen para estos dos casos son diferentes. En un intervalo de predicción el margen de error es mayor.

Intervalo de confianza para el valor medio de y

Con la ecuación de regresión estimada se obtiene una estimación puntual del valor medio de y que corresponde a un valor dado de x. Para desarrollar un intervalo de confianza se usa la notación siguiente.

x_p = valor particular o determinado de la variable independiente x
y_p = valor de la variable dependiente y que corresponde al valor dado x_p
$E(y_p)$ = valor medio o valor esperado de la variable dependiente y que corresponde al valor dado x_p
$\hat{y}_p = b_0 + b_1 x_p$ = estimación puntual de $E(y_p)$ cuando $x = x_p$

Empleando esta notación para estimar la media de las ventas de todos los restaurantes de Armand's que se encuentran cerca de un campus con 10 000 estudiantes, tenemos que $x_p = 10$, y $E(y_p)$ denota el valor medio desconocido de las ventas de todos los restaurantes para los que $x_p = 10$. La estimación puntual de $E(y_p)$ está dada por $\hat{y}_p = 60 + 5(10) = 110$.

En general, no se puede esperar que $\hat{y}_p$ sea exactamente igual a $E(y_p)$. Para hacer una inferencia acerca de qué tan cerca está $\hat{y}_p$ de la media verdadera $E(y_p)$, es necesario estimar la varianza de $\hat{y}_p$. La fórmula para estimar la varianza de $\hat{y}_p$ para un x_p dado, se denota como $s^2_{\hat{y}_p}$, y es

$$s^2_{\hat{y}_p} = s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right] \tag{14.22}$$

La estimación de la desviación estándar de $\hat{y}_p$ está dada por la raíz cuadrada de la ecuación (14.22).

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \tag{14.23}$$

En los resultados calculados en la sección 14.5 para el ejemplo de Armand's Pizza Parlors se tiene $s = 13.829$. Como $x_p = 10$, $\bar{x} = 14$ y $\Sigma(x_i - \bar{x})^2 = 568$, usando la ecuación (14.23) se obtiene

$$s_{\hat{y}_p} = 13.829\sqrt{\frac{1}{10} + \frac{(10 - 14)^2}{568}}$$
$$= 13.829\sqrt{0.1282} = 4.95$$

A continuación se presenta la fórmula general para obtener un intervalo de confianza.

INTERVALO DE CONFIANZA PARA $E(y_p)$

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p} \tag{14.24}$$

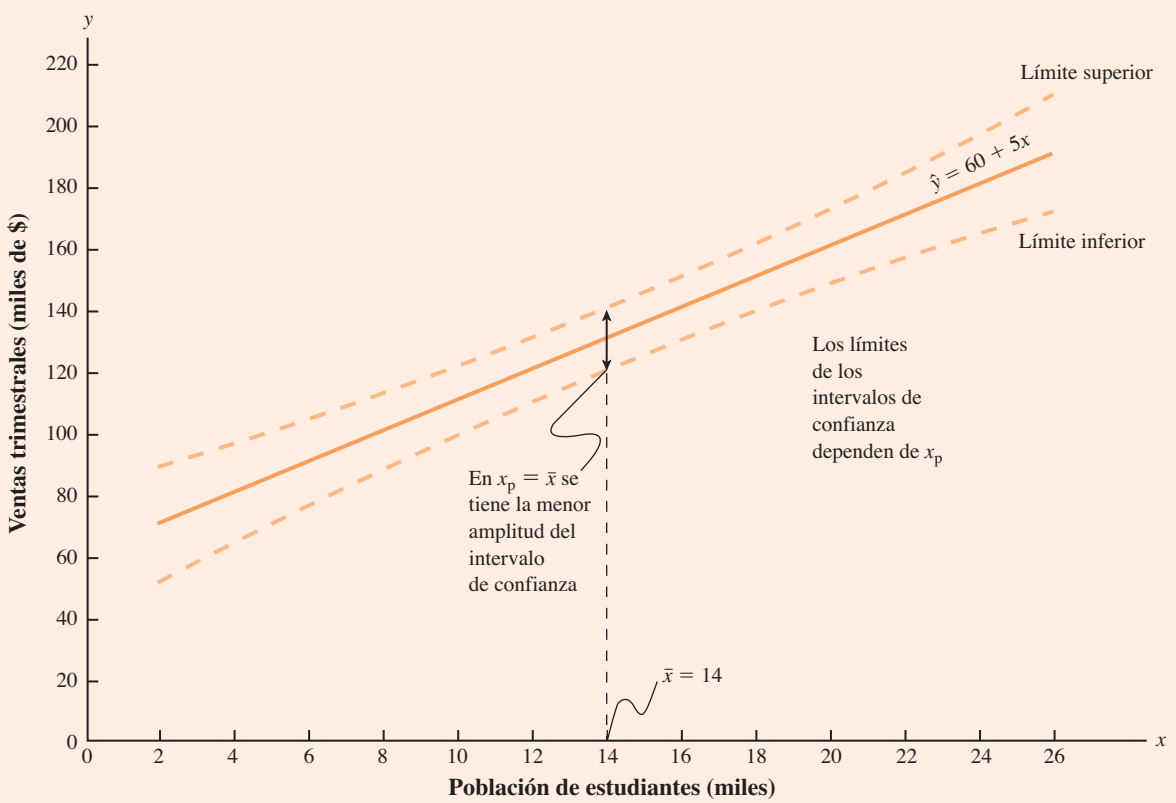
donde el coeficiente de confianza es $1 - \alpha$ y $t_{\alpha/2}$ se basa en una distribución t con $n - 2$ grados de libertad.

El margen de error asociado con esta estimación de intervalo es $t_{\alpha/2} s_{\hat{y}_p}$.

Para obtener, con la fórmula (14.24), un intervalo de 95% de confianza para la media de las ventas trimestrales de todos los restaurantes Armand's ubicados cerca de campus con 10 000 estudiantes, se necesita el valor de t para $\alpha/2 = 0.025$ y $n - 2 = 10 - 2 = 8$ grados de libertad. En la tabla 2 del apéndice B se encuentra $t_{0.025} = 2.306$. Por tanto, como $\hat{y}_p = 110$ y el margen de error de $t_{\alpha/2} s_{\hat{y}_p} = 2.306(4.95) = 11.415$, la estimación del intervalo de 95% de confianza es

$$110 \pm 11.415$$

FIGURA 14.8 Intervalos de confianza para la media de las ventas y correspondientes a valores dados de la población de estudiantes x



En dólares, el intervalo de 95% de confianza para la media de las ventas trimestrales de todos los restaurantes que se encuentran cerca de un campus con 10 000 estudiantes es \$110 000 ± \$11 415. Así, el intervalo de confianza de 95% para la media de las ventas trimestrales cuando el tamaño de la población es 10 000 es de \$98 585 a \$121 415.

Observe que la desviación estándar estimada de $\hat{y}_p$ dada por la ecuación (14.23) es menor cuando $x_p = \bar{x}$ y la cantidad $x_p - \bar{x} = 0$. En este caso, la desviación estándar estimada de $\hat{y}_p$ se convierte en

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(\bar{x} - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} = s\sqrt{\frac{1}{n}}$$

Este resultado implica que se obtiene la mejor o más precisa estimación del valor medio de y cuando $x_p = \bar{x}$. De hecho, entre más alejado esté x_p de $\bar{x}$, mayor será $x_p - \bar{x}$. Como resultado, los intervalos de confianza para el valor medio de y son más amplios a medida que x_p se aleja de $\bar{x}$. En la figura 14.8 se muestra gráficamente este patrón.

Intervalo de predicción para un solo valor de y

Suponga que en lugar del valor medio de las ventas de todos los restaurantes Armand's que se encuentran cerca de campus con 10 000 estudiantes, se busque estimar las ventas de un solo restaurante que se encuentra cerca de Talbot College, una escuela de 10 000 alumnos. Como ya se indicó, la estimación puntual de y_p, el valor de y que corresponde a un valor dado x_p, se obtiene

mediante la ecuación de regresión estimada $\hat{y}_p = b_0 + b_1 x_p$. En el caso del establecimiento de Talbot College, como $x_p = 10$, las ventas trimestrales pronosticadas serán $\hat{y}_p = 60 + 5(10) = 110$ o $110 000. Observe que este valor es el mismo que el obtenido como estimación puntual de la media de las ventas en todos los restaurantes que se encuentran cerca de campus con 10 000 estudiantes.

Para obtener un intervalo de predicción, es necesario determinar primero la varianza correspondiente al uso de $\hat{y}_p$ como estimación de un valor individual de y cuando a $x = x_p$. Esta varianza está formada por la suma de los dos componentes siguientes.

1. La varianza de los valores individuales de y respecto de la media $E(y_p)$, para la cual una estimación está dada por s^2
2. La varianza correspondiente al uso de $\hat{y}_p$ para estimar $E(y_p)$, para la cual una estimación está dada por $s^2_{\hat{y}_p}$

La fórmula para estimar la varianza de un valor individual de y_p, que se denota como s^2_{ind}, es

$$\begin{aligned} s^2_{\text{ind}} &= s^2 + s^2_{\hat{y}_p} \\ &= s^2 + s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right] \\ &= s^2\left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right] \end{aligned} \qquad \textbf{(14.25)}$$

Por tanto, una estimación de la desviación estándar de un solo valor de y_p está dada por

$$s_{\text{ind}} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \qquad \textbf{(14.26)}$$

En el ejemplo de Armand's Pizza Parlors, la desviación estándar estimada que corresponde a la predicción de las ventas de un determinado restaurante ubicado cerca de un campus con 10 000 estudiantes se calcula como sigue

$$\begin{aligned} s_{\text{ind}} &= 13.829\sqrt{1 + \frac{1}{10} + \frac{(10 - 14)^2}{568}} \\ &= 13.829\sqrt{1.1282} \\ &= 14.69 \end{aligned}$$

La fórmula general para un intervalo de predicción es la siguiente.

INTERVALO DE PREDICCIÓN PARA y_p

$$\hat{y}_p \pm t_{\alpha/2} s_{\text{ind}} \qquad \textbf{(14.27)}$$

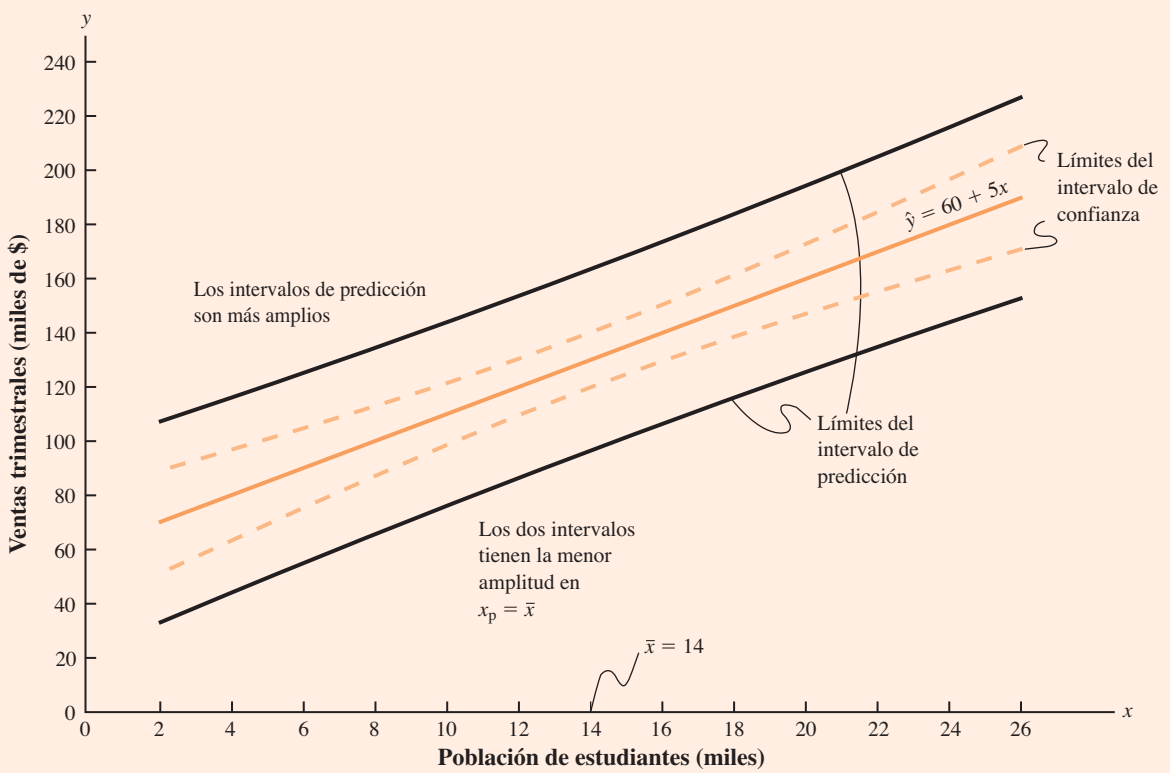
donde el coeficiente de confianza es $1 - \alpha$ y $t_{\alpha/2}$ se basa en una distribución t con $n - 2$ grados de libertad.

El margen de error asociado con este intervalo de estimación es $t_{\alpha/2} s_{\text{ind}}$.

El intervalo de predicción de 95% de las ventas trimestrales del restaurante de Talbot College se encuentra usando $t_{0.025} = 2.306$ y $s_{\text{ind}} = 14.69$. Por tanto, con $\hat{y}_p = 110$ y un margen de error de $t_{\alpha/2} s_{\text{ind}} = 2.306(14.69) = 33.875$, el intervalo de predicción de 95% es

$$110 \pm 33.875$$

FIGURA 14.9 Intervalos de confianza y de predicción para las ventas y que corresponden a valores dados de la población de estudiantes x



En dólares, este intervalo de predicción es de \$110 000 ± \$33 875 o de \$76 125 a \$143 875. Observe que el intervalo de predicción para un solo restaurante que se encuentre cerca de un campus con 10 000 estudiantes es más amplio que el intervalo de confianza para la media de las ventas de todos los restaurantes ubicados cerca de campus con 10 000 estudiantes. Esta diferencia refleja el hecho de que se puede estimar con más precisión la media de y que un solo valor de y.

En general, tanto las líneas de los límites para los intervalos de confianza como las de los límites para los intervalos de predicción presentan cierta curvatura.

Tanto las estimaciones mediante un intervalo de confianza como mediante un intervalo de predicción son más precisas cuando el valor de la variable independiente es $x_p = \bar{x}$. En la figura 14.9 se muestra la forma general de los intervalos de confianza y de predicción que son más anchos.

Ejercicios

Métodos

AUTOevaluación

32. Los datos siguientes provienen del ejercicio 1.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a) Use la ecuación (14.23) para estimar la desviación estándar de $\hat{y}_p$ cuando $x = 4$.
b) Con la expresión (14.24) obtenga un intervalo de confianza de 95% para el valor esperado de y cuando $x = 4$.

c) Use la ecuación (14.26) para estimar la desviación estándar de un valor individual de y cuando $x = 4$.

d) Utilice la expresión (14.27) a efecto de obtener un intervalo de predicción de 95% para y cuando $x = 4$.

33. Los datos siguientes provienen del ejercicio 2.

x_i	3	12	6	20	14
y_i	55	40	55	10	15

a) Estime la desviación estándar de $\hat{y}_p$ cuando $x = 8$.

b) Obtenga un intervalo de 95% de confianza para el valor esperado de y cuando $x = 8$.

c) Estime la desviación estándar de un valor individual de y cuando $x = 8$.

d) Obtenga un intervalo de predicción de 95% para y cuando $x = 8$.

34. Los datos siguientes corresponden al ejercicio 3.

x_i	2	6	9	13	20
y_i	7	18	9	26	23

Obtenga los intervalos de confianza y de predicción de 95% cuando $x = 12$. Explique por qué son diferentes estos dos intervalos.

Aplicaciones

AUTOevaluación

35. En el ejercicio 18, con los datos sobre los promedios de calificaciones x y los sueldos mensuales y se obtuvo la ecuación de regresión estimada $\hat{y} = 1790.5 + 581.1x$.

a) Proporcione un intervalo de 95% de confianza para el sueldo medio inicial de todos los estudiantes cuyo promedio fue 3.0.

b) Desarrolle un intervalo de 95% de predicción para el sueldo medio inicial de Joe Heller, un estudiante cuyo promedio fue 3.0.

WEB archivo

SleepingBags

36. En el ejercicio 8, con los datos $x =$ rango de temperatura (°F) y $y =$ precio ($) de 11 *sleeping bags* fabricados por Bergans of Noway, se obtuvo la ecuación de regresión $\hat{y} = 359.2668 - 5.2772x$. Para estos datos $s = 37.9372$.

a) Obtenga una estimación puntual del precio de un *sleeping bag* cuyo rango de temperatura sea 30.

b) Desarrolle un intervalo de 95% de confianza para la temperatura global media de todos los *sleeping bags* cuyo rango de temperatura sea 30.

c) Suponga que Bergans crea un nuevo modelo cuyo rango de temperatura es 30. Obtenga un intervalo de predicción de 95% para el precio de este nuevo modelo.

d) Explique la diferencia entre sus respuestas a los incisos *b*) y *c*).

37. En el ejercicio 13 se proporcionaron datos sobre el ingreso bruto ajustado x y el monto de las deducciones declaradas por los contribuyentes. Los datos se reportaron en miles de dólares. Como la ecuación de regresión estimada es $\hat{y} = 4.68 + 0.16x$, el punto estimado de un nivel razonable del total de las deducciones declaradas para un contribuyente cuyo ingreso bruto ajustado sea $52 500 es $13 080.

a) Obtenga un intervalo de 95% de confianza para el monto medio del total de las deducciones declaradas de todos los contribuyentes cuyo ingreso bruto ajustado sea $52 500.

b) Obtenga un intervalo de predicción de 95% para el monto del total de las deducciones declaradas de un contribuyente en particular cuyo ingreso bruto ajustado sea $52 500.

c) Si el contribuyente del inciso *b*) solicita un total de $20 400 por deducciones declaradas, ¿se justificaría que un agente fiscal lo requiriera para aplicarle una auditoría?

d) Retome su respuesta al inciso *b*) para dar al agente fiscal una guía acerca del monto del total de las deducciones declaradas que puede solicitar un contribuyente cuyo ingreso bruto ajustado sea $52 500 antes de que sea recomendable una auditoría.

38. Remítase al ejercicio 21, donde se utilizaron los datos de volumen de producción x y costos totales y de una determinada operación de manufactura para obtener la ecuación de regresión estimada $\hat{y} = 1246.67 + 7.6x$.

a) El plan de producción de la empresa muestra que el próximo mes deberán fabricarse 500 unidades. ¿Cuál es la estimación puntual de los costos totales para ese mes?

b) Obtenga un intervalo de predicción de 99% para el costo total de producción del siguiente mes.

c) Si al final del próximo mes el informe de costos de un contador indica que en ese lapso los costos reales de producción fueron de $6 000, ¿deberían preocuparse los gerentes por haber incurrido en costos totales tan altos en ese mes? Analice.

39. En Estados Unidos, casi todo el sistema de tranvías usa vagones eléctricos que corren sobre vías a nivel de la calle. La Federal Transit Administration afirma que el tranvía es uno de los medios de transporte más seguros, ya que arroja una tasa de 0.99 accidentes por millón de millas-pasajero en comparación con 2.29 en los autobuses. En los datos siguientes se proporcionan las millas de vía y la cantidad de pasajeros transportados en los días laborables, en miles, de seis sistemas de tranvías (*USA Today,* 7 de enero de 2003).

Ciudad	Millas de vía	Pasajeros transportados (miles)
Cleveland	15	15
Denver	17	35
Portland	38	81
Sacramento	21	31
San Diego	47	75
San José	31	30
St. Louis	34	42

a) Use estos datos para obtener la ecuación de regresión estimada que podría emplearse para predecir la cantidad de pasajeros, dadas las millas de vía.

b) ¿ La ecuación de regresión estimada proporciona un buen ajuste? Explique.

c) Obtenga un intervalo de 95% de confianza para la media de la cantidad de pasajeros transportados en los días laborables en los sistemas de tranvías con 30 millas de vía.

d) Suponga que Charlotte está considerando la construcción de un sistema de tranvía de 30 millas de vía. Obtenga un intervalo de predicción de 95% para la cantidad de pasajeros transportada en un día laborable con el sistema de Charlotte. ¿Considera usted que el intervalo de predicción que desarrolló pueda ser útil para los encargados de la planeación en Chalotte a fin de anticipar la cantidad de pasajeros en un día laborable en su sistema de tranvías? Explique.

14.7 Solución por computadora

Realizar los cálculos del análisis de regresión sin la ayuda de una computadora puede requerir mucho tiempo. En esta sección se verá cómo puede minimizarse la complicación de tantos cálculos usando software de Minitab.

Los datos sobre población de estudiantes y ventas de Armand's Pizza Parlors se han ingresado en la hoja de cálculo de Minitab. A la variable independiente se le ha llamado Pop y a la variable dependiente Sales (ventas), para facilitar la interpretación de los resultados que proporciona la computadora. Usando Minitab para el ejemplo de Armand's se obtuvieron los resultados que se muestran en la figura 14.10.[2] A continuación se explica cómo interpretarlos.

1. Minitab muestra la ecuación de regresión estimada como Sales = 60.0 + 5.00 Pop.
2. Presenta también una tabla en la que indica el valor de los coeficientes b_0 y b_1, la desviación estándar de cada coeficiente, el valor *t* obtenido al dividir cada coeficiente entre su desviación estándar y el valor-*p* correspondiente a la prueba *t*. Como el valor-*p* es cero (a tres posiciones decimales), los resultados muestrales indican que debe rechazarse la hipótesis nula (H_0: $\beta_1 = 0$). O bien, se puede comparar 8.62 (que aparece en la columna *t*) con el valor crítico apropiado. Este procedimiento para la prueba *t* se describió en la sección 14.5.

[2] En el apéndice 14.3 se explican los pasos a seguir con Minitab para obtener estos resultados.

FIGURA 14.10 Resultados en Minitab para el problema de Armand's Pizza Parlors

```
The regression equation is
Sales = 60.0 + 5.00 Pop

Predictor      Coef   SE Coef      T       p
Constant     60.000     9.226   6.50   0.000
Pop          5.0000    0.5803   8.62   0.000

S = 13.8293     R-sq = 90.3%     R-sq(adj) = 89.1%

Analysis of Variance

SOURCE           DF      SS      MS       F       p
Regression        1   14200   14200   74.25   0.000
Residual Error    8    1530     191
Total             9   15730

Predicted Values for New Observations

New
Obs      Fit   SE Fit        95% C.I.          95% P.I.
  1   110.00     4.95   (98.58, 121.42)   (76.13, 143.87)
```

Ecuación de regresión estimada

Tabla de ANOVA

Estimaciones de intervalo

3. Minitab muestra el error estándar de estimación, $s = 13.8293$, así como información acerca de la bondad de ajuste. Observe que "R-sq = 90.3%" es el coeficiente de determinación expresado como porcentaje. El valor "R-sq(adj) = 89.1%" se verá en el capítulo 15.
4. La tabla ANOVA se presenta bajo el encabezado Analysis of Variance. Minitab usa la etiqueta Residual Error para la fuente de variación del error. Observe que DF son las siglas de *degrees of freedom* (grados de libertad) y que el CMR está dado como 14 200 y ECM como 191. El cociente de estos dos valores proporciona el valor F, que es 74.25, y el correspondiente valor-p 0.000. Como el valor-p es cero (a tres posiciones decimales), la relación entre ventas (Sales) y población (Pop) se considera estadísticamente significativa.
5. La estimación de las ventas esperadas mediante un intervalo de confianza de 95% y la estimación de las ventas de un determinado restaurante cercano a un campus de 10 000 estudiantes mediante un intervalo de predicción de 95% se presentan abajo de la tabla ANOVA. El intervalo de confianza es (98.58, 121.42) y el intervalo de predicción es (76.13, 143.88), como se mostró en la sección 14.6.

Ejercicios

Aplicaciones

AUTOevaluación

40. La división comercial de una firma inmobiliaria realiza un análisis de regresión sobre la relación entre x, rentas brutas anuales (en miles de dólares) y y, precio de venta (en miles de dólares) de edificios de departamentos. Se recabaron datos sobre varias propiedades vendidas últimamente, y con la computadora se obtuvieron los resultados que se muestran enseguida.
a) ¿Cuántos edificios de departamentos había en la muestra?

```
The regression equation is
Y = 20.0 + 7.21 X

Predictor      Coef    SE Coef       T
Constant     20.000     3.2213    6.21
X             7.210     1.3626    5.29

Analysis of Variance

SOURCE            DF         SS
Regression         1    41587.3
Residual Error     7
Total              8    51984.1
```

b) Escriba la ecuación de regresión estimada.

c) ¿Cuál es el valor de s_{b_1}?

d) Use el estadístico *F* para probar la significancia de la relación empleando 0.05 como nivel de significancia.

e) Estime el precio de venta de un edificio de departamentos cuyas rentas anuales brutas son de $50 000.

41. A continuación se presenta una parte de los resultados por computadora de un análisis de regresión en el que se relaciona y = gastos de mantenimiento (dólares por mes) con x = uso (horas por semana) para una marca determinada de terminal de computadora.

a) Escriba la ecuación de regresión estimada.

```
The regression equation is
Y = 6.1092 + .8951 X

Predictor       Coef     SE Coef
Constant      6.1092      0.9361
X             0.8951      0.1490

Analysis of Variance

SOURCE            DF         SS         MS
Regression         1    1575.76    1575.76
Residual Error     8     349.14      43.64
Total              9    1924.90
```

b) Utilice una prueba *t* para determinar si los gastos mensuales de mantenimiento están relacionados con el uso; maneje 0.05 como nivel de significancia.

c) Utilice la ecuación de regresión estimada para predecir los gastos mensuales de mantenimiento de una terminal que se usa 25 horas por semana.

42. Un modelo de regresión que relaciona *x,* el número de vendedores en una sucursal, con *y,* las ventas anuales en esa sucursal (en miles de dólares), proporcionó el resultado de computadora, que se muestra a continuación, empleando análisis de regresión de los datos.

a) Escriba la ecuación de regresión estimada.

```
The regression equation is
Y = 80.0 + 50.00 X

Predictor       Coef    SE Coef       T
Constant        80.0     11.333    7.06
X               50.0      5.482    9.12

Analysis of Variance

SOURCE             DF         SS         MS
Regression          1     6828.6     6828.6
Residual Error     28     2298.8       82.1
Total              29     9127.4
```

b) ¿Cuántas sucursales participaron en el estudio?

c) Calcule el estadístico F y pruebe la significancia de la relación empleando 0.05 como nivel de significancia.

d) Pronostique las ventas anuales de la sucursal de Memphis. En esta sucursal operan 12 vendedores.

43. Los expertos en salud recomiendan que los corredores beban 4 onzas de agua por cada 15 minutos que corran. Aunque las botellas de plástico son una buena alternativa para la mayoría de los corredores, cuando se corre todo un día a campo traviesa se requieren sistemas de hidratación que se lleven en la cintura o sobre la espalda. Además de agua, estos sistemas permiten llevar también alimento o ropa. Por supuesto, a medida que aumenta la capacidad de estos sistemas, aumenta también su peso y precio. En la lista siguiente se proporciona el peso (Weight) en onzas y el precio (Price) de 26 modelos (Model) de sistemas de hidratación (*Trail Runner Gear Guide,* 2003).

WEB archivo
Hydration1

Model	Weight (oz)	Price ($)
Fastdraw	3	10
Fastdraw Plus	4	12
Fitness	5	12
Access	7	20
Access Plus	8	25
Solo	9	25
Serenade	9	35
Solitaire	11	35
Gemini	21	45
Shadow	15	40
SipStream	18	60
Express	9	30
Lightning	12	40
Elite	14	60
Extender	16	65
Stinger	16	65
GelFlask Belt	3	20
GelDraw	1	7
GelFlask Clip-on Holster	2	10
GelFlask Holster SS	1	10
Strider (W)	8	30

Model	Weight (oz)	Price ($)
Walkabout (W)	14	40
Solitude I.C.E.	9	35
Getaway I.C.E.	19	55
Profile I.C.E.	14	50
Traverse I.C.E.	13	60

a) Con estos datos obtenga una ecuación de regresión estimada que pueda ser utilizada para predecir el precio de un sistema de hidratación en función de su peso.
b) Pruebe la significancia de la relación empleando 0.05 como nivel de significancia.
c) ¿Proporciona un buen ajuste la ecuación de regresión estimada? Explique su respuesta.
d) Suponga que la ecuación de regresión estimada obtenida en el inciso *a*) también puede aplicarse a sistemas de hidratación fabricados por otras empresas. Obtenga un intervalo de confianza de 95% para estimar el precio de todos los sistemas de hidratación que pesan 10 onzas.
e) Suponga que la ecuación de regresión estimada obtenida en el inciso *a*) también puede usarse para sistemas de hidratación producidos por otras empresas. Obtenga un intervalo de predicción de 95% para estimar el precio del sistema Back Draft creado por Eastern Mountain Sports y que pesa 10 onzas.

44. Las carreras de autos, las escuelas de manejo de alto desempeño y los programas de educación para conducir realizados por clubes de automovilistas siguen creciendo en popularidad. Todas esas actividades requieren que el participante utilice un casco certificado por la Snell Memorial Foundation, una organización sin fines de lucro dedicada a la investigación, educación, prueba y desarrollo de estándares de seguridad de los cascos. Snell "SA" (Sports Application) evaluó que los cascos profesionales estén diseñados para carreras de autos y proporcionen resistencia a impactos extremos y una alta protección contra el fuego. Uno de los factores clave en la selección de un casco es el peso, ya que los más ligeros generan menos tensión en el cuello. Los datos siguientes muestran peso (Weight) y precio (Price) de 18 cascos (Helmet) SA (sitio web de SoloRacer, 20 de abril de 2008).

Helmet	Weight (oz)	Price ($)
Pyrotect Pro Airflow	64	248
Pyrotect Pro Airflow Graphics	64	278
RCi Full Face	64	200
RaceQuip RidgeLine	64	200
HJC AR-10	58	300
HJC Si-12	47	700
HJC HX-10	49	900
Impact Racing Super Sport	59	340
Zamp FSA-1	66	199
Zamp RZ-2	58	299
Zamp RZ-2 Ferrari	58	299
Zamp RZ-3 Sport	52	479
Zamp RZ-3 Sport Painted	52	479
Bell M2	63	369
Bell M4	62	369
Bell M4 Pro	54	559
G Force Pro Force 1	63	250
G Force Pro Force 1 Grafx	63	280

WEB archivo
RaceHelmets

a) Trace un diagrama de dispersión usando el peso como variable independiente.
b) ¿Parece haber alguna relación entre las dos variables?

c) Obtenga la ecuación de regresión estimada que pueda utilizarse para predecir el precio de acuerdo con el peso.
d) Pruebe la significancia de la relación en un nivel de significancia de 0.05.
e) ¿La ecuación de regresión estimada proporciona un buen ajuste? Explique.

14.8 Análisis de residuales: confirmación de los supuestos del modelo

El **análisis de residuales** *es la herramienta principal para determinar si el modelo de regresión empleado es apropiado.*

Como ya se indicó, el *residual* de la observación *i* es la diferencia entre el valor observado de la variable dependiente (y_i) y el valor estimado de la variable dependiente ($\hat{y}_i$).

RESIDUAL DE LA OBSERVACIÓN *i*

$$y_i - \hat{y}_i \tag{14.28}$$

donde:

y_i = valor observado de la variable dependiente
$\hat{y}_i$ = valor estimado de la variable dependiente

En otras palabras, el residual *i*ésimo es el error que resulta de usar la ecuación de regresión estimada para predecir el valor de la variable dependiente. En la tabla 14.7 se calculan los residuales correspondientes a los datos del ejemplo de Armand's Pizza Parlors. En la segunda columna de la tabla se presentan los valores observados de la variable dependiente, y en la tercera los valores estimados de la variable dependiente obtenidos con la ecuación de regresión estimada $\hat{y} = 60 + 5x$. Un análisis de los residuales correspondientes, en la cuarta columna de la tabla, ayuda a determinar si los supuestos acerca del modelo de regresión son adecuados.

A continuación se revisan los supuestos de regresión en el ejemplo de Armand's Pizza Parlors. Se supuso un modelo de regresión lineal simple.

$$y = \beta_0 + \beta_1 x + \epsilon \tag{14.29}$$

TABLA 14.7 Residuales en el ejemplo de Armand's Pizza Parlors

Población de estudiantes x_i	Ventas y_i	Ventas estimadas $\hat{y}_i = 60 + 5x_i$	Residuales $y_i - \hat{y}_i$
2	58	70	−12
6	105	90	15
8	88	100	−12
8	118	100	18
12	117	120	−3
16	137	140	−3
20	157	160	−3
20	169	160	9
22	149	170	−21
26	202	190	12

Este modelo indica que se supone que las ventas trimestrales (y) son la función lineal del tamaño de la población de estudiantes (x) más un término del error ϵ. En la sección 14.4 se plantearon los siguientes supuestos para el término del error ϵ.

1. $E(\epsilon) = 0$.
2. La varianza de ϵ, que se denota σ^2, es la misma para todos los valores de x.
3. Los valores de ϵ son independientes.
4. El término del error ϵ tiene una distribución normal.

Estos supuestos constituyen la base teórica para las pruebas t y F que se usan para determinar si la relación entre x y y es significativa, y para las estimaciones de los intervalos de confianza y de predicción presentadas en la sección 14.6. Si los supuestos sobre el término del error ϵ son dudosos, quizá las pruebas de hipótesis acerca de la significancia de la relación de regresión y los resultados de la estimación por intervalo no sean válidos.

Los residuales proporcionan la mejor información de ϵ; por tanto, su análisis es muy importante para determinar si los supuestos planteados acerca de ϵ son apropiados. Gran parte del análisis residual se basa en examinar gráficas. En esta sección se estudiarán las siguientes gráficas de residuales.

1. Una gráfica de residuales contra los valores de la variable independiente x.
2. Una gráfica de residuales contra los valores pronosticados para la variable dependiente $\hat{y}$.
3. Una gráfica de residuales estandarizada.
4. Una gráfica de probabilidad normal.

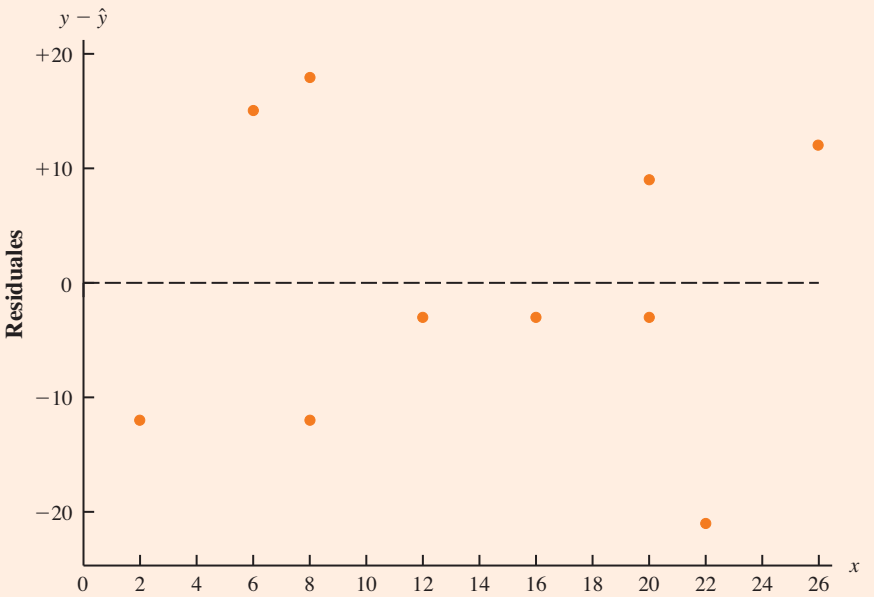
Gráfica de residuales contra x

En la **gráfica de residuales** contra la variable independiente x, los valores de esta variable se representan en el eje horizontal y los valores de los residuales correspondientes se presentan en el eje vertical. Para cada residual se grafica un punto. La primera coordenada de cada punto está dada por el valor x_i y la segunda, por el correspondiente valor del residual $y_i - \hat{y}_i$. En la gráfica de residuales contra x obtenida con los datos de Armand's Pizza Parlors de la tabla 14.7, las coordenadas del primer punto son $(2, -12)$, que corresponden a $x_1 = 2$ y $y_1 - \hat{y}_1 = -12$; las coordenadas del segundo punto son $(6, 15)$, que corresponden a $x_2 = 6$ y $y_2 - \hat{y}_2 = 15$, y así sucesivamente. En la figura 14.11 se muestra la gráfica de residuales obtenida.

Antes de interpretar los resultados se considerarán algunas formas generales que pueden adoptar las gráficas de residuales. En la figura 14.12 se muestran tres ejemplos. Si el supuesto de que la varianza de ϵ es el mismo para todos los valores de x y el modelo de regresión empleado representa adecuadamente la relación entre las variables, el aspecto general de la gráfica de residuales será el de una banda horizontal de puntos como en A. Pero si la varianza de ϵ no es la misma para todos los valores x (por ejemplo, si la variabilidad respecto de la línea de regresión es mayor para valores de x mayores) el aspecto de la gráfica puede ser como en B. En este caso se viola el supuesto de que ϵ tiene una varianza constante. En C se muestra otra forma que puede tomar la gráfica de residuales. En este caso, se concluye que el modelo de regresión empleado no representa adecuadamente la relación entre las variables, y deberá considerarse un modelo de regresión curvilíneo o múltiple.

Regresando a la gráfica de los residuales del ejemplo de Armand's Pizza Parlors de la figura 14.11, estos residuales parecen tener una forma que se aproxima a la de la banda horizontal de la gráfica A de la figura 14.12. Por tanto, concluimos que esta gráfica no muestra evidencias de que los supuestos formulados para el modelo de regresión de Armand's puedan ser dudosos. Concluimos que el modelo de regresión lineal simple empleado es válido.

FIGURA 14.11 Gráfica de residuales contra la variable independiente x para Armand's Pizza Parlors



Para la adecuada interpretación de las gráficas de residuales, la experiencia y el criterio son muy importantes. Es raro que estas gráficas tengan exactamente la forma de uno de los patrones mostrados en la figura 14.12. Sin embargo, los analistas que realizan frecuentemente estudios de regresión y gráficas de residuales se vuelven expertos en reconocer las diferencias entre las formas razonables y las que indican que se puede dudar de los supuestos del modelo. Una gráfica de residuales proporciona una técnica para evaluar la validez de los supuestos en un modelo de regresión.

Gráfica de residuales contra $\hat{y}$

En otras gráficas de residuales los valores pronosticados para la variable dependiente $\hat{y}$ se representan en el eje horizontal y los valores de los residuales en el eje vertical. A cada residual corresponde un punto en la gráfica. La primera coordenada de cada uno de los puntos está dada por $\hat{y}_i$ y la segunda es el valor correspondiente del residual *i*ésimo, $y_i - \hat{y}_i$. Con los datos de Armand's de la tabla 14.7, las coordenadas del primer punto son (70, −12), que corresponden a $\hat{y}_1 = 70$ y $y_1 - \hat{y}_1 = -12$; las coordenadas del segundo punto son (90, 15), y así sucesivamente. En la figura 14.13 se presenta esta gráfica de residuales. Observe que su forma es igual a la de la gráfica de residuales contra la variable independiente x. Éste no es un patrón que pudiera llevar a dudar de los supuestos del modelo. En la regresión lineal simple, tanto la gráfica de residuales contra x como la gráfica de residuales contra $\hat{y}$ tienen la misma forma. En el análisis de regresión múltiple, la gráfica de residuales contra $\hat{y}$ se usa más debido a que se tiene más de una variable independiente.

Residuales estandarizados

Muchas gráficas de residuales que se obtienen con software de computadora utilizan una versión estandarizada de los residuales. Como se demostró en el capítulo anterior, una variable aleatoria se estandariza al sustraerle su media y dividir el resultado entre su desviación estándar. Cuando se emplea el método de mínimos cuadrados, la media de los residuales es cero. Por

FIGURA 14.12 Gráfica de residuales de tres estudios de regresión

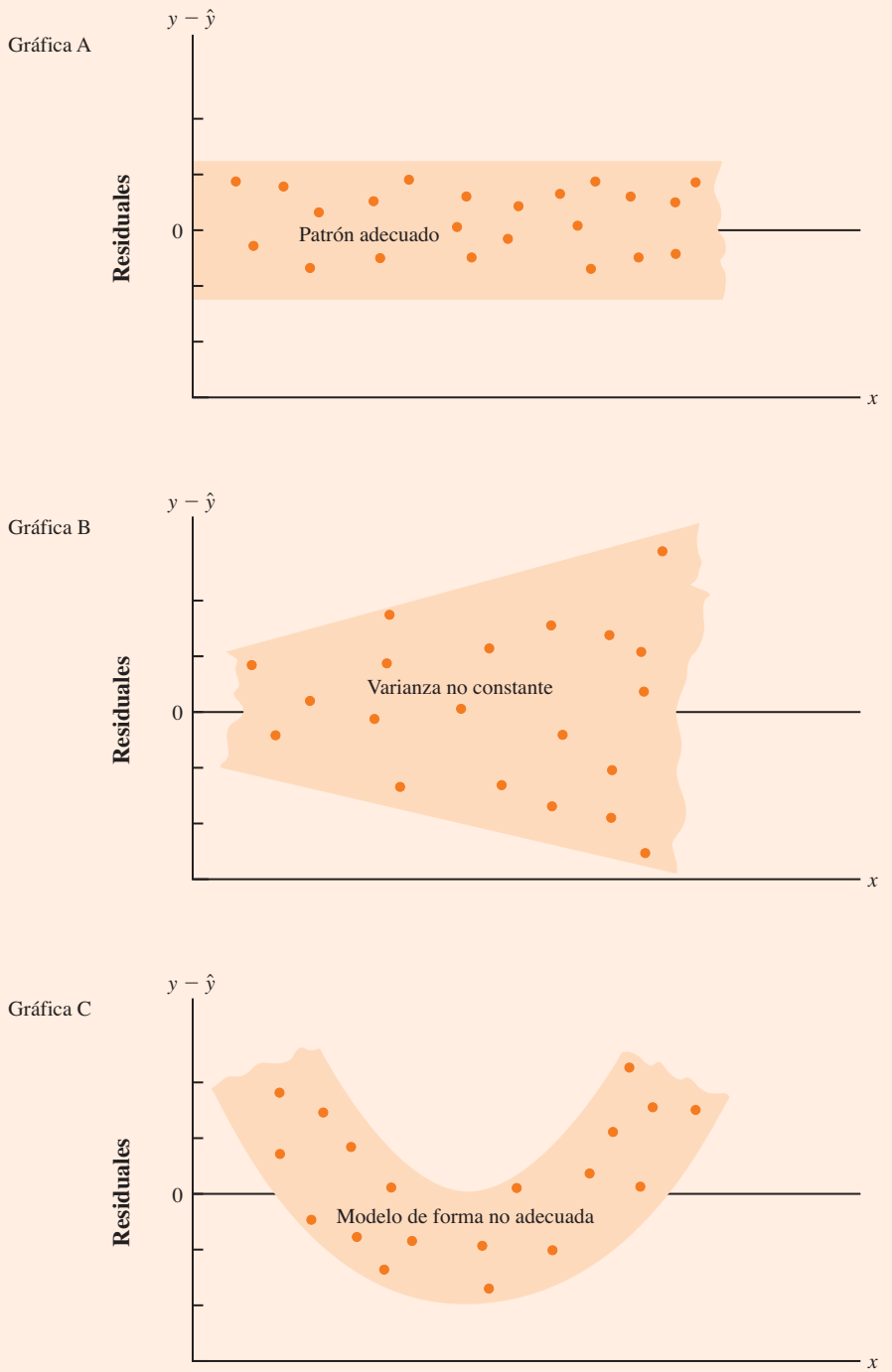
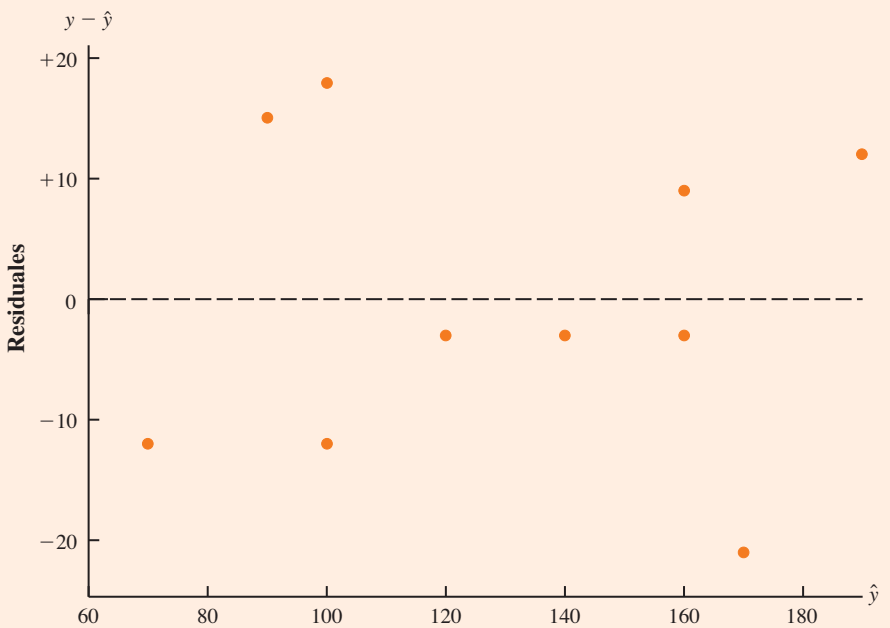


FIGURA 14.13 Gráfica de residuales contra los valores pronosticados de $\hat{y}$ para Armand's Pizza Parlors



consiguiente, para obtener el **residual estandarizado** sólo es necesario dividir cada residual entre su desviación estándar.

Se puede demostrar que la desviación estándar del residual i depende del error estándar de la estimación s y del valor correspondiente de la variable independiente x_i.

DESVIACIÓN ESTÁNDAR DEL RESIDUAL *i*ésimo[3]

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \tag{14.30}$$

donde:

$s_{y_i - \hat{y}_i}$ = desviación estándar del residual i

s = error estándar de estimación

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} \tag{14.31}$$

Observe que la ecuación (14.30) indica que la desviación estándar del residual *i*ésimo depende de x_i debido a la presencia de h_i en la fórmula.[4] Una vez determinada la desviación estándar de cada uno de los residuales, se pueden calcular los residuales estandarizados al dividir cada residual entre sus desviaciones estándar correspondientes.

[3] En realidad, esta ecuación proporciona una estimación de la desviación estándar del residual *i*ésimo, ya que se usa s en lugar de σ.

[4] A h_i se le conoce como el valor de *influencia* de la observación i. El valor de influencia se abordará en la sección 14.9, cuando se consideren las observaciones influyentes.

TABLA 14.8 Cálculo de los residuales estandarizados del ejemplo de Armand's Pizza Parlors

Restaurante i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}$	h_i	$s_{y_i - \hat{y}_i}$	$y_i - \hat{y}_i$	Residuales estandarizados
1	2	−12	144	0.2535	0.3535	11.1193	−12	−1.0792
2	6	−8	64	0.1127	0.2127	12.2709	15	1.2224
3	8	−6	36	0.0634	0.1634	12.6493	−12	−0.9487
4	8	−6	36	0.0634	0.1634	12.6493	18	1.4230
5	12	−2	4	0.0070	0.1070	13.0682	−3	−0.2296
6	16	2	4	0.0070	0.1070	13.0682	−3	−0.2296
7	20	6	36	0.0634	0.1634	12.6493	−3	−0.2372
8	20	6	36	0.0634	0.1634	12.6493	9	0.7115
9	22	8	64	0.1127	0.2127	12.2709	−21	−1.7114
10	26	12	144	0.2535	0.3535	11.1193	12	1.0792
		Total	568					

Nota. En la tabla 14.7 se calcularon los valores de los residuales.

RESIDUAL ESTANDARIZADO DE LA OBSERVACIÓN i

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \qquad \textbf{(14.32)}$$

Pequeñas desviaciones de la normalidad no tienen un gran efecto sobre las pruebas estadísticas utilizadas en el análisis de regresión.

En la tabla 14.8 se presentan los cálculos de los residuales estandarizados con el ejemplo de Armand's Pizza Parlors. Recuerde que ya en cálculos previos se obtuvo $s = 13.829$. La figura 14.14 es la gráfica de los residuales estandarizados contra la variable independiente x.

Esta gráfica permite ver si es correcto el supuesto de que el término del error ϵ tiene distribución normal. Si este supuesto se satisface, debe parecer que la distribución de los residuales estandarizados proviene de una distribución de probabilidad normal estándar.[5] Por tanto, al observar la gráfica de los residuales estandarizados se espera encontrar que aproximadamente 95% de ellos esté entre −2 y +2. En la figura 14.14 vemos que en el ejemplo de Armand's todos los residuales estandarizados se encuentran entre −2 y +2. Así, con base en los residuales estandarizados, esta gráfica no da razones para dudar del supuesto de que ϵ tiene una distribución normal.

Debido al esfuerzo que significa calcular los valores estimados de $\hat{y}$, los residuales y los residuales estandarizados, la mayoría de los paquetes para estadística proporcionan, de manera opcional, estos datos como parte de los resultados de la regresión. Por tanto, las gráficas de residuales se pueden obtener con facilidad. Tratándose de problemas grandes, el software de computadora es la única opción práctica para obtener las gráficas de residuales analizadas en esta sección.

Gráfica de probabilidad normal

Otro enfoque para determinar la validez del supuesto de que el término del error tiene una distribución normal es la **gráfica de probabilidad normal**. Para mostrar cómo se elabora, se presenta el concepto de *puntos normales*.

Suponga que se toman aleatoriamente 10 valores de una distribución de probabilidad normal donde la media es cero y la desviación estándar es uno, y este proceso de muestreo se repite una y otra vez con los 10 valores de cada muestra ordenados de menor a mayor. Por ahora,

[5] Como en la fórmula (14.30) se usa s en lugar de σ, la distribución de probabilidad de los residuales estandarizados no es técnicamente normal. Sin embargo, en la mayoría de los estudios de regresión el tamaño de la muestra es suficientemente grande para que una aproximación normal sea muy buena.

FIGURA 14.14 Gráfica de residuales estandarizados contra la variable independiente x, obtenida con los datos de Armand's Pizza Parlors.

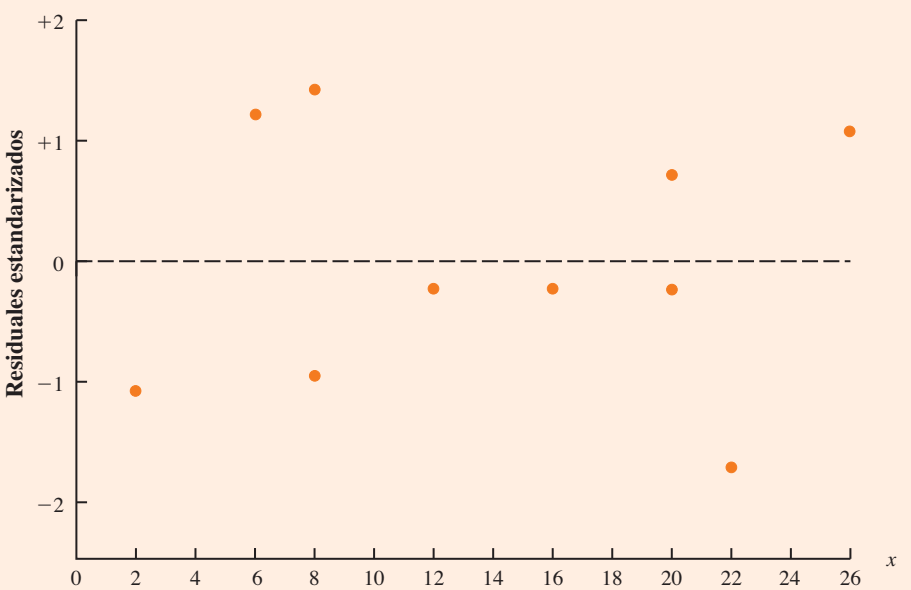


TABLA 14.9

Punto normal para $n = 10$

Estadístico de orden	Punto normal
1	−1.55
2	−1.00
3	−0.65
4	−0.37
5	−0.12
6	0.12
7	0.37
8	0.65
9	1.00
10	1.55

TABLA 14.10

Puntos normales y residuales estandarizados ordenados de Armand's Pizza Parlors

Puntos normales	Residuales estandarizados ordenados
−1.55	−1.7114
−1.00	−1.0792
−0.65	−0.9487
−0.37	−0.2372
−0.12	−0.2296
0.12	−0.2296
0.37	0.7115
0.65	1.0792
1.00	1.2224
1.55	1.4230

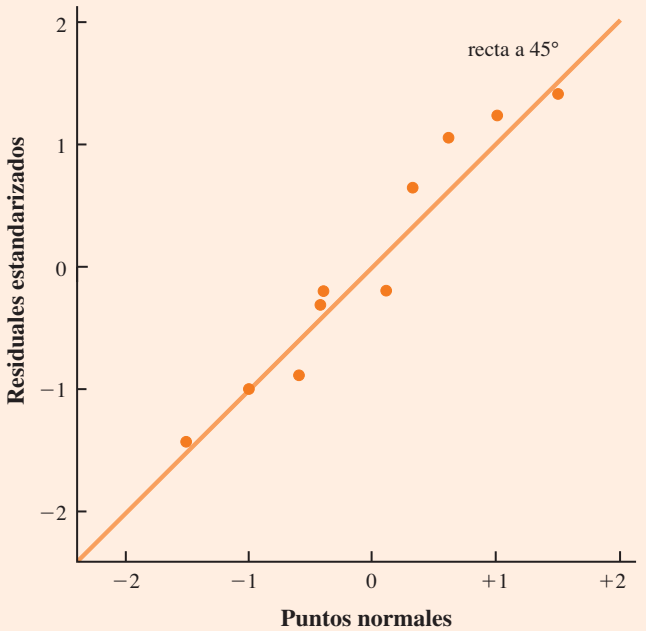
considere únicamente el valor menor de cada muestra. A la variable aleatoria que representa el valor menor de estos diversos muestreos se le conoce como *estadístico de primer orden*.

Los expertos en estadística han demostrado que, en muestras de tamaño 10 tomadas de una distribución de probabilidad normal estándar, el valor esperado del estadístico de primer orden es −1.55. A este valor esperado se le conoce como *punto normal*. En el caso de una muestra de tamaño $n = 10$, hay 10 estadísticos de orden y 10 puntos normales (vea la tabla 14.9). En general, un conjunto de datos que conste de n observaciones tendrá n estadísticos de orden y por tanto n puntos normales.

A continuación vemos el uso de estos 10 puntos normales para determinar si los residuales estandarizados de Armand's Pizza Parlors aparentemente provienen de una distribución de probabilidad normal estándar. Para empezar, se ordenan los 10 residuales estandarizados de la tabla 14.8. En la tabla 14.10 se presentan juntos los 10 puntos normales y los residuales estandarizados ordenados. Si se satisface el supuesto de normalidad, el menor residual estandarizado deberá tener un valor parecido al del menor punto normal, el siguiente residual deberá tener un valor similar al del siguiente punto normal, y así sucesivamente. En el caso en que los residuales estandarizados se encuentren distribuidos de una manera aproximadamente normal, en una gráfica en la que los puntos normales correspondan al eje horizontal y los residuales estandarizados al eje vertical, los puntos estarán situados cerca de una línea recta de 45 grados que pase por el origen. A esta gráfica se le conoce como *gráfica de probabilidad normal.*

La figura 14.15 ilustra la gráfica de probabilidad normal del ejemplo de Armand's Pizza Parlors. Para determinar si el patrón observado se desvía lo suficiente de la recta como para concluir que los residuales estandarizados no provienen de una distribución de probabilidad normal, habrá que emplear el propio criterio. En la figura, todos los puntos se agrupan cerca de esta recta. Se concluye, por tanto, que el supuesto de que los términos del error tienen una distribución de probabilidad normal es razonable. En general, entre más cerca de la recta a 45° se agrupen los puntos, más fuerte es la evidencia a favor del supuesto de normalidad. Cualquier curvatura sustancial en la gráfica es evidencia de que los residuales no provienen de una distribución normal. Tanto los puntos normales como la correspondiente gráfica de probabilidad normal pueden obtenerse fácilmente empleando software como Minitab.

FIGURA 14.15 Gráfica de probabilidad normal obtenida con los datos de Armand's Pizza Parlors



NOTAS Y COMENTARIOS

1. Las gráficas de residuales y de probabilidad normal se usan para confirmar los supuestos de un modelo de regresión. Si en esta revisión se encuentra que uno o más supuestos son dudosos, habrá que considerar un modelo de regresión diferente o una transformación de los datos. Cuando se violan los supuestos, las medidas a tomar deben basarse en un criterio adecuado; las recomendaciones de un experto en estadística pueden ser útiles.
2. El análisis de residuales es el principal método estadístico para verificar que los supuestos del modelo de regresión sean válidos. Aun cuando no se encuentre ninguna violación, esto no necesariamente implica que el modelo vaya a proporcionar buenas predicciones. Pero si además existen otras pruebas estadísticas que favorezcan la conclusión de significancia y el coeficiente de determinación es grande, deberá ser posible obtener buenas estimaciones y predicciones empleando la ecuación de regresión estimada.

Ejercicios

Métodos

AUTOevaluación

45. Dados los datos de las dos variables x y y.

x_i	6	11	15	18	20
y_i	6	8	12	20	30

a) Obtenga una ecuación de regresión estimada para estos datos.
b) Calcule los residuales.

c) Trace una gráfica de residuales contra la variable independiente *x*. ¿Los supuestos acerca de los términos del error parecen satisfacerse?
d) Calcule los residuales estandarizados.
e) Elabore una gráfica de residuales estandarizados contra $\hat{y}$. ¿Qué conclusiones puede formular de esta gráfica?

46. En un estudio de regresión se emplearon los datos siguientes.

Observación	x_i	y_i	Observación	x_i	y_i
1	2	4	6	7	6
2	3	5	7	7	9
3	4	4	8	8	5
4	5	6	9	9	11
5	7	4			

a) Obtenga una ecuación de regresión estimada para estos datos.
b) Trace una gráfica de residuales. ¿Los supuestos acerca del término del error parecen satisfacerse?

Aplicaciones

AUTOevaluación

47. A continuación se presentan datos sobre los gastos en publicidad y los ingresos (en miles de dólares) del restaurante Four Seasons.

Gastos en publicidad	Ingresos
1	19
2	32
4	44
6	40
10	52
14	53
20	54

a) Sea *x* igual a gastos en publicidad y *y* igual a ingresos. Utilice el método de mínimos cuadrados para obtener una recta que aproxime la relación entre las dos variables.
b) Empleando como nivel de significancia 0.05, pruebe si los ingresos y los gastos en publicidad están relacionados.
c) Elabore una gráfica de residuales de $y - \hat{y}$ contra $\hat{y}$. Use el resultado del inciso *a*) para obtener los valores de $\hat{y}$.
d) ¿Qué conclusiones se pueden formular del análisis de residuales? ¿Se puede aplicar este modelo o se debe buscar uno mejor?

48. Remítase al ejercicio 7, donde se obtuvo una ecuación de regresión estimada que relaciona los años de experiencia con las ventas anuales.
a) Calcule los residuales y trace una gráfica de residuales para este problema.
b) A la luz de la gráfica, ¿Los supuestos acerca de los términos del error parecen razonables?

49. Las ventas recientes de casas familiares en San Antonio proporcionan los datos que se listan a continuación acerca de la extensión en pies cuadrados (Square Footage) y precio (Price) de los inmuebles (sitio web de San Antonio Realty Watch, noviembre de 2008).

WEB archivo
HomePrices

Square Footage	Price ($)
1580	142 500
1572	145 000
1352	115 000
2224	155 900
1556	95 000
1435	128 000
1438	100 000
1089	55 000
1941	142 000
1698	115 000
1539	115 000
1364	105 000
1979	155 000
2183	132 000
2096	140 000
1400	85 000
2372	145 000
1752	155 000
1386	80 000
1163	100 000

a) Obtenga una ecuación de regresión estimada que pueda utilizarse para pronosticar los precios de venta dada la extensión en pies cuadrados.
b) Construya una gráfica de residuales estandarizados contra la variable independiente.
c) A la luz de la gráfica, ¿los supuestos acerca de los términos del error y de la forma del modelo parecen razonables?

14.9

Análisis de residuales: observaciones atípicas y observaciones influyentes

En la sección 14.8 se mostró cómo emplear el análisis de residuales para determinar violaciones a los supuestos del modelo de regresión. En esta sección se retoma este análisis para identificar observaciones que se pueden clasificar como atípicas o como especialmente influyentes sobre la ecuación de regresión estimada. También se analizan algunos pasos que deben seguirse cuando se presentan tales observaciones.

Detección de observaciones atípicas

La figura 14.16 es un diagrama de dispersión de un conjunto de datos que contiene una **observación atípica**, un dato (una observación) que no sigue la tendencia del resto de los datos. Las observaciones atípicas representan observaciones sospechosas que requieren un análisis cuidadoso. Pueden consistir de datos erróneos; si es así, deben ser corregidos. Puede tratarse de una violación a los supuestos del modelo; si es el caso, habrá que considerar otro modelo. Por último, puede tratarse, simplemente, de valores inusuales que se presenten por casualidad. En ese caso, esos valores deberán conservarse.

Para ilustrar el proceso de detección de las observaciones atípicas, considere el conjunto de datos de la tabla 14.11; la figura 14.17 muestra el diagrama de dispersión respectivo. Con excepción de la observación 4 ($x_4 = 3, y_4 = 75$), estos datos parecen seguir un patrón que indica una relación lineal negativa. En efecto, dado el patrón del resto de los datos, se esperaría que y_4 fuera mucho más pequeño, por lo que a esta observación se le considera atípica. En el caso de la regresión lineal simple, las observaciones atípicas pueden detectarse mediante un simple examen del diagrama de dispersión.

Para detectar observaciones atípicas también se usan los residuales estandarizados. Si una observación se aleja mucho del patrón del resto de los datos (por ejemplo, la observación atípica de la figura 14.16), el valor absoluto del correspondiente residual estandarizado será grande.

FIGURA 14.16 Conjunto de datos con una observación atípica

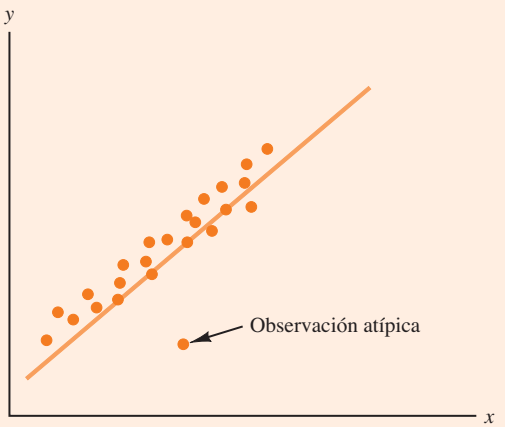


TABLA 14.11
Conjunto de datos para ilustrar el efecto de una observación atípica

x_i	y_i
1	45
1	55
2	50
3	75
3	40
3	45
4	30
4	35
5	25
6	15

Mucho del software identifica de manera automática las observaciones cuyos residuales tienen un valor absoluto grande. En la figura 14.18 se presentan los resultados de Minitab para el análisis de regresión de los datos de la tabla 14.11. En la penúltima fila se lee que el residual estandarizado de la observación 4 es 2.67. Minitab proporciona una lista de todas las observaciones cuyo residual estandarizado sea menor a -2 o mayor a $+2$ en la sección Unusual Observations de la pantalla; en tales casos la observación aparece en una fila aparte con una R al lado del residual estandarizado, como se observa en la figura 14.18. Si los errores están distribuidos normalmente, sólo 5% de los residuales estandarizados se encontrará fuera de estos límites.

Para decidir qué hacer con una observación atípica, primero hay que verificar si es correcta. Puede ser que se trate de un error incurrido al anotar los datos o al ingresarlos a la computadora. Suponga, por ejemplo, que al verificar la observación atípica de la figura 14.17, se encuentra que hubo un error; el valor correcto de la observación 4 es $x_4 = 3$, $y_4 = 30$. En la figura 14.19 se presenta el resultado que proporciona Minitab una vez corregido el valor de y_4.

FIGURA 14.17 Diagrama de dispersión de un conjunto de datos con observación atípica

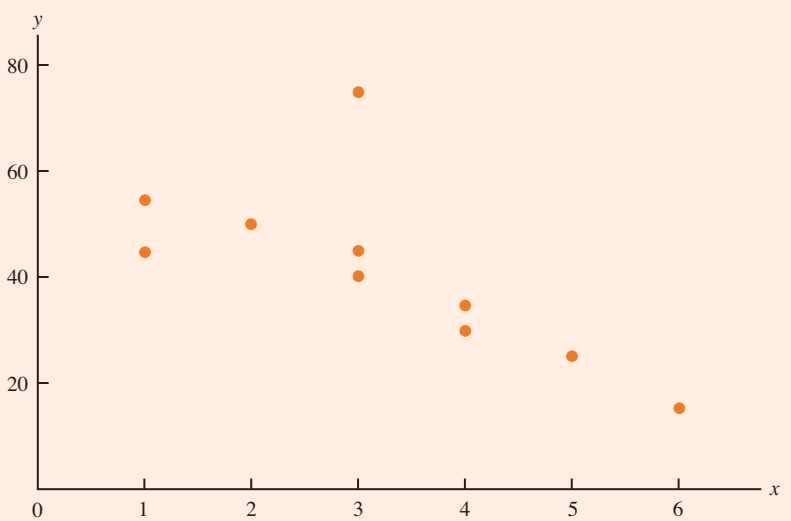


FIGURA 14.18 Resultado de Minitab para el análisis de regresión de un conjunto de datos con una observación atípica

```
The regression equation is
y = 65.0 - 7.33 x

Predictor    Coef  SE Coef      T      p
Constant   64.958    9.258   7.02  0.000
X          -7.331    2.608  -2.81  0.023

S = 12.6704   R-sq = 49.7%   R-sq(adj) = 43.4%

Analysis of Variance

SOURCE          DF      SS      MS     F      p
Regression       1  1268.2  1268.2  7.90  0.023
Residual Error   8  1284.3   160.5
Total            9  2552.5

Unusual Observations
Obs     x      y    Fit  SE Fit  Residual  St Resid
  4  3.00  75.00  42.97    4.04     32.03      2.67R

R denotes an observation with a large standardized residual.
```

FIGURA 14.19 Resultados de Minitab para un conjunto de datos con una observación atípica ya corregida

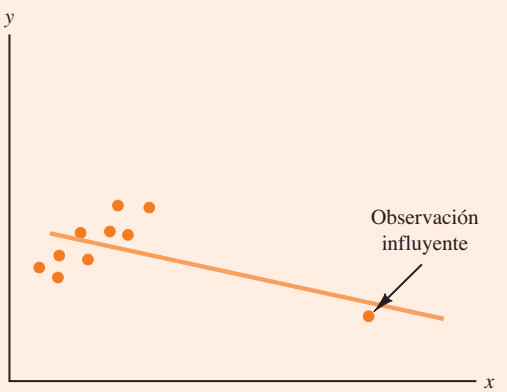
```
The regression equation is
Y = 59.2 - 6.95 X

Predictor    Coef  SE Coef      T      p
Constant   59.237    3.835  15.45  0.000
X          -6.949    1.080  -6.43  0.000

S = 5.24808   R-sq = 83.8%   R-sq(adj) = 81.8%

Analysis of Variance

SOURCE          DF      SS      MS      F      p
Regression       1  1139.7  1139.7  41.38  0.000
Residual Error   8   220.3    27.5
Total            9  1360.0
```

Se observa que el dato incorrecto afecta de forma significativa la bondad de ajuste. Con el dato correcto, el valor de R-sq aumenta de 49.7% a 83.8%, y el de b_0 disminuye de 64.958 a 59.237. La pendiente de la recta cambia de -7.331 a -6.949. La identificación de los datos atípicos permite corregir errores y mejora los resultados de la regresión.

Detección de observaciones influyentes

Algunas veces una o más observaciones tienen una influencia fuerte sobre los resultados. En la figura 14.20 se muestra un ejemplo de una **observación influyente** en una regresión lineal simple. La recta de regresión estimada tiene pendiente negativa, pero si la observación influ-

FIGURA 14.20 Conjunto de datos con una observación influyente



yente se elimina del conjunto de datos, la pendiente cambia a positiva y la intersección con el eje y es menor. Es claro que esta sola observación tiene mucha más influencia sobre la recta de regresión estimada que cualquiera otra; el efecto de la eliminación de cualquiera de las otras observaciones sobre la ecuación de regresión estimada es muy pequeño.

Cuando sólo se tiene una variable independiente, las observaciones influyentes pueden identificarse mediante un diagrama de dispersión. Una observación de este tipo puede ser atípica (cuyo valor de y se desvía sustancialmente de la tendencia general), puede ser un valor de x muy alejado de la media (por ejemplo, vea la figura 14.20) o tratarse de la combinación de estos dos factores (un valor de y algo fuera de la tendencia y un valor de x un poco extremo).

Las observaciones influyentes deben examinarse con cuidado, dado el gran efecto que tienen sobre la ecuación de regresión estimada. Lo primero que hay que hacer es verificar que no se haya cometido algún error al recolectar los datos. Si se cometió, se corrige y se obtiene una nueva ecuación de regresión estimada. Si la observación es válida, podemos considerarnos afortunados. Tal dato, cuando es correcto, contribuye a una mejor comprensión del modelo adecuado y conduce a una mejor ecuación de regresión estimada. En la figura 14.20, la presencia de la observación influyente, si es correcta, llevará a tratar de obtener datos con valores x intermedios que permitan comprender mejor la relación entre x y y.

Las observaciones en las que la variable independiente toma valores extremos se denominan **puntos (datos, observaciones) de gran influencia**. La observación influyente de la figura 14.20 es un punto de gran influencia. La influencia de una observación depende de qué tan lejos está el valor de la variable independiente de su media. En el caso de una sola variable independiente, la influencia de la observación i, que se denota h_i, se calcula mediante la ecuación (14.33).

TABLA 14.12
Conjunto de datos con una observación de gran influencia

x_i	y_i
10	125
10	130
15	120
20	115
20	120
25	110
70	100

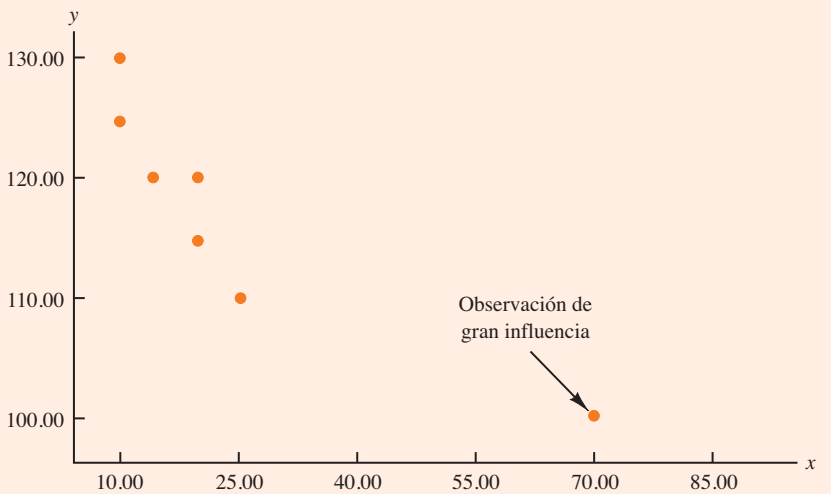
INFLUENCIA DE LA OBSERVACIÓN i

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} \qquad \textbf{(14.33)}$$

Con base en esta fórmula, es claro que entre más alejada se encuentre x_i de su media $\bar{x}$, mayor será la influencia de la observación i.

Mucho del software para estadística identifica automáticamente los puntos de gran influencia como parte de los resultados de regresión estándar. Para ilustrar cómo Minitab identifica los puntos de gran influencia, se considerará el conjunto de datos de la tabla 14.12.

FIGURA 14.21 Diagrama de dispersión del conjunto de datos con un punto de gran influencia



Al revisar la figura 14.21, que es el diagrama de dispersión del conjunto de datos presentado en la tabla 14.12, vemos que la observación 7 ($x = 70$, $y = 100$) tiene un valor extremo de x. Por tanto, es de esperarse que sea identificado como un punto de gran influencia. La influencia de esta observación se calcula usando la ecuación (14.33) como sigue.

$$h_7 = \frac{1}{n} + \frac{(x_7 - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} = \frac{1}{7} + \frac{(70 - 24.286)^2}{2621.43} = 0.94$$

El software es esencial para efectuar los cálculos que permiten identificar las observaciones influyentes. Aquí se analiza la regla de selección que emplea Minitab.

En el caso de la regresión lineal simple, Minitab identifica como observaciones de gran influencia aquellas para las que $h_i > 6/n$ o 0.99, lo que sea menor. En el conjunto de datos de la tabla 14.12, $6/n = 6/7 = 0.86$. Como $h_7 = 0.94 > 0.86$. Minitab identificará la observación 7 como una observación cuyo valor de x tiene una gran influencia. En la figura 14.22 se presenta el resultado que proporciona Minitab para el análisis de regresión de este conjunto de datos. A la 7 ($x = 70$, $y = 100$) la identifica como una observación de gran influencia y la presenta en una fila especial en la parte inferior de los resultados con una X en el margen derecho.

Las observaciones influyentes debido a la interacción de una observación de gran influencia y de residuales grandes suelen ser difíciles de detectar. Existen procedimientos de diagnóstico que toman en cuenta ambos aspectos para determinar si una observación es influyente. En el capítulo 15 se estudiará uno de estos procedimientos, el estadístico D de Cook.

NOTAS Y COMENTARIOS

Una vez identificada una observación como potencialmente influyente debido a que tiene un residual grande o por ser de gran influencia, su impacto sobre la ecuación de regresión estimada debe ser evaluado. En libros más avanzados se presentan los métodos de diagnóstico apropiados. Pero cuando no se está familiarizado con el material más avanzado, un procedimiento sencillo es realizar el análisis de regresión con y sin esa observación. Este método permite apreciar el impacto que tiene la observación potencialmente influyente sobre el resultado.

FIGURA 14.22 Resultado de Minitab para el conjunto de datos con una observación de gran influencia

```
The regression equation is
y = 127 - 0.425 x

Predictor       Coef  SE Coef      T      p
Constant     127.466    2.961  43.04  0.000
X           -0.42507  0.09537  -4.46  0.007

S = 4.88282   R-sq = 79.9%   R-sq(adj) = 75.9%

Analysis of Variance

SOURCE          DF      SS      MS      F      p
Regression       1  473.65  473.65  19.87  0.007
Residual Error   5  119.21   23.84
Total            6  592.86

Unusual Observations
Obs     x       y    Fit  SE Fit  Residual  St Resid
  7  70.0  100.00  97.71    4.73      2.29      1.91 X

X denotes an observation whose X value gives it large influence.
```

Ejercicios

Métodos

AUTOevaluación

50. Considere los datos siguientes para las variables x y y.

x_i	135	110	130	145	175	160	120
y_i	145	100	120	120	130	130	110

a) Calcule los residuales estandarizados de estos datos. ¿Éstos incluyen alguna observación atípica? Explique.

b) Trace una gráfica de residuales estandarizados contra $\hat{y}$. ¿Se observa en esta gráfica alguna observación atípica?

c) Con estos datos elabore un diagrama de dispersión. ¿Se aprecia en este diagrama alguna observación atípica? En general, ¿qué consecuencias tienen estos hallazgos para la regresión lineal simple?

51. Considere los datos siguientes para las variables x y y.

x_i	4	5	7	8	10	12	12	22
y_i	12	14	16	15	18	20	24	19

a) Calcule los residuales estandarizados de estos datos. ¿Identifica alguna observación atípica? Explique.

b) Calcule los valores de influencia para estos datos. ¿Parece haber alguna observación influyente? Explique.

c) Elabore un diagrama de dispersión con estos datos. ¿Se advierte alguna observación atípica? Explique.

Aplicaciones

AUTO evaluación

52. Los datos siguientes muestran los gastos en medios (Media Expenditures) en millones de $ y los embarques (Shipments) en millones de barriles de 10 importantes marcas (Brand) de cerveza.

WEB archivo
Beer

Brand	Media Expenditures ($ millions)	Shipments
Budweiser	120.0	36.3
Bud Light	68.7	20.7
Miller Lite	100.1	15.9
Coors Light	76.6	13.2
Busch	8.7	8.1
Natural Light	0.1	7.1
Miller Genuine Draft	21.5	5.6
Miller High Life	1.4	4.4
Busch Light	5.3	4.3
Milwaukee's Best	1.7	4.3

a) Obtenga una ecuación de regresión estimada con estos datos.
b) Emplee el análisis residual para hallar observaciones atípicas u observaciones influyentes. Resuma brevemente sus hallazgos y conclusiones.

53. Los expertos en salud recomiendan que los corredores beban 4 onzas de agua por cada 15 minutos que corran. Las personas que corren de tres a ocho horas requieren sistemas de hidratación que se llevan en la cintura o sobre la espalda. En los datos siguientes se proporciona el volumen en onzas de fluido (Volume/fl oz) y el precio (Price) de 26 modelos (Model) de sistemas de hidratación que se llevan en la cintura o sobre la espalda (*Trail Runner Gear Guide,* 2003).

WEB archivo
Hydration2

Model	Volume (fl oz)	Price ($)
Fastdraw	20	10
Fastdraw Plus	20	12
Fitness	20	12
Access	20	20
Access Plus	24	25
Solo	20	25
Serenade	20	35
Solitaire	20	35
Gemini	40	45
Shadow	64	40
SipStream	96	60
Express	20	30
Lightning	28	40
Elite	40	60
Extender	40	65
Stinger	32	65
GelFlask Belt	4	20
GelDraw	4	7
GelFlask Clip-on Holster	4	10
GelFlask Holster SS	4	10
Strider (W)	20	30
Walkabout (W)	230	40
Solitude I.C.E.	20	35
Getaway I.C.E.	40	55
Profile I.C.E.	64	50
Traverse I.C.E.	64	60

a) Desarrolle una ecuación de regresión estimada para pronosticar el precio de un sistema de hidratación, dado su volumen de líquido.

b) Use el análisis residual para determinar si hay observaciones atípicas u observaciones influyentes. Resuma sus hallazgos y conclusiones.

54. En la tabla siguiente se presentan datos del ingreso (Revenue) anual y el valor (Value) estimado del equipo (Team) en millones de $ para los 32 equipos de la National Football League (sitio web de Forbes, febrero de 2009).

WEB archivo
NFLValues

Team	Revenue ($ millions)	Value ($ millions)
Arizona Cardinals	203	914
Atlanta Falcons	203	872
Baltimore Ravens	226	1 062
Buffalo Bills	206	885
Carolina Panthers	221	1 040
Chicago Bears	226	1 064
Cincinnati Bengals	205	941
Cleveland Browns	220	1 035
Dallas Cowboys	269	1 612
Denver Broncos	226	1 061
Detroit Lions	204	917
Green Bay Packers	218	1 023
Houston Texans	239	1 125
Indianapolis Colts	203	1 076
Jacksonville Jaguars	204	876
Kansas City Chiefs	214	1 016
Miami Dolphins	232	1 044
Minnesota Vikings	195	839
New England Patriots	282	1 324
New Orleans Saints	213	937
New York Giants	214	1 178
New York Jets	213	1 170
Oakland Raiders	205	861
Philadelphia Eagles	237	1 116
Pittsburgh Steelers	216	1 015
San Diego Chargers	207	888
San Francisco 49ers	201	865
Seattle Seahawks	215	1 010
St. Louis Rams	206	929
Tampa Bay Buccaneers	224	1 053
Tennessee Titans	216	994
Washington Redskins	327	1 538

a) Trace un diagrama de dispersión con el ingreso en el eje horizontal y el valor en el eje vertical. Inspeccione el diagrama: ¿parece que hay observaciones atípicas u observaciones influyentes en los datos?

b) Obtenga una ecuación de regresión estimada para predecir el valor de un equipo, dado el valor del ingreso anual.

c) Use el análisis residual para determinar si hay observaciones atípicas u observaciones influyentes. Resuma brevemente sus hallazgos y conclusiones.

Resumen

En este capítulo se estudió el análisis de regresión para determinar cómo es la relación entre una variable dependiente y y una variable independiente x. En la regresión lineal simple, el modelo de regresión es $y = \beta_0 + \beta_1 x + \epsilon$. La ecuación de regresión lineal simple $E(y) = \beta_0 + \beta_1 x$ describe la relación de la media o valor esperado de y con x. Para obtener la ecuación de regresión estimada $\hat{y} = b_0 + b_1 x$ se emplearon datos muestrales y el método de mínimos cuadrados.

En efecto, b_0 y b_1 son estadísticos muestrales útiles para estimar los parámetros desconocidos del modelo, β_0 y β_1.

El coeficiente de determinación se presentó como una medida de la bondad de ajuste para la ecuación de regresión estimada, y se puede interpretar como la proporción de la variación en la variable dependiente y que puede ser explicada por la ecuación de regresión estimada. Asimismo, se revisó la correlación como una medida descriptiva de la intensidad de la relación lineal entre dos variables.

Se analizaron los supuestos acerca del modelo de regresión y su correspondiente término del error, y se presentaron las pruebas t y F, basadas en esos supuestos, como un medio para determinar si la relación entre dos variables es estadísticamente significativa. Se mostró cómo usar la ecuación de regresión estimada para obtener estimaciones de intervalos de confianza para el valor medio de y y estimaciones por medio de intervalos de predicción para valores individuales de y.

El capítulo concluyó con una sección sobre soluciones por computadora de los problemas de regresión y dos secciones sobre el uso del análisis residual para validar los supuestos del modelo e identificar las observaciones atípicas e influyentes.

Glosario

Análisis residual Análisis de los residuales que se usa para determinar si parecen ser válidos los supuestos planteados acerca del modelo de regresión. También se utiliza para identificar observaciones atípicas y observaciones influyentes.

Coeficiente de correlación Medida de la intensidad de la relación lineal entre dos variables (ya estudiado en el capítulo 3).

Coeficiente de determinación Medida de la bondad de ajuste de la ecuación de regresión estimada. Se interpreta como la proporción de la variabilidad de la variable dependiente y que es explicada por la ecuación de regresión estimada.

Diagrama de dispersión Gráfica de datos bivariados en la que la variable independiente se ubica en el eje horizontal y la variable dependiente en el eje vertical.

Ecuación de regresión Ecuación que describe cómo está relacionada la media o valor esperado de la variable dependiente con la variable independiente; en la regresión lineal simple, $E(y) = \beta_0 + \beta_1 x$.

Ecuación de regresión estimada Estimación de la ecuación de regresión obtenida a partir de datos muestrales empleando el método de mínimos cuadrados. En la regresión lineal simple, la ecuación de regresión estimada es $\hat{y} = b_0 + b_1 x$.

Error cuadrado medio Estimación insesgada de la varianza del término del error σ^2. Se denota como ECM o s^2.

Error estándar de estimación Raíz cuadrada del error cuadrado medio; se denota como s. Es una estimación de σ, la desviación estándar del término del error ϵ.

Gráfica de probabilidad normal Gráfica en la que los residuales estandarizados se grafican contra los puntos normales. Ayuda a determinar si parece ser válido el supuesto de que los términos del error tienen una distribución de probabilidad normal.

Gráfica de residuales Representación gráfica de los residuales. Se usa para determinar si parecen ser válidos los supuestos planteados acerca del modelo de regresión.

Intervalo de confianza Estimación por intervalo del valor medio de y para un valor dado de x.

Intervalo de predicción Estimación por intervalo de un solo valor de y para un valor dado de x.

Método de mínimos cuadrados Procedimiento para obtener la ecuación de regresión estimada. El objetivo es minimizar $\Sigma(y_i - \hat{y}_i)^2$.

Modelo de regresión Ecuación que describe cómo están relacionadas y y x, más un término del error. En la regresión lineal simple, el modelo de regresión es $y = \beta_0 + \beta_1 x + \epsilon$.

Observación atípica Dato u observación que no sigue la tendencia del resto de los datos.

Observación influyente Observación que tiene una fuerte influencia o efecto en los resultados de regresión.

Puntos de gran influencia Observaciones en las que la variable independiente presenta valores extremos.
Regresión lineal simple Análisis de regresión en el que participan una variable independiente y una variable dependiente, y la relación entre estas variables se aproxima mediante una línea recta.
Residual estandarizado Valor obtenido al dividir un residual entre su desviación estándar.
Residual *i*ésimo Diferencia que existe entre el valor observado de la variable dependiente y el valor pronosticado empleando la ecuación de regresión estimada; para la observación *i*ésima, el residual *i*ésimo es $y_i - \hat{y}_i$.
Tabla de ANOVA En el análisis de varianza, tabla que se usa para resumir los cálculos asociados con la prueba F de significancia.
Variable dependiente Variable que se predice o explica. Se denota por y.
Variable independiente Variable que predice o explica. Se denota por x.

Fórmulas clave

Modelo de regresión lineal simple

$$y = \beta_0 + \beta_1 x + \epsilon \tag{14.1}$$

Ecuación de regresión lineal simple

$$E(y) = \beta_0 + \beta_1 x \tag{14.2}$$

Ecuación de regresión lineal simple estimada

$$\hat{y} = b_0 + b_1 x \tag{14.3}$$

Criterio de mínimos cuadrados

$$\min \Sigma(y_i - \hat{y}_i)^2 \tag{14.5}$$

Pendiente e intersección con el eje *y* de la ecuación de regresión estimada

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \tag{14.6}$$

$$b_0 = \bar{y} - b_1 \bar{x} \tag{14.7}$$

Suma de cuadrados debido al error

$$\text{SCE} = \Sigma(y_i - \hat{y}_i)^2 \tag{14.8}$$

Suma total de cuadrados

$$\text{STC} = \Sigma(y_i - \bar{y})^2 \tag{14.9}$$

Suma de cuadrados debido a la regresión

$$\text{SCR} = \Sigma(\hat{y}_i - \bar{y})^2 \tag{14.10}$$

Relación entre STC, SCR y SCE

$$\text{STC} = \text{SCR} + \text{SCE} \tag{14.11}$$

Coeficiente de determinación

$$r^2 = \frac{\text{SCR}}{\text{STC}} \tag{14.12}$$

Coeficiente de correlación muestral

$$r_{xy} = (\text{signo de } b_1)\sqrt{\text{coeficiente de determinación}} = (\text{signo de } b_1)\sqrt{r^2} \tag{14.13}$$

Error cuadrado medio (estimación de σ^2)

$$s^2 = \text{ECM} = \frac{\text{SCE}}{n-2} \tag{14.15}$$

Error estándar de estimación

$$s = \sqrt{\text{ECM}} = \sqrt{\frac{\text{SCE}}{n-2}} \tag{14.16}$$

Desviación estándar de b_1

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\Sigma(x_i - \bar{x})^2}} \tag{14.17}$$

Desviación estándar estimada de b_1

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} \tag{14.18}$$

Estadístico de prueba t

$$t = \frac{b_1}{s_{b_1}} \tag{14.19}$$

Cuadrado medio de la regresión

$$\text{CMR} = \frac{\text{SCR}}{\text{número de variables independientes}} \tag{14.20}$$

Estadístico de prueba F

$$F = \frac{\text{CMR}}{\text{ECM}} \tag{14.21}$$

Desviación estándar estimada de $\hat{y}_p$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \tag{14.23}$$

Intervalo de confianza para $E(y_p)$

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p} \tag{14.24}$$

Desviación estándar estimada para un solo valor

$$s_{\text{ind}} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \tag{14.26}$$

Intervalo de predicción para y_p

$$\hat{y}_p \pm t_{\alpha/2} s_{\text{ind}} \tag{14.27}$$

Residual de la observación *i*

$$y_i - \hat{y}_i \tag{14.28}$$

Desviación estándar del residual *i*ésimo

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \tag{14.30}$$

Residual estandarizado de la observación *i*

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \tag{14.32}$$

Influencia de la observación *i*

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} \tag{14.33}$$

Ejercicios complementarios

55. ¿Un valor alto de r^2 implica que entre las dos variables hay una relación de causa y efecto? Explique.

56. Describe la diferencia entre estimación por intervalo del valor medio de las y para un valor dado de x y estimación por intervalo de un valor de y para una x dada.

57. ¿Qué objeto tiene probar si $\beta_1 = 0$? Si se rechaza que $\beta_1 = 0$, ¿eso significa un buen ajuste?

58. En la tabla siguiente se proporciona el número de acciones en venta (Shares Selling) en millones y el precio esperado (Expected Price), es decir, el promedio del precio mínimo y del precio máximo proyectado, de 10 acciones de oferta pública inicial (IPO, por sus siglas en inglés).

WEB archivo
IPO

Company	Shares Selling (millions)	Expected Price ($)
American Physician	5.0	15
Apex Silver Mines	9.0	14
Dan River	6.7	15
Franchise Mortgage	8.75	17
Gene Logic	3.0	11
International Home Foods	13.6	19
PRT Group	4.6	13
Rayovac	6.7	14
RealNetworks	3.0	10
Software AG Systems	7.7	13

a) Obtenga la ecuación de regresión estimada donde la cantidad de acciones en venta sea la variable independiente y el precio esperado sea la variable dependiente.
b) Empleando 0.05 como nivel de significancia, ¿existe una relación significativa entre las dos variables?
c) ¿La ecuación de regresión estimada proporciona un buen ajuste? Explique.
d) Empleando la ecuación de regresión estimada, determine el precio esperado en una empresa que considera una oferta pública inicial de 6 millones de acciones.

59. Los datos siguientes muestran la estimación del valor justo (Fair Value) de Morningstar y el precio por acción (Share Price) de 28 empresas (Company). El valor justo es una estimación del valor de una empresa por acción que considera pronósticos de crecimiento, rentabilidad, riesgo y otros factores de las compañías para los próximos cinco años (*Morningstar Stocks 500,* edición 2008).

WEB archivo
Stocks500

Company	Fair Value ($)	Share Price ($)
Air Products and Chemicals	80	98.63
Allied Waste Industries	17	11.02
America Mobile	83	61.39
AT&T	35	41.56
Bank of America	70	41.26
Barclays PLC	68	40.37
Citigroup	53	29.44
Costco Wholesale Corp.	75	69.76
Covidien, Ltd.	58	44.29
Darden Restaurants	52	27.71
Dun & Bradstreet	87	88.63
Equifax	42	36.36
Gannett Co.	38	39.00
Genuine Parts	48	46.30
GlaxoSmithKline PLC	57	50.39
Iron Mountain	33	37.02
ITT Corporation	83	66.04
Johnson & Johnson	80	66.70
Las Vegas Sands	98	103.05
Macrovision	23	18.33
Marriott International	39	34.18
Nalco Holding Company	29	24.18
National Interstate	25	33.10
Portugal Telecom	15	13.02
Qualcomm	48	39.35
Royal Dutch Shell Ltd.	87	84.20
SanDisk	60	33.17
Time Warner	42	27.60

a) Obtenga una ecuación de regresión estimada para calcular el precio por acción dado el valor justo.

b) Empleando 0.05 como nivel de significancia, ¿existe una relación significativa entre las dos variables?

c) Utilice la ecuación de regresión estimada para calcular el precio por acción de una empresa que tiene un valor justo de $50.

d) ¿Cree que la ecuación de regresión estimada proporcione una buena predicción del precio por acción? Emplee r^2 para justificar su respuesta.

60. Uno de los más grandes cambios en educación superior de los años recientes ha sido el crecimiento de las universidades en línea. La Online Education Database es una organización independiente cuya misión es construir una lista amplia de las universidades en línea más acreditadas. La tabla siguiente muestra los porcentajes (%) de la tasa de retención (Retention Rate) y la tasa de graduación (Graduation Rate) de 29 universidades (College) en línea (sitio web de Online Education Database, enero de 2009).

College	Retention Rate (%)	Graduation Rate (%)
Western International University	7	25
South University	51	25
University of Phoenix	4	28
American InterContinental University	29	32
Franklin University	33	33
Devry University	47	33

WEB archivo
OnlineEdu

College	Retention Rate (%)	Graduation Rate (%)
Tiffin University	63	34
Post University	45	36
Peirce College	60	36
Everest University	62	36
Upper Iowa University	67	36
Dickinson State University	65	37
Western Governors University	78	37
Kaplan University	75	38
Salem International University	54	39
Ashford University	45	41
ITT Technical Institute	38	44
Berkeley College	51	45
Grand Canyon University	69	46
Nova Southeastern University	60	47
Westwood College	37	48
Everglades University	63	50
Liberty University	73	51
LeTourneau University	78	52
Rasmussen College	48	53
Keiser University	95	55
Herzing College	68	56
National University	100	57
Florida National College	100	61

a) Trace el diagrama de dispersión con la tasa de retención como variable independiente. ¿Qué indica el diagrama respecto de la relación entre las dos variables?
b) Obtenga la ecuación de regresión estimada.
c) Pruebe la significancia de la relación. Use $\alpha = 0.05$.
d) ¿La ecuación de regresión estimada proporciona un buen ajuste? Explique.
e) Suponga que usted es el rector de South University. Luego de revisar los resultados, ¿estaría usted preocupado por el desempeño de su institución comparada con otras universidades en línea?
f) Ahora suponga que es el rector de la Universidad de Phoenix. Luego de revisar los resultados, ¿estaría usted preocupado por el desempeño de su institución comparada con otras universidades en línea?

61. Jensen Tire & Auto está en el proceso de decidir si firma un contrato de mantenimiento para su nueva máquina de alineación y balanceo de neumáticos por computadora. Los gerentes piensan que los gastos de mantenimiento deberán estar relacionados con el uso, y recabaron los datos siguientes acerca del uso semanal en horas (Weekly Usage/hours) y los gastos de mantenimiento anuales (Annual Maintenance Expense) en cientos de dólares.

WEB archivo
Jensen

Weekly Usage (hours)	Annual Maintenance Expense
13	17.0
10	22.0
20	30.0
28	37.0
32	47.0
17	30.5
24	32.5
31	39.0
40	51.5
38	40.0

a) Obtenga la ecuación de regresión estimada que relaciona gastos anuales de mantenimiento con el uso semanal.
b) Pruebe la significancia de la relación del inciso *a*) con 0.05 como nivel de significancia.
c) Jensen espera usar la nueva máquina 30 horas a la semana. Obtenga un intervalo de predicción de 95% para los gastos anuales de mantenimiento de la empresa.
d) Si el precio del contrato de mantenimiento es por $3 000 anuales, ¿recomendaría firmarlo? ¿Por qué?

62. En un determinado proceso de manufactura, se cree que la velocidad (pies por minuto) de la línea de ensamble afecta el número de partes defectuosas halladas en el proceso de inspección. Para probar esto, los gerentes han ideado un procedimiento en el que la misma cantidad de partes por lote se examina visualmente a diferentes velocidades de la línea. Se obtienen los datos siguientes.

Velocidad de la línea	Número de partes defectuosas encontradas
20	21
20	19
40	15
30	16
60	14
40	17

a) Obtenga la ecuación de regresión estimada que relaciona la velocidad de la línea de producción con el número de partes defectuosas encontradas.
b) Empleando el nivel de significancia 0.05, determine si la velocidad de la línea y el número de partes defectuosas están relacionados.
c) ¿La ecuación de regresión estimada proporciona un buen ajuste a los datos?
d) Obtenga un intervalo de confianza de 95% para predecir el número medio de partes defectuosas si la velocidad de la línea es de 50 pies por minuto.

63. Un hospital de una ciudad grande contrató a un sociólogo para que investigara la relación entre el número de días de ausencia (Number of Days Absent) sin autorización de los empleados por año y la distancia en millas entre su casa y el trabajo (Distance to Work/miles). Se tomó una muestra de 10 empleados y se obtuvieron los datos siguientes.

WEB archivo
Absent

Distance to Work (miles)	Number of Days Absent
1	8
3	5
4	8
6	7
8	6
10	3
12	5
14	2
14	4
18	2

a) Elabore un diagrama de dispersión con estos datos. ¿Aparenta ser razonable una relación lineal? Explique.
b) Obtenga la ecuación de regresión estimada por mínimos cuadrados.
c) ¿Existe una relación significativa entre las dos variables? Use $\alpha = 0.05$.
d) ¿La ecuación de regresión estimada proporciona un buen ajuste? Explique.
e) Utilice la ecuación de regresión estimada obtenida en el inciso *b*) para calcular un intervalo de confianza de 95% para el número esperado de días de ausencia de los empleados que viven a 5 millas de la empresa.

64. La autoridad de tránsito local de una zona metropolitana importante desea determinar si hay relación entre la antigüedad de un autobús (Age of Bus/years) y los gastos anuales derivados de su mantenimiento (Maintenance Cost). En una muestra de 10 autobuses se obtuvieron los datos siguientes.

WEB archivo
AgeCost

Age of Bus (years)	Maintenance Cost ($)
1	350
2	370
2	480
2	520
2	590
3	550
4	750
4	800
5	790
5	950

a) Con el método de mínimos cuadrados, obtenga la ecuación de regresión estimada.
b) Realice una prueba para determinar si las dos variables están relacionadas de manera significativa con $\alpha = 0.05$.
c) ¿La recta de mínimos cuadrados proporciona un buen ajuste a los datos observados? Explique.
d) Calcule un intervalo de predicción de 95% para los gastos de mantenimiento de un determinado autobús cuya antigüedad es de 4 años.

65. Un profesor de marketing del Givens College desea saber cuál es la relación entre las horas destinadas al estudio (Hours Spent Studying) y la calificación ganada (Total Points Earned) en un curso. A continuación se presentan los datos obtenidos de 10 estudiantes que tomaron el curso el trimestre pasado.

WEB archivo
HoursPts

Hours Spent Studying	Total Points Earned
45	40
30	35
90	75
60	65
105	90
65	50
90	90
80	80
55	45
75	65

a) Obtenga la ecuación de regresión estimada que indica la relación entre la calificación obtenida y las horas dedicadas al estudio.
b) Empleando $\alpha = 0.05$, pruebe la significancia del modelo.
c) Pronostique la calificación total que obtendrá Mark Sweeney, quien dedicó 95 horas al estudio.
d) Calcule un intervalo de predicción de 95% para la calificación de Mark Sweeney.

66. Reuters reportó que la beta del mercado de Xerox era de 1.22 (sitio web de Reuters, 30 de enero de 2009). La beta del mercado de cada acción se determina mediante regresión lineal simple. En cada caso, la variable dependiente es la rentabilidad porcentual trimestral (revalorización del capital más dividendos), menos el rendimiento porcentual que se hubiera obtenido en una inversión libre de riesgo (como tasa libre de riesgo se usó la Treasury Bill). La variable independiente es la rentabilidad porcentual trimestral para el mercado de valores (S&P 500), menos la rentabilidad porcentual de una inversión libre de riesgo. A partir de los datos trimestrales se desarrolla la ecuación de regresión estimada, y la beta del mercado de la acción en cuestión es la pendiente de esta ecuación (b_1). El valor de la beta del mercado suele interpretarse como

una medida del riesgo asociado con la acción bursátil. Si la beta del mercado es mayor de 1, la volatilidad de la acción es mayor al promedio en el mercado; si es menor de 1, la volatilidad de la acción es menor al promedio en el mercado. Suponga que las cifras siguientes son diferencias entre rentabilidad porcentual y rentabilidad libre de riesgo a lo largo de 10 trimestres para S&P 500 y Horizon Technology.

WEB archivo
MktBeta

S&P 500	Horizon
1.2	−0.7
−2.5	−2.0
−3.0	−5.5
2.0	4.7
5.0	1.8
1.2	4.1
3.0	2.6
−1.0	2.0
0.5	−1.3
2.5	5.5

a) Obtenga la ecuación de regresión estimada que sirve para determinar la beta del mercado de Horizon Technology. ¿Cuál es la beta del mercado de esta empresa?
b) Empleando 0.05 como nivel de significancia, pruebe la significancia de la relación.
c) ¿La ecuación de regresión estimada proporciona un buen ajuste? Explique.
d) Utilice las betas del mercado de Xerox y de Horizon Techology para comparar los riesgos asociados con estas dos acciones.

67. La Transactional Records Access Clearinghouse de la Universidad de Syracuse publica datos que muestran las probabilidades de una auditoría del Departamento del Tesoro de Estados Unidos. En la tabla siguiente se muestra la media del ingreso bruto ajustado (Adjusted Gross Income) y el porcentaje de declaraciones que fueron auditadas (Percent Audited) en 20 distritos.

WEB archivo
IRSAudit

District	Adjusted Gross Income ($)	Percent Audited
Los Angeles	36 664	1.3
Sacramento	38 845	1.1
Atlanta	34 886	1.1
Boise	32 512	1.1
Dallas	34 531	1.0
Providence	35 995	1.0
San Jose	37 799	0.9
Cheyenne	33 876	0.9
Fargo	30 513	0.9
New Orleans	30 174	0.9
Oklahoma City	30 060	0.8
Houston	37 153	0.8
Portland	34 918	0.7
Phoenix	33 291	0.7
Augusta	31 504	0.7
Albuquerque	29 199	0.6
Greensboro	33 072	0.6
Columbia	30 859	0.5
Nashville	32 566	0.5
Buffalo	34 296	0.5

a) Obtenga la ecuación de regresión estimada que sirve para pronosticar el porcentaje de auditorías dado el promedio del ingreso bruto ajustado reportado.
b) Empleando como nivel de significancia 0.05, determine si hay relación entre el ingreso bruto ajustado y el porcentaje de auditorías.
c) ¿La ecuación de regresión estimada proporciona un buen ajuste? Explique.

d) Con la ecuación de regresión estimada del inciso *a*) calcule un intervalo de 95% de confianza para el porcentaje esperado de auditorías en un distrito donde el promedio del ingreso bruto ajustado es $35 000.

68. El State of the Service Report 2002-2003, de la Comisión de Servicio Público de Australia, reportó evaluaciones sobre la satisfacción laboral. Una parte la encuesta consistía en elegir (de una lista) los cinco factores principales que contribuyen a la satisfacción laboral. Después se solicitaba a los encuestados que indicaran su nivel de satisfacción con cada uno de esos cinco factores. En la tabla siguiente se presentan los porcentajes de personas para los que el factor indicado fue uno de los cinco principales, junto con una evaluación obtenida utilizando el porcentaje de empleados que consideró el factor como uno de los cinco principales y que estaban "muy satisfechos" o "satisfechos" con éste en su actual lugar de trabajo (*www.apsc.gov.au/stateoftheservice*). Workplace Factor indica el factor de satisfacción laboral, Top Five lista los cinco valores principales y Satisfaction Rating indica la evaluación de la satisfacción.

WEB archivo
JobSat

Workplace Factor	Top Five (%)	Satisfaction Rating (%)
Carga de trabajo apropiada	30	49
Oportunidad de ser creativo/innovador	38	64
Oportunidad para hacer contribuciones últiles a la sociedad	40	67
Derechos/expectativas claras	40	69
Sistemas de trabajo flexibles	55	86
Buenas relaciones de trabajo	60	85
Trabajo interesante constante	48	74
Oportunidad para desarrollarme en mi carrera	33	43
Oportunidad para desarrollar mis habilidades	46	66
Condiciones para utilizar mis capacidades	50	70
Retroalimentación regular/reconocimiento al esfuerzo	42	53
Salario	47	62
Ver los resultados tangibles de mi trabajo	42	69

a) Elabore un diagrama de dispersión colocando en el eje horizontal los cinco factores principales (%) y en el eje vertical el nivel de satisfacción (%).
b) ¿Qué indica el diagrama elaborado en el inciso *a*) respecto de la relación entre las dos variables?
c) Obtenga la ecuación de regresión estimada que sirva para pronosticar el nivel de satisfacción (%) dados los cinco factores principales (%).
d) Empleando como nivel de significancia 0.05, realice una prueba para determinar la significancia de la relación.
e) ¿La ecuación de regresión estimada proporciona un buen ajuste? Explique.
f) ¿Cuál es el valor del coeficiente de correlación muestral?

Caso a resolver 1 Medición del riesgo en el mercado bursátil

Una medida del riesgo o volatilidad de una acción bursátil es la desviación estándar del rendimiento total (revalorización del capital más dividendos) durante varios periodos. Aunque la desviación estándar es fácil de calcular, no considera la variación del precio de una acción en función de un índice estándar del mercado, como el S&P 500. Por esta razón, muchos analistas financieros prefieren emplear otra medida, conocida como *beta,* para medir el riesgo.

La beta de una acción individual se determina mediante regresión lineal simple. La variable dependiente es la rentabilidad total de la acción de que se trate y la variable independiente es la rentabilidad total del mercado de valores.* En este caso a resolver se utilizará el índice S&P

* Distintas fuentes emplean diferentes métodos para calcular las betas. Por ejemplo, algunas fuentes, antes de calcular la ecuación de regresión estimada, restan de las variables tanto dependiente como independiente la rentabilidad que podría haberse obtenido con una inversión libre de riesgo (por ejemplo, letras del Tesoro o T-bills). Otras emplean diversos índices para la rentabilidad total del mercado de valores; por ejemplo, *Value Line* calcula las betas utilizando el índice compuesto de la bolsa de Nueva York.

WEB archivo
Beta

500 como medida de la rentabilidad total del mercado de valores y se obtendrá una ecuación de regresión estimada usando datos mensuales. La beta de una acción es la pendiente en la ecuación de regresión estimada (b_1). Los datos en el archivo Beta proporcionan la rentabilidad total (revalorización del capital más dividendos) de ocho acciones comunes muy conocidas y la del S&P 500 a lo largo de 36 meses.

El valor beta del mercado de valores siempre será 1; por tanto, una acción que tienda a aumentar o a disminuir con el mercado de valores tendrá también una beta cercana a 1. Betas mayores a 1 corresponden a acciones más volátiles que el mercado y betas menores a 1 corresponden a acciones menos volátiles que el mercado. Por ejemplo, si la beta de una acción es 1.4, esta acción es 40% *más* volátil que el mercado, y si es 0.4, la acción es 60% *menos* volátil que el mercado.

Informe gerencial

Se le ha encomendado la tarea de analizar las características del riesgo de estas acciones. Elabore un informe que comprenda los puntos siguientes, sin limitarse sólo a ellos.

a) Calcular los estadísticos descriptivos de cada una de las acciones y del S&P 500. Comente los resultados. ¿Qué acción es la más volátil?
b) Calcular la beta de cada acción. ¿Cuál de estas acciones se esperaría que se comportara mejor en un mercado de alta calidad? ¿Cuál conservaría mejor su valor en un mercado para el sector popular?
c) Comente qué tanto de la rentabilidad de cada una de las acciones es explicada por el mercado.

Caso a resolver 2

Departamento de Transporte de Estados Unidos

Como parte de un estudio sobre seguridad en el transporte, el Departamento de Transporte de Estados Unidos recabó datos de una muestra de 42 ciudades sobre el número de accidentes fatales por cada 1 000 licencias (Fatal Accidents per 1 000 Licenses) autorizadas, y sobre el porcentaje de licencias de conductores menores de 21 años (Percent Under 21). A continuación se presentan los datos obtenidos en el lapso de un año, los cuales se encuentran en el archivo *Safety*.

WEB archivo
Safety

Percent Under 21	Fatal Accidents per 1000 Licenses	Percent Under 21	Fatal Accidents per 1 000 Licenses
13	2.962	17	4.100
12	0.708	8	2.190
8	0.885	16	3.623
12	1.652	15	2.623
11	2.091	9	0.835
17	2.627	8	0.820
18	3.830	14	2.890
8	0.368	8	1.267
13	1.142	15	3.224
8	0.645	10	1.014
9	1.028	10	0.493
16	2.801	14	1.443
12	1.405	18	3.614
9	1.433	10	1.926
10	0.039	14	1.643
9	0.338	16	2.943
11	1.849	12	1.913
12	2.246	15	2.814
14	2.855	13	2.634
14	2.352	9	0.926
11	1.294	17	3.256

Informe gerencial

1. Presente resúmenes numéricos y gráficos de los datos.
2. Emplee el análisis de regresión para investigar la relación entre el número de accidentes fatales y el porcentaje de conductores menores de 21 años. Analice sus hallazgos.
3. ¿Qué conclusión y qué recomendaciones puede deducir de su análisis?

Caso a resolver 3 Donaciones de exalumnos

Las donaciones de exalumnos son una importante fuente de ingresos para las universidades. Si los gerentes pudieran determinar los factores que influyen en el aumento del porcentaje de exalumnos que hace donaciones, podrían poner en marcha políticas que llevaran a ganancias mayores. Las investigaciones indican que estudiantes más satisfechos de la relación con sus profesores tienen más probabilidad de titularse, lo que a su vez puede llevar al aumento del porcentaje que hace donaciones. En la tabla 14.13 se muestran datos de 48 universidades de Estados Unidos (*American's Best Colleges,* edición 2000). La columna titulada "% of Classes Under 20" (% de grupos con menos de 20) muestra el porcentaje de grupos con menos de 20 alumnos. La columna titulada "Student/Faculty Ratio" (tasa de estudiantes/profesor) proporciona el número de estudiantes inscritos, dividido entre el número total de docentes. Por último, la columna titulada "Alumni Giving Rate" (tasa de exalumnos donantes) presenta el porcentaje de egresados que ha efectuado alguna donación a la universidad.

Informe gerencial

1. Presente resúmenes numéricos y gráficos de los datos.
2. Emplee el análisis de regresión para obtener una ecuación de regresión estimada que sirva para pronosticar el porcentaje de exalumnos que hace donaciones, dado el porcentaje de grupos con menos de 20 estudiantes.
3. Use el análisis de regresión para obtener una ecuación de regresión estimada que sirva para pronosticar el porcentaje de exalumnos donantes dada la proporción de estudiantes por facultad.
4. ¿Cuál de las dos ecuaciones de regresión estimada muestra un mejor ajuste? Con esa ecuación de regresión estimada realice un análisis de residuales y discuta sus hallazgos y conclusiones.
5. ¿Qué conclusiones y recomendaciones puede derivar de este análisis?

Caso a resolver 4 Estadísticas del PGA Tour

La Asociación de Golfistas Profesionales (PGA, por sus siglas en inglés) mantiene datos sobre desempeño y ganancias de los participantes en el PGA Tour. Los 125 mejores jugadores con base en las ganancias totales en los eventos del PGA Tour son exentos para la siguiente temporada. La lista de los 125 con más dinero es importante, pues un jugador "exento" califica para ser miembro de tiempo completo del PGA Tour en la siguiente temporada.

Durante los años recientes en el PGA Tour se han dado avances significativos en tecnología de pelotas y clubes de golf, y esta tecnología ha sido una de las principales razones para el incremento en la distancia promedio de tiro de los jugadores del PGA Tour. En 1992, la distancia promedio de tiro fue de 260 yardas, pero en 2003 se incrementó a 286 yardas. Los profesionales están lanzando la pelota más lejos que nunca, pero ¿qué tan importante es la distancia de tiro en términos del desempeño de un jugador? y ¿qué efecto ha tenido el incremento de esa distancia en la precisión de los jugadores? Para responder estas interrogantes, en el archivo PGA

WEB archivo
PGATour

TABLA 14.13 Datos de 48 universidades nacionales

WEB archivo
Alumni

	% of Classes Under 20	Student/Faculty Ratio	Alumni Giving Rate
Boston College	39	13	25
Brandeis University	68	8	33
Brown University	60	8	40
California Institute of Technology	65	3	46
Carnegie Mellon University	67	10	28
Case Western Reserve University	52	8	31
College of William and Mary	45	12	27
Columbia University	69	7	31
Cornell University	72	13	35
Dartmouth College	61	10	53
Duke University	68	8	45
Emory University	65	7	37
Georgetown University	54	10	29
Harvard University	73	8	46
Johns Hopkins University	64	9	27
Lehigh University	55	11	40
Massachusetts Inst. of Technology	65	6	44
New York University	63	13	13
Northwestern University	66	8	30
Pennsylvania State University	32	19	21
Princeton University	68	5	67
Rice University	62	8	40
Stanford University	69	7	34
Tufts University	67	9	29
Tulane University	56	12	17
U. of California–Berkeley	58	17	18
U. of California–Davis	32	19	7
U. of California–Irvine	42	20	9
U. of California–Los Angeles	41	18	13
U. of California–San Diego	48	19	8
U. of California–Santa Barbara	45	20	12
U. of Chicago	65	4	36
U. of Florida	31	23	19
U. of Illinois–Urbana Champaign	29	15	23
U. of Michigan–Ann Arbor	51	15	13
U. of North Carolina–Chapel Hill	40	16	26
U. of Notre Dame	53	13	49
U. of Pennsylvania	65	7	41
U. of Rochester	63	10	23
U. of Southern California	53	13	22
U. of Texas–Austin	39	21	13
U. of Virginia	44	13	28
U. of Washington	37	12	12
U. of Wisconsin–Madison	37	13	13
Vanderbilt University	68	9	31
Wake Forest University	59	11	38
Washington University–St. Louis	73	7	33
Yale University	77	7	50

Tour (sitio web del PGA Tour, 2009) se encuentran los datos del desempeño al final del año de 125 jugadores con las ganancias totales más altas en los eventos del PGA Tour de 2008. Cada fila del conjunto de datos corresponde a un jugador y los datos se han ordenado con base en el total de ganancias. Las descripciones de los datos son las siguientes.

Money (*dinero*). Ganancias totales en eventos del PGA Tour.

Scoring Average (*puntuación promedio*). Número promedio de golpes por ronda completa.

DrDist (*distancia del tiro*). Número promedio de yardas por tiro medido. En el PGA Tour, la distancia de tiro se mide en dos hoyos por ronda. Se tiene cuidado en seleccionar dos hoyos orientados en direcciones opuestas para contrarrestar el efecto del viento. Los golpes se miden hasta el punto donde se detiene sin importar que esté o no dentro del recorrido.

DrAccu (*precisión del tiro*). Porcentaje de veces en que un tiro de salida se detiene dentro del recorrido (sin importar el club). La precisión del tiro se mide en cada hoyo, excluyendo los que sean par 3.

GIR (*greens en regulación*). Porcentaje de veces que un jugador logra alcanzar el *green* en regulación. Se considera un *green* alcanzado en regulación si cualquier parte de la pelota toca la superficie de *putting* luego de que se tomó el golpe GIR, el cual se determina al restar 2 del par (primer golpe en un par 3, segundo en un par 4, tercero en un par 5). En otras palabras, se considera un *green* alcanzado en regulación si el jugador llegó a la superficie de *putting* en par menos dos golpes.

Informe gerencial

1. Presente resúmenes numéricos y gráficos de los datos.
2. Utilice el análisis de regresión para investigar la relación entre puntuación promedio y distancia del tiro. ¿En apariencia, los jugadores que lanzan la pelota más lejos tienen puntuaciones promedio más bajas?
3. Con base en el análisis de regresión investigue la relación entre puntuación promedio y precisión del tiro. ¿En apariencia, los jugadores más precisos en golpear dentro del recorrido tienen puntuaciones promedio más bajas?
4. Utilice el análisis de regresión para investigar la relación entre puntuación promedio y *greens* en regulación. ¿En apariencia los jugadores más precisos en alcanzar el *green* en regulación tienen puntuaciones promedio más bajas?
5. ¿Cuál de las tres variables (DrDist, DrAccu y GIR) aparenta ser el factor más significativo en términos de la puntuación promedio de un jugador?
6. Tomando DrDist como la variable independiente y DrAccu como la variable dependiente, investigue la relación entre distancia y precisión del tiro.

Apéndice 14.1 Deducción de la fórmula de mínimos cuadrados basada en el cálculo

Como ya se indicó en este capítulo, el método de mínimos cuadrados se usa para determinar los valores de b_0 y b_1 que minimicen la suma de los cuadrados de los residuales, la cual está dada por

$$\Sigma(y_i - \hat{y}_i)^2$$

Al sustituir $\hat{y}_i = b_0 + b_1 x_i$, obtenemos

$$\Sigma(y_i - b_0 - b_1 x_i)^2 \qquad \textbf{(14.34)}$$

como expresión que se debe minimizar.

Para minimizar la fórmula (14.34), se obtienen las derivadas parciales respecto de b_0 y b_1, se igualan a cero y se despeja. Haciendo esto obtenemos

$$\frac{\partial \Sigma (y_i - b_0 - b_1 x_i)^2}{\partial b_0} = -2\Sigma (y_i - b_0 - b_1 x_i) = 0 \qquad \textbf{(14.35)}$$

$$\frac{\partial \Sigma (y_i - b_0 - b_1 x_i)^2}{\partial b_1} = -2\Sigma x_i (y_i - b_0 - b_1 x_i) = 0 \qquad \textbf{(14.36)}$$

Al dividir la ecuación (14.35) entre dos y hacer las sumas por separado, obtenemos

$$-\Sigma y_i + \Sigma b_0 + \Sigma b_1 x_i = 0$$

Al llevar Σy_i al otro lado del signo igual y observar que $\Sigma b_0 = nb_0$, conseguimos

$$nb_0 + (\Sigma x_i) b_1 = \Sigma y_i \qquad \textbf{(14.37)}$$

Simplificaciones algebraicas similares aplicadas a la ecuación (14.36) producen

$$(\Sigma x_i) b_0 + (\Sigma x_i^2) b_1 = \Sigma x_i y_i \qquad \textbf{(14.38)}$$

A las expresiones (14.37) y (14.38) se les conoce como *ecuaciones normales.* Al despejar b_0 en la (14.37) obtenemos

$$b_0 = \frac{\Sigma y_i}{n} - b_1 \frac{\Sigma x_i}{n} \qquad \textbf{(14.39)}$$

Utilizando la ecuación (14.39) para sustituir a b_0 en la (14.38) tenemos

$$\frac{\Sigma x_i \Sigma y_i}{n} - \frac{(\Sigma x_i)^2}{n} b_1 + (\Sigma x_i^2) b_1 = \Sigma x_i y_i \qquad \textbf{(14.40)}$$

Al reordenar los términos de la ecuación (14.40) obtenemos

$$b_1 = \frac{\Sigma x_i y_i - (\Sigma x_i \Sigma y_i)/n}{\Sigma x_i^2 - (\Sigma x_i)^2/n} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2} \qquad \textbf{(14.41)}$$

Como $\bar{y} = \Sigma y_i / n$ y $\bar{x} = \Sigma x_i / n$, la fórmula (14.39) se puede reescribir como

$$b_0 = \bar{y} - b_1 \bar{x} \qquad \textbf{(14.42)}$$

Las ecuaciones (14.41) y (14.42) son las fórmulas (14.6) y (14.7) usadas en este capítulo para calcular los coeficientes de la ecuación de regresión estimada.

Apéndice 14.2 Prueba de significancia usando correlación

Al emplear el coeficiente de correlación muestral r_{xy} también se puede determinar si la relación lineal entre x y y es significativa mediante la siguiente prueba de hipótesis acerca del coeficiente de correlación poblacional ρ_{xy}.

$$H_0: \rho_{xy} = 0$$
$$H_a: \rho_{xy} \neq 0$$

Si H_0 es rechazada, se concluye que el coeficiente de correlación poblacional no es igual a cero y que la relación lineal entre las dos variables es significativa. A continuación se presenta esta prueba de significancia.

PRUEBA DE SIGNIFICANCIA USANDO CORRELACIÓN

$$H_0: \rho_{xy} = 0$$
$$H_a: \rho_{xy} \neq 0$$

ESTADÍSTICO DE PRUEBA

$$t = r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}} \qquad (14.43)$$

REGLA DE RECHAZO

Método del valor-p: Rechazar H_0 si el valor-$p \leq \alpha$

Método del valor crítico: Rechazar H_0 si $t \leq -t_{\alpha/2}$ o si $t \geq t_{\alpha/2}$

donde $t_{\alpha/2}$ pertenece a la distribución t con $n - 2$ grados de libertad.

En la sección 14.3 se encontró que con una muestra de $n = 10$ el coeficiente de correlación muestral para la población de estudiantes y las ventas trimestrales era $r_{xy} = 0.9501$. El estadístico de prueba es

$$t = r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}} = 0.9501\sqrt{\frac{10-2}{1-(0.9501)^2}} = 8.61$$

En la tabla de la distribución t se encuentra que para $n - 2 = 10 - 2 = 8$ grados de libertad, $t = 3.355$ proporciona un área de 0.005 en la cola superior. Por tanto, al área en la cola superior que corresponde al estadístico de prueba $t = 8.61$ debe ser menor de 0.005. Como ésta es una prueba de dos colas, se duplica este valor y se concluye que el valor-p que corresponde a $t = 8.61$ debe ser menor a $2(0.005) = 0.01$. Con Excel o con Minitab se obtiene el valor-$p = 0.000$. Como el valor-p es menor a $\alpha = 0.01$, H_0 es rechazada y se concluye que ρ_{xy} no es igual a cero. Esta evidencia es suficiente para concluir que entre la población de estudiantes y las ventas trimestrales existe una relación lineal significativa.

Observe que, excepto por el redondeo, el valor del estadístico de prueba t y la conclusión sobre la significancia de la relación son idénticos a los resultados obtenidos en la prueba t de la sección 14.5, donde se usó la ecuación de regresión estimada $\hat{y} = 60 + 5x$. El análisis de regresión permite obtener una conclusión sobre la relación significativa entre las variables x y y, además de la ecuación que indica cuál es la relación entre las variables. Por consiguiente, la mayoría de los analistas emplea software moderno para realizar el análisis de regresión y encuentra que el uso de la correlación como prueba de significancia es innecesario.

Apéndice 14.3 Análisis de regresión con Minitab

WEB archivo
Armand's

En la sección 14.7 se estudió la solución de los problemas de regresión en computadora mostrando los resultados que genera Minitab para el problema de Armand's Pizza Parlors. En este apéndice se describen los pasos necesarios para generar la solución de computadora en Minitab. Primero se ingresan los datos en una hoja de trabajo. Los datos de las poblaciones de estudiantes se ingresan en la columna Cl y los de las ventas trimestrales en la columna C2. Los nombres de las variables Pop y Sales se ingresan como encabezados de esas columnas. En los pasos subsecuentes, para referirse a los datos se emplearán los nombres de las variables o los indicadores

de las columnas Cl y C2. Los pasos siguientes describen cómo usar Minitab para obtener los resultados del análisis de regresión que se muestran en la figura 14.10.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Seleccione el menú **Regression**.
Paso 3. Elija **Regression**.
Paso 4. Cuando el cuadro de diálogo Regression aparezca:
Ingrese Sales en el cuadro **Response**.
Ingrese Pop en el cuadro **Predictors**.
Haga clic en el botón **Options**.
Cuando el cuadro de diálogo Regression-Options aparezca:
Ingrese 10 en el cuadro **Prediction intervals for new observations**.
Haga clic en **OK**.
Cuando el cuadro de diálogo Regression aparezca:
Haga clic en **OK**.

El cuadro de diálogo Regression tiene otras posibilidades más que se pueden aprovechar seleccionando las opciones deseadas. Por ejemplo, para obtener una gráfica de residuales en la que los valores pronosticados de la variable dependiente $\hat{y}$ aparezcan en el eje horizontal y los valores de los residuales estandarizados en el eje vertical, el paso 4 deberá ser el siguiente:

Paso 4 Cuando el cuadro de diálogo Regression aparezca:
Ingrese Sales en el cuadro **Response**.
Ingrese Pop en el cuadro **Predictors**.
Haga clic en el botón **Graphs**.
Cuando el cuadro de diálogo Regression-Graphs aparezca:
Seleccione **Standardized** en Residuals for Plots.
Elija **Residuals versus fits** en Residual Plots.
Haga clic en **OK**.
Cuando el cuadro de diálogo Regression aparezca:
Haga clic en **OK**.

Apéndice 14.4 Análisis de regresión con Excel

WEB archivo
Armand's

En este apéndice se ilustra el uso de la herramienta Regression de Excel para realizar los cálculos del análisis de regresión empleando el problema de Armand's Pizza Parlors. Consulte la figura 14.23 para seguir la descripción de los pasos. En las celdas A1:C1 de la hoja de cálculo se ingresan los rótulos Restaurant, Population y Sales. Para identificar cada una de las 10 observaciones, se introducen los números del 1 al 10 en las celdas A2:A11. Los datos muestrales se ingresan en las celdas B2:C11. Los pasos siguientes indican cómo usar Excel para obtener los resultados del análisis de regresión.

Paso 1. Haga clic en la ficha **Data** en la cinta.
Paso 2. En el grupo **Analysis**, haga clic en **Data Analysis**.
Paso 3. Elija **Regression** en la lista de Analysis Tools.
Paso 4. Haga clic en **OK**.
Paso 5. Cuando aparezca el cuadro de diálogo Regression:
Ingrese C1:C11 en el cuadro **Input Y Range**.
Introduzca B1:B11 en el cuadro **Input X Range**.
Seleccione **Labels**.
Elija **Confidence Level**.
Ingrese 99 en el cuadro **Confidence Level**.
Seleccione **Output Range**.
Ingrese Al3 en el cuadro **Output Range**.
(También se puede ingresar cualquier celda de la esquina superior izquierda para indicar dónde deberán empezar los resultados.)
Haga clic en **OK**.

FIGURA 14.23 Solución con Excel del problema de Armand's Pizza Parlors

	A	B	C	D	E	F	G	H	I	J
1	**Restaurant**	**Population**	**Sales**							
2	1	2	58							
3	2	6	105							
4	3	8	88							
5	4	8	118							
6	5	12	117							
7	6	16	137							
8	7	20	157							
9	8	20	169							
10	9	22	149							
11	10	26	202							
12										
13	SUMMARY OUTPUT									
14										
15	*Regression Statistics*									
16	Multiple R	0.9501								
17	R Square	0.9027								
18	Adjusted R Square	0.8906								
19	Standard Error	13.8293								
20	Observations	10								
21										
22	ANOVA									
23		*df*	*SS*	*MS*	*F*	*Significance F*				
24	Regression	1	14 200	14 200	74.2484	2.55E-05				
25	Residual	8	1 530	191.25						
26	Total	9	15 730							
27										
28		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*	
29	Intercept	60	9.2260	6.5033	0.0002	38.7247	81.2753	29.0431	90.9569	
30	Population	2	0.5803	8.6167	2.55E-05	3.6619	6.3381	3.0530	6.9470	
31										
32										
33										
34										

La primera sección del resultado, titulada *Regression Statistics,* contiene resúmenes estadísticos como el coeficiente de determinación (R Square). La segunda sección titulada ANOVA, contiene la tabla del análisis de varianza. La última sección, que no tiene ningún título, incluye los coeficientes de regresión estimados e información relacionada con ellos. A continuación se proporciona la interpretación de los resultados de la regresión empezando con la información contenida en las celdas A28:I30.

Interpretación de los resultados de la ecuación de regresión estimada

La intersección de la recta de regresión con el eje *y,* $b_0 = 60$, se presenta en la celda B29, y la pendiente de la recta de regresión estimada, $b_1 = 5$, aparece en la celda B30. Las etiquetas Intercept en la celda A29 y Population en la celda A30 sirven para identificar estos dos valores.

En la sección 14.5 se mostró que la desviación estándar estimada de b_1 es $s_{b_1} = 0.5803$. Observe que el valor de la celda C30 es 0.5803. El rótulo Standard Error que aparece en la celda C28 es la manera en que Excel indica que el valor de la celda C30 es el error estándar o la desviación estándar de b_1. Recuerde que en la prueba *t* de significancia de la relación fue necesario calcular el estadístico *t,* $t = b_1/s_{b_1}$. Empleando los datos de Armand's, el valor calculado fue $t = 5/0.5803 = 8.62$. El rótulo *t Stat* de la celda D28 sirve para recordar que en la celda D30 se encuentra el valor del estadístico de prueba *t.*

El valor en la celda E30 es el valor-p que corresponde a la prueba t de significancia. El valor-p que despliega Excel en la celda E30 está en notación científica. Para obtenerlo en notación decimal se recorre el punto decimal 5 lugares a la izquierda, con lo que resulta 0.0000255. Dado que valor-p = 0.0000255 el $\alpha = 0.01$, H_0 es rechazada y se concluye que existe una relación significativa entre la población de estudiantes y las ventas trimestrales.

La información de las celdas F28:I30 se puede emplear para obtener estimaciones por intervalos de confianza para la intersección con el eje y y la pendiente de la ecuación de regresión estimada. Excel proporciona siempre los límites inferior y superior de un intervalo de 95% de confianza. Como en el paso 4 se seleccionó Confidence Level y se ingresó 99 en el cuadro respectivo, la herramienta de Excel para la regresión proporciona también los límites inferior y superior de un intervalo de 99% de confianza. El valor en la celda H30 es el límite inferior de la estimación por intervalo del 99% de confianza de β_1 y el valor en la celda I30 es el límite superior. Por tanto, una vez redondeada, el intervalo de 99% de confianza para estimar β_1 va de 3.05 a 6.95. Los valores en las celdas F30 a G30 proporcionan los límites inferior y superior del intervalo de 95% de confianza, por lo que éste va de 3.66 a 6.34.

Interpretación de los resultados del ANOVA

La información en las celdas A22:F26 es un resumen de los cálculos del análisis de varianza. Las tres fuentes de variación están rotuladas como Regression, Residual y Total. La etiqueta *df* en la celda B23 representa los grados de libertad, la etiqueta *SS* en la celda C23 presenta la suma de los cuadrados y *MS* en la celda D23 representa el cuadrado de la media.

En la sección 14.5 se dijo que el error cuadrado medio, que se obtiene al dividir el error o la suma de cuadrados del residual entre sus grados de libertad, proporciona una estimación de σ^2. El valor en la celda D25, 191.25, es el error cuadrado medio de los resultados de regresión para el problema de Armand's. En la sección 14.5 también se mostró que se puede usar una prueba F como prueba de significancia en la regresión. El valor en la celda F24, 0.0000255, es el valor-p que corresponde a la prueba F de significancia. Dado que el valor-p = 0.0000255 $<$ $\alpha = 0.01$, H_0 es rechazada y se concluye que se tiene una relación significativa entre la población de estudiantes y las ventas trimestrales. En la celda F23, el rótulo que emplea Excel para identificar el valor-p de la prueba F de significancia es *Significance F*.

La etiqueta Significance F se entiende mejor si se considera el valor en la celda F24 como el nivel de significancia de la prueba F.

Interpretación de los resultados de los estadísticos de regresión

El coeficiente de determinación, 0.9027, aparece en la celda B17; el rótulo correspondiente, R Square, se presenta en la celda A17. La raíz cuadrada del coeficiente de determinación es el coeficiente de correlación muestral, 0.9501, que aparece en la celda B16. Observe que para identificar este valor, Excel emplea como rótulo Multiple R (celda A16). En la celda A19, el rótulo Standard Error se usa para identificar el valor del error estándar de estimación que aparece en la celda B19. Así que el error estándar de estimación es 13.8293. Hay que considerar que en los resultados de Excel, el rótulo Standard Error está en dos lugares: en la sección de los resultados titulada Regression Statistics, Standard Error se refiere a la estimación de σ, y en la sección de los resultados correspondiente a la ecuación de regresión estimada se refiere a s_{b_1}, la desviación estándar de la distribución muestral de b_1.

Apéndice 14.5 Análisis de regresión con StatTools

WEB archivo
Armand's

En este apéndice se muestra el uso de StatTools para realizar los cálculos del análisis de regresión empleando el problema de Armand's Pizza Parlors. Se inicia con el uso del Data Set Manager a efecto de crear un conjunto de datos de StatTools para los datos utilizados en el procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes describen el uso de StatTools para proporcionar los resultados de la regresión.

Paso 1. Haga clic en la ficha **StatTools** en la cinta.
Paso 2. En el grupo **Analyses**, haga clic en **Regression and Classification**.
Paso 3. Elija la opción **Regression**.
Paso 4. Cuando el cuadro de diálogo StatTools-Regression aparezca:
Seleccione **Multiple** en el cuadro **Regression Type**.
En la sección **Variables:**
Haga clic en el botón **Format** y seleccione **Unstacked**.
En la columna etiquetada **I** seleccione **Population**.
En la columna etiquetada **D** seleccione **Sales**.
Haga clic en **OK**.

El resultado del análisis de regresión aparecerá en una nueva hoja de trabajo.

Observe que en el paso 4 se seleccionó Multiple en el cuadro Regression Type. En StatTools, esta opción se utiliza tanto para la regresión lineal simple como para la regresión múltiple. El cuadro de diálogo StatTools-Regression contiene varias de las opciones más avanzadas para desarrollar estimaciones de intervalos de predicción y gráficas de residuales. La herramienta StatTools Help proporciona información sobre el uso de todas esas opciones.

CAPÍTULO 15

Regresión múltiple

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: dunnhumby

15.1 MODELO DE REGRESIÓN MÚLTIPLE
- Modelo de regresión y ecuación de regresión
- Ecuación de regresión múltiple estimada

15.2 MÉTODO DE MÍNIMOS CUADRADOS
- Un ejemplo: Butler Trucking Company
- Nota sobre la interpretación de los coeficientes

15.3 COEFICIENTE DE DETERMINACIÓN MÚLTIPLE

15.4 SUPUESTOS DEL MODELO

15.5 PRUEBA DE SIGNIFICANCIA
- Prueba *F*
- Prueba *t*
- Multicolinealidad

15.6 USO DE LA ECUACIÓN DE REGRESIÓN ESTIMADA PARA ESTIMACIONES Y PREDICCIONES

15.7 VARIABLES INDEPENDIENTES CUALITATIVAS
- Un ejemplo: Johnson Filtration Inc.
- Interpretación de los parámetros
- Variables cualitativas más complejas

15.8 ANÁLISIS RESIDUAL
- Detección de observaciones atípicas
- Residuales estudentizados eliminados y observaciones atípicas
- Observaciones influyentes
- Uso de la medida de la distancia de Cook para identificar observaciones influyentes

15.9 REGRESIÓN LOGÍSTICA
- Ecuación de regresión logística
- Estimación de la ecuación de regresión logística
- Prueba de significancia
- Uso en la administración
- Interpretación de la ecuación de regresión logística
- Transformación logit

ESTADÍSTICA *en* LA PRÁCTICA

dunnhumby*

LONDRES, INGLATERRA

dunnhumby, fundada en 1989 por el equipo de esposos Clive Humby (matemático) y Edwina Dunn (experto en marketing), combina probadas habilidades naturales con grandes ideas para encontrar claves y patrones de lo que los consumidores compran y por qué. La empresa convierte esas señales en estrategias viables que generan notorio crecimiento y lealtad sostenible y mejoran, en última instancia, el valor de la marca y la experiencia del cliente.

Con una nómina de más de 950 personas en Europa, Asia y América, dunnhumby proporciona servicio a una lista de prestigiosas empresas, que incluye Kroger, Tesco, Coca-Cola, General Mills, Kimberly-Clark, PepsiCo, Procter & Gamble y Home Depot. dunnhumbyUSA es una empresa conjunta (*joint venture*) entre Kroger y dunnhumby, con oficinas en Nueva York, Chicago, Atlanta, Minneapolis, Cincinnati y Portland.

Los estudios de dunnhumby inician con la obtención de datos de los consumidores de la empresa que la contrata. Los datos se toman de registros de compras con tarjetas de descuento o recompensas para el cliente, operaciones en los puntos de venta electrónicos e investigación de mercados tradicional. El análisis de los datos con frecuencia se traduce de miles de millones de puntos de datos en ideas detalladas acerca del comportamiento, preferencias y estilos de vida de los clientes. Tales ideas conducen a poner en acción efectivos programas de comercialización, que incluyen recomendaciones estratégicas acerca de fijación de precios, promoción, publicidad y decisiones sobre surtido de productos.

Los investigadores utilizan una técnica de regresión múltiple llamada *regresión logística* como ayuda en sus análisis de datos basados en el cliente. Al utilizar la regresión logística se desarrolla una ecuación de regresión múltiple estimada de la siguiente forma.

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_px_p$$

La variable dependiente $\hat{y}$ es una estimación de la probabilidad de que un cliente pertenezca a un grupo específico de consumidores. Las variables independientes x_1, x_2, x_3, . . . , x_p son medidas del comportamiento de compra real del consumidor y pueden incluir el artículo específico adquirido, la cantidad adquirida, monto de la compra, día de la semana, hora del día, y así sucesivamente. El análisis ayuda a identificar las variables independientes que son más relevantes para predecir el grupo del consumidor y proporciona una mejor comprensión de la población de clientes, posibilitando análisis posteriores con una confiabilidad mucho mayor. El enfoque del análisis se dirige a la comprensión del cliente hasta el punto de desarrollar programas de comercialización, marketing y marketing directo que maximicen la pertinencia y el servicio al grupo de consumidores.



dunnhumby utiliza la regresión logística para predecir el comportamiento de compra del consumidor.
© Ariel Skelley/Blend Images/Jupiter Images

En este capítulo se analizará la regresión múltiple y cómo los conceptos de la regresión lineal simple estudiados en el capítulo 14 pueden aplicarse al caso de la regresión múltiple. Además, se mostrará el uso del software respectivo. En la última sección del capítulo se presenta la regresión logística con un ejemplo que ilustra cómo utilizar esta técnica en una aplicación de investigación de mercados.

* Los autores agradecen a Paul Hunter, vicepresidente senior de Solutions for dunnhumby, por proporcionar este material para *Estadística en la práctica*.

En el capítulo 14 se presentó la regresión lineal simple y su uso en la obtención de una ecuación de regresión estimada que describe la relación entre dos variables. Recuerde que la variable que se predice o explica es la variable dependiente y, mientras que la independiente se utiliza para predecir o explicar la variable dependiente. En este capítulo se continúa con el estudio del análisis de regresión considerando las situaciones en las que intervienen dos o más variables independientes. Este estudio, al que se le conoce como **análisis de regresión múltiple**, permite tomar más factores en consideración y, por tanto, obtener mejores estimaciones que las que ofrece la regresión lineal simple.

15.1 Modelo de regresión múltiple

El análisis de regresión múltiple estudia la relación de una variable dependiente y con dos o más variables independientes. En términos generales, suele utilizarse p para denotar el número de estas últimas.

Modelo de regresión y ecuación de regresión

Los conceptos de modelo de regresión y ecuación de regresión estudiados en el capítulo anterior son aplicables al caso de la regresión múltiple. A la ecuación que describe cómo está relacionada la variable dependiente y con las variables independientes $x_1, x_2, \ldots, x_p$ y un término de error se le conoce como **modelo de regresión múltiple**. Se inicia con el supuesto de que este modelo toma la forma siguiente.

MODELO DE REGRESIÓN MÚLTIPLE

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \qquad (15.1)$$

En el modelo de regresión múltiple, $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, son los parámetros y el término de error ϵ (la letra griega épsilon) es una variable aleatoria. Examinando con atención este modelo vemos que y es una función lineal de $x_1, x_2, \ldots, x_p$ (la parte de $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$) más el término de error ϵ. Este último corresponde a la variabilidad en y que no puede ser explicada por el efecto lineal de las p variables independientes.

En la sección 15.4 se discutirán los supuestos para el modelo de regresión múltiple y para ϵ. Uno de los supuestos indica que el valor medio o esperado de ϵ es cero. Una consecuencia de este supuesto implica que el valor medio o esperado de y, que se denota $E(y)$, es igual a $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$. A la ecuación que describe cómo está relacionado el valor medio de y con $x_1, x_2, \ldots, x_p$ se le conoce como **ecuación de regresión múltiple**.

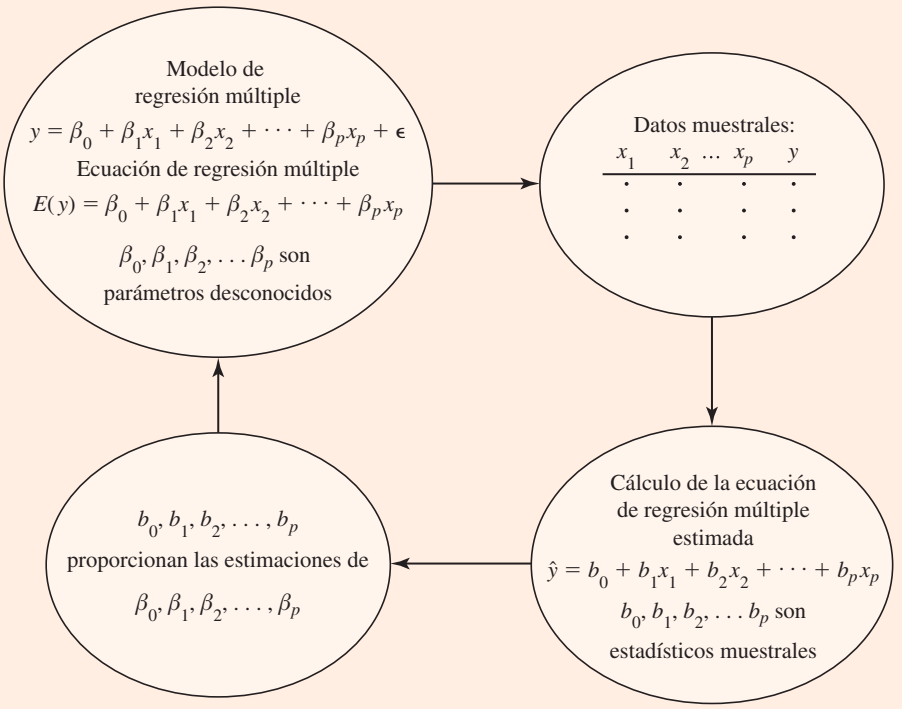
ECUACIÓN DE REGRESIÓN MÚLTIPLE

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (15.2)$$

Ecuación de regresión múltiple estimada

Si se conocieran los valores de $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, se podría usar la ecuación (15.2) a efecto de calcular el valor medio de y para valores dados de $x_1, x_2, \ldots, x_p$. Desafortunadamente, los valores de estos parámetros suelen en general no conocerse y es necesario estimarlos a partir de datos muestrales. Para calcular los valores de los estadísticos muestrales $b_0, b_1, b_2, \ldots, b_p$, que se usan como estimadores puntuales de los parámetros $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, se emplea una

FIGURA 15.1 Proceso de estimación en la regresión múltiple



En la regresión lineal simple, b_0 y b_1 son los estadísticos muestrales utilizados para estimar los parámetros β_0 y β_1. En la regresión múltiple, en el proceso de inferencia estadística análogo, $b_0, b_1, b_2, \ldots, b_p$ denotan los estadísticos muestrales utilizados para estimar los parámetros $\beta_0, \beta_1, \beta_2, \ldots \beta_p$.

muestra aleatoria simple. Con los estadísticos muestrales se obtiene la siguiente **ecuación de regresión múltiple estimada**.

ECUACIÓN DE REGRESIÓN MÚLTIPLE ESTIMADA

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \qquad \textbf{(15.3)}$$

donde:

$b_0, b_1, b_2, \ldots, b_p$ son las estimaciones de $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$
$\hat{y}$ = valor estimado de la variable dependiente

Este proceso de estimación para la regresión múltiple se muestra en la figura 15.1.

15.2 Método de mínimos cuadrados

En el capítulo 14 se usó el **método de mínimos cuadrados** para obtener la ecuación de regresión estimada que permitía aproximar mejor la relación lineal entre las variables dependiente e independiente. Con este método también se desarrolla la ecuación de regresión múltiple estimada. El criterio en el método de mínimos cuadrados, como ya se dijo, es el siguiente.

CRITERIO DE MÍNIMOS CUADRADOS

$$\min \Sigma(y_i - \hat{y}_i)^2 \qquad \textbf{(15.4)}$$

donde:

y_i = valor observado de la variable dependiente para la observación *i*ésima
$\hat{y}_i$ = valor estimado de la variable dependiente para la observación *i*ésima

Los valores estimados de la variable dependiente se calculan con la ecuación de regresión múltiple estimada

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p$$

Como indica la expresión (15.4), el método de mínimos cuadrados usa datos muestrales para obtener los valores de $b_0, b_1, b_2, \ldots, b_p$ que hacen que la suma de los cuadrados de los residuales [las desviaciones entre los valores observados de la variable dependiente (y_i) y los valores estimados de la variable dependiente ($\hat{y}_i$)] sea un mínimo.

En el capítulo 14 se proporcionaron las fórmulas para calcular los estimadores b_0 y b_1 para la ecuación de regresión lineal simple estimada $\hat{y} = b_0 + b_1x$ empleando el método de mínimos cuadrados. Con conjuntos de datos relativamente pequeños fue posible usar esas fórmulas para obtener b_0 y b_1 mediante cálculos manuales. En la regresión múltiple, en cambio, las fórmulas para los coeficientes de regresión $b_0, b_1, b_2, \ldots, b_p$ utilizan álgebra matricial y quedan fuera del alcance de este libro. Por esta razón, el estudio de la regresión múltiple centrará la atención en el uso de software para obtener la ecuación de regresión estimada y alguna otra información. Se hará énfasis en la interpretación de los resultados que proporciona este software y no en cómo efectuar los cálculos para la regresión múltiple.

Un ejemplo: Butler Trucking Company

Para ilustrar el análisis de regresión múltiple se utilizará el problema de Butler Trucking Company, una empresa que se dedica al transporte de objetos y mercancías en el sur de California. Su actividad principal es hacer entregas en su área local. Para mejorar el horario de trabajo, los gerentes deseaban estimar el tiempo total de recorrido diario necesario para efectuar las entregas.

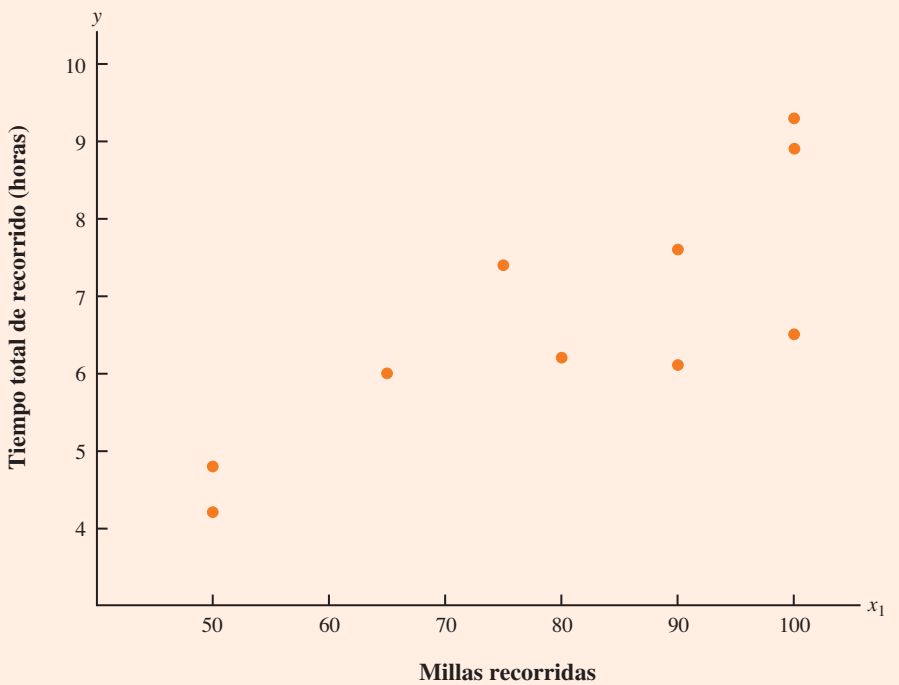
Al principio, los gerentes creyeron que el tiempo total de recorrido diario estaba estrechamente relacionado con el número de millas recorridas para realizar las entregas. A partir de una muestra aleatoria simple de 10 repartidores con asignación de recorrido (Driving Assignment) se obtuvieron los datos que se presentan en la tabla 15.1 y en el diagrama de dispersión de la figura 15.2. Después de observar este diagrama, los gerentes consideraron que, para describir la relación entre tiempo total de recorrido y en horas (Travel Time/hours) y el número de millas recorridas x_1 (Miles Traveled), podía emplearse el modelo de regresión lineal simple

TABLA 15.1 Datos preliminares para Butler Trucking

Driving Assignment	x_1 = Miles Traveled	y = Travel Time (hours)
1	100	9.3
2	50	4.8
3	100	8.9
4	100	6.5
5	50	4.2
6	80	6.2
7	75	7.4
8	65	6.0
9	90	7.6
10	90	6.1

WEB archivo
Butler

FIGURA 15.2 Diagrama de dispersión de los datos preliminares del ejemplo de Butler Trucking



$y = \beta_0 + \beta_1 x_1 + \epsilon$. Para estimar los parámetros β_0 y β_1 se empleó el método de mínimos cuadrados y se obtuvo la ecuación de regresión estimada.

$$\hat{y} = b_0 + b_1 x_1 \qquad \textbf{(15.5)}$$

En la figura 15.3 se presentan los resultados obtenidos con Minitab aplicando la regresión lineal simple a los datos de la tabla 15.1. La ecuación de regresión estimada es

$$\hat{y} = 1.27 + 0.0678 x_1$$

Con 0.05 como nivel de significancia, el valor F de 15.81 y su correspondiente valor-p de 0.004 indican que la relación es significativa; es decir, que H_0: $\beta_1 = 0$ puede ser rechazada debido a que el valor-p es menor que $\alpha = 0.05$. Observe que utilizando el valor t de 3.98, y su valor-p asociado de 0.004, se llega a la misma conclusión. Por tanto, podemos concluir que la relación entre el tiempo total de recorrido y el número de millas recorridas es significativa; trayectos de más duración corresponden a cantidades mayores de millas recorridas. Como el coeficiente de determinación (expresado como porcentaje) es R-sq = 66.4%, vemos que 66.4% de la variabilidad en el tiempo de recorrido podemos explicarla por el efecto lineal del número de millas recorridas. Este descubrimiento es bastante satisfactorio; sin embargo, los gerentes deseaban considerar otra variable independiente más para explicar parte de la variabilidad restante de la variable dependiente.

Al tratar de identificar otra variable independiente, los gerentes encontraron que el número de entregas podía contribuir también a la duración total del recorrido. En la tabla 15.2 se presentan los datos de Butler Trucking después de agregar el número de entregas. En la figura 15.4 se expone el resultado que provee Minitab al considerar variables independientes, tanto el número de millas recorridas (x_i) como el número de entregas (x_2) realizadas. La ecuación de regresión estimada es

$$\hat{y} = 0.869 + 0.0611 x_1 + 0.923 x_2 \qquad \textbf{(15.6)}$$

FIGURA 15.3 Resultados de Minitab para el problema de Butler Trucking con una variable independiente

Los nombres de las variables Miles (*millas*) *y* Time (*tiempo*) *que aparecen en los resultados de Minitab fueron ingresados en la hoja de trabajo como encabezados de las columnas correspondientes; por tanto,* x_1 = Miles y y = Time.

```
The regression equation is
Time = 1.27 + 0.0678 Miles

Predictor      Coef  SE Coef     T      p
Constant      1.274    1.401  0.91  0.390
Miles       0.06783  0.01706  3.98  0.004

S = 1.00179   R-sq = 66.4%   R-sq(adj) = 62.2%

Analysis of Variance

SOURCE           DF      SS      MS      F      p
Regression        1  15.871  15.871  15.81  0.004
Residual Error    8   8.029   1.004
Total             9  23.900
```

En la siguiente sección se analizará el uso del coeficiente de determinación múltiple para medir qué tan buen ajuste proporciona la ecuación de regresión estimada. Antes se examinarán con más cuidado los valores de $b_1 = 0.0611$ y $b_2 = 0.923$ en la ecuación (15.6).

Nota sobre la interpretación de los coeficientes

En este punto es útil hacer una observación sobre la relación entre la ecuación de regresión estimada en la que la única variable independiente es el número de millas recorridas y la ecuación en la que, como segunda variable independiente, se incluye el número de entregas x_2 (Number of Deliveries). El valor de b_1 no es igual en ambos casos. En la regresión lineal simple, b_1 se interpreta como una estimación del cambio en y debido al cambio en una unidad de la variable independiente. En el análisis de regresión múltiple esta interpretación cambia ligeramente. Es decir, en este tipo de análisis cada uno de los coeficientes de regresión se interpreta como sigue: b_1 representa la estimación del cambio en y debido a un cambio en una unidad en x_i, mientras todas las demás variables independientes permanecen constantes. En el ejemplo de Butler Trucking con dos variables independientes, $b_1 = 0.0611$. Por tanto, 0.0611 horas es la estimación del aumento esperado en el tiempo de recorrido que corresponde al incremento de una milla en

TABLA 15.2 Datos de Butler Trucking con millas recorridas (x_1) y cantidad de entregas (x_2) como variables independientes

WEB archivo
Butler

Driving Assignment	x_1 = Miles Traveled	x_2 = Number of Deliveries	y = Travel Time (hours)
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	6.5
5	50	2	4.2
6	80	2	6.2
7	75	3	7.4
8	65	4	6.0
9	90	3	7.6
10	90	2	6.1

FIGURA 15.4 Resultados de Minitab para el problema de Butler Trucking con dos variables independientes

```
The regression equation is
Time = - 0.869 + 0.0611 Miles + 0.923 Deliveries

Predictor       Coef   SE Coef      T      p
Constant     -0.8687    0.9515  -0.91  0.392
Miles       0.061135  0.009888   6.18  0.000
Deliveries    0.9234    0.2211   4.18  0.004

S = 0.573142   R-sq = 90.4%   R-sq(adj) = 87.6%

Analysis of Variance

SOURCE          DF      SS      MS      F      p
Regression       2  21.601  10.800  32.88  0.000
Residual Error   7   2.299   0.328
Total            9  23.900
```

Los nombres de las variables Miles *(millas),* Deliveries (*entregas*) *y* Time (*tiempo*) *que aparecen en los resultados de Minitab fueron ingresados en la hoja de trabajo como encabezados de columna; por tanto,* x_1 = Miles, x_2 = Deliveries y y = Time.

la distancia recorrida cuando el número de entregas permanece constante. De manera similar, como $b_2 = 0.923$, una estimación del incremento esperado en el tiempo de recorrido que corresponde al aumento de una entrega cuando el número de millas recorridas permanece constante es 0.923 horas.

Ejercicios

Nota al lector. Los ejercicios de esta sección y las siguientes en los que se proporcionan datos están diseñados para ser resueltos mediante software.

Métodos

1. A continuación se proporciona la ecuación de regresión estimada obtenida a partir de 10 observaciones para un modelo con dos variables independientes.

$$\hat{y} = 29.1270 + 0.5906x_1 + 0.4980x_2$$

   a) Interprete los coeficientes b_1 y b_2 de esta ecuación de regresión estimada.
   b) Estime y para $x_1 = 180$ y $x_2 = 310$.

AUTOevaluación

2. Considere los datos siguientes que corresponden a la variable dependiente y y las dos variables independientes x_1 y x_2.

WEB archivo
Exer2

x_1	x_2	y
30	12	94
47	10	108
25	17	112
51	16	178
40	5	94
51	19	175
74	7	170

(*Continúa*)

x_1	x_2	y
36	12	117
59	13	142
76	16	211

a) Obtenga una ecuación de regresión estimada que relacione y con x_1. Estime y si $x_1 = 45$.
b) Desarrolle una ecuación de regresión estimada que relacione y con x_2. Estime y si $x_2 = 15$.
c) Obtenga una ecuación de regresión estimada que relacione y con x_1 y x_2. Calcule y si $x_1 = 45$ y $x_2 = 15$.

3. En un análisis de regresión se emplean 30 observaciones y se obtiene la siguiente ecuación de regresión estimada.

$$\hat{y} = 17.6 + 3.8x_1 - 2.3x_2 + 7.6x_3 + 2.7x_4$$

a) Interprete los coeficientes b_1, b_2, b_3 y b_4 de esta ecuación de regresión estimada.
b) Estime y para $x_1 = 10$; $x_2 = 5$; $x_3 = 1$, y $x_4 = 2$.

Aplicaciones

4. Una zapatería obtuvo la siguiente ecuación de regresión estimada en la que se relacionan las ventas contra la inversión en inventario y los gastos en publicidad.

$$\hat{y} = 25 + 10x_1 + 8x_2$$

donde

x_1 = inversión en inventario (en miles de $)
x_2 = gasto en publicidad (en miles de $)
y = ventas (en miles de $)

a) Estime las ventas resultantes si la inversión en inventario es de $15 000 y el presupuesto para publicidad es de $10 000.
b) Interprete b_1 y b_2 en esta ecuación de regresión estimada.

AUTO evaluación

5. El propietario de Showtime Movie Theaters, Inc. desea estimar el ingreso bruto semanal (Weekly Gross Revenue) en función de los gastos en publicidad por televisión (Television Advertising) y en periódicos (Newspaper Advertising). A continuación se presentan los datos históricos de ocho semanas en miles de dólares ($1 000s).

WEB archivo
Showtime

Weekly Gross Revenue ($1 000s)	Television Advertising ($1 000s)	Newspaper Advertising ($1 000s)
96	5.0	1.5
90	2.0	2.0
95	4.0	1.5
92	2.5	2.5
95	3.0	3.3
94	3.5	2.3
94	2.5	4.2
94	3.0	2.5

a) Obtenga una ecuación de regresión estimada en la que el monto gastado en publicidad por televisión sea la variable independiente.
b) Desarrolle una ecuación de regresión estimada en la que los montos gastados en publicidad por televisión y periódicos sean las variables independientes.
c) ¿Es el coeficiente correspondiente a los gastos de publicidad en televisión de la ecuación de regresión estimada del inciso *a*) igual al del inciso *b*)? Interprete este coeficiente en cada caso.

d) ¿Cuál es la estimación del ingreso semanal bruto en una semana en la que se gastan $3 500 en publicidad en televisión y $1 800 en publicidad en periódicos?

6. En el beisbol, el éxito de un equipo suele valorarse en función del desempeño en bateo y en lanzamiento. Una medida del desempeño en el bateo es la cantidad de jonrones que anota el equipo mientras que en lanzamiento es el promedio de carreras permitidas por el equipo que lanza. En general, se cree que los equipos que anotan más jonrones y tienen un promedio menor de carreras permitidas ganan un mayor porcentaje de juegos. Los datos siguientes muestran la proporción de juegos ganados (Proportion Won), la cantidad de jonrones (HR, home runs) del equipo (Team) y el promedio de carreras permitidas (ERA, earned run average) de 16 equipos de la Liga Nacional que participaron en la temporada de las Grandes Ligas de Beisbol de 2003 (sitio web de *USA Today*, 7 de enero de 2004).

WEB archivo
MLB

Team	Proportion Won	HR	ERA	Team	Proportion Won	HR	ERA
Arizona	0.519	152	3.857	Milwaukee	0.420	196	5.058
Atlanta	0.623	235	4.106	Montreal	0.512	144	4.027
Chicago	0.543	172	3.842	New York	0.410	124	4.517
Cincinnati	0.426	182	5.127	Philadelphia	0.531	166	4.072
Colorado	0.457	198	5.269	Pittsburgh	0.463	163	4.664
Florida	0.562	157	4.059	San Diego	0.395	128	4.904
Houston	0.537	191	3.880	San Francisco	0.621	180	3.734
Los Ángeles	0.525	124	3.162	St. Louis	0.525	196	4.642

a) Obtenga la ecuación de regresión estimada para predecir la proporción de juegos ganados en función de la cantidad de jonrones.

b) Desarrolle la ecuación de regresión estimada para predecir la proporción de juegos ganados dado el promedio de carreras permitidas por los miembros del equipo que lanza.

c) Obtenga la ecuación de regresión estimada para predecir la proporción de juegos ganados en función de la cantidad de jonrones y del promedio de carreras permitidas por los miembros del equipo que lanza.

d) En la temporada 2003, San Diego ganó sólo 39.5% de sus juegos, el más bajo de la Liga Nacional. Para mejorar el récord del año siguiente, el equipo buscó nuevos jugadores que incrementaran la cantidad de jonrones a 180 y disminuyera el promedio de carreras permitidas por el equipo que lanza a 4.0. Use la ecuación de regresión estimada obtenida en el inciso *c*) para estimar el porcentaje de juegos que ganaría San Diego si tuviera 180 jonrones y su promedio de carreras permitidas fuera de 4.0.

7. *PC World* valoró cuatro componentes singulares de 10 computadoras laptop ultraportátiles: características, desempeño, diseño y precio. Cada elemento fue valorado con una escala de puntos de 0 a 100. Luego se desarrolló un índice general, llamado índice *PCW World*, para cada laptop. La siguiente tabla muestra el índice de desempeño (Performance), el índice de características (Features) y el índice (Rating) *PCW World* para las 10 computadoras (sitio web de PC World, 5 de febrero de 2009).

WEB archivo
Laptop

Model	Performance	Features	PCW Rating
Thinkpad X200	77	87	83
VGN-Z598U	97	85	82
U6V	83	80	81
Elitebook 2530P	77	75	78
X360	64	80	78
Thinkpad X300	56	76	78
Ideapad U110	55	81	77
Micro Express JFT2500	76	73	75
Toughbook W7	46	79	73
HP Voodoo Envy133	54	68	72

a) Determine la ecuación de regresión estimada para predecir el índice *PCW World*, utilizando el índice de desempeño como variable independiente.
b) Obtenga la ecuación de regresión estimada para predecir el índice *PCW World*, utilizando tanto el índice de desempeño como el de características.
c) Prediga cuál será el índice *PCW World* de una computadora laptop con un índice de desempeño de 80 y un índice de características de 70.

8. ¿Se esperaría mayor confiabilidad y desempeño en autos que cuestan más? *Consu-mer Reports* dio a conocer índices de confiabilidad (Reliability), calificaciones generales en pruebas de carretera (Road-Test Score) y precios (Price) de automóviles sedán familiares económicos, de precio medio y precio alto, incluyendo fabricante y modelo (Make and Model) (*Consumer Reports*, febrero de 2008). A continuación se proporciona una parte de los datos. La confiabilidad se valoró en una escala de 5 puntos, desde deficiente (1) hasta excelente (5). La calificación en pruebas de carretera se valoró en una escala de 100 puntos, donde los valores más altos indican un mejor desempeño. Los datos completos se encuentran en el archivo Sedans.

WEB archivo
Sedans

Make and Model	Road-Test Score	Reliability	Price ($)
Nissan Altima 2.5 S	85	4	22705
Honda Accord LX-P	79	4	22795
Kia Optima EX (4-cyl.)	78	4	22795
Toyota Camry LE	77	4	21080
Hyundai Sonata SE	76	3	22995
⋮	⋮	⋮	⋮
Chrysler 300 Touring	60	2	30255
Dodge Charger SXT	58	4	28860

a) Obtenga la ecuación de regresión estimada para predecir el precio del automóvil, dado el índice de confiabilidad. Pruebe la significancia con $\alpha = 0.05$.
b) Considere la incorporación de la calificación general en pruebas de carretera como variable independiente. Obtenga la ecuación de regresión estimada para predecir el precio del automóvil, dados la calificación en pruebas de carretera y el índice de confiabilidad.
c) Estime el precio de un automóvil con una calificación en pruebas de carretera de 80 y un índice de confiabilidad de 4.

9. El waterskiing y el wakeboarding son dos populares deportes acuáticos. Ya se trate de uno o de otro, o de simple navegación, hallar el equipo que mejor se ajuste a las necesidades puede resultar una ardua tarea. La revista *WaterSki* realizó amplias pruebas en 88 botes y proporcionó una amplia variedad de información como ayuda para los consumidores. A continuación se presenta una parte de los datos que publicó sobre 20 lanchas, incluyendo fabricante y modelo (Make and Model), de entre 20 y 22 pies de longitud (*WaterSki,* enero/febrero de 2006). La manga (Beam) es el ancho máximo del bote en pulgadas, la potencia del motor se mide en caballos de fuerza (HP) y la velocidad máxima (TopSpeed) es la que puede alcanzar el bote en millas por hora (mph).

WEB archivo
Boats

Make and Model	Beam	HP	TopSpeed
Calabria Cal Air Pro V-2	100	330	45.3
Correct Craft Air Nautique 210	91	330	47.3
Correct Craft Air Nautique SV-211	93	375	46.9
Correct Craft Ski Nautique 206 Limited	91	330	46.7
Gekko GTR 22	96	375	50.1
Gekko GTS 20	83	375	52.2
Malibu Response LXi	93.5	340	47.2
Malibu Sunsetter LXi	98	400	46
Malibu Sunsetter 21 XTi	98	340	44

Make and Model	Beam	HP	TopSpeed
Malibu Sunscape 21 LSV	98	400	47.5
Malibu Wakesetter 21 XTi	98	340	44.9
Malibu Wakesetter VLX	98	400	47.3
Malibu vRide	93.5	340	44.5
Malibu Ride XTi	93.5	320	44.5
Mastercraft ProStar 209	96	350	42.5
Mastercraft X-1	90	310	45.8
Mastercraft X-2	94	310	42.8
Mastercraft X-9	96	350	43.2
MB Sports 190 Plus	92	330	45.3
Svfara SVONE	91	330	47.7

a) Con estos datos, obtenga la ecuación de regresión estimada que relaciona la velocidad máxima con la manga y los caballos de fuerza del bote.

b) El Svfara SV609 tiene una manga de 85 pulgadas y motor de 330 HP. Utilice la ecuación de regresión estimada del inciso *a*) para estimar la velocidad máxima de este modelo.

10. La Asociación Nacional de Basquetbol (NBA, por sus siglas en inglés) registra diversos datos estadísticos de cada equipo. Cuatro de estos datos indican la proporción de juegos ganados (PCT), el porcentaje de anotaciones de campo (FG%), la proporción de tiros de tres puntos logrados por el equipo contrario (Opp 3 Pt%) y la cantidad de pérdidas de balón del equipo adversario (Opp TO). La siguiente tabla muestra los valores de estos datos estadísticos para los 29 equipos (Team) de la NBA en parte de la temporada 2004 (sitio web de la NBA, 3 de enero de 2004).

WEB archivo
NBA

Team	PCT	FG%	Opp 3 Pt%	Opp TO	Team	PCT	FG%	Opp 3 Pt%	Opp TO
Atlanta	0.265	0.435	0.346	13.206	Minnesota	0.677	0.473	0.348	13.839
Boston	0.471	0.449	0.369	16.176	New Jersey	0.563	0.435	0.338	17.063
Chicago	0.313	0.417	0.372	15.031	New Orleans	0.636	0.421	0.330	16.909
Cleveland	0.303	0.438	0.345	12.515	New York	0.412	0.442	0.330	13.588
Dallas	0.581	0.439	0.332	15.000	Orlando	0.242	0.417	0.360	14.242
Denver	0.606	0.431	0.366	17.818	Philadelphia	0.438	0.428	0.364	16.938
Detroit	0.606	0.423	0.262	15.788	Phoenix	0.364	0.438	0.326	16.515
Golden State	0.452	0.445	0.384	14.290	Portland	0.484	0.447	0.367	12.548
Houston	0.548	0.426	0.324	13.161	Sacramento	0.724	0.466	0.327	15.207
Indiana	0.706	0.428	0.317	15.647	San Antonio	0.688	0.429	0.293	15.344
L.A. Clippers	0.464	0.424	0.326	14.357	Seattle	0.533	0.436	0.350	16.767
L.A. Lakers	0.724	0.465	0.323	16.000	Toronto	0.516	0.424	0.314	14.129
Memphis	0.485	0.432	0.358	17.848	Utah	0.531	0.456	0.368	15.469
Miami	0.424	0.410	0.369	14.970	Washington	0.300	0.411	0.341	16.133
Milwaukee	0.500	0.438	0.349	14.750					

a) Desarrolle una ecuación de regresión estimada para predecir la proporción de juegos ganados, dada la proporción de anotaciones de campo del equipo.

b) Interprete la pendiente de la ecuación de regresión estimada obtenida con el inciso *a*).

c) Obtenga una ecuación de regresión estimada para predecir la proporción de juegos ganados dada la proporción de anotaciones de campo del equipo, el porcentaje de tiros de tres puntos del equipo contrario y el número de pérdidas de balón del equipo adversario.

d) Analice las implicaciones prácticas de la ecuación obtenida en el inciso *c*).

e) Estime la proporción de juegos ganados por un equipo para el que los valores de las tres variables independientes son: FG% = 0.45; Opp 3 Pt% = 0.34, y Opp TO = 17.

15.3 Coeficiente de determinación múltiple

En la regresión lineal simple se mostró que la suma de cuadrados se podía dividir o particionar en dos componentes: la suma de cuadrados debido a la regresión y la suma de cuadrados debido al error. Podemos aplicar el mismo procedimiento a la suma de cuadrados de la regresión múltiple.

RELACIÓN ENTRE STC, SCR Y SCE

$$\text{STC} = \text{SCR} + \text{SCE} \tag{15.7}$$

donde

$$\text{STC} = \text{suma total de cuadrados} = \Sigma(y_i - \bar{y})^2$$

$$\text{SCR} = \text{suma de cuadrados debido a la regresión} = \Sigma(\hat{y}_i - \bar{y})^2$$

$$\text{SCE} = \text{suma de cuadrados debido al error} = \Sigma(y_i - \hat{y}_i)^2$$

Dado lo complejo de los cálculos de estas tres sumas de cuadrados, es necesario emplear un software para realizarlos. En los resultados de Minitab de la figura 15.4, en la parte del análisis de varianza, se presentan estos tres valores para el problema de Butler Trucking con dos variables independientes: STC = 23.900, SCR = 21.601 y SCE = 2.299. Cuando se emplea una sola variable independiente (número de millas recorridas) en los resultados de Minitab de la figura 15.3 observamos que STC = 23.900, SCR = 15.871 y SCE = 8.029. El valor de la STC es el mismo en ambos casos debido a que no depende de $\hat{y}$, pero al agregar otra variable (el número de entregas), SCR aumenta y SCE disminuye. Esto tiene como consecuencia que la ecuación de regresión estimada tenga un mejor ajuste para los datos observados.

En el capítulo 14 se empleó el coeficiente de determinación, $r^2 = \text{SCR/STC}$, para medir la bondad de ajuste de la ecuación de regresión estimada. El mismo concepto es válido en la regresión múltiple. El término **coeficiente de determinación múltiple** indica que se mide la bondad de ajuste de la ecuación de regresión múltiple estimada. El coeficiente de determinación múltiple, que se denota R^2, se calcula como sigue.

COEFICIENTE DE DETERMINACIÓN MÚLTIPLE

$$R^2 = \frac{\text{SCR}}{\text{STC}} \tag{15.8}$$

Este coeficiente puede interpretarse como la proporción de la variabilidad en la variable dependiente que es explicada por la ecuación de regresión estimada. Por tanto, el producto de este coeficiente por 100 se interpreta como el porcentaje de la variabilidad en y que es explicada por la ecuación de regresión estimada.

Cuando se emplean dos variables independientes en el ejemplo de Butler Trucking, como SCR = 21.601 y STC = 23.900, tenemos

$$R^2 = \frac{21.601}{23.900} = 0.904$$

Por tanto, 90.4% de la variabilidad en el tiempo de recorrido y se explica por la ecuación de regresión estimada en la que las variables independientes son las millas recorridas y el número de entregas. En la figura 15.4 observamos que en el resultado proporcionado por Minitab aparece también el coeficiente de determinación múltiple, que se denota R-sq = 90.4%.

Al aumentar el número de variables independientes los errores de predicción se hacen más pequeños, con lo que se reduce la suma de cuadrados debido al error, SCE. *Como* SCR = STC − SCE, *cuando* SCE *disminuye,* SCR *aumenta, lo cual ocasiona que* R^2 = SCR/STC *aumente.*

En la figura 15.3 el valor de R-sq para la ecuación de regresión estimada con una sola variable, número de millas recorridas (x_1), es 66.4%. Por tanto, al agregar el número de entregas como una variable independiente más, el porcentaje de variabilidad en el tiempo de recorrido explicado por la ecuación de regresión estimada aumenta de 66.4 a 90.4%. En general, siempre que se añade una variable independiente al modelo, R^2 aumenta.

Muchos analistas prefieren ajustar R^2 al número de variables independientes para evitar sobreestimar el efecto que tiene agregar una variable independiente sobre la cantidad de la variabilidad explicada por la ecuación de regresión estimada. Siendo n el número de observaciones y p el número de variables independientes, el **coeficiente de determinación múltiple ajustado** se calcula como sigue.

Cuando se agrega una variable al modelo, R^2 *se vuelve más grande, aun cuando esta variable no sea estadísticamente significativa. El coeficiente de determinación múltiple ajustado compensa el número de variables independientes en el modelo.*

COEFICIENTE DE DETERMINACIÓN MÚLTIPLE AJUSTADO

$$R_a^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1} \qquad (15.9)$$

En el ejemplo de Butler Trucking con $n = 10$ y $p = 2$ tenemos

$$R_a^2 = 1 - (1 - 0.904)\frac{10 - 1}{10 - 2 - 1} = 0.88$$

Por tanto, una vez que el coeficiente de determinación múltiple se ha ajustado a dos variables independientes, su valor es de 0.88. En los resultados de Minitab de la figura 15.4 este valor (expresado como porcentaje) se presenta como R-sq(adj) = 87.6%; el valor obtenido arriba difiere porque en los cálculos se empleó un valor redondeado de R^2.

NOTAS Y COMENTARIOS

Si el valor de R^2 es pequeño y el número de variables independientes en el modelo es grande, el coeficiente de determinación ajustado puede tomar un valor negativo; en tales casos, Minitab establece el cero como coeficiente de determinación ajustado.

Ejercicios

Métodos

11. En el ejercicio 1 se presentó la siguiente ecuación de regresión estimada basada en 10 observaciones.

$$\hat{y} = 29.1270 + 0.5906x_1 + 0.4980x_2$$

Los valores de STC y SCR son 6724.125 y 6216.375, respectivamente.

a) Determine SCE.
b) Calcule R^2.
c) Calcule R_a^2.
d) Analice la bondad de ajuste.

AUTOevaluación

12. En el ejercicio 2 se presentaron 10 observaciones con los valores de la variable dependiente y y dos variables independientes x_1 y x_2; con estos datos, STC = 15 182.9 y SCR = 14 052.2.

a) Calcule R^2.
b) Determine R_a^2.
c) ¿La ecuación de regresión estimada explica una proporción grande de la variabilidad de los datos? Explique.

13. En el ejercicio 3 se presentó la siguiente ecuación de regresión estimada basada en 30 observaciones.

$$\hat{y} = 17.6 + 3.8x_1 - 2.3x_2 + 7.6x_3 + 2.7x_4$$

Los valores de STC y SCR son 1805 y 1760, respectivamente.

a) Calcule R^2.
b) Calcule R_a^2.
c) Analice la bondad de ajuste.

Aplicaciones

14. En el ejercicio 4 se proporcionó la siguiente ecuación de regresión estimada que relaciona las ventas contra la inversión en inventario y los gastos de publicidad.

$$\hat{y} = 25 + 10x_1 + 8x_2$$

Los datos para desarrollar este modelo provienen de 10 tiendas; con esta información, la STC = 16 000 y la SCR = 12 000.

a) Calcule R^2 para la ecuación de regresión estimada.
b) Calcule R_a^2.
c) ¿Este modelo parece explicar gran parte de la variabilidad de los datos? Explique.

AUTO evaluación

15. En el ejercicio 5 el propietario de Showtime Movie Theaters Inc. empleó el análisis de regresión múltiple para predecir el ingreso bruto (y) en función de la publicidad en televisión (x_1) y la publicidad en los periódicos (x_2). La ecuación de regresión estimada es

$$\hat{y} = 83.2 + 2.29x_1 - 1.30x_2$$

WEB archivo
Showtime

La solución obtenida con software proporciona STC = 25.2 y SCR = 23.435.

a) Calcule e interprete R^2 y R_a^2.
b) Cuando la publicidad en televisión es la variable independiente, $R^2 = 0.653$ y $R_a^2 = 0.595$. ¿Prefiere los resultados de la regresión múltiple? Explique.

WEB archivo
MLB

16. En el ejercicio 6 se presentaron los datos sobre la proporción de juegos ganados, la cantidad de jonrones del equipo y el promedio de carreras permitidas por el equipo lanzador para los 16 equipos de la Liga Nacional que participaron en la temporada de las Grandes Ligas de Beisbol de 2003 (sitio web de *USA Today*, 7 de enero de 2004).

a) ¿La ecuación de regresión estimada aporta un buen ajuste para predecir la proporción de juegos ganados si tiene como única variable independiente la cantidad de jonrones? Explique.
b) Analice la ventaja de usar tanto la cantidad de jonrones como el promedio de carreras ganadas para predecir la proporción de juegos ganados.

WEB archivo
Boats

17. En el ejercicio 9 se obtuvo una ecuación de regresión estimada que relaciona la velocidad máxima de un bote con la manga y los caballos de fuerza del motor.

a) Calcule e interprete R^2 y R_a^2.
b) ¿Esta ecuación de regresión estimada proporciona un buen ajuste? Explique.

WEB archivo
NBA

18. Remítase al ejercicio 10, en el que se presentaron varios datos estadísticos de 29 equipos de la NBA en parte de la temporada 2004 (sitio web de la NBA, 3 de enero de 2004).

a) En el inciso *c*) del ejercicio 10 se obtuvo una ecuación de regresión estimada que arrojó la proporción de juegos ganados dado el porcentaje de anotaciones de campo del equipo, la proporción de tiros de tres puntos del conjunto contrario y la cantidad de recuperaciones de balón del equipo adversario. ¿Cuáles son los valores de R^2 y R_a^2?
b) ¿Esta ecuación de regresión estimada proporciona un buen ajuste a los datos? Explique.

15.4 Supuestos del modelo

En la sección 15.1 se presentó el siguiente modelo de regresión múltiple.

MODELO DE REGRESIÓN MÚLTIPLE

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \qquad (15.10)$$

Los supuestos acerca del término del error ϵ en el modelo de regresión múltiple son análogos a los supuestos en el modelo de regresión lineal simple.

SUPUESTOS SOBRE EL TÉRMINO DEL ERROR ϵ EN EL MODELO DE REGRESIÓN MÚLTIPLE

$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$

1. El término del error ϵ es una variable aleatoria cuya media o valor esperado es cero, es decir, $E(\epsilon) = 0$
   Consecuencia. Para los valores dados de $x_1, x_2, \ldots, x_p$, el valor esperado o valor promedio de y está dado por

   $$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (15.11)$$

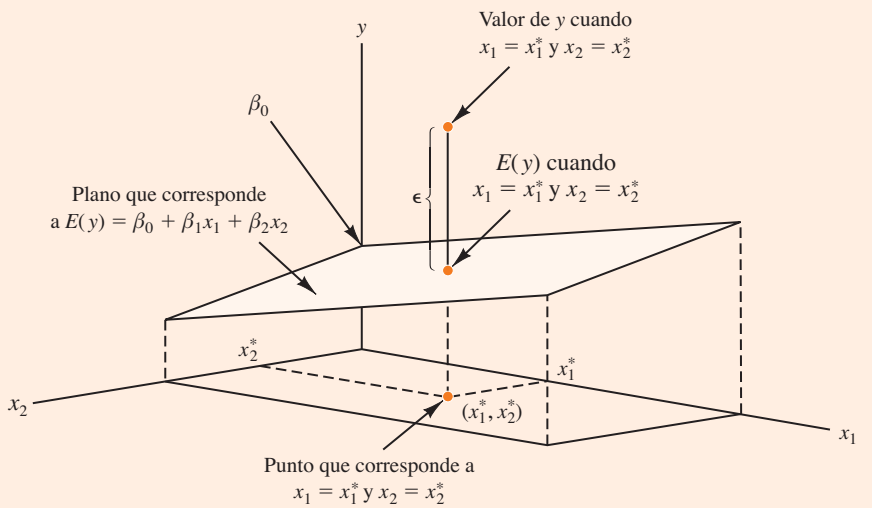
   Ésta es la ecuación de regresión múltiple presentada en la sección 15.1. En dicha expresión, $E(y)$ representa el promedio de todos los valores que puede tomar y para valores dados de $x_1, x_2, \ldots, x_p$.
2. La varianza de ϵ se denota σ^2 y es la misma para todos los valores de las variables independientes $x_1, x_2, \ldots, x_p$.
   Consecuencia. La varianza de y respecto de la línea de regresión es σ^2 y es la misma para todos los valores de $x_1, x_2, \ldots, x_p$.
3. Los valores de ϵ son independientes.
   Consecuencia. El valor de ϵ para un determinado conjunto de valores de las variables independientes no está relacionado con el valor de ϵ de ningún otro conjunto de valores.
4. El término del error ϵ es una variable aleatoria distribuida normalmente y refleja la desviación entre el valor de y y su valor esperado dado por $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.
   Consecuencia. Como $\beta_0, \beta_1, \ldots, \beta_p$ son constantes para los valores dados de $x_1, x_2, \ldots, x_p$, la variable dependiente y es también una variable aleatoria distribuida normalmente.

Para entender mejor la forma de la relación dada por la ecuación (15.11), considere la siguiente ecuación de regresión múltiple con dos variables independientes.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

La gráfica de esta ecuación es un plano en el espacio tridimensional. La figura 15.5 es un ejemplo de gráfica de este tipo. Observe que, como se indica, el valor de ϵ es la diferencia entre el verdadero valor de y y su valor esperado, $E(y)$, cuando $x_1 = x_1^*$ y $x_2 = x_2^*$.

FIGURA 15.5 Gráfica de la ecuación de regresión empleada en el análisis de regresión múltiple con dos variables independientes



En el análisis de regresión suele emplearse el término *variable de respuesta* en lugar de *variable dependiente*. Además, como la ecuación de regresión múltiple genera un plano o superficie, a su gráfica se le llama *superficie de respuesta*.

15.5 Prueba de significancia

En esta sección se estudia cómo realizar una prueba de significancia para una relación de regresión múltiple. Las pruebas de significancia utilizadas en la regresión lineal simple fueron la prueba *t* y la prueba *F*, y ambas llevan a la misma conclusión; es decir, si se rechaza la hipótesis nula, se concluye que $\beta_1 \neq 0$. En la regresión múltiple, la prueba *t* y la prueba *F* tienen propósitos diferentes.

1. La prueba *F* determina si existe una relación de significancia entre la variable dependiente y el conjunto de todas las variables independientes; a esta prueba *F* se le llama prueba de *significancia global*.
2. Si la prueba *F* indica que hay significancia global, se usa la prueba *t* para identificar si cada una de las variables individuales es significativa. Para cada una de las variables independientes del modelo se realiza una prueba *t*, y a cada cual se le refiere como prueba de *significancia individual*.

A continuación se explican las pruebas *F* y *t*, y su aplicación al ejemplo de Butler Trucking Company.

Prueba *F*

El modelo de regresión múltiple que se definió en la sección 15.4 es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

La hipótesis de la prueba *F* comprende los parámetros del modelo de regresión múltiple.

H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$
H_a: uno o más de los parámetros es distinto de cero

Cuando H_0 es rechazada, la prueba proporciona evidencia estadística suficiente para concluir que uno o más de los parámetros es diferente de cero y que la relación global entre y y el conjunto de variables independientes $x_1, x_2, \ldots, x_p$ es significativa. En cambio, si H_0 no puede ser rechazada, no se tiene evidencia suficiente para concluir que existe una relación significativa.

Antes de describir los pasos de la prueba F es necesario revisar el concepto de *cuadrado medio*. Éste es la suma de cuadrados dividida entre sus correspondientes grados de libertad. En el caso de la regresión múltiple, la suma de cuadrados del total tiene $n - 1$ grados de libertad, la suma de cuadrados debido a la regresión (SCR) tiene p grados de libertad y la suma de cuadrados debido al error tiene $n - p - 1$ grados de libertad. Por tanto, el cuadrado medio debido a la regresión (CMR) es SCR/p y el cuadrado medio debido al error (CME) es SCE/$(n - p - 1)$.

$$\text{CMR} = \frac{\text{SCR}}{p} \tag{15.12}$$

y

$$\text{CME} = \frac{\text{SCE}}{n - p - 1} \tag{15.13}$$

Como se vio en el capítulo 14, CME proporciona una estimación insesgada de σ^2, la varianza del término del error ϵ. Si H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ es verdadera, CMR también provee un estimador insesgado de σ^2, y el valor de CMR/CME será cercano a 1. Pero si H_0 es falsa, el CMR sobreestima σ^2 y el valor de CMR/CME será mayor. Para determinar qué tan grande debe ser este valor para que H_0 sea rechazada, se retoma el hecho de que si H_0 es verdadera y los supuestos acerca del modelo de regresión múltiple son válidos, la distribución muestral de CMR/CME es una distribución F con p grados de libertad en el numerador y $n - p - 1$ en el denominador. A continuación se presenta un resumen de la prueba F de significancia para la regresión múltiple.

PRUEBA F DE SIGNIFICANCIA GLOBAL

H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$
H_a: uno o más de los parámetros no son iguales a cero

ESTADÍSTICO DE PRUEBA

$$F = \frac{\text{CMR}}{\text{CME}} \tag{15.14}$$

REGLA DE RECHAZO

Método del valor-p: Rechazar H_0 si el valor-$p \leq \alpha$
Método del valor crítico: Rechazar H_0 si $F \geq F_\alpha$

donde F_α pertenece a la distribución F con p grados de libertad en el numerador y $n - p - 1$ grados de libertad en el denominador.

Enseguida se presenta la aplicación de la prueba F al problema de regresión múltiple de Butler Trucking Company. Como se tienen dos variables independientes, las hipótesis se expresan como sigue.

H_0: $\beta_1 = \beta_2 = 0$
H_a: β_1 y/o β_2 no es igual a cero

FIGURA 15.6 Resultado de Minitab para el ejemplo de Butler Trucking con dos variables independientes, millas recorridas (x_1) y número de entregas (x_2)

```
The regression equation is
Time = - 0.869 + 0.0611 Miles + 0.923 Deliveries

Predictor       Coef   SE Coef      T      p
Constant     -0.8687    0.9515  -0.91  0.392
Miles       0.061135  0.009888   6.18  0.000
Deliveries    0.9234    0.2211   4.18  0.004

S = 0.573142   R-sq = 90.4%   R-sq(adj) = 87.6%

Analysis of Variance

SOURCE          DF      SS      MS      F      p
Regression       2  21.601  10.800  32.88  0.000
Residual Error   7   2.299   0.328
Total            9  23.900
```

En la figura 15.6 se presentan los resultados de Minitab para el modelo de regresión múltiple con dos variables independientes: millas recorridas (x_1) y número de entregas (x_2). En la parte que corresponde al análisis de varianza, vemos que CMR = 10.8 y CME = 0.328. Con la ecuación (15.14) obtenemos el valor del estadístico de prueba.

$$F = \frac{10.8}{0.328} = 32.9$$

Observe que el valor de F en los resultados de Minitab es $F = 32.88$; este valor difiere del estimado aquí debido a que en los cálculos se emplearon los valores redondeados de CMR y CME. Con $\alpha = 0.01$, el valor-$p = 0.000$ que aparece en la última columna de la tabla del análisis de varianza (figura 15.6) indica que H_0: $\beta_1 = \beta_2 = 0$ puede ser rechazada debido a que el valor-p es menor que $\alpha = 0.01$. De manera alterna, en la tabla 4 del apéndice B observamos que con 2 grados de libertad en el numerador y 7 en el denominador, $F_{0.01} = 9.55$. Como $32.9 > 9.55$, H_0: $\beta_1 = \beta_2 = 0$ es rechazada, y se concluye que existe una relación significativa entre el tiempo de recorrido y y las dos variables independientes, millas recorridas y número de entregas.

Como ya se indicó, el error cuadrado medio proporciona un estimador insesgado de σ^2, la varianza del término del error ϵ. En la figura 15.6 vemos que la estimación de σ^2 es CME = 0.328. La raíz cuadrada del CME es la estimación de la desviación del término del error. Como se definió en la sección 14.5, esta desviación es el error estándar de estimación que se denota s. Por tanto, tenemos que $s = \sqrt{\text{CME}} = \sqrt{0.328} = 0.573$. Observe que este valor del error estándar de estimación aparece en los resultados de Minitab de la figura 15.6.

La tabla 15.3 es la tabla general para el análisis de varianza (ANOVA) que proporciona los resultados de la prueba F para un modelo de regresión múltiple. El valor del estadístico de prueba F aparece en la última columna y debe compararse con F_α con p grados de libertad en el numerador y $n - p - 1$ grados de libertad en el denominador para obtener la conclusión de la prueba de hipótesis. Revisando los resultados de Minitab para el ejemplo de Butler Trucker Company de la figura 15.6, vemos que la tabla del análisis de varianza contiene esta información. Además, Minitab también proporciona el respectivo valor-p al estadístico de prueba F.

TABLA 15.3 Tabla ANOVA para el modelo de regresión múltiple con p variables independientes

Fuente	Suma de cuadrados	Grados de libertad	Cuadrado medio	F
Regresión	SCR	p	$\text{CMR} = \dfrac{\text{SCR}}{p}$	$F = \dfrac{\text{CMR}}{\text{CME}}$
Error	SCE	$n - p - 1$	$\text{CME} = \dfrac{\text{SCE}}{n - p - 1}$	
Total	STC	$n - 1$		

Prueba t

Si la prueba F indica que la relación de regresión múltiple es significativa, entonces podemos realizar una prueba t para determinar la significancia de cada uno de los parámetros. A continuación se presenta la prueba t de significancia para el efecto.

PRUEBA t DE SIGNIFICANCIA PARA CADA UNO DE LOS PARÁMETROS

Para cualquier parámetro β_i

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

ESTADÍSTICO DE PRUEBA

$$t = \frac{b_i}{s_{b_i}} \qquad (15.15)$$

REGLA DE RECHAZO

Método del valor-p: Rechazar H_0 si el valor-$p \leq \alpha$

Método del valor crítico: Rechazar H_0 si $t \leq -t_{\alpha/2}$ o si $t \geq t_{\alpha/2}$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n - p - 1$ grados de libertad.

En el estadístico de prueba, s_{b_i} es la estimación de la desviación estándar de b_i. El software proporciona el valor de s_{b_i}.

A continuación se realiza la prueba t para el problema de regresión de Butler Trucking. Remítase a la sección de la figura 15.6 en la que se proporcionan los resultados de Minitab para el cálculo del cociente t. Los valores de b_1, b_2, s_{b_1} y s_{b_2} son los siguientes.

$$b_1 = 0.061135 \qquad s_{b_1} = 0.009888$$
$$b_2 = 0.9234 \qquad s_{b_2} = 0.2211$$

Con la ecuación (15.15) obtenemos el estadístico de prueba para las hipótesis en que intervienen β_1 y β_2.

$$t = 0.061135/0.009888 = 6.18$$
$$t = 0.9234/0.2211 = 4.18$$

Observe que los valores de estas dos razones-t y sus correspondientes valores-p aparecen en la figura 15.6. Usando $\alpha = 0.01$, los valores-p 0.000 y 0.004 en los resultados de Minitab indican que H_0: $\beta_1 = 0$ y H_0: $\beta_2 = 0$ pueden ser rechazadas. Así, ambos parámetros son estadísticamente significativos. También en la tabla 2 del apéndice B se encuentra que con $n - p - 1 = 10 - 2 - 1 = 7$ grados de libertad, $t_{0.005} = 3.499$. Como $6.18 > 3.499$, H_0: $\beta_1 = 0$ es rechazada. De manera similar, como $4.18 > 3.499$, H_0: $\beta_2 = 0$ es rechazada.

Multicolinealidad

En el análisis de regresión se recurre al término *variable independiente* para referirse a cualquier variable utilizada para predecir o explicar el valor de la variable dependiente. Sin embargo, este término no significa que tales variables sean independientes entre ellas en sentido estadístico. Al contrario, en un problema de regresión múltiple la mayoría de las variables independientes están, en cierto grado, correlacionadas unas con otras. En el ejemplo de Butler Trucking con dos variables independientes x_1 (millas recorridas) y x_2 (número de entregas), las millas recorridas pueden tratarse como la variable dependiente y el número de entregas como la variable independiente para determinar si ambas están relacionadas entre sí. Después se calcula el coeficiente de correlación muestral $r_{x_1x_2}$ para determinar la magnitud de tal relación. Con esto obtenemos $r_{x_1x_2} = 0.16$. Por tanto, se encuentra que existe cierto grado de relación lineal entre estas dos variables independientes. En el análisis de regresión múltiple, la **multicolinealidad** expresa la correlación entre las variables independientes.

Para tener una mejor perspectiva de los problemas potenciales de la multicolinealidad, se considerará una modificación al ejemplo de Butler Trucking. En lugar de que x_2 sea el número de entregas, denotará el número de galones de gasolina consumidos. Es claro que x_1 (las millas recorridas) y x_2 están relacionadas, es decir, se sabe que el número de galones de gasolina consumidos depende del número de millas recorridas. Por tanto, se concluirá que x_1 y x_2 son variables independientes fuertemente correlacionadas.

Suponga que se obtiene la ecuación $\hat{y} = b_0 + b_1x_1 + b_2x_2$ y que la prueba F indica que esta relación es significativa. Después suponga que se realiza la prueba t para β_1 a efecto de determinar si $\beta_1 \neq 0$ y H_0: $\beta_1 = 0$ no puede ser rechazada. ¿Esto significa que el tiempo de recorrido no está relacionado con las millas recorridas? No necesariamente. Lo que probablemente significa es que estando x_2 en el modelo, x_1 no tiene una contribución significativa en la determinación del valor de y. En el presente ejemplo esta interpretación parece razonable; conociendo la cantidad de gasolina consumida, no se gana más información para la predicción de y conociendo el número de millas recorridas. De manera similar, una prueba t puede llevar a la conclusión de que $\beta_2 = 0$ con base en que, cuando x_1 está en el modelo, no se gana mucho al conocer la cantidad de gasolina consumida.

Valores del coeficiente de correlación muestral mayores que +0.7 o menores que −0.7 para dos variables independientes, es un regla general que alerta sobre problemas potenciales de multicolinealidad.

En resumen, en las pruebas t para la significancia de cada uno de los parámetros, la dificultad ocasionada por la multicolinealidad lleva a concluir que ninguno de los parámetros es significativamente distinto de cero cuando la prueba F sobre la ecuación de regresión múltiple general indica que hay una relación significante. Este problema se evita cuando existe poca correlación entre las variables independientes.

Se han desarrollado diversas pruebas a efecto de determinar si la multicolinealidad es lo suficientemente alta para ocasionar problemas. Con base en una regla práctica, la multicolinealidad es un problema potencial si el valor absoluto del coeficiente de correlación muestral es mayor de 0.7 para cualquier par de variables independientes. Otros tipos de pruebas son más avanzados y quedan fuera del alcance de este libro.

Cuando las variables independientes están fuertemente correlacionadas, es imposible determinar por separado el efecto de cada una de las variables independientes sobre la variable dependiente.

Siempre que sea posible, debe evitarse incluir variables independientes fuertemente correlacionadas. Sin embargo, en la práctica, la estricta adherencia a esta conducta raramente es posible. Cuando las personas que toman las decisiones tienen razones para creer que existe una multicolinealidad importante, se darán cuenta de que es difícil separar los efectos de cada una de las variables independientes sobre la variable dependiente.

NOTAS Y COMENTARIOS

Por lo general, la multicolinealidad no afecta la manera en que se realiza el análisis de regresión o en que se interpretan los resultados de un estudio. Pero si es severa —esto es, cuando dos o más variables independientes están altamente correlacionados una con otra—, podemos tener dificultades al interpretar los resultados de las pruebas t acerca de cada uno de los parámetros. Además del tipo de problemas ilustrados en esta sección, se ha demostrado que los casos severos de multicolinealidad dan como resultado estimacio-nes por mínimos cuadrados con signo erróneo. Esto es, en estudios simulados en los que los investigadores crearon el modelo de regresión subyacente y después aplicaron el método de mínimos cuadrados para obtener estimaciones de β_0, β_1, β_2, etc., se ha demostrado que en condiciones de fuerte multicolinealidad, las estimaciones obtenidas por mínimos cuadrados pueden tener signo opuesto al del parámetro que se estima. Por ejemplo, b_2 puede ser en realidad $+10$ y β_2, su estimación, resulta ser -2. Por tanto, si existe una fuerte multicolinealidad, podrá tenerse poca confianza en los coeficientes.

Ejercicios

Métodos

AUTOevaluación

19. En el ejercicio 1 se presentó la siguiente ecuación de regresión estimada basada en 10 observaciones.

$$\hat{y} = 29.1270 + 0.5906x_1 + 0.4980x_2$$

donde STC $= 6\,724.125$, SCR $= 6\,216.375$, $s_{b_1} = 0.0813$ y $s_{b_2} = 0.0567$.

a) Calcule CMR y CME.
b) Determine F y realice la prueba F adecuada. Use $\alpha = 0.05$.
c) Realice una prueba t para la significancia de β_1. Utilice $\alpha = 0.05$.
d) Efectúe una prueba t para la significancia de β_2. Utilice $\alpha = 0.05$.

20. Remítase a los datos presentados en el ejercicio 2. La ecuación de regresión estimada de estos datos es

$$\hat{y} = -18.37 + 2.01x_1 + 4.74x_2$$

Donde STC $= 15\,182.9$, SCR $= 14\,052.2$, $s_{b_1} = 0.2471$ y $s_{b_2} = 0.9484$.

a) Realice una prueba para determinar si hay una relación significativa entre x_1, x_2 y y. Use $\alpha = 0.05$.
b) ¿Es significativo β_1? Considere $\alpha = 0.05$.
c) ¿Es significativo β_2? Use $\alpha = 0.05$.

21. Se obtuvo la siguiente ecuación de regresión estimada para un modelo con dos variables independientes.

$$\hat{y} = 40.7 + 8.63x_1 + 2.71x_2$$

Después de eliminar x_2 del modelo, se empleó el método de mínimos cuadrados para obtener una ecuación de regresión estimada con una sola variable independiente, x_1.

$$\hat{y} = 42.0 + 9.01x_1$$

a) Proporcione la interpretación del coeficiente de x_1 en ambos modelos.
b) ¿La multicolinealidad podría explicar por qué el coeficiente de x_1 es diferente en los dos modelos? De ser así, ¿cómo ocurriría?

Aplicaciones

22. En el ejercicio 4 se proporcionó la siguiente ecuación de regresión estimada que relaciona las ventas contra la inversión en inventario y los gastos de publicidad.

$$\hat{y} = 25 + 10x_1 + 8x_2$$

Los datos utilizados para obtener el modelo provinieron de un estudio realizado a 10 tiendas; para estos datos, STC $= 16\,000$ y SCR $= 12\,000$.

a) Calcule SCE, CME y CMR.

b) Use la prueba F y 0.05 como nivel de significancia para determinar si existe una relación entre las variables.

AUTOevaluación

23. Remítase al ejercicio 5.

a) Use $\alpha = 0.01$ para probar las hipótesis

$$H_0: \beta_1 = \beta_2 = 0$$
$$H_a: \beta_1 \text{ y/o } \beta_2 \text{ no son iguales a cero}$$

en el modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, donde

$x_1 =$ publicidad en televisión (en miles de dólares)
$x_2 =$ publicidad en periódicos (en miles de dólares)

b) Tome $\alpha = 0.05$ para probar la significancia de β_1; ¿debe ser eliminada x_1 del modelo?

c) Use $\alpha = 0.05$ para probar la significancia de β_2; ¿debe ser eliminada x_2 del modelo?

24. *The Wall Street Journal* realizó un estudio acerca de los gastos que realizan las mejores universidades en el basquetbol. Una parte de los datos se lista a continuación e incluye algunas escuelas (School), los ingresos (Revenue) en millones de $, el porcentaje de victorias (% Wins) y el sueldo del entrenador (Salary) en millones de $ de 39 de los mejores programas de basquetbol de Estados Unidos (*The Wall Street Journal*, 11-12 de marzo de 2006).

WEB archivo
Basketball

School	Revenue	%Wins	Salary
Alabama	6.5	61	1.00
Arizona	16.6	63	0.70
Arkansas	11.1	72	0.80
Boston College	3.4	80	0.53
⋮	⋮	⋮	⋮
Washington	5.0	83	0.89
West Virginia	4.9	67	0.70
Wichita State	3.1	75	0.41
Wisconsin	12.0	66	0.70

a) Desarrolle la ecuación de regresión estimada para predecir el sueldo del entrenador dados los ingresos generados por el programa y el porcentaje de victorias.

b) Use la prueba F para determinar la significancia global de la relación. ¿Cuál es su conclusión empleando 0.05 como nivel de significancia?

c) Utilice la prueba t para determinar la significancia de cada una de las variables independientes. ¿Cuál es su conclusión con un nivel de significancia de 0.05?

25. *Barron's* realiza revisiones anuales de los corredores de bolsa en línea, en las que se incluyen tanto aquellos a los que se puede contactar vía un navegador de Internet, como corredores que tienen acceso directo y colocan al cliente en contacto directo con el servidor de una red de agentes bursátiles. La oferta y el desempeño de cada corredor se evalúan en seis áreas, con una escala de 0 a 5 para cada categoría. Los resultados se ponderan para obtener una evaluación

general, y a cada agente se le asigna una clasificación final que va de cero a cinco estrellas. Tres de las áreas evaluadas son ejecución de la operación, facilidad de uso y gama de ofertas. Un valor de 5 en la primera significa que la llegada del pedido y el proceso de ejecución fluyeron con facilidad de un paso al siguiente. En la segunda área, un valor de 5 significa que el sitio es de manejo accesible y que podemos ajustar para identificar lo que le interesa ver al usuario. Por último, un valor de 5 en gama de ofertas indica que todas las transacciones pueden realizarse en línea. En los datos siguientes se presentan las puntuaciones obtenidas en la ejecución de la operación (Trade Execution), facilidad de uso (Use), rango de ofertas (Range) y clasificación por estrellas (Rating) obtenidas por los integrantes de una muestra de 10 corredores de bolsa (Broker) (*Barron's,* 10 de marzo de 2003).

WEB archivo
Brokers

Broker	Trade Execution	Use	Range	Rating
Wall St. Access	3.7	4.5	4.8	4.0
E*TRADE (Power)	3.4	3.0	4.2	3.5
E*TRADE (Standard)	2.5	4.0	4.0	3.5
Preferred Trade	4.8	3.7	3.4	3.5
my Track	4.0	3.5	3.2	3.5
TD Waterhouse	3.0	3.0	4.6	3.5
Brown & Co.	2.7	2.5	3.3	3.0
Brokerage America	1.7	3.5	3.1	3.0
Merrill Lynch Direct	2.2	2.7	3.0	2.5
Strong Funds	1.4	3.6	2.5	2.0

a) Defina una ecuación de regresión estimada para predecir la clasificación por estrellas dados los valores de ejecución, facilidad de uso y rango de ofertas.
b) Use la prueba *F* para determinar la significancia global de la relación. ¿Cuál es su conclusión empleando 0.05 como nivel de significancia?
c) Utilice la prueba *t* para determinar la significancia de cada variable independiente. ¿Cuál es su conclusión utilizando 0.05 como nivel de significancia?
d) Elimine cualquiera de las variables independientes que no sea significativa para la ecuación de regresión estimada. ¿Cuál ecuación de regresión estimada recomienda? Compare R^2 con el valor de R^2 del inciso *a*). Analice las diferencias.

WEB archivo
NBA

26. En el ejercicio 10 se obtuvo una ecuación de regresión estimada relacionada con la proporción de juegos ganados cuando se conocía la proporción de anotaciones de campo del equipo, la proporción de tiros de tres puntos del conjunto contrario y la cantidad de pérdidas de balón del equipo adversario.
a) Use la prueba *F para* determinar la significancia global de la relación. ¿Cuál es su conclusión utilizando 0.05 como nivel de significancia?
b) Considere la prueba *t para* determinar la significancia de cada una de las variables independientes. ¿Cuál es su conclusión utilizando 0.05 como nivel de significancia?

15.6 Uso de la ecuación de regresión estimada para estimaciones y predicciones

Los procedimientos que se siguen en la regresión múltiple para estimar el valor medio de y y predecir un solo valor de y son similares a los utilizados en el análisis de regresión para una sola variable independiente. Recuerde, primero, que en el capítulo 14 se mostró que la estimación puntual del valor esperado de y para un valor dado de x es la misma que la estimación puntual de un solo valor de y. En ambos casos se usó como estimación puntual $\hat{y} = b_0 + b_1x$.

En la regresión múltiple se sigue el mismo procedimiento, es decir, los valores dados de $x_1, x_2, \ldots, x_p$ se sustituyen en la ecuación de regresión y como estimación puntual se usa el correspondiente valor de $\hat{y}$. Suponga que en el ejemplo de Butler Truking se desea manejar la

TABLA 15.4 Intervalos de 95% de confianza y de predicción para el ejemplo de Butler Trucking

Valor de x_1	Valor de x_2	Intervalo de confianza		Intervalo de predicción	
		Límite inferior	Límite superior	Límite inferior	Límite superior
50	2	3.146	4.924	2.414	5.656
50	3	4.127	5.789	3.368	6.548
50	4	4.815	6.948	4.157	7.607
100	2	6.258	7.926	5.500	8.683
100	3	7.385	8.645	6.520	9.510
100	4	8.135	9.742	7.362	10.515

ecuación de regresión estimada con x_1 (millas recorridas) y x_2 (número de entregas) para obtener dos estimaciones por intervalo:

1. Un *intervalo de confianza* para la media del tiempo de recorrido de todos los camiones que recorren 100 millas y efectúan dos entregas.
2. Un *intervalo de predicción* para el tiempo de recorrido de *un determinado* camión que recorre 100 millas y efectúa dos entregas.

Utilizando la ecuación de regresión estimada $\hat{y} = -0.869 + 0.0611x_1 + 0.923x_2$ con $x_1 = 100$ y $x_2 = 2$, obtenemos el siguiente valor de $\hat{y}$.

$$\hat{y} = -0.869 + 0.0611(100) + 0.923(2) = 7.09$$

Por tanto, en ambos casos la estimación puntual del tiempo de recorrido es aproximadamente de 7 horas.

Para obtener las estimaciones por intervalo del valor medio de y y un solo valor de y se utilizan procedimientos similares a los aplicados en el análisis de regresión con una sola variable independiente. Las fórmulas que se necesitan quedan fuera del alcance de este libro, sin embargo, el software para el análisis de regresión múltiple suele proporcionar intervalos de confianza una vez que el usuario especifica los valores de $x_1, x_2, \ldots, x_p$. En la tabla 15.4 se presentan los intervalos de 95% de confianza y de predicción para algunos valores de x_1 y x_2 seleccionados del ejemplo de Butler Trucking. Estos valores se obtuvieron usando Minitab. Observe que las estimaciones por intervalo para un solo valor de y proporcionan valores más amplios que las estimaciones por intervalo para el valor esperado de y. Esta diferencia refleja simplemente que, dados los valores x_1 y x_2, podemos estimar con mayor precisión el tiempo medio de recorrido de todos los camiones, que predecir el de un determinado camión.

Ejercicios

Métodos

27. En el ejercicio 1 se presentó la siguiente ecuación de regresión estimada basada en 10 observaciones.

$$\hat{y} = 29.1270 + 0.5906x_1 + 0.4980x_2$$

a) Desarrolle una estimación puntual del valor medio de y para $x_1 = 180$ y $x_2 = 310$.
b) Obtenga una estimación puntual para un solo valor de y cuando $x_1 = 180$ y $x_2 = 310$.

AUTOevaluación

28. Remítase al ejercicio 2. La ecuación de regresión estimada de los datos es

$$\hat{y} = -18.4 + 2.01x_1 + 4.74x_2$$

a) Desarrolle un intervalo de 95% de confianza para el valor medio de y cuando $x_1 = 45$ y $x_2 = 15$.

b) Obtenga un intervalo de 95% de predicción para el valor de y cuando $x_1 = 45$ y $x_2 = 15$.

Aplicaciones

AUTO evaluación

29. En el ejercicio 5, el propietario de Showtime Movie Theater, Inc. empleó el análisis de regresión múltiple para predecir el ingreso bruto (y) en función de la publicidad en televisión (x_1) y de la publicidad en periódicos (x_2). La ecuación de regresión estimada fue

$$\hat{y} = 83.2 + 2.29x_1 + 1.30x_2$$

a) ¿Cuál será el ingreso bruto esperado en una semana en la que se gastan $3 500 en publicidad en televisión ($x_1 = 3.5$) y $1 800 en publicidad en periódicos ($x_2 = 1.8$)?

b) Proporcione un intervalo de 95% de confianza para el ingreso medio de todas las semanas en las que los gastos son los indicados en el inciso *a*).

c) Determine un intervalo de 95% de predicción para la media del ingreso de una semana asumiendo que los gastos son los indicados en el inciso *a*).

WEB archivo
Boats

30. En el ejercicio 9 se obtuvo una ecuación de regresión estimada que relacionaba la velocidad máxima de un bote con su manga y sus caballos de fuerza.

a) Proporcione un intervalo de 95% de confianza para la media de la velocidad máxima de un bote cuya manga es de 85 pulgadas y cuyo motor tiene 330 caballos de fuerza.

b) La Svfara SV609 tiene una manga de 85 pulgadas y un motor de 330 caballos de fuerza. Desarrolle un intervalo de 95% de confianza para la media de la velocidad máxima de la Svfara SV609.

31. La sección Guía para el usuario del sitio web de la revista *Car and Driver* proporciona información sobre pruebas de carretera de automóviles, camiones, SUV y vans. Las puntuaciones incluyen calidad general (Overall), estilo de vehículo, frenado, manejo (Handling), economía de combustible, confort interior, aceleración, confiabilidad (Dependability), ajuste y acabado (Fit and Finish), y transmisión atribuidos a diversos vehículos con una escala de 1 (lo peor) a 10 (lo mejor). Aquí se presenta una parte de los datos de 14 automóviles deportivos/GT (sitio web de Car and Driver, 7 de enero de 2004).

WEB archivo
SportsCar

Sports/GT	Overall	Handling	Dependability	Fit and Finish
Acura 3.2CL	7.80	7.83	8.17	7.67
Acura RSX	9.02	9.46	9.35	8.97
Audi TT	9.00	9.58	8.74	9.38
BMW 3-Series/M3	8.39	9.52	8.39	8.55
Chevrolet Corvette	8.82	9.64	8.54	7.87
Ford Mustang	8.34	8.85	8.70	7.34
Honda Civic Si	8.92	9.31	9.50	7.93
Infiniti G35	8.70	9.34	8.96	8.07
Mazda RX-8	8.58	9.79	8.96	8.12
Mini Cooper	8.76	10.00	8.69	8.33
Mitsubishi Eclipse	8.17	8.95	8.25	7.36
Nissan 350Z	8.07	9.35	7.56	8.21
Porsche 911	9.55	9.91	8.86	9.55
Toyota Celica	8.77	9.29	9.04	7.97

a) Proporcione una ecuación de regresión estimada usando manejo, confiabilidad y ajuste y acabado para predecir la calidad general.

b) Otro de los vehículos deportivos/GT evaluados por *Car and Driver* es el Honda Accord. Las evaluaciones de manejo, confiabilidad y ajuste y acabado atribuidas a este automóvil fueron 8.28, 9.06 y 8.07, respectivamente. Estime la evaluación general conferida a este modelo.

c) Determine un intervalo de 95% de confianza para la calidad general de todos los automóviles deportivos y GT con las características listadas en el inciso *b*).

d) Desarrolle un intervalo de predicción de 95% para la calidad general del Honda Accord descrito en el inciso *b*).

e) La evaluación general de *Car and Driver* para el Honda Accord fue 8.65. Compare esta calificación con las estimaciones obtenidas en los incisos *b*) y *d*).

15.7 Variables independientes cualitativas

Las variables independientes pueden ser cualitativas o cuantitativas.

En los ejemplos considerados hasta ahora se han manejado variables independientes cuantitativas, por ejemplo, población de estudiantes, distancia recorrida y número de entregas. Sin embargo, en muchas situaciones se tiene que trabajar con **variables independientes cualitativas**, como género (masculino o femenino), modo de pago (efectivo, tarjeta de crédito, cheque), etc. En esta sección el objetivo es mostrar cómo se manejan las variables cualitativas independientes en el análisis de regresión. Para ilustrar su uso e interpretación se considerará el caso de Johnson Filtration, Inc.

Un ejemplo: Johnson Filtration, Inc.

Esta empresa provee servicio de mantenimiento a los sistemas de filtración de agua en el sur de Florida. Los clientes llaman a Johnson Filtration, Inc. solicitando un servicio de mantenimiento para sus sistemas. A efecto de estimar el tiempo que se requerirá para el servicio y su costo, los directivos de Johnson desean predecir este tiempo para cada solicitud. Por tanto, el tiempo requerido en horas para la reparación es la variable dependiente. Se cree que este tiempo está relacionado con dos factores: meses transcurridos desde el último servicio de mantenimiento y tipo de problema (mecánico o eléctrico). En la tabla 15.5 se presentan los datos de una muestra de 10 solicitudes de servicio.

Sea y el tiempo de reparación en horas y x_1 los meses transcurridos desde el último servicio de mantenimiento. El modelo de regresión en el que sólo se usa x_1 para predecir y es

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

Usando Minitab para obtener la ecuación de regresión estimada se tienen los resultados de la figura 15.7. La ecuación de regresión estimada es

$$\hat{y} = 2.15 + 0.304 x_1 \qquad \textbf{(15.16)}$$

Con 0.05 como nivel de significancia, el valor-p de 0.016 para la prueba t (o F) indica que el número de meses transcurridos desde el último servicio está relacionado significativamente con el tiempo que se requiere para la reparación. R-sq = 53.4% indica que x_1 explica sólo 53.4% de la variabilidad en el tiempo necesario para una reparación.

TABLA 15.5 Datos para el ejemplo de Johnson Filtration

Solicitud de servicio	Meses desde el último servicio	Tipo de reparación	Tiempo de reparación en horas
1	2	Eléctrica	2.9
2	6	Mecánica	3.0
3	8	Eléctrica	4.8
4	3	Mecánica	1.8
5	2	Eléctrica	2.9
6	7	Eléctrica	4.9
7	9	Mecánica	4.2
8	8	Mecánica	4.8
9	4	Eléctrica	4.4
10	6	Eléctrica	4.5

FIGURA 15.7 Resultado de Minitab para el problema de Johnson Filtration con (x_1), número de meses desde el último servicio, como variable independiente

Los nombres de las variables Months (*meses*) *y* Time (*tiempo*) *que aparecen en los resultados de Minitab fueron ingresados en la hoja de trabajo como títulos de columna; por tanto,* x_1 = Months y y = Time.

```
The regression equation is
Time = 2.15 + 0.304 Months

Predictor     Coef  SE Coef     T      p
Constant    2.1473   0.6050  3.55  0.008
Months      0.3041   0.1004  3.03  0.016

S = 0.781022   R-sq = 53.4%   R-sq(adj) = 47.6%

Analysis of Variance

SOURCE          DF       SS      MS     F      p
Regression       1   5.5960  5.5960  9.17  0.016
Residual Error   8   4.8800  0.6100
Total            9  10.4760
```

Para incluir el tipo de reparación en el modelo de regresión, se define la variable siguiente.

$$x_2 = \begin{cases} 0 \text{ si el tipo de reparación es mecánica} \\ 1 \text{ si el tipo de reparación es eléctrica} \end{cases}$$

En el análisis de regresión a x_2 se le llama **variable ficticia** o variable *indicadora*. Al utilizar esta variable, el modelo de regresión múltiple se expresa como sigue.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

En la tabla 15.6 se presentan los datos de la tabla 15.5, más los valores de la variable ficticia. Customer indica cliente; Months Since Last Service, meses desde el último servicio; Type of Repair, tipo de reparación, y Repair Time in Hours, tiempo de reparación en horas. Con Minitab y los datos de la tabla 15.6 se obtienen estimaciones para los parámetros del modelo. En el resultado de Minitab de la figura 15.8 se puede ver que la ecuación de regresión múltiple estimada es

$$\hat{y} = 0.93 + 0.388x_1 + 1.26x_2 \qquad \textbf{(15.17)}$$

TABLA 15.6 Datos para el ejemplo de Johnson Filtration con el tipo de reparación indicado por una variable ficticia ($x_2 = 0$ si es mecánica; $x_2 = 1$ si es eléctrica)

WEB archivo
Johnson

Customer	Months Since Last Service (x_1)	Type of Repair (x_2)	Repair Time in Hours (y)
1	2	1	2.9
2	6	0	3.0
3	8	1	4.8
4	3	0	1.8
5	2	1	2.9
6	7	1	4.9
7	9	0	4.2
8	8	0	4.8
9	4	1	4.4
10	6	1	4.5

FIGURA 15.8 Resultado de Minitab para el ejemplo de Johnson Filtration con (x_1), meses desde el último servicio, y (x_2), tipo de reparación, como variables independientes

Los nombres de las variables Months, Type (*tipo*) *y Time que aparecen en los resultados de Minitab fueron ingresados en la hoja de trabajo como títulos de columna; por tanto,* x_1 = Months, x_2 = Type y y = Time.

```
The regression equation is
Time = 0.930 + 0.388 Months + 1.26 Type

Predictor      Coef  SE Coef      T      p
Constant     0.9305   0.4670   1.99  0.087
Months      0.38762  0.06257   6.20  0.000
Type         1.2627   0.3141   4.02  0.005

S = 0.459048   R-sq = 85.9%   R-sq(adj) = 81.9%

Analysis of Variance

SOURCE           DF       SS      MS      F      p
Regression        2   9.0009  4.5005  21.36  0.001
Residual Error    7   1.4751  0.2107
Total             9  10.4760
```

Con 0.05 como nivel de significancia, el valor-p correspondiente al estadístico de prueba F ($F = 21.36$) es 0.001, lo cual indica que la relación de regresión es significativa. En la figura 15.8, en la parte de los resultados de Minitab que corresponde a la prueba t, observamos que tanto meses transcurridos desde el último servicio (valor-p = 0.000) como tipo de reparación (valor-p = 0.005) son estadísticamente significativos. Además, R-sq = 85.9% y R-sq(adj) = 81.9% indican que la ecuación de regresión estimada explica adecuadamente la variabilidad en el tiempo de reparación. Por tanto, la ecuación (15.17) sí es útil para estimar el tiempo necesario para la reparación relacionada con las diversas solicitudes de servicio.

Interpretación de los parámetros

La ecuación de regresión múltiple para el ejemplo de Johnson Filtration es

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad \textbf{(15.18)}$$

Para entender cómo interpretar los parámetros β_0, β_1 y β_2 cuando hay una variable cualitativa, considérese el caso en que $x_2 = 0$ (reparación mecánica). Usando $E(y \mid \text{mecánica})$ para denotar la media o valor esperado del tiempo necesario para una reparación *dado* que ésta es mecánica, tenemos

$$E(y \mid \text{mecánica}) = \beta_0 + \beta_1 x_1 + \beta_2(0) = \beta_0 + \beta_1 x_1 \qquad \textbf{(15.19)}$$

De manera similar, en el caso de una reparación eléctrica ($x_2 = 1$) tenemos

$$\begin{aligned} E(y \mid \text{eléctrica}) &= \beta_0 + \beta_1 x_1 + \beta_2(1) = \beta_0 + \beta_1 x_1 + \beta_2 \\ &= (\beta_0 + \beta_2) + \beta_1 x_1 \end{aligned} \qquad \textbf{(15.20)}$$

Al comparar las ecuaciones (15.19) y (15.20) vemos que la media del tiempo requerido para efectuar una reparación es función lineal de x_1, tanto cuando es de tipo mecánico como eléctrico. La pendiente en ambas ecuaciones es β_1, pero la intersección con el eje y varía. En la ecuación (15.19) para las reparaciones mecánicas, la intersección con el eje y es β_0, y en la ecuación (15.20) para las reparaciones eléctricas, la intersección es $(\beta_0 + \beta_2)$. La interpretación de β_2 indica la diferencia entre las medias del tiempo requerido para una reparación eléctrica y una reparación mecánica.

Si β_2 es positiva, la media del tiempo necesario para una reparación eléctrica será mayor que para una mecánica; si β_2 es negativa, la media del tiempo requerido para una reparación eléctrica será menor que para la mecánica. Por último, si $\beta_2 = 0$, no hay diferencia entre las medias del tiempo que se necesita para ambos trabajos, y el tipo de reparación no está relacionado con el tiempo necesario para efectuarla.

Utilizando la ecuación de regresión múltiple estimada $\hat{y} = 0.93 + 0.388x_1 + 1.26x_2$, vemos que 0.93 es la estimación de β_0 y la estimación de β_2 es 1.26. Por tanto, cuando $x_2 = 0$ (reparación mecánica)

$$\hat{y} = 0.93 + 0.388x_1 \tag{15.21}$$

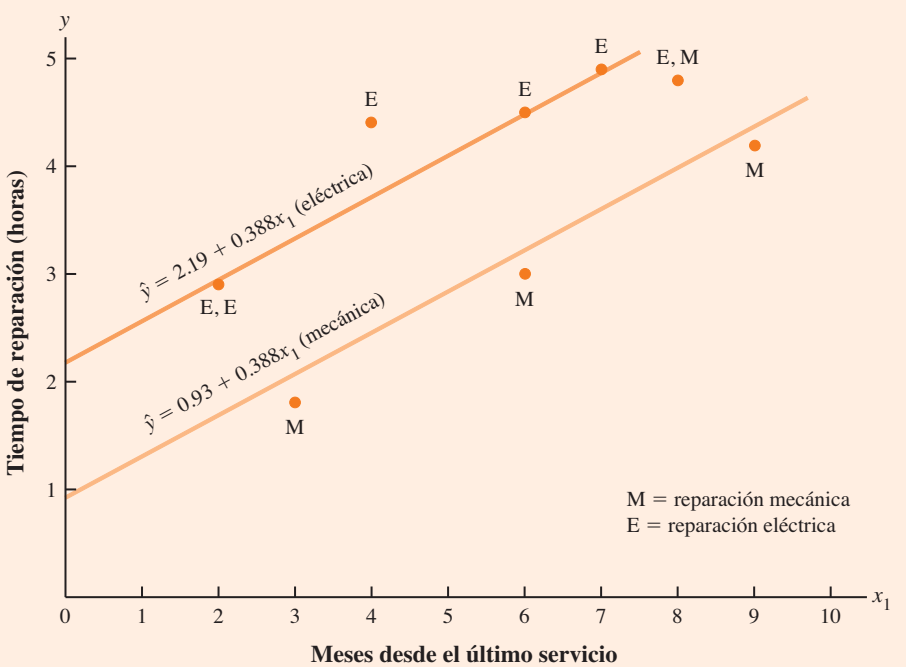
y cuando $x_2 = 1$ (reparación eléctrica)

$$\begin{aligned}\hat{y} &= 0.93 + 0.388x_1 + 1.26(1) \\ &= 2.19 + 0.388x_1\end{aligned} \tag{15.22}$$

De esta manera, el uso de una variable ficticia para el tipo de reparación proporciona dos ecuaciones para predecir el tiempo requerido para efectuarla: una ecuación corresponde a las reparaciones mecánicas y la otra a las eléctricas. Además, como $b_2 = 1.26$, sabemos que, en promedio, en las reparaciones eléctricas se necesitan 1.26 horas más que en las mecánicas.

En la figura 15.9 se presenta una gráfica con los datos de la tabla 15.6. El tiempo de reparación en horas (y) se representa en el eje vertical y los meses transcurridos desde el último servicio (x_1) en el eje horizontal. Los puntos que corresponden a una reparación mecánica se indican con una M y los que corresponden a una reparación eléctrica con una E. En esta gráfica se representan también las ecuaciones (15.21) y (15.22) con objeto de mostrar las dos ecuaciones que sirven para predecir el tiempo que se requerirá para una y otra reparación.

FIGURA 15.9 Diagrama de dispersión para los datos de reparación de Johnson Filtration de la tabla 15.6



Variables cualitativas más complejas

Para modelar una variable cualitativa con k niveles se requieren $k - 1$ variables ficticias. Se debe ser cuidadoso al definir e interpretar estas variables.

En el ejemplo de Johnson Filtration, como la variable cualitativa tenía dos niveles (mecánica y eléctrica), fue fácil definirla empleando 0 para indicar una reparación mecánica y 1 para indicar una reparación eléctrica. Sin embargo, cuando una variable cualitativa tiene más de dos niveles, habrá que tener cuidado tanto al definir como al interpretar estas variables ficticias. Como se verá a continuación, si una variable cualitativa tiene k niveles, se necesitan $k - 1$ variables ficticias, cada una de las cuales tomará el valor 0 o 1.

Suponga, por ejemplo, que un fabricante de fotocopiadoras divide un estado en tres regiones de ventas: A, B y C. Sus gerentes desean aplicar el análisis de regresión para predecir las ventas semanales. Utilizando como variable dependiente el número de fotocopiadoras vendidas, están considerando diversas variables independientes (número de vendedores, gastos en publicidad, etc.). Suponga que los gerentes piensan que la región de ventas puede ser también un factor importante en la predicción del número de unidades vendidas. Como región es una variable cualitativa que tiene tres niveles, A, B y C, para representar la región de ventas se necesitarán $3 - 1 = 2$ variables ficticias, cada una de las cuales tomará los valores 0 o 1.

$$x_1 = \begin{cases} 1 \text{ si la región de ventas es B} \\ 0 \text{ si no es el caso} \end{cases}$$

$$x_2 = \begin{cases} 1 \text{ si la región de ventas es C} \\ 0 \text{ si no es el caso} \end{cases}$$

Con base en esta definición, para x_1 y x_2 tenemos los valores siguientes

Región	x_1	x_2
A	0	0
B	1	0
C	0	1

En las observaciones correspondientes a la región A se codificará $x_1 = 0, x_2 = 0$; en las observaciones para la región B se tendrá $x_1 = 1, x_2 = 0$, y en las observaciones de la región C se tendrá $x_1 = 0, x_2 = 1$.

La ecuación de regresión que relaciona el valor esperado del número de fotocopiadoras vendidas, $E(y)$, con las variables ficticias se expresa como sigue.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Para interpretar los parámetros β_0, β_1 y β_2, considere las siguientes tres variaciones de la ecuación de regresión.

$$E(y \mid \text{región A}) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$$
$$E(y \mid \text{región B}) = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$$
$$E(y \mid \text{región C}) = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$$

Por tanto, β_0 es la media o valor esperado de las ventas en la región A; β_1 es la diferencia entre las medias del número de unidades vendidas entre la región B y la región A, y β_2 es la diferencia entre las medias del número de unidades vendidas en la región C y la región A.

Se necesitaron dos variables ficticias debido a que la región de ventas es una variable cualitativa con tres niveles. Sin embargo, la asignación de $x_1 = 0, x_2 = 0$ para identificar la región A; $x_1 = 1, x_2 = 0$ para identificar la región B, y $x_1 = 0, x_2 = 1$ para la región C fue arbitraria.

De igual manera se podría haber elegido, por ejemplo, $x_1 = 1, x_2 = 0$ para identificar la región A, $x_1 = 0, x_2 = 0$ para la B y $x_1 = 0, x_2 = 1$ para la C. En ese caso, β_1 se habría interpretado como la media de la diferencia entre las regiones A y B, y β_2 como la media de la diferencia entre C y B.

Es importante recordar que en el análisis de regresión múltiple, cuando una variable cualitativa tiene k niveles, se requieren $k - 1$ variables ficticias. Entonces, si en nuestro ejemplo hubiera una cuarta región, D, se necesitarían tres variables ficticias, las cuales se pueden codificar como sigue.

$$x_1 = \begin{cases} 1 \text{ si la región de ventas es B} \\ 0 \text{ si no es el caso} \end{cases} \qquad x_2 = \begin{cases} 1 \text{ si la región de ventas es C} \\ 0 \text{ si no es el caso} \end{cases} \qquad x_3 = \begin{cases} 1 \text{ si la región de ventas es D} \\ 0 \text{ si no es el caso} \end{cases}$$

Ejercicios

Métodos

AUTOevaluación

32. Considere un estudio de regresión en el que intervienen una variable dependiente y, una variable independiente cualitativa x_1 y una variable cualitativa de dos niveles (nivel 1 y nivel 2).
    a) Escriba la ecuación de regresión múltiple que relacione x_1 con la variable cualitativa y.
    b) ¿Cuál es el valor esperado de y que corresponde al nivel 1 de la variable cualitativa?
    c) ¿Cuál es el valor esperado de y que corresponde al nivel 2?
    d) Interprete los parámetros de la ecuación de regresión.

33. Considere un estudio de regresión en el que intervienen una variable dependiente y, una variable independiente cuantitativa x_1 y una variable cualitativa de tres niveles (nivel 1, nivel 2 y nivel 3).
    a) ¿Cuántas variables ficticias se requieren para representar la variable cualitativa?
    b) Proporcione una ecuación de regresión múltiple que relacione x_1 con la variable cualitativa y.
    c) Interprete los parámetros de la ecuación de regresión.

Aplicaciones

AUTOevaluación

34. La gerencia propuso el siguiente modelo de regresión para predecir las ventas en un punto de venta de comida rápida.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

donde

$x_1 =$ número de competidores a no más de una milla

$x_2 =$ población a no más de una milla (en miles)

$$x_3 = \begin{cases} 1 \text{ si tiene ventanilla para conductores} \\ 0 \text{ si no es el caso} \end{cases}$$

$y =$ ventas (en miles de $)

Se obtuvo la siguiente ecuación de regresión estimada con los datos de 20 puntos de venta.

$$\hat{y} = 10.1 - 4.2x_1 + 6.8x_2 + 15.3x_3$$

a) ¿Cuál es la cantidad esperada de ventas atribuible a la ventanilla para conductores?
b) Pronostique las ventas de un negocio que tiene dos competidores, una población de 8 000 a no más de una milla y ventanilla para conductores.
c) Prediga las ventas de un negocio que tiene un competidor, una población de 3 000 a no más de una milla y ventanilla para conductores.

35. Remítase al problema de Johnson Filtration presentado en esta sección. Suponga que además de la información sobre los meses transcurridos desde el último servicio (Months Since Last Service), del tipo de reparación (Type of Repair), mecánica (Mechanical) o eléctrica (Electrical), los gerentes presentan una lista con los técnicos (Repairperson) que realizaron el servicio. A continuación se exponen los nuevos datos.

WEB archivo
Repair

Repair Time in Hours	Months Since Last Service	Type of Repair	Repairperson
2.9	2	Electrical	Dave Newton
3.0	6	Mechanical	Dave Newton
4.8	8	Electrical	Bob Jones
1.8	3	Mechanical	Dave Newton
2.9	2	Electrical	Dave Newton
4.9	7	Electrical	Bob Jones
4.2	9	Mechanical	Bob Jones
4.8	8	Mechanical	Bob Jones
4.4	4	Electrical	Bob Jones
4.5	6	Electrical	Dave Newton

a) Por ahora ignore los meses transcurridos desde el último servicio (x_1) y el técnico asignado. Obtenga la ecuación de regresión lineal simple estimada para predecir el tiempo que se requiere para la reparación (y) dado el tipo de reparación (x_2). Recuerde que $x_2 = 0$ si ésta es mecánica y $x_2 = 1$ si es eléctrica.

b) ¿La ecuación obtenida en el inciso *a*) proporciona un buen ajuste a los datos observados? Explique.

c) Por ahora ignore los meses transcurridos desde el último servicio y el tipo de reparación. Obtenga la ecuación de regresión lineal simple estimada para predecir el tiempo necesario para la reparación dado el técnico que realizó el servicio. Sea $x_3 = 0$ si éste fue realizado por Bob Jones, y $x_3 = 1$ si lo realizó Dave Newton.

d) ¿La ecuación obtenida en el inciso *c*) proporciona un buen ajuste a los datos observados? Explique.

36. Este problema es una extensión de la situación descrita en el ejercicio 35.

a) Obtenga la ecuación de regresión estimada para predecir el tiempo que requiere una reparación dados los meses transcurridos desde la última efectuada, el tipo de reparación y el técnico que realizó el servicio.

b) Con un nivel de significancia de 0.05, realice una prueba para ver si la ecuación de regresión estimada obtenida en el inciso *a*) representa una relación significativa entre las variables independientes y la variable dependiente.

c) ¿Es estadísticamente significativo agregar la variable x_3, el técnico que realizó el servicio? Use $\alpha = 0.05$. ¿Qué explicación puede dar para los resultados observados?

37. El estudio de satisfacción del cliente sobre restaurantes de *Consumer Reports* se basa en más de 148 599 visitas a diferentes cadenas de restaurantes de servicio completo (sitio web de Consumer Reports, 11 de febrero de 2009). Suponga que los siguientes datos son representativos de los resultados reportados. La variable tipo (Type), indica si el restaurante es de comida italiana (Italian), o de mariscos/carnes (Seafood/Steakhouse). Precio (Price) indica la cantidad promedio que paga una persona por la comida y la bebida, menos la propina. Puntuación (Score) refleja la satisfacción general de los comensales, con los valores más altos indicando mayor satisfacción general. Podemos interpretar como "muy satisfecho" una puntuación de 80.

WEB archivo
RestaurantRatings

Restaurant	Type	Price ($)	Score
Bertucci's	Italian	16	77
Black Angus Steakhouse	Seafood/Steakhouse	24	79
Bonefish Grill	Seafood/Steakhouse	26	85

Restaurant	Type	Price ($)	Score
Bravo! Cucina Italiana	Italian	18	84
Buca di Beppo	Italian	17	81
Bugaboo Creek Steak House	Seafood/Steakhouse	18	77
Carrabba's Italian Grill	Italian	23	86
Charlie Brown's Steakhouse	Seafood/Steakhouse	17	75
Il Fornaio	Italian	28	83
Joe's Crab Shack	Seafood/Steakhouse	15	71
Johnny Carino's Italian	Italian	17	81
Lone Star Steakhouse & Saloon	Seafood/Steakhouse	17	76
LongHorn Steakhouse	Seafood/Steakhouse	19	81
Maggiano's Little Italy	Italian	22	83
McGrath's Fish House	Seafood/Steakhouse	16	81
Olive Garden	Italian	19	81
Outback Steakhouse	Seafood/Steakhouse	20	80
Red Lobster	Seafood/Steakhouse	18	78
Romano's Macaroni Grill	Italian	18	82
The Old Spaghetti Factory	Italian	12	79
Uno Chicago Grill	Italian	16	76

a) Obtenga una ecuación de regresión estimada que muestre cómo se relaciona la satisfacción general del cliente con la variable independiente precio promedio de la comida.

b) Utilizando como nivel de significancia 0.05, pruebe si la ecuación de regresión estimada obtenida en el inciso *a*) indica una relación significativa entre la satisfacción general del cliente y el precio promedio de la comida.

c) Proporcione una variable ficticia para el tipo de restaurante (italiano o de mariscos/carnes).

d) Obtenga una ecuación de regresión estimada que muestre cómo se relaciona la satisfacción general del cliente con el precio promedio de los alimentos y el tipo de restaurante.

e) ¿Es el tipo de restaurante un factor significativo en la satisfacción general del cliente?

f) Estime la puntuación de satisfacción del cliente de *Consumer Reports* para un restaurante de mariscos/carnes con un precio promedio de $20 por comida. ¿Cuánto cambiaría la puntuación estimada para un restaurante italiano?

38. Un estudio realizado durante 10 años por la Asociación Estadounidense de Cardiología proporcionó datos sobre la relación que guardan la edad (Age), la presión sanguínea (Pressure) y el hábito de fumar sobre el riesgo de sufrir un infarto. Los datos que se listan a continuación se obtuvieron como parte de este estudio. El riesgo (Risk) se interpreta como la probabilidad (multiplicada por 100) de que el paciente sufra un infarto en los próximos 10 años. Para la variable fumador (Smoker), defina una variable ficticia que tome los valores 1 si el individuo es fumador y 0 si no lo es.

WEB archivo
Stroke

Risk	Age	Pressure	Smoker
12	57	152	No
24	67	163	No
13	58	155	No
56	86	177	Yes
28	59	196	No
51	76	189	Yes
18	56	155	Yes
31	78	120	No
37	80	135	Yes
15	78	98	No
22	71	152	No
36	70	173	Yes

(*Continúa*)

Risk	Age	Pressure	Smoker
15	67	135	Yes
48	77	209	Yes
15	60	199	No
36	82	119	Yes
8	66	166	No
34	80	125	Yes
3	62	117	No
37	59	207	Yes

a) Obtenga la ecuación de regresión estimada que relacione el riesgo de infarto con la edad, la presión sanguínea y si la persona fuma o no.
b) ¿Fumar es un factor significativo para el riesgo de infarto? Explique. Use $\alpha = 0.05$.
c) ¿Cuál es la probabilidad de que Art Speen sufra un infarto en la próxima década si tiene 68 años, fuma y su presión sanguínea es de 175? ¿Qué recomendará el médico a este paciente?

15.8 Análisis residual

En el capítulo 14 se indicó que los residuales estandarizados suelen utilizarse en las gráficas de residuales y en la identificación de observaciones atípicas. A continuación se presenta la fórmula general para obtener el residual estandarizado de la observación *i*.

RESIDUAL ESTANDARIZADO DE LA OBSERVACIÓN *i*

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \qquad (15.23)$$

donde

$s_{y_i - \hat{y}_i}$ = desviación estándar del residual i

La fórmula general para obtener la desviación estándar del residual *i* está definida como se indica a continuación.

DESVIACIÓN ESTÁNDAR DEL RESIDUAL *i*

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \qquad (15.24)$$

donde

s = error estándar de la estimación
h_i = influencia de la observación *i*

Como se dijo en el capítulo 14, la **influencia** de una observación está en función de qué tan lejos de sus medias están los valores de las variables independientes. En el análisis de regresión múltiple, calcular h_i y $s_{y_i - \hat{y}_i}$, y por tanto el residual estandarizado de la observación *i*,

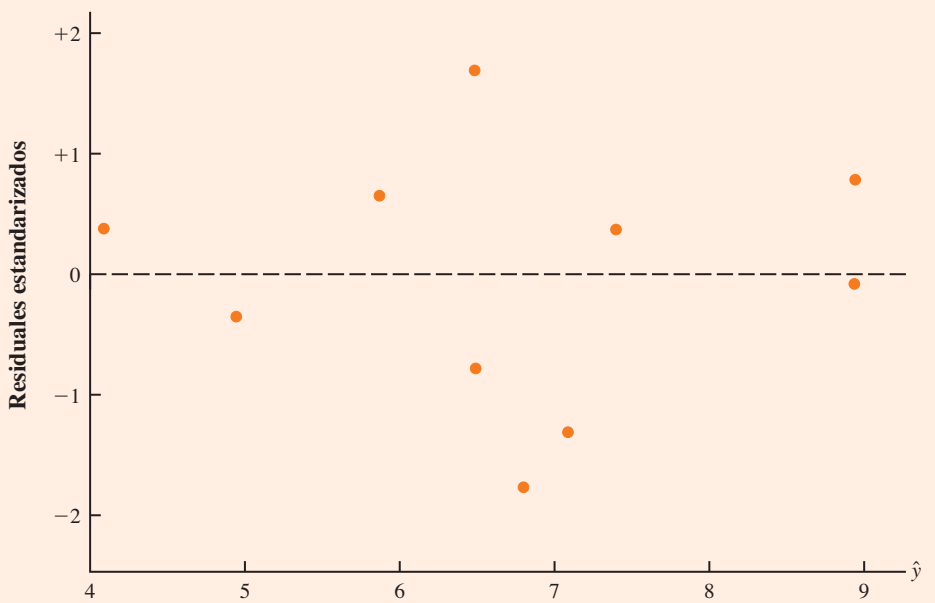
TABLA 15.7 Residuales y residuales estandarizados correspondientes al análisis de regresión de Butler Trucking

Millas recorridas (x_1)	Entregas (x_2)	Tiempo de recorrido (y)	Valor pronosticado ($\hat{y}$)	Residual ($y - \hat{y}$)	Residual estandarizado
100	4	9.3	8.93846	0.361541	0.78344
50	3	4.8	4.95830	−0.158304	−0.34962
100	4	8.9	8.93846	−0.038460	−0.08334
100	2	6.5	7.09161	−0.591609	−1.30929
50	2	4.2	4.03488	0.165121	0.38167
80	2	6.2	5.86892	0.331083	0.65431
75	3	7.4	6.48667	0.913331	1.68917
65	4	6.0	6.79875	−0.798749	−1.77372
90	3	7.6	7.40369	0.196311	0.36703
90	2	6.1	6.48026	−0.380263	−0.77639

es muy complicado como para efectuarlo a mano. Sin embargo, los residuales estandarizados se obtienen fácilmente con el software para estadística. En la tabla 15.7 se listan valores pronosticados, residuales y residuales estandarizados con los datos del ejemplo de Butler Trucking presentado previamente en este capítulo; estos valores se obtuvieron mediante Minitab. Los valores pronosticados de la tabla están basados en la ecuación de regresión estimada $\hat{y} = -0.869 + 0.0611x_1 + 0.923x_2$.

Los residuales estandarizados y los valores pronosticados de y de la tabla 15.7 se utilizaron en la figura 15.10, la gráfica de residuales estandarizados para el ejemplo de regresión múltiple de Butler Trucking. En esta gráfica no se observa ninguna anormalidad. Además, todos los residuales estandarizados se encuentran entre -2 y $+2$; por tanto, no hay ninguna razón para cuestionar el supuesto de que el término del error ϵ esté distribuido normalmente. Así, se concluye que los supuestos del modelo son razonables.

FIGURA 15.10 Gráfica de residuales estandarizados para el ejemplo de Butler Trucking



Para determinar si la distribución de ϵ parece ser normal, también se usa una gráfica de probabilidad normal. En la sección 14.8 se discutió el procedimiento y la interpretación de ese tipo de gráfica. Ese mismo procedimiento es adecuado para la regresión múltiple. La gráfica de probabilidad normal se puede obtener con un software para estadística que realice los cálculos.

Detección de observaciones atípicas

Una **observación atípica** es una observación inusual en relación con el conjunto de los da-tos; en otras palabras, no sigue el patrón del resto de los datos. En el capítulo 14 se mostró un ejemplo con una observación atípica y se vio el uso de los residuales estandarizados para detectarla. Minitab clasifica una observación como atípica si el valor de su residual estandarizado es menor que -2 o mayor que $+2$. Aplicando esta regla a los residuales estandarizados del ejemplo de Butler Trucking (tabla 15.7), en este conjunto de datos no se detecta ninguna observación atípica.

En general, la presencia de una o más observaciones atípicas en un conjunto de datos tiende a incrementar s, el error estándar de estimación y, por tanto, a incrementar $s_{y_i - \hat{y}_i}$, la desviación estándar del residual i. Dado que $s_{y_i - \hat{y}_i}$ aparece como denominador en la fórmula (15.23) del residual estandarizado, el tamaño de éste disminuirá a medida que s aumente. Esto da como resultado que aun cuando un residual sea inusualmente grande, el denominador de la fórmula (15.23), que será grande, hará que la regla del residual estandarizado falle para identificar una observación atípica. Es posible sortear esta dificultad empleando una forma de los residuales estandarizados conocida como **residuales estudentizados eliminados**.

Residuales eliminados estudentizados y observaciones atípicas

Suponga que del conjunto de datos se elimina la observación $i$ésima y que de las $n - 1$ observaciones restantes se obtiene una nueva ecuación de regresión estimada. Sea $s_{(i)}$ el error estándar de estimación basado en el conjunto de datos en los que se ha eliminado la observación $i$ésima. Si se calcula la desviación estándar del residual i usando $s_{(i)}$ en lugar de s, y después se calcula el residual estandarizado de la observación i con el nuevo valor de $s_{y_i - \hat{y}_i}$, al residual estandarizado que se obtiene se le llama *residual estudentizado eliminado*. Si la observación $i$ésima es atípica, $s_{(i)}$ será menor que s. Por tanto, el valor absoluto del residual estudentizado eliminado $i$ésimo será mayor que el valor absoluto del residual estandarizado. De esta manera, los residuales eliminados estudentizados pueden reconocer observaciones atípicas que los residuales estandarizados no detectan.

Mucho del software para estadística proporciona una opción para obtener residuales estudentizados eliminados. Para el ejemplo de Butler Trucking, Minitab proporciona los resultados que se presentan en la siguiente tabla. Para determinar si los residuales estudentizados elimi-

TABLA 15.8 Residuales eliminados estudentizados de Butler Trucking

Millas recorridas (x_1)	Entregas (x_2)	Tiempo recorrido (y)	Residual estandarizado	Residual estudentizado eliminado
100	4	9.3	0.78344	0.75939
50	3	4.8	−0.34962	−0.32654
100	4	8.9	−0.08334	−0.07720
100	2	6.5	−1.30929	−1.39494
50	2	4.2	0.38167	0.35709
80	2	6.2	0.65431	0.62519
75	3	7.4	1.68917	2.03187
65	4	6.0	−1.77372	−2.21314
90	3	7.6	0.36703	0.34312
90	2	6.1	−0.77639	−0.75190

TABLA 15.9 Influencia y distancia de Cook's correspondientes al ejemplo de Butler Trucking

Millas recorridas (x_1)	Entregas (x_2)	Tiempo recorrido (y)	Influencia (h_i)	D Cook (D_i)
100	4	9.3	0.351704	0.110994
50	3	4.8	0.375863	0.024536
100	4	8.9	0.351704	0.001256
100	2	6.5	0.378451	0.347923
50	2	4.2	0.430220	0.036663
80	2	6.2	0.220557	0.040381
75	3	7.4	0.110009	0.117562
65	4	6.0	0.382657	0.650029
90	3	7.6	0.129098	0.006656
90	2	6.1	0.269737	0.074217

nados indican la presencia de observaciones atípicas, se emplea la distribución t. Recuerde que p denota el número de variables independientes y n el número de observaciones. Por tanto, si se elimina la observación i, el número de observaciones en el nuevo conjunto de datos es $n - 1$; en este caso, la suma de cuadrados del error tiene $(n - 1) - p - 1$ grados de libertad. Como en el ejemplo de Butler Trucking $n = 10$ y $p = 2$, los grados de libertad para la suma de cuadrados del error con la observación $i$ésima eliminada es $9 - 2 - 1 = 6$. Utilizando como nivel de significancia 0.05, en la distribución t (tabla 2 del apéndice B) para 6 grados de libertad obtenemos, $t_{0.025} = 2.447$. Se concluye que la observación $i$ésima es atípica si el residual estudentizado eliminado es menor que -2.447 o mayor que $+2.447$. En la tabla 15.8 observamos que los residuales eliminados estudentizados no se encuentran fuera de estos límites; por tanto, se concluye que en este conjunto de datos no hay observaciones atípicas.

Observaciones influyentes

En la sección 14.9 se estudió cómo utilizar la influencia de una observación para identificar observaciones cuyo valor de la variable independiente puede tener una fuerte influencia en los resultados de la regresión. Como se indicó respecto de los residuales estandarizados, la influencia de una observación, que se denota h_i, mide qué tan lejos de sus medias se encuentran los valores de las variables independientes. Los valores de influencia se obtienen como parte de los resultados que proporciona el software para estadística. Minitab calcula estos valores, y para detectar **observaciones influyentes** emplea la regla $h_i > 3(p + 1)/n$. En el ejemplo de Butler Trucking, como hay $p = 2$ variables independientes y $n = 10$ observaciones, el valor crítico para la influencia es $3(2 + 1)/10 = 0.9$. En la tabla 15.9 se presentan los valores de influencia correspondientes al ejemplo de Butler Trucking obtenidos con Minitab. Como ninguno de los valores h_i es mayor que 0.9, en este conjunto de datos no se detectan observaciones influyentes.

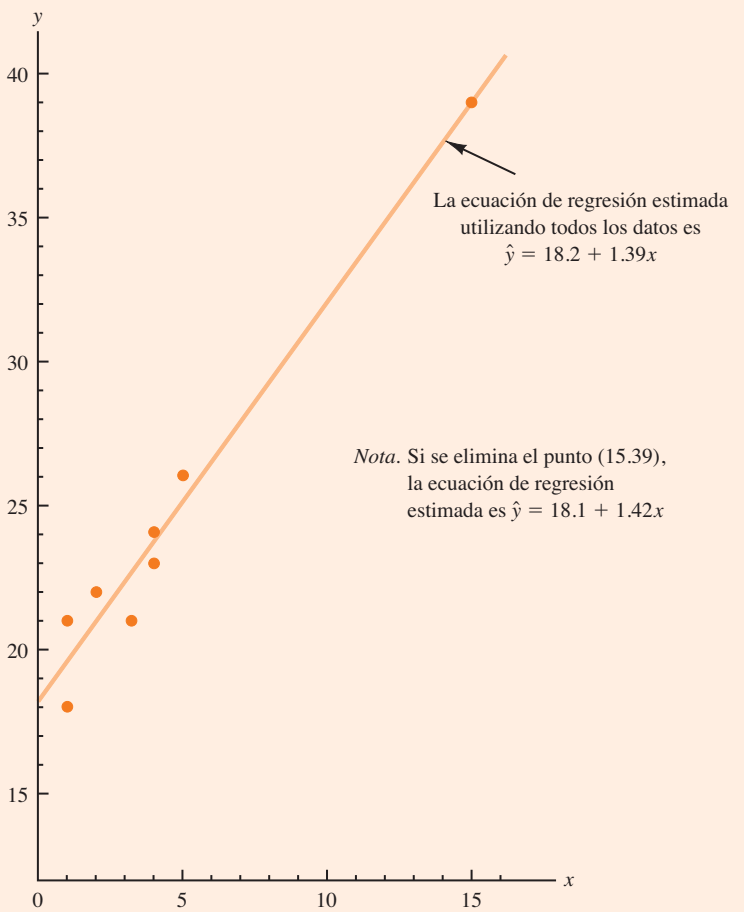
TABLA 15.10
Conjunto de datos que ilustran problemas potenciales usando el criterio de influencia

x_i	y_i	Influencia h_i
1	18	0.204170
1	21	0.204170
2	22	0.164205
3	21	0.138141
4	23	0.125977
4	24	0.125977
5	26	0.127715
15	39	0.909644

Uso de la medida de la distancia de Cook para identificar observaciones influyentes

Un problema potencial al usar la influencia para identificar observaciones influyentes es la posibilidad de identificar una observación como fuertemente influyente sin que necesariamente lo sea en términos de la ecuación de regresión estimada que se obtiene. Por ejemplo, en la tabla 15.10 se presenta un conjunto de datos que consta de ocho observaciones y sus correspondientes valores de influencia (obtenidos con Minitab). Como la influencia de la última observación es $0.91 > 0.75$ (el valor de influencia crítico), se le identificará como observación influyente. Sin embargo, antes de aceptar una conclusión final, considere la situación desde una perspectiva diferente.

FIGURA 15.11 Diagrama de dispersión obtenido con el conjunto de datos de la tabla 15.10



En la figura 15.11 se presenta el diagrama de dispersión que corresponde al conjunto de datos de la tabla 15.10. A partir de estos datos se obtuvo la siguiente ecuación de regresión estimada con Minitab.

$$\hat{y} = 18.2 + 1.39x$$

La línea recta que se observa en la figura 15.11 es la gráfica de esta ecuación. Ahora, si de este conjunto de datos se elimina la observación $x = 15$, $y = 39$, con las siete observaciones restantes se obtiene una nueva ecuación de regresión estimada:

$$\hat{y} = 18.1 + 1.42x$$

En la nueva ecuación se observa que la intersección con el eje y y la pendiente no tienen valores significativamente diferentes a los de la ecuación en la que se usan todos los datos. A pesar de que con el criterio de influencia se identificó la octava observación como influyente, es claro que tiene poca influencia en los resultados obtenidos. Por tanto, hay casos en los que emplear únicamente la influencia para identificar las observaciones influyentes puede llevar a conclusiones erróneas.

La **medida de la distancia de Cook** utiliza tanto la influencia de la observación i, h_i, como el residual de la observación i, $(y_i - \hat{y}_i)$, para determinar si una observación es influyente.

MEDIDA DE LA DISTANCIA DE COOK

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(p+1)s^2}\left[\frac{h_i}{(1-h_i)^2}\right] \quad \textbf{(15.25)}$$

donde

D_i = medida de la distancia de Cook para la observación i
$y_i - \hat{y}_i$ = residual de la observación i
h_i = influencia de la observación i
p = número de variables independientes
s = error estándar de estimación

Si el residual o la influencia son grandes, la medida de la distancia de Cook será grande e indicará una observación influyente. Como regla general, se acepta que si $D_i > 1$, la observación *i*ésima es influyente y debe ser analizada con más detenimiento. En la última columna de la tabla 15.9 se presentan las medidas de la distancia de Cook correspondientes al problema Butler Trucking obtenidas con Minitab. La observación con mayor influencia es la 8, para la que $D_i = 0.650029$. Sin embargo, al aplicar la regla $D_i > 1$ vemos que no hay por qué preocuparse acerca de la presencia de observaciones influyentes en el conjunto de datos de Butler Trucking.

NOTAS Y COMENTARIOS

1. Los procedimientos para detectar observaciones atípicas e influyentes permiten estar alerta acerca de los efectos potenciales de algunas observaciones en los resultados de la regresión. Cada observación atípica o influyente justifica un examen cuidadoso. Si se encuentran errores en los datos, podemos corregirlos y repetir el análisis de regresión. En general, las observaciones atípicas y las influyentes no deben ser eliminadas del conjunto de datos a menos que haya una evidencia clara que indique que no provienen de elementos de la población en estudio y que no tenían que ser incluidos en el conjunto de datos original.
2. Para determinar si el valor de una medida de la distancia de Cook D_i es lo suficientemente grande como para concluir que la observación *i*ésima es influyente, también puede compararse el valor de D_i con el percentil 50 de una distribución F (denotado $F_{0.50}$) con $p + 1$ grados de libertad en el numerador y $n - p - 1$ grados de libertad en el denominador. Para esta prueba se necesita contar con tablas F a un nivel de significancia de 0.50. La regla práctica proporcionada antes ($D_i > 1$) se basa en el hecho de que en muchos casos los valores en la tabla son cercanos a 1.

Ejercicios

Métodos

AUTOevaluación

39. A continuación se proporcionan datos para las variables x y y.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a) Obtenga una ecuación de regresión estimada para estos datos.
b) Grafique los residuales estandarizados contra $\hat{y}$. ¿Parece haber alguna observación atípica en este conjunto de datos? Explique.
c) Calcule los residuales eliminados estudentizados de estos datos. Empleando como nivel de significancia 0.05, ¿puede clasificarse cualquiera de estas observaciones como atípica? Explique.

40. A continuación se proporcionan datos para las variables x y y.

x_i	22	24	26	28	40
y_i	12	21	31	35	70

a) Obtenga una ecuación de regresión estimada para estos datos.
b) Calcule los residuales eliminados estudentizados de estos datos. Empleando como nivel de significancia 0.05, ¿puede clasificarse cualquiera de estas observaciones como atípica? Explique.
c) Determine los valores de influencia. ¿Parece haber alguna observación influyente? Explique.
d) Calcule la medida de la distancia de Cook de estos datos. ¿Alguna de las observaciones es influyente? Explique.

Aplicaciones

AUTO evaluación

41. En el ejercicio 5 se presentaron los datos siguientes sobre el ingreso bruto semanal (Weekly Gross Revenue) y la publicidad tanto en televisión (Television Advertising) como en periódicos (Newspaper Advertising) de Showtime Movie Theaters.

WEB archivo
Showtime

Weekly Gross Revenue ($1 000s)	Television Advertising ($1 000s)	Newspaper Advertising ($1 000s)
96	5.0	1.5
90	2.0	2.0
95	4.0	1.5
92	2.5	2.5
95	3.0	3.3
94	3.5	2.3
94	2.5	4.2
94	3.0	2.5

a) Proporcione una ecuación de regresión estimada que relacione el ingreso bruto semanal con los gastos en publicidad en televisión y periódicos.
b) Grafique los residuales estandarizados contra $\hat{y}$. ¿La gráfica respalda los supuestos acerca de ϵ? Explique.
c) Revise que no haya observaciones atípicas en estos datos. ¿A qué conclusión llega?
d) ¿Hay alguna observación influyente? Explique.

42. En los datos siguientes se presenta peso en vacío (Curb Weight), caballos de fuerza (Horsepower) y velocidad en ¼ de milla (Speed at ¼ Mile) de 16 populares automóviles deportivos y de gran turismo (Sports & GT Car). Suponga que se tiene también el precio (Price) de cada uno de estos vehículos. Todo el conjunto de datos es el siguiente.

WEB archivo
Auto2

Sports & GT Car	Price ($1 000s)	Curb Weight (lb)	Horsepower	Speed at ¼ Mile (mph)
Acura Integra Type R	25.035	2577	195	90.7
Acura NSX-T	93.758	3066	290	108.0
BMW Z3 2.8	40.900	2844	189	93.2
Chevrolet Camaro Z28	24.865	3439	305	103.2
Chevrolet Corvette Convertible	50.144	3246	345	102.1
Dodge Viper RT/10	69.742	3319	450	116.2
Ford Mustang GT	23.200	3227	225	91.7
Honda Prelude Type SH	26.382	3042	195	89.7
Mercedes-Benz CLK320	44.988	3240	215	93.0
Mercedes-Benz SLK230	42.762	3025	185	92.3
Mitsubishi 3000GT VR-4	47.518	3737	320	99.0

Sports & GT Car	Price ($1 000s)	Curb Weight (lb)	Horsepower	Speed at ¼ Mile (mph)
Nissan 240SX SE	25.066	2862	155	84.6
Pontiac Firebird Trans Am	27.770	3455	305	103.2
Porsche Boxster	45.560	2822	201	93.2
Toyota Supra Turbo	40.989	3505	320	105.0
Volvo C70	41.120	3285	236	97.0

a) Obtenga la ecuación de regresión estimada en la que se emplee precio y caballos de fuerza para predecir la velocidad en ¼ de milla.
b) Grafique los residuales estandarizados contra $\hat{y}$. ¿La gráfica respalda los supuestos respecto de ϵ? Explique.
c) Verifique si hay observaciones atípicas. ¿A qué conclusión llega?
d) ¿Hay alguna observación influyente? Explique.

WEB archivo
LPGA

43. La Asociación de Mujeres Golfistas Profesionales (LPGA, por sus siglas en inglés) lleva estadísticas sobre el desempeño y las ganancias de sus miembros en el LPGA Tour. El archivo LPGA contiene las estadísticas de final de año sobre el desempeño de las 30 jugadoras que tuvieron las mejores ganancias en el LPGA Tour de 2005 (sitio web LPGA, 2006). Earnings ($1 000) son las ganancias totales en miles de dólares; Scoring Avg es el número promedio de golpes en todo el evento; Greens in Reg. es el porcentaje de las veces que una jugadora logra un *green* en regulación, y Putting Avg es el promedio de golpes cortos por *green* en regulación. Un *green* se considera un golpe en regulación si alguna parte de la bola está en contacto con la superficie de *putting* y la diferencia entre el valor del par de hoyos y el número de golpes que lleva a golpear el *green* es menor de 2.
a) Desarrolle una ecuación de regresión estimada para predecir la puntuación promedio de todos los eventos dado el porcentaje de las veces que una jugadora logra un *green* en regulación y el promedio de golpes cortos por *green* en regulación.
b) Grafique los residuales estandarizados contra $\hat{y}$. ¿Esta gráfica confirma los supuestos hechos acerca de ϵ?
c) Verifique si existen observaciones atípicas. ¿A qué conclusión llega?
d) ¿Hay alguna observación influyente? Explique.

15.9 Regresión logística

En muchas aplicaciones de la regresión la variable dependiente asume sólo dos valores discretos. Por ejemplo, en un banco suele necesitarse una ecuación de regresión estimada para predecir si a una persona se le aprobará su solicitud de tarjeta de crédito. A esta variable dependiente pueden dársele los valores $y = 1$ si la solicitud es aprobada, y $y = 0$ si es rechazada. Con la regresión logística, dado un conjunto particular de valores de las variables independientes elegidas, se estima la probabilidad de que el banco apruebe la solicitud de tarjeta de crédito.

A continuación se considera una aplicación de la regresión logística. La empresa Simmons Stores, una cadena nacional de ropa para dama, realizará una promoción por correo. Ha ordenado imprimir 5 000 copias de su costoso catálogo de productos a cuatro tintas, y en cada uno incluye un cupón de $50 de descuento en compras por $200 o más. Como el catálogo es costoso, Simmons desea enviarlo sólo a aquellos clientes que tengan mayor probabilidad de usar el cupón.

La gerencia considera que la cantidad gastada anualmente por el cliente en las tiendas Simmons, así como si posee o no una tarjeta de crédito de la tienda, son dos variables útiles para predecir si ese cliente usará el cupón. La empresa realiza un estudio piloto con una muestra aleatoria de 50 clientes con tarjeta de crédito de Simmons y 50 sin ella. Por ende, envió los

catálogos a cada uno de estos 100 clientes elegidos. Al final del periodo de prueba, Simmons anota si los clientes han usado o no el cupón. En la tabla 15.11 se presentan los datos muestrales de las 10 primeras personas (Customer) que recibieron el catálogo, y se incluye la siguiente información: cantidad en miles de dólares gastada por el cliente en las tiendas Simmons durante el año anterior (Annual Spending), y la tarjeta de crédito de Simmons (Simmons Card) codificada como 1 si el cliente la tiene y como 0 si no la tiene. En la columna correspondiente al cupón (Coupon), 1 significa que el cliente usó el cupón y 0 indica que no lo usó.

Para ayudar a Simmons a predecir si las personas que reciban el catálogo usarán o no el cupón, se podría pensar en construir un modelo de regresión múltiple con los datos de la tabla 15.11. Las variables independientes serían cantidad gastada anualmente en Simmons Stores y tarjeta de crédito, en tanto que el cupón sería la variable dependiente. Sin embargo, el modelo común de regresión múltiple no es aplicable porque la variable dependiente sólo puede tomar los valores 0 y 1. Con este ejemplo se ilustra el tipo de situación para la cual fue creada la regresión logística. A continuación se verá cómo utilizarla para ayudar a Simmons Stores a pronosticar qué tipo de clientes es más probable que aproveche su promoción.

Ecuación de regresión logística

La regresión logística se parece en muchos aspectos a la regresión común. Se necesita una variable dependiente y, y una o varias variables independientes. En el análisis de regresión múltiple, a la media o valor esperado de y se le conoce como ecuación de regresión múltiple.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad \textbf{(15.26)}$$

En la regresión logística, tanto la teoría como la práctica estadística han demostrado que la relación existente entre $E(y)$ y $x_1, x_2, \ldots, x_p$, queda mejor descrita por medio de la siguiente ecuación no lineal.

ECUACIÓN DE REGRESIÓN LOGÍSTICA

$$E(y) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}} \qquad \textbf{(15.27)}$$

Como los dos valores de la variable dependiente y son codificados como 0 y 1, el valor de $E(y)$ en la ecuación (15.27) proporcionará *la probabilidad* de que $y = 1$ para un conjunto dado

TABLA 15.11 Datos muestrales parciales del ejemplo de Simmons Stores

WEB archivo
Simmons

Customer	Annual Spending ($1 000)	Simmons Card	Coupon
1	2.291	1	0
2	3.215	1	0
3	2.135	1	0
4	3.924	0	0
5	2.528	1	0
6	2.473	0	1
7	2.384	0	0
8	7.076	0	0
9	1.182	1	1
10	3.345	0	0

de valores de las variables independientes $x_1, x_2, \ldots, x_p$. Dado que $E(y)$ se interpreta como una probabilidad, la **ecuación de regresión logística** suele expresarse de la siguiente manera.

INTERPRETACIÓN DE $E(y)$ COMO UNA PROBABILIDAD EN LA REGRESIÓN LOGÍSTICA

$$E(y) = P(y = 1|x_1, x_2, \ldots, x_p) \tag{15.28}$$

Para entender mejor las características de la ecuación de regresión logística, suponga que el modelo sólo involucra una variable independiente x y que los valores de los parámetros del modelo son $\beta_0 = -7$ y $\beta_1 = 3$. La ecuación de regresión logística correspondiente a estos valores de los parámetros es

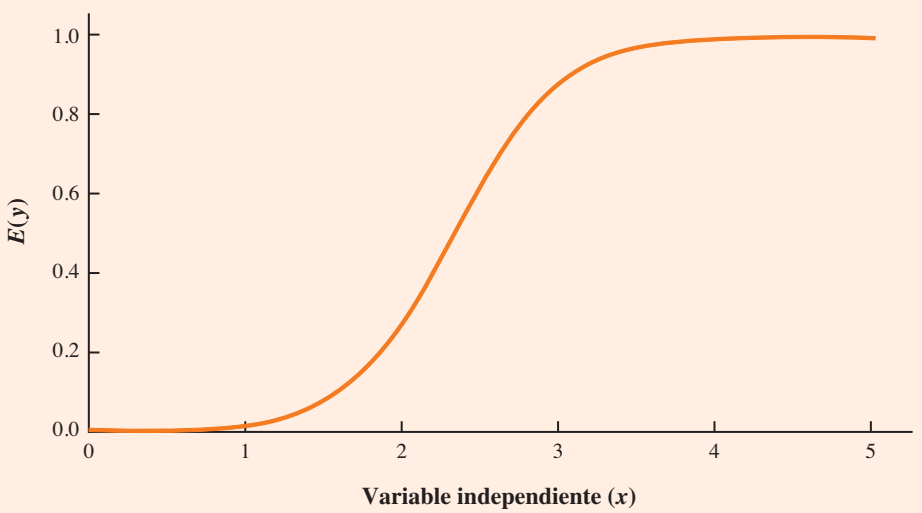
$$E(y) = P(y = 1|x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} = \frac{e^{-7+3x}}{1 + e^{-7+3x}} \tag{15.29}$$

En la figura 15.12 se muestra la gráfica de la ecuación (15.29). Observe que tiene forma de S. El valor de $E(y)$ va desde 0 hasta 1, aproximándose gradualmente a medida que el valor de x aumenta, y a 0 a medida que el valor de x disminuye. Observe también que el valor de $E(y)$, que representa la probabilidad, se incrementa rápidamente al aumentar x de 2 a 3. El hecho de que los valores de $E(y)$ vayan de 0 a 1 y que la curva tenga forma de S hacen la ecuación (15.29) ideal para modelar la probabilidad de que la variable dependiente sea igual a 1.

Estimación de la ecuación de regresión logística

En la regresión lineal simple y en la regresión múltiple, el método de mínimos cuadrados permite calcular las estimaciones $b_0, b_1, \ldots, b_p$ de los parámetros $(\beta_0, \beta_1, \ldots, \beta_p)$ del modelo. Debido a la forma no lineal de la ecuación de regresión logística, el método para calcular estas estimaciones es más complejo y queda fuera del alcance de este libro. Para obtener estas estimaciones se empleará software. La **ecuación de regresión logística estimada** se indica enseguida.

FIGURA 15.12 Ecuación de regresión logística en la que $\beta_0 = -7$ y $\beta_1 = 3$



ECUACIÓN DE REGRESIÓN LOGÍSTICA ESTIMADA

$$\hat{y} = \text{estimación de } P(y = 1|x_1, x_2, \ldots, x_p) = \frac{e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}}{1 + e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}} \quad \textbf{(15.30)}$$

Aquí $\hat{y}$ es una estimación de la probabilidad de que $y = 1$ para un determinado conjunto de valores de las variables independientes.

De vuelta al ejemplo de Simmons Stores, las variables en este estudio están definidas como sigue.

$$y = \begin{cases} 0 \text{ si el cliente no usó el cupón} \\ 1 \text{ si el cliente usó el cupón} \end{cases}$$

$x_1 =$ cantidad anual gastada en Simmons Stores (en miles de $)

$$x_2 = \begin{cases} 0 \text{ si el cliente no tiene tarjeta de crédito de Simmons} \\ 1 \text{ si el cliente tiene tarjeta de crédito de Simmons} \end{cases}$$

Por tanto, se elige una ecuación de regresión logística con dos variables independientes.

$$E(y) = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2}}{1 + e^{\beta_0+\beta_1x_1+\beta_2x_2}} \quad \textbf{(15.31)}$$

En el apéndice 15.3 se explica cómo usar Minitab para generar el resultado de la figura 15.13.

Para calcular las estimaciones de los parámetros β_0, β_1 y β_2 del modelo se aplicó el procedimiento de regresión logística binaria de Minitab a los datos muestrales de la tabla 15.11. En la figura 15.13 se muestra parte de los resultados obtenidos. Como vemos, $b_0 = -2.14637$, $b_1 = 0.341643$ y $b_2 = 1.09873$. Así, la ecuación de regresión logística estimada es

$$\hat{y} = \frac{e^{b_0+b_1x_1+b_2x_2}}{1 + e^{b_0+b_1x_1+b_2x_2}} = \frac{e^{-2.14637+0.341643x_1+1.09873x_2}}{1 + e^{-2.14637+0.341643x_1+1.09873x_2}} \quad \textbf{(15.32)}$$

Ahora, con la ecuación (15.32) se estima la probabilidad de que un determinado tipo de clientes use el cupón. Por ejemplo, para estimar la probabilidad de que aquellos que tienen un gasto anual de $2000 en Simmons Stores y que no tienen tarjeta de crédito de la tienda usen el cupón, en la ecuación (15.32) se sustituyen $x_1 = 2$ y $x_2 = 0$.

FIGURA 15.13 Resultado parcial de la regresión logística para el ejemplo de las tiendas Simmons

En los resultados de Minitab, x_1 = Spending (cantidad gastada) y x_2 = Card (tarjeta de crédito).

```
Logistic Regression Table

                                                        Odds    95%  CI
Predictor      Coef    SE Coef      Z       p   Ratio  Lower  Upper
Constant   -2.14637   0.577245  -3.72   0.000
Spending   0.341643   0.128672   2.66   0.008    1.41   1.09   1.81
Card        1.09873   0.444696   2.47   0.013    3.00   1.25   7.17

Log-Likelihood = -60.487
Test that all slopes are zero: G = 13.628, DF = 2, P-Value = 0.001
```

$$\hat{y} = \frac{e^{-2.14637+0.341643(2)+1.09873(0)}}{1 + e^{-2.14637+0.341643(2)+1.09873(0)}} = \frac{e^{-1.4631}}{1 + e^{-1.4631}} = \frac{0.2315}{1.2315} = 0.1880$$

Por tanto, la probabilidad estimada de que este tipo de clientes use el cupón es de 0.19. De manera similar, la probabilidad de que lo usen aquellos que tienen un gasto anual de $2 000 en Simmons Stores y tarjeta de crédito de la tienda se estima sustituyendo $x_1 = 2$ y $x_2 = 1$ en la ecuación (15.32).

$$\hat{y} = \frac{e^{-2.14637+0.341643(2)+1.09873(1)}}{1 + e^{-2.14637+0.341643(2)+1.09873(1)}} = \frac{e^{-0.3644}}{1 + e^{-0.3644}} = \frac{0.6946}{1.6946} = 0.4099$$

Como se ve, la probabilidad de que los clientes de este grupo usen el cupón es aproximadamente de 0.41. Parece ser que quienes manejan tarjeta de crédito de Simmons tienen mayor probabilidad de usar el cupón. Pero antes de llegar a una conclusión, es necesario evaluar la significancia estadística de este modelo.

Prueba de significancia

La prueba de significancia en la regresión logística es similar a la que se aplica en la regresión múltiple. Primero se prueba la significancia global. En el ejemplo de Simmons Stores, las hipótesis para probar la significancia global son las siguientes.

H_0: $\beta_1 = \beta_2 = 0$
H_a: uno o los dos parámetros son distintos de cero

La prueba de significancia global del modelo se basa en el valor del estadístico de prueba G. Si la hipótesis nula es verdadera, la distribución muestral de G es una distribución ji-cuadrada con grados de libertad igual al número de variables independientes en el modelo. El cálculo de G queda fuera del alcance de este libro, pero este valor y su correspondiente valor-p se obtienen como parte del resultado de regresión logística binaria que proporciona Minitab. En la última línea de la figura 15.13 se encuentra que el valor de G es 13.628, sus grados de libertad son 2 y su correspondiente valor-p es 0.001. Por tanto, cualquier nivel de significancia $\alpha \geq 0.001$, nos llevará a rechazar la hipótesis nula y a concluir que el modelo global es significativo.

Una vez que la prueba G ha indicado que sí existe una significancia global, suele realizarse una prueba z para determinar si la contribución de cada una de las variables independientes al modelo es significativa. Para cada una de las variables independientes x_i las hipótesis son:

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

Si la hipótesis nula es verdadera, el valor del coeficiente estimado dividido entre su error estándar seguirá una distribución de probabilidad normal estándar. En el resultado de Minitab, en la columna titulada Z, se presentan los valores de $z_i = b_i/s_{b_1}$ para cada uno de los coeficientes estimados, y en la columna denominada p se encuentran sus valores-p correspondientes. Suponga que en el modelo de Simmons se emplea $\alpha = 0.05$ para probar la significancia de las variables independientes. Para x_1 el valor z es 2.66 y su correspondiente valor-p es 0.008. Por tanto, para el nivel de significancia 0.05 podemos rechazar H_0: $\beta_1 = 0$. De la misma manera se rechaza H_0: $\beta_2 = 0$, dado que el valor-p correspondiente a $z = 2.47$ es 0.013. Como se ve, empleando como nivel de significancia 0.05, ambas variables son estadísticamente significativas.

Uso en la administración

Ya se describió cómo obtener la ecuación de regresión logística estimada y cómo probar su significancia. Ahora se podrá hacer una recomendación para la decisión que se tomará en Simmons Stores sobre la promoción de su catálogo. Ya se calcularon $P(y = 1|x_1 = 2, x_2 = 1) = 0.4099$ y $P(y = 1|x_1 = 2, x_2 = 0) = 0.1880$. Con base en estas probabilidades, vemos que entre aquellos clientes cuyo gasto anual en Simmons Stores es de $2000, los que cuentan con una tarjeta de crédito de la tienda tienen mayor probabilidad de utilizar el cupón. En la tabla 15.12 se presentan las probabilidades estimadas correspondientes a clientes tanto con tarjeta de crédito como sin ella cuyos desembolsos anuales en Simmons Stores van de $1000 hasta $7000. ¿Cómo puede utilizar Simmons esta información para elegir a los clientes a los que dirigirá la nueva promoción? Suponga que desea enviar este catálogo únicamente a clientes cuya probabilidad de utilizar el cupón sea 0.40 o mayor. Con base en las probabilidades estimadas que aparecen en la tabla 15.12, la estrategia en esta promoción de Simmons sería la siguiente.

Clientes con tarjeta de crédito de Simmons. Enviar el catálogo a todos aquellos que durante el pasado año gastaron $2000 o más.

Clientes sin tarjeta de crédito de Simmons. Enviar el catálogo a todos aquellos que durante el pasado año gastaron $6000 o más.

Sin embargo, al observar con más detalle las probabilidades estimadas, vemos que la probabilidad de que usen el cupón aquellos clientes sin tarjeta de crédito de Simmons que gastaron $5000 en un año es de 0.3922. Por tanto, será conveniente que la tienda reconsidere su estrategia e incluya a clientes que no tienen tarjeta de crédito pero que gastaron en Simmons $5000 o más el año pasado.

Interpretación de la ecuación de regresión logística

Para interpretar una ecuación de regresión es necesario relacionar las variables independientes con la cuestión de negocios a la que se trata de dar respuesta con esa ecuación. En la regresión logística, debido a que la ecuación de regresión logística no es lineal, es difícil interpretar directamente la relación entre las variables independientes y la probabilidad de que $y = 1$. Sin embargo, se ha demostrado que podemos interpretar indirectamente esta relación mediante un concepto llamado *cociente de posibilidades* (*odds ratio*).

Las **posibilidades a favor de que ocurra un evento** se definen como la probabilidad de que el evento ocurra, dividida entre la probabilidad de que no ocurra. En la regresión logística el evento de interés siempre es $y = 1$. Dado un determinado conjunto de valores de las variables independientes, las posibilidades a favor de $y = 1$ se calculan como sigue.

$$\text{odds} = \frac{P(y = 1|x_1, x_2, \ldots, x_p)}{P(y = 0|x_1, x_2, \ldots, x_p)} = \frac{P(y = 1|x_1, x_2, \ldots, x_p)}{1 - P(y = 1|x_1, x_2, \ldots, x_p)} \qquad \textbf{(15.33)}$$

El **cociente de posibilidades** mide el efecto que tiene sobre estas posibilidades el aumento en una unidad en una sola de las variables independientes. Es, por ende, la probabilidad de que $y = 1$ cuando una de las variables independientes es incrementada en una unidad (odds_1)

TABLA 15.12 Probabilidades estimadas para Simmons Stores

		Gasto anual						
		$1000	$2000	$3000	$4000	$5000	$6000	$7000
Tarjeta de crédito	Sí	0.3305	0.4099	0.4943	0.5791	0.6594	0.7315	0.7931
	No	0.1413	0.1880	0.2457	0.3144	0.3922	0.4759	0.5610

dividida entre las posibilidades de que $y = 1$ dado que no ha habido cambio en los valores de las variables independientes (odds_0).

COCIENTE DE POSIBILIDADES

$$\text{Cociente de posibilidades} = \frac{\text{odds}_1}{\text{odds}_0} \tag{15.34}$$

Por ejemplo, suponga que se desea comparar las posibilidades de que use el cupón un cliente que gasta $2 000 anuales y tiene tarjeta de crédito de Simmons ($x_1 = 2$ y $x_2 = 1$) con las posibilidades de que lo use otro que gasta $2 000 anuales y no tiene tarjeta de crédito de Simmons ($x_1 = 2$ y $x_2 = 0$). Lo que interesa es interpretar el efecto que tiene un incremento de una unidad en la variable independiente x_2. En este caso,

$$\text{odds}_1 = \frac{P(y = 1 | x_1 = 2, x_2 = 1)}{1 - P(y = 1 | x_1 = 2, x_2 = 1)}$$

y

$$\text{odds}_0 = \frac{P(y = 1 | x_1 = 2, x_2 = 0)}{1 - P(y = 1 | x_1 = 2, x_2 = 0)}$$

Como ya se demostró, la estimación de la probabilidad de que $y = 1$ cuando $x_1 = 2$ y $x_2 = 1$ es 0.4099, y la estimación de la probabilidad que $y = 1$ cuando $x_1 = 2$ y $x_2 = 0$ es 0.1880. Por tanto,

$$\text{estimación de odds}_1 = \frac{0.4099}{1 - 0.4099} = 0.6946$$

y

$$\text{estimación de odds}_0 = \frac{0.1880}{1 - 0.1880} = 0.2315$$

La estimación resultante es

$$\text{Estimación del cociente de posibilidades} = \frac{0.6946}{0.2315} = 3.00$$

Por consiguiente, podemos concluir que las posibilidades estimadas de que usen el cupón los clientes que gastaron $2 000 el año pasado y tienen tarjeta de crédito de Simmons son tres veces mayores que las de quienes gastaron $2 000 el año pasado y no tienen tarjeta de crédito de Simmons.

El cociente de posibilidades de cada una de las variables independientes se calcula al mantener constantes todas las demás variables independientes. Sin embargo, no tiene importancia qué valores constantes se usen para todas las demás variables. Por ejemplo, si se calcula el cociente de posibilidades para la variable tarjeta de crédito de Simmons (x_2) utilizando $3 000 en lugar de $2 000 como valor de la variable cantidad de gasto anual (x_1), el valor obtenido para el cociente de posibilidad estimado será el mismo (3.00). Por tanto, se concluye que las posibilidades estimadas de que use el cupón un cliente con tarjeta de crédito de Simmons son tres veces mayores que las posibilidades estimadas de que lo use un cliente sin la tarjeta de crédito.

El cociente de posibilidades es un resultado estándar para la regresión logística en los programas de software. Remítase a los resultados de Minitab de la figura 15.13. En la columna titulada Odds Ratio aparecen los cocientes de posibilidad estimados correspondientes a cada una de las variables independientes. Para x_1 es 1.41 y para x_2 es 3.00. Ya se indicó antes cómo

interpretar el coeficiente de posibilidad estimado en el caso de la variable binaria independiente x_2. Ahora se considerará la interpretación de este cociente en el caso de la variable continua independiente x_1.

El valor 1.41 en la columna Odds Ratio de los resultados de Minitab indica que la posibilidad estimada de que use el cupón un cliente que gastó \$3 000 durante el año pasado es 1.41 veces mayor que la probabilidad estimada de que lo use un cliente que gastó \$2 000. Más aún, esta interpretación es correcta para cualquier cambio en una unidad de x_1. Por ejemplo, las posibilidades estimadas de que use el cupón alguien que gastó el año pasado \$5 000 son 1.41 veces mayores que las de que lo use un cliente cuyo gasto anual fue de \$4 000. Pero suponga que interesa la variación en las posibilidades cuando hay un incremento de más de una unidad en cualquiera de las variables independientes. Observe que x_1 toma valores desde 1 hasta 7. El cociente de posibilidades presentado en los resultados de Minitab no responde esta pregunta. Para contestarla es necesario explorar la relación entre el cociente de posibilidades y los coeficientes de regresión.

Existe una relación única entre el cociente de posibilidades de una variable y su correspondiente coeficiente de regresión. Podemos demostrar que para toda variable independiente de una ecuación de regresión logística

$$\text{Cociente de posibilidades} = e^{\beta_i}$$

Para ilustrar esta relación con el ejemplo de Simmons Stores, considere la variable independiente x_1. El cociente de posibilidades estimado para x_1 es

$$\text{Cociente de posibilidades estimado} = e^{b_1} = e^{0.341643} = 1.41$$

De manera similar para x_2

$$\text{Cociente de posibilidades estimado} = e^{b_2} = e^{1.09873} = 3.00$$

Esta relación entre el cociente de posibilidades y los coeficientes de las variables independientes facilitan el cálculo del primero una vez obtenidas las estimaciones de los parámetros del modelo. Además, también permite investigar cambios en el cociente de posibilidades cuando se presentan variaciones mayores o menores a una unidad en una de las variables independientes continuas.

El cociente de posibilidades de una variable independiente representa la variación en las posibilidades de un cambio de una unidad en ella, permaneciendo constantes todas las demás variables independientes. Suponga que se desea conocer el efecto de una variación de más de una unidad, por ejemplo de c unidades. Digamos que, en el ejemplo de Simmons, queremos comparar las posibilidades de que use el cupón un cliente que gasta \$5 000 anuales ($x_1 = 5$) con las posibilidades de que lo use un cliente que gasta \$2 000 anuales ($x_1 = 2$). En este caso $c = 5 - 2 = 3$, y el correspondiente cociente de posibilidades es

$$e^{cb_1} = e^{3(0.341643)} = e^{1.0249} = 2.79$$

Esto indica que las posibilidades estimadas de que usen el cupón los clientes cuyo gasto anual es de \$5 000 son 2.79 veces mayores que las de quienes gastan \$2 000. En otras palabras, el cociente de posibilidades estimado para un aumento de \$3 000 en los gastos anuales es 2.79.

En general, el cociente de posibilidades permite comparar las posibilidades de dos eventos diferentes. Si el valor de este cociente es 1, los dos eventos tienen las mismas posibilidades. Por tanto, si la variable independiente que se considera (como el estatus respecto de la tarjeta de crédito de Simmons) tiene efecto positivo sobre la probabilidad de que el evento ocurra, el cociente de posibilidades correspondiente será mayor que 1. La mayoría del software para estadística también proporciona un intervalo de confianza para el cociente de posibilidades. En la

figura 15.13 los resultados de Minitab indican un intervalo de 95% de confianza para cada uno de los cocientes. Por ejemplo, la estimación puntual del cociente de posibilidad de x_1 es 1.41 y el intervalo de 95% de confianza va de 1.09 a 1.81. Como este intervalo no contiene el valor 1, se concluye que x_1 tiene un efecto significativo sobre el cociente de posibilidades estimado. De manera similar, el intervalo de 95% de confianza para el cociente de posibilidades de x_2 va de 1.25 a 7.17, y como tampoco contiene el valor 1, también concluimos que x_2 tiene un efecto significativo sobre el cociente de posibilidades.

Transformación logit

Entre las posibilidades a favor de $y = 1$ y el exponente de e en la ecuación de regresión logística, observamos una interesante relación. Podemos demostrar que

$$\ln(\text{odds}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Esta ecuación indica que el logaritmo natural de las posibilidades a favor de $y = 1$ es una función lineal de las variables independientes. A esta función lineal se le llama **logit**. Para denotar el logit usamos la notación $g(x_1, x_2, \ldots, x_p)$.

LOGIT

$$g(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad \textbf{(15.35)}$$

Al sustituir en la ecuación (15.27) $\beta_1 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ por $g(x_1, x_2, \ldots, x_p)$, podemos expresar la ecuación de regresión logística como

$$E(y) = \frac{e^{g(x_1, x_2, \ldots, x_p)}}{1 + e^{g(x_1, x_2, \ldots, x_p)}} \qquad \textbf{(15.36)}$$

Una vez estimados los parámetros de la ecuación de regresión logística, calculamos una estimación del logit. Con $\hat{g}(x_1, x_2, \ldots, x_p)$ para denotar el **logit estimado** tenemos

LOGIT ESTIMADO

$$\hat{g}(x_1, x_2, \ldots, x_p) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \qquad \textbf{(15.37)}$$

Por consiguiente, en términos del logit estimado, la ecuación de regresión estimada es

$$\hat{y} = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}} = \frac{e^{\hat{g}(x_1, x_2, \ldots, x_p)}}{1 + e^{\hat{g}(x_1, x_2, \ldots, x_p)}} \qquad \textbf{(15.38)}$$

En el ejemplo de Simmons Stores, el logit estimado es

$$\hat{g}(x_1, x_2) = -2.14637 + 0.341643 x_1 + 1.09873 x_2$$

y la ecuación de regresión estimada es

$$\hat{y} = \frac{e^{\hat{g}(x_1, x_2)}}{1 + e^{\hat{g}(x_1, x_2)}} = \frac{e^{-2.14637 + 0.341643 x_1 + 1.09873 x_2}}{1 + e^{-2.14637 + 0.341643 x_1 + 1.09873 x_2}}$$

Por tanto, debido a la relación única que existe entre el logit estimado y la ecuación de regresión logística estimada, podemos calcular las probabilidades estimadas para Simmons Stores dividiendo $e^{\hat{g}(x_1, x_2)}$ entre $1 + e^{\hat{g}(x_1, x_2)}$.

NOTAS Y COMENTARIOS

1. Debido a la relación única que existe entre los coeficientes estimados del modelo y los correspondientes cocientes de posibilidades, la prueba general de significancia basada en el estadístico G es también una prueba general de significancia para los cocientes de posibilidades. Además, la prueba z para la significancia de cada uno de los parámetros del modelo también es una prueba estadística de significancia para los correspondientes cocientes de posibilidades.
2. En las regresiones simple y múltiple se usa el coeficiente de determinación para medir la bondad de ajuste. En la regresión logística no hay una sola medida que tenga una interpretación similar. El estudio de la bondad de ajuste queda fuera del alcance de esta introducción a la regresión logística.

Ejercicios

Aplicaciones

WEB archivo
Simmons

44. Remítase al ejemplo de Simmons Stores presentado en esta sección. La variable dependiente es $y = 1$ si el cliente usó el cupón y $y = 0$ si no lo usó. Suponga que la única información de que se dispone para predecir si un cliente usará o no el cupón es su estatus respecto de la posesión de una tarjeta de crédito de la empresa, que es $x = 1$ si el cliente cuenta con ella y $x = 0$ si no es así.
    a) Proporcione la ecuación de regresión logística que relaciona x y y.
    b) ¿Cuál es la interpretación de $E(y)$ cuando $x = 0$?
    c) Con los datos de Simmons presentados en la tabla 15.11, use Minitab para calcular el logit estimado.
    d) Con el logit estimado del inciso *c*) obtenga una estimación de la probabilidad de que usen el cupón los clientes que no tienen tarjeta de crédito de Simmons y una estimación de la probabilidad de que lo usen quienes tienen la tarjeta.
    e) Proporcione la estimación del cociente de posibilidades. ¿Cuál es su interpretación?

45. En la tabla 15.12 se presentaron estimaciones de las probabilidades de uso del cupón en la promoción por catálogo de Simmons Stores. Para cada combinación de valores de las variables independientes se obtuvo un valor diferente.
    a) Calcule las posibilidades de que use el cupón un cliente cuyo gasto anual en Simmons es de $4 000 y que no tiene tarjeta de crédito de la tienda ($x_1 = 4, x_2 = 0$).
    b) Use la información de la tabla 15.12 y el inciso *a*) para calcular el cociente de posibilidades para la variable tarjeta de crédito de Simmons $x_2 = 0$, manteniendo constantes los gastos anuales en $x_1 = 4$.
    c) En el libro, el cociente de posibilidades para la variable tarjeta de crédito se calculó con la información presentada en la columna $2 000 de la tabla 15.12. ¿Obtuvo la misma información para el valor del cociente de posibilidades en el inciso *b*)?

46. El Community Bank desea aumentar la cantidad de clientes a los que se les deposita directamente su sueldo. La gerencia está considerando una campaña que requerirá que cada gerente de sucursal llame a cada cliente que no reciba su sueldo por depósito. Como incentivo para que acepten esta propuesta, se les ofrecerá consultas de saldo gratis durante dos años. Debido al tiempo y a los costos de esta campaña, la gerencia desea que se dirija a clientes que tengan la mayor probabilidad de aceptar recibir su sueldo por depósito. La gerencia piensa que el saldo promedio mensual en la cuenta de cheques del cliente puede ser un predictor útil para determinar si aceptará o no recibir su sueldo por depósito. Para investigar la relación entre estas dos variables, Community Bank prueba la nueva campaña con una muestra de cuentas de cheques de 50 clientes que actualmente no reciben directamente su sueldo por depósito. En los datos muestrales se presenta el saldo mensual (Monthly Balance) promedio en la cuenta de cheques (en miles de dólares) y si el cliente (Customer) aceptó recibir su sueldo por depósito directo (Direct Deposit) 1 significa que aceptó y 0 que no aceptó. Esta información se encuentra en el conjunto de datos nombrado Bank; a continuación se presenta una parte.

WEB archivo
Bank

Customer	x = Monthly Balance	y = Direct Deposit
1	1.22	0
2	1.56	0
3	2.10	0
4	2.25	0
5	2.89	0
6	3.55	0
7	3.56	0
8	3.65	1
.	.	.
.	.	.
.	.	.
48	18.45	1
49	24.98	0
50	26.05	1

a) Escriba la ecuación de regresión logística que relaciona x con y.
b) Con base en los datos de Community Bank, use Minitab para calcular la ecuación de regresión logística estimada.
c) Realice una prueba de significancia con el estadístico de prueba G. Use $\alpha = 0.05$.
d) Estime la probabilidad de que los clientes cuyo saldo mensual promedio es de $1 000 acepten recibir directamente el depósito de su nómina.
e) Suponga que Community Bank desea contactar únicamente a los clientes para los que la probabilidad de que acepten recibir directamente su nómina es de 0.50 o mayor. ¿Cuál es el saldo promedio mensual requerido para tener este nivel de probabilidad?
f) Proporcione la estimación del cociente de posibilidades. ¿Cuál es su interpretación?

47. En Lakeland College aumentó el porcentaje de estudiantes que abandonaron sus estudios después del primer año. El año pasado la institución inició un programa voluntario de orientación de una semana para ayudar a los de primer ingreso a que se adapten a la vida del campus. Si Lakeland College demuestra que ese programa tiene resultados positivos, se considerará la posibilidad de que sea obligatorio para todos los alumnos de primer año. La administración del colegio supone que los estudiantes que obtienen una calificación baja en el examen GPA tienen mayor probabilidad de abandonar los estudios al final del primer año. Con objeto de investigar la relación de estas variables con la permanencia en la escuela, Lakeland College tomó una muestra aleatoria de 100 estudiantes (Student) de primer año. La información se encuentra en el conjunto de datos Lakeland. A continuación se reproduce parte de esos datos. En Program se listan los valores asignados al programa y en Return, los de la permanencia de los alumnos.

WEB archivo
Lakeland

Student	GPA	Program	Return
1	3.78	1	1
2	2.38	0	1
3	1.30	0	0
4	2.19	1	0
5	3.22	1	1
6	2.68	1	1
.	.	.	.
.	.	.	.
.	.	.	.
98	2.57	1	1
99	1.70	1	1
100	3.85	1	1

La variable dependiente toma el valor $y = 1$ si el estudiante permanece en la escuela y $y = 0$ si la abandona. Las dos variables independientes son:

$$x_1 = \text{GPA al final del primer semestre}$$

$$x_2 = \begin{cases} 0 \text{ si el estudiante participa en el programa de orientación} \\ 1 \text{ si el estudiante no participa en el programa de orientación} \end{cases}$$

a) Escriba la ecuación de regresión logística que relaciona x_1 y x_2 con y.
b) ¿Cuál es la interpretación de $E(y)$ cuando $x_2 = 0$?
c) Use las dos variables independientes y Minitab para calcular el logit estimado.
d) Realice una prueba de significancia global empleando $\alpha = 0.05$.
e) Utilizando $\alpha = 0.05$, determine si cada una de las variables independientes es significativa.
f) Con el logit estimado del inciso *c*) obtenga una estimación de la probabilidad de que un estudiante cuyo GPA es 2.5 y que no participó en el programa de orientación permanezca en la escuela. ¿Cuál es la estimación de esta probabilidad para un estudiante que participó en el programa de orientación cuyo GPA es 2.5?
g) Proporcione la estimación del cociente de posibilidades para el programa de orientación. Interprétela.
h) ¿Recomendaría convertir el programa de orientación en un curso obligatorio? ¿Por qué?

48. *Consumer Reports* realizó una prueba de sabor sobre 19 marcas de chocolates empacados. En los datos siguientes se proporciona el precio (Price) por porción con base en el tamaño de porción de la FDA que es de 1.4 onzas, así como una evaluación (Rating) de la calidad de los 19 chocolates, con su respectivo fabricante (Manufacturer), considerados para la prueba. Los niveles de calificación son: Excellent (excelente), Very Good (muy bueno), Good (bueno) y Fair (regular). (*Consumer Reports,* febrero de 2002).

WEB archivo
Chocolate

Manufacturer	Price	Rating
Bernard Callebaut	3.17	Very Good
Candinas	3.58	Excellent
Fannie May	1.49	Good
Godiva	2.91	Very Good
Hershey's	0.76	Good
L.A. Burdick	3.70	Very Good
La Maison du Chocolate	5.08	Excellent
Leonidas	2.11	Very Good
Lindt	2.20	Good
Martine's	4.76	Excellent
Michael Recchiuti	7.05	Very Good
Neuchatel	3.36	Good
Neuchatel Sugar Free	3.22	Good
Richard Donnelly	6.55	Very Good
Russell Stover	0.70	Good
See's	1.06	Very Good
Teuscher Lake of Zurich	4.66	Very Good
Whitman's	0.70	Fair
Whitman's Sugar Free	1.21	Fair

Suponga que desea determinar si los productos más caros tienen la calificación más alta en calidad. Para los propósitos de este ejercicio, use la siguiente variable binaria dependiente.

$y = 1$ si la evaluación de la calidad fue excelente o muy buena, y 0 si fue buena o regular.

a) Escriba la ecuación de regresión logística que relaciona $x =$ precio por porción con y.
b) Use Minitab para calcular el logit estimado.
c) Con base en el logit estimado, desarrolle una estimación de la probabilidad de que la evaluación de un chocolate cuyo precio por porción es de $4.00 sea muy bueno o excelente.
d) ¿Cuál la estimación del cociente de posibilidades? ¿Cuál es su interpretación?

Resumen

En este capítulo se presentó la regresión múltiple como extensión del análisis de regresión lineal simple expuesto en el capítulo 14. El análisis de regresión múltiple permite entender cómo está relacionada una variable dependiente con dos o más variables independientes. La ecuación

de regresión múltiple $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ indica que el valor esperado o media de la variable dependiente y, denotado como $E(y)$, está relacionado con los valores de las variables independientes $x_1, x_2, \ldots, x_p$. Para obtener la ecuación de regresión múltiple estimada $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$ se emplean los datos muestrales y el método de mínimos cuadrados. En efecto, $b_0, b_1, b_2, \ldots, b_p$ son estadísticos muestrales utilizados para estimar los parámetros desconocidos $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ del modelo. A lo largo del capítulo se emplearon salidas de pantalla de computadora para hacer énfasis en el hecho de que el software para estadística es el único medio realista para realizar los numerosos cálculos que se requieren en el análisis de regresión múltiple.

El coeficiente de determinación múltiple se presentó como una medida de la bondad de ajuste de la ecuación de regresión estimada. Este coeficiente determina la proporción de la variación en y que puede ser explicada por la ecuación de regresión estimada. El coeficiente de determinación múltiple ajustado es una medida similar de bondad de ajuste que se adapta al número de variables independientes evitando, de esta manera, sobreestimar el efecto de adicionar más variables independientes.

Como un medio para determinar estadísticamente si la relación entre las variables era significativa, se presentaron una prueba *F* y una prueba *t*. La primera permite determinar si existe una relación global significativa entre la variable dependiente y el conjunto de todas las variables independientes, y la segunda se usa para determinar si existe una relación significativa entre la variable dependiente y una determinada variable independiente del modelo de regresión. También se trató la relación entre las variables independientes, a lo cual se le llama *multicolinealidad*.

En la sección sobre variables cualitativas independientes se mostró el uso de variables ficticias para incorporar datos cualitativos en el análisis de regresión múltiple. En la sección sobre análisis residual se estudió el análisis residual, el cual permite confirmar los supuestos del modelo, detectar observaciones atípicas e identificar observaciones influyentes. Se estudiaron asimismo los residuales estandarizados, la influencia, los residuales eliminados estudentizados y la medida de la distancia de Cook. El capítulo concluye con una sección sobre el uso de la regresión logística para modelar situaciones en las que la variable dependiente sólo puede asumir dos valores.

Glosario

Análisis de regresión múltiple Análisis de regresión que involucra dos o más variables independientes.

Cociente de posibilidades Cociente que se obtiene al dividir la posibilidad de que $y = 1$ dado que una de las variables independientes aumentó en una unidad (odds_1), entre la posibilidad de que $y = 1$ dado que no hay alguna variación en los valores de las variables independientes (odds_0); es decir, cociente de posibilidades (odds ratio) = $\text{odds}_1/\text{odds}_0$.

Coeficiente de determinación múltiple Medida de la bondad de ajuste de la ecuación de regresión múltiple estimada. Se puede interpretar como la proporción en la variabilidad de la variable dependiente que es explicada por la ecuación de regresión estimada.

Coeficiente de determinación múltiple ajustado Medida de la bondad de ajuste de la ecuación de regresión múltiple estimada que se modifica con base en el número de variables independientes en el modelo, y por tanto evita sobreestimar el efecto de agregar más variables independientes.

Ecuación de regresión logística Ecuación matemática que relaciona $E(y)$, la probabilidad de que $y = 1$, con los valores de las variables independientes; es decir $E(y) = P(y = 1|x_1, x_2, \ldots, x_p) = \dfrac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$.

Ecuación de regresión logística estimada Estimación de la ecuación de regresión logística que se basa en datos muestrales; es decir $\hat{y}$ = estimación de $P(y = 1|x_1, x_2, \ldots, x_p) = \dfrac{e^{b_0+b_1 x_1+b_2 x_2+\cdots+b_p x_p}}{1 + e^{b_0+b_1 x_1+b_2 x_2+\cdots+b_p x_p}}$.

Ecuación de regresión múltiple Ecuación matemática que relaciona el valor esperado o valor medio de la variable dependiente con los valores de las variables independientes; es decir, $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.

Ecuación de regresión múltiple estimada Estimación de la ecuación de regresión múltiple basada en datos muestrales y en el método de mínimos cuadrados; es decir, $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p$.

Influencia Mide qué tan lejos se encuentran de su media los valores de las variables independientes.

Logit Logaritmo natural de las posibilidades a favor de $y = 1$; es decir, $g(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p$.

Logit estimado Estimación del logit basado en datos muestrales; es decir, $\hat{g}(x_1, x_2, \ldots, x_p) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p$.

Medida de la distancia de Cook Medida de la influencia de una observación basada tanto en la influencia (*leverage*) de la observación *i* como en el residual de la observación *i*.

Método de mínimos cuadrados Método empleado para obtener la ecuación de regresión estimada. Permite minimizar la suma de los cuadrados de los residuales (las desviaciones entre los valores observados de la variable dependiente, y_i, y los valores estimados de la variable dependiente $\hat{y}_i$).

Modelo de regresión múltiple Ecuación matemática que describe cómo está relacionada la variable dependiente y con las variables independientes $x_1, x_2, \ldots, x_p$ y con el término del error ϵ.

Multicolinealidad Término que describe la correlación entre las variables independientes.

Observación atípica Es aquella que se sale del patrón que sigue el resto de las observaciones.

Observación influyente Es aquella que tiene una gran influencia en los resultados de la regresión.

Posibilidades a favor de la ocurrencia de un evento Probabilidad de que ocurra el evento dividida entre la probabilidad de que no ocurra.

Residuales eliminados estudentizados Residuales estandarizados que se basan en un error estándar de estimación corregido, el cual se obtuvo eliminando la observación *i* del conjunto de datos y realizando después el análisis de regresión y los cálculos.

Variable cualitativa independiente Variable independiente con datos cualitativos.

Variable ficticia Se utiliza para modelar el efecto de las variables cualitativas independientes. Las variables ficticias sólo toman los valores 0 y 1.

Fórmulas clave

Modelo de regresión múltiple

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p + \epsilon \quad (15.1)$$

Ecuación de regresión múltiple

$$E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p \quad (15.2)$$

Ecuación de regresión múltiple estimada

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p \quad (15.3)$$

Criterio de mínimos cuadrados

$$\min \Sigma(y_i - \hat{y}_i)^2 \quad (15.4)$$

Relación entre STC, SCR y SCE

$$\text{STC} = \text{SCR} + \text{SCE} \quad (15.7)$$

Coeficiente de determinación múltiple

$$R^2 = \frac{\text{SCR}}{\text{STC}} \tag{15.8}$$

Coeficiente de determinación múltiple ajustado

$$R_a^2 = 1 - (1 - R^2) = \frac{n - 1}{n - p - 1} \tag{15.9}$$

Cuadrado medio debido a la regresión

$$\text{CMR} = \frac{\text{SCR}}{p} \tag{15.12}$$

Cuadrado medio debido al error

$$\text{CME} = \frac{\text{SCE}}{n - p - 1} \tag{15.13}$$

Estadístico de prueba *F*

$$F = \frac{\text{CMR}}{\text{CME}} \tag{15.14}$$

Estadístico de prueba *t*

$$t = \frac{b_i}{s_{b_i}} \tag{15.15}$$

Residual estandarizado de la observación *i*

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \tag{15.23}$$

Desviación estándar del residual *i*

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \tag{15.24}$$

Medida de la distancia de Cook

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(p + 1)s^2}\left[\frac{h_i}{(1 - h_i)^2}\right] \tag{15.25}$$

Ecuación de regresión logística

$$E(y) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}} \tag{15.27}$$

Ecuación de regresión logística estimada

$$\hat{y} = \text{estimación de } P(y = 1 | x_1, x_2, \ldots, x_p) = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}} \tag{15.30}$$

Cociente de posibilidades (*odds ratio*)

$$\text{Cociente de posibilidades} = \frac{\text{odds}_1}{\text{odds}_0} \tag{15.34}$$

Logit

$$g(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{15.35}$$

Logit estimado

$$\hat{g}(x_1, x_2, \ldots, x_p) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \tag{15.37}$$

Ejercicios complementarios

49. El departamento de admisión de Clearwater College obtuvo la siguiente ecuación de regresión estimada que relaciona el promedio final obtenido en la universidad (GPA) con la puntuación del estudiante en el área de matemáticas del examen de admisión a la universidad (SAT) y con su promedio final (GPA) en bachillerato.

$$\hat{y} = -1.41 + 0.0235x_1 + 0.00486x_3$$

donde

x_1 = promedio final en el bachillerato
x_2 = puntuación en el área de matemáticas del SAT
y = promedio final en la universidad

a) Interprete los coeficientes de esta ecuación de regresión estimada.
b) Estime el promedio final (GPA) en la universidad para un estudiante cuyo promedio en el bachillerato es 84 y cuya puntuación en el área de matemáticas del examen de admisión (SAT) es 540.

50. El director de personal de Electronics Associates desarrolló la siguiente ecuación de regresión estimada que relaciona la puntuación obtenida por un empleado en un examen de satisfacción laboral con su antigüedad y nivel salarial.

$$\hat{y} = 14.4 - 8.69x_1 + 13.5x_3$$

donde

x_1 = antigüedad (años)
x_2 = nivel salarial (dólares)
y = puntuación en el examen sobre satisfacción laboral (puntuaciones más altas indican mayor satisfacción laboral)

a) Interprete los coeficientes de esta ecuación de regresión estimada.
b) Estime la puntuación obtenida en el examen sobre satisfacción laboral de un empleado que gana $6.50 por hora y cuya antigüedad es de cuatro años.

51. A continuación se presenta una parte del resultado obtenido con software para el análisis de regresión.

```
The regression equation is
Y = 8.103 + 7.602 X1 + 3.111 X2

Predictor             Coef       SE Coef           T
Constant           _______         2.667      _____
X1                 _______         2.105      _____
X2                 _______         0.613      _____

S = 3.335      R-sq = 92.3%     R-sq(adj) = _____%

Analysis of Variance

SOURCE              DF          SS          MS          F
Regression      ______        1612      ______      _____
Residual Error      12      ______      ______
Total           ______      ______
```

a) Calcule las entradas que faltan en esta pantalla.
b) Use la prueba F y $\alpha = 0.05$ para identificar si existe una relación significativa.
c) Utilice la prueba t y $\alpha = 0.05$ para demostrar H_0: $\beta_1 = 0$ y H_0: $\beta_2 = 0$.
d) Calcule R_a^2.

52. Remítase al ejercicio 49, donde se vio que el departamento de admisión de Clearwater College obtuvo la siguiente ecuación de regresión estimada que relaciona el promedio final obtenido por un estudiante en la universidad (GPA) con la puntuación alcanzada en el área de matemáticas del examen de admisión a la universidad (SAT) y su promedio final (GPA) en bachillerato.

$$\hat{y} = -1.41 + 0.0235x_1 + 0.00486x_3$$

donde

x_1 = promedio final en el bachillerato
x_2 = puntuación en el área de matemáticas del SAT
y = promedio final en la universidad

A continuación se presenta una parte del resultado obtenido con Minitab.

```
The regression equation is
Y = -1.41 + .0235 X1 + .00486 X2

Predictor             Coef        SE Coef           T
Constant           -1.4053         0.4848      _____
X1                0.023467       0.008666      _____
X2                 _______       0.001077      _____

S = 0.1298      R-sq = _______   R-sq(adj) = _______

Analysis of Variance

SOURCE              DF           SS          MS          F
Regression      _____      1.76209      _____      _____
Residual Error  _____      _______      _____
Total               9      1.88000
```

a) Calcule las entradas que faltan en esta pantalla.
b) Use la prueba *F* y 0.05 como nivel de significancia para saber si existe una relación significativa.
c) Utilice la prueba *t* y $\alpha = 0.05$ para probar $H_0: \beta_1 = 0$ y $H_0: \beta_2 = 0$.
d) ¿La ecuación de regresión estimada proporciona un buen ajuste a los datos? Explique.

53. Remítase al ejercicio 50, donde el director de personal de Electronics Associates obtuvo la siguiente ecuación de regresión estimada que relaciona la puntuación obtenida por un empleado en un examen sobre satisfacción laboral con su antigüedad y nivel salarial.

$$\hat{y} = 14.4 - 8.69x_1 + 13.5x_3$$

donde

x_1 = antigüedad (años)
x_2 = nivel salarial (dólares)
y = puntuación en el examen sobre satisfacción laboral (puntuaciones más altas indican mayor satisfacción laboral)

A continuación se presenta una parte del resultado obtenido con Minitab.

```
The regression equation is
Y = 14.4 - 8.69 X1 + 13.52 X2

Predictor             Coef        SE Coef           T
Constant            14.448          8.191        1.76
X1                  ______          1.555       _____
X2                  13.517          2.085       _____

S = 3.773       R-sq = _______%  R-sq(adj) = _______%

Analysis of Variance

SOURCE             DF           SS         MS         F
Regression          2       ______      _____     _____
Residual Error  _____         71.17      _____
Total               7         720.0
```

a) Calcule las entradas que faltan en esta pantalla.
b) Calcule *F*. Con un nivel de significancia de 0.05, pruebe si la relación es significativa.
c) ¿La ecuación de regresión estimada proporciona un buen ajuste a los datos? Explique.
d) Use la prueba *t* y $\alpha = 0.05$ para probar $H_0: \beta_1 = 0$ y $H_0: \beta_2 = 0$.

54. Tire Rack, principal distribuidor en línea de neumáticos en Estados Unidos, realiza pruebas exhaustivas a efecto de ofrecer a sus clientes productos adecuados para su vehículo, estilo de conducción y condiciones de manejo. Además, aplica una encuesta independiente entre los consumidores para ayudar a los conductores a auxiliarse mutuamente, compartiendo sus experiencias a largo plazo con los neumáticos. Los siguientes datos muestran los índices de la encuesta (en una escala de 1 a 10, donde la calificación más alta es 10) de 18 neumáticos de verano de alto desempeño (sitio web de Tire Rack, 3 de febrero de 2009). La variable Steering califica la respuesta de los neumáticos a la conducción, Trade Wear evalúa la rapidez del desgaste con base en las expectativas del conductor y Buy Again califica la satisfacción del conductor con el neumático y el deseo de comprar nuevamente uno similar.

WEB archivo
TireRack

Tire	Steering	Tread Wear	Buy Again
Goodyear Assurance TripleTred	8.9	8.5	8.1
Michelin HydroEdge	8.9	9.0	8.3
Michelin Harmony	8.3	8.8	8.2
Dunlop SP 60	8.2	8.5	7.9
Goodyear Assurance ComforTred	7.9	7.7	7.1
Yokohama Y372	8.4	8.2	8.9
Yokohama Aegis LS4	7.9	7.0	7.1
Kumho Power Star 758	7.9	7.9	8.3
Goodyear Assurance	7.6	5.8	4.5
Hankook H406	7.8	6.8	6.2
Michelin Energy LX4	7.4	5.7	4.8
Michelin MX4	7.0	6.5	5.3
Michelin Symmetry	6.9	5.7	4.2
Kumho 722	7.2	6.6	5.0
Dunlop SP 40 A/S	6.2	4.2	3.4
Bridgestone Insignia SE200	5.7	5.5	3.6
Goodyear Integrity	5.7	5.4	2.9
Dunlop SP20 FE	5.7	5.0	3.3

a) Proporcione una ecuación de regresión estimada para predecir la calificación de Buy Again con base en la puntuación de Steering. Con un nivel de significancia de 0.05, pruebe si la relación es significativa.

b) ¿La ecuación obtenida en el inciso *a*) proporciona un buen ajuste a los datos? Explique.

c) Proporcione una ecuación de regresión estimada para predecir la puntuación de Buy Again con base en la puntuación de Steering y de Trade Wear.

d) ¿Es significativa la incorporación de la variable independiente Trade Wear? Use $\alpha = 0.05$.

55. *Consumer Reports* examinó y presentó las evaluaciones de 24 caminadoras. A cada equipo se le dio una calificación general basada principalmente en su facilidad de uso, ergonomía, gama de ejercicio y calidad. En general, una mejor calificación corresponde a un mejor desempeño. En la información siguiente se presenta el precio (Price), la evaluación de la calidad (Quality) y la puntuación general (Score) de las 24 caminadoras, incluyendo marca y modelo (Brand and Model) (*Consumer Reports,* febrero de 2006). Los niveles de calidad son Excellent (excelente) y Very Good (muy bueno).

WEB archivo
Treadmills

Brand and Model	Price	Quality	Score
Landice L7	2 900	Excellent	86
NordicTrack S3000	3 500	Very good	85
SportsArt 3110	2 900	Excellent	82
Precor	3 500	Excellent	81
True Z4 HRC	2 300	Excellent	81
Vision Fitness T9500	2 000	Excellent	81
Precor M 9.31	3 000	Excellent	79
Vision Fitness T9200	1 300	Very Good	78
Star Trac TR901	3 200	Very Good	72
Trimline T350HR	1 600	Very Good	72
Schwinn 820p	1 300	Very Good	69
Bowflex 7-Series	1 500	Excellent	83
NordicTrack S1900	2 600	Very Good	83
Horizon Fitness PST8	1 600	Very Good	82
Horizon Fitness 5.2T	1 800	Very Good	80
Evo by Smooth Fitness FX30	1 700	Very Good	75
ProForm 1000S	1 600	Very Good	75
Horizon Fitness CST4.5	1 000	Very Good	74

(*Continúa*)

Brand & Model	Price	Quality	Score
Keys Fitness 320t	1 200	Very good	73
Smooth Fitness 7.1HR Pro	1 600	Very good	73
NordicTrack C2300	1 000	Good	70
Spirit Inspire	1 400	Very good	70
ProForm 750	1 000	Good	67
Image 19.0 R	600	Good	66

a) Con estos datos obtenga una ecuación de regresión estimada para calcular la calificación general cuando se conoce el precio.

b) Use $\alpha = 0.05$ para probar la significancia general.

c) Para incorporar el efecto de la calidad, una variable cualitativa de tres niveles, se emplearon dos variables ficticias: calidad-E y calidad-MB. Cada variable toma los valores 0 y 1 como sigue.

$$\text{Calidad-E} = \begin{cases} 1 \text{ si evaluación de la calidad es excelente} \\ 0 \text{ si no es el caso} \end{cases}$$

$$\text{Calidad-MB} = \begin{cases} 1 \text{ si evaluación de la calidad es muy buena} \\ 0 \text{ si no es el caso} \end{cases}$$

Obtenga una ecuación de regresión estimada para determinar la puntuación general cuando se conoce el precio y la evaluación de la calidad.

d) Pruebe la significancia general de la ecuación de regresión estimada obtenida en el inciso *c*) utilizando $\alpha = 0.10$.

e) Con la prueba *t* determine la significancia de cada una de las variables independientes de la ecuación de regresión estimada obtenida en el inciso *c*). Use $\alpha = 0.10$.

f) Proporcione la gráfica de los residuales estandarizados. ¿La forma de la gráfica parece razonable?

g) ¿Hay en estos datos alguna observación atípica o alguna observación influyente?

h) Estime la calificación general para una caminadora cuyo precio es de $2 000 y que obtuvo una evaluación de calidad buena. ¿Cuánto varía esta estimación si la evaluación de la calidad es muy buena? Explique.

56. A continuación se presenta un conjunto de datos con información de 2008 acerca de 45 fondos de inversión que forman parte del *Morningstar Funds 500*. El conjunto de datos completo está disponible en el archivo MutualFunds e incluye las siguientes cinco variables.

Fund Type (tipo de fondo). Se etiquetan como DE (capital nacional), IE (capital internacional) y FI (de renta fija).

Net Asset Value (valor neto del activo). Precio de cierre por acción al 31 de diciembre de 2007.

5-Year Average Return (rendimiento promedio de 5 años). Rendimiento anual promedio del fondo después de 5 años.

Expense Ratio (tasa de gastos). Porcentaje que se deduce de los activos cada año fiscal para fondo de gastos.

Morningstar Rank (calificación Morningstar). Puntuación con estrellas del riesgo ajustado de cada fondo. La calificación Morningstar va desde baja de 1 estrella hasta alta de 5 estrellas.

WEB archivo
MutualFunds

Fund Name	Fund Type	Net Asset Value ($)	5-Year Average Return (%)	Expense Ratio (%)	Morningstar Rank
Amer Cent Inc & Growth Inv	DE	28.88	12.39	0.67	2-Star
American Century Intl. Disc	IE	14.37	30.53	1.41	3-Star
American Century Tax-Free Bond	FI	10.73	3.34	0.49	4-Star

Fund Name	Fund Type	Net Asset Value ($)	5-Year Average Return (%)	Expense Ratio (%)	Morningstar Rank
American Century Ultra	DE	24.94	10.88	0.99	3-Star
Ariel	DE	46.39	11.32	1.03	2-Star
Artisan Intl Val	IE	25.52	24.95	1.23	3-Star
Artisan Small Cap	DE	16.92	15.67	1.18	3-Star
Baron Asset	DE	50.67	16.77	1.31	5-Star
Brandywine	DE	36.58	18.14	1.08	4-Star
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

a) Proporcione una ecuación de regresión estimada para predecir el rendimiento promedio de 5 años cuando se conoce el tipo de fondo. Realice una prueba de significancia con $\alpha = 0.05$.

b) ¿La ecuación obtenida en el inciso *a*) proporciona un buen ajuste a los datos? Explique.

c) Determine la ecuación de regresión estimada para predecir el rendimiento promedio de 5 años cuando se conoce el tipo de fondo, el valor neto de los activos y el porcentaje de gastos. Con un nivel de significancia de 0.05, pruebe para saber si hay una relación significativa. ¿Considera que se deberían eliminar algunas variables de la ecuación de regresión estimada? Explique.

d) Morningstar Rank es una variable cualitativa. Como el conjunto de datos contiene sólo fondos con cuatro categorías (de dos a cinco estrellas), utilice las siguientes variables ficticias: categoría de 3 estrellas = 1 para un fondo de 3 estrellas, 0 para cualquier otro; categoría de 4 estrellas = 1 para un fondo de 4 estrellas, 0 para cualquier otro; categoría de 5 estrellas = 1 para un fondo de 5 estrellas, 0 para cualquier otro. Obtenga una ecuación de regresión estimada para predecir el rendimiento promedio de 5 años cuando se conoce el tipo de fondo, el porcentaje de gastos y la calificación Morningstar. Utilizando $\alpha = 0.05$, elimine cualesquiera variables dependientes que no sean significativas.

e) Utilice la ecuación de regresión estimada establecida en el inciso *d*) para estimar el rendimiento promedio de 5 años de un fondo de capital nacional con un porcentaje de gastos de 1.05% y la calificación Morningstar de 3 estrellas.

57. El Departamento de Energía de Estados Unidos publica *Fuel Economy Guide*, que proporciona datos de eficiencia de combustible de automóviles y camiones (sitio web del U.S. Department of Energy, 22 de febrero de 2008). A continuación se proporciona parte de los datos de 311 automóviles compactos, medianos y grandes. En la columna respectiva, Class identifica el tamaño del automóvil (Compact/compacto, Midsize/mediano o Large/grande); Displacement muestra el desplazamiento del motor en litros; Fuel Type indica el tipo de combustible que utiliza el automóvil, premium (P) o regular (R), y Hwy MPG indica el nivel de eficiencia del combustible para el manejo en carretera en términos de millas por galón. El conjunto de datos completo se encuentra en el archivo FuelData.

WEB archivo
FuelData

Car	Class	Displacement	Fuel Type	Hwy MPG
1	Compact	3.1	P	25
2	Compact	3.1	P	25
3	Compact	3	P	25
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
161	Midsize	2.4	R	30
162	Midsize	2	P	29
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
310	Large	3	R	25
311	Large	3	R	25

a) Obtenga la ecuación de regresión estimada a efecto de predecir la eficiencia de combustible para el manejo en carretera conociendo el desplazamiento del motor. Pruebe la significancia usando $\alpha = 0.05$.

b) Considere la incorporación de las variables ficticias ClassMidsize y ClassLarge. El valor de la primera es 1 si el automóvil es mediano y 0 en cualquier otro caso; el valor de la segunda es 1 si el vehículo es grande y 0 en cualquier otro caso. Así, en un automóvil compacto, el valor de ClassMidsize y el de ClassLarge es 0. Obtenga la ecuación de regresión estimada para predecir la eficiencia de combustible para el manejo en carretera conociendo el desplazamiento del motor y las variables ficticias ClassMidsize y ClassLarge.

c) Utilice $\alpha = 0.05$ para determinar si el agregado de las variables ficticias en el inciso *b*) es significativo.

d) Considere la incorporación de la variable ficticia FuelPremium, cuyo valor es 1 si el automóvil usa combustible premium y 0 si usa gasolina regular. Obtenga la ecuación de regresión estimada a efecto de predecir la eficiencia de combustible para el manejo en carretera dados el desplazamiento del motor, las variables ficticias ClassMidsize, ClassLarge y FuelPremium.

e) Para la ecuación de regresión estimada obtenida en el inciso *d*), pruebe si hay significancia global y significancia individual utilizando $\alpha = 0.05$.

Caso a resolver 1 Consumer Research, Inc.

Consumer Research, Inc. es una agencia independiente que realiza estudios para diversas empresas sobre las actitudes y el comportamiento de los consumidores. Uno de sus clientes solicitó un estudio sobre las características de los consumidores para predecir los montos que cargan a sus tarjetas de crédito. De una muestra de 50 sujetos se obtuvieron datos sobre ingreso (Income) anual, tamaño de la familia (Household Size) y cargos (Amount Charged) anuales aplicados a las tarjetas de crédito. Los datos que se presentan a continuación provienen del archivo Consumer.

WEB archivo
Consumer

Income ($1 000s)	Household Size	Amount Charged ($)	Income ($1 000s)	Household Size	Amount Charged ($)
54	3	4 016	54	6	5 573
30	2	3 159	30	1	2 583
32	4	5 100	48	2	3 866
50	5	4 742	34	5	3 586
31	2	1 864	67	4	5 037
55	2	4 070	50	2	3 605
37	1	2 731	67	5	5 345
40	2	3 348	55	6	5 370
66	4	4 764	52	2	3 890
51	3	4 110	62	3	4 705
25	3	4 208	64	2	4 157
48	4	4 219	22	3	3 579
27	1	2 477	29	4	3 890
33	2	2 514	39	2	2 972
65	3	4 214	35	1	3 121
63	4	4 965	39	4	4 183
42	6	4 412	54	3	3 730
21	2	2 448	23	6	4 127
44	1	2 995	27	2	2 921
37	5	4 171	26	7	4 603
62	6	5 678	61	2	4 273
21	3	3 623	30	2	3 067
55	7	5 301	22	4	3 074
42	2	3 020	46	5	4 820
41	7	4 828	66	4	5 149

Informe gerencial

1. Utilice los métodos de la estadística descriptiva para resumir estos datos. Comente sus hallazgos.
2. Obtenga ecuaciones de regresión estimada usando como variables independientes primero el ingreso anual y después el tamaño de la familia. ¿Cuál de estas variables es mejor predictor de los cargos anuales a las tarjetas de crédito? Analice sus hallazgos.
3. Obtenga una ecuación de regresión estimada en la que ingreso anual y tamaño de la familia sean las variables independientes. Analice sus hallazgos.
4. ¿Cuál es el monto del cargo anual en tarjetas de crédito que podemos predecir para un hogar de tres personas con ingreso anual de $40 000?
5. Analice la necesidad de agregar otras variables independientes al modelo. ¿Cuáles serían útiles?

Caso a resolver 2 Aportaciones de exalumnos

Los donativos de los exalumnos son una importante fuente de ingresos para institutos y universidades. Si los directivos pudieran determinar los factores que influyen sobre el aumento de la cantidad de alumnos que aportan donativos, podrían poner en marcha políticas que llevarían a ganancias mayores. Las investigaciones indican que los estudiantes más satisfechos de la relación con sus profesores tienen más probabilidad de titularse. Como resultado, se podría intuir que clases más pequeñas y una menor tasa alumno/docente podrán favorecer un mayor porcentaje de graduados satisfechos, lo que a su vez conduciría al incremento de la cantidad de alumnos que realizan donaciones. En la tabla 15.13 se muestran datos de 48 universidades de Estados Unidos (*(America's Best College,* edición 2000). La columna de tasa de titulados se refiere al porcentaje de alumnos graduados respecto de los inicialmente inscritos. La columna porcentaje de clases con menos de 20 muestra la proporción de grupos con menos de 20 alumnos. La columna tasa de estudiantes por profesor proporciona la cantidad total de estudiantes inscritos dividida entre el número total de profesores. Por último, la columna tasa de alumnos que donan indica el porcentaje de egresados que ha efectuado alguna aportación a la universidad.

Informe gerencial

1. Resuma estos datos con los métodos de la estadística descriptiva.
2. Obtenga una ecuación de regresión estimada para predecir la tasa de donativos de los exalumnos, dada la cantidad de estudiantes que se titula. Analice sus hallazgos.
3. Obtenga una ecuación de regresión estimada para predecir la tasa de donativos de los exalumnos utilizando los datos proporcionados.
4. ¿Qué conclusiones y recomendaciones puede obtener de su análisis?

Caso a resolver 3 Estadísticas del PGA Tour

La Asociación de Golfistas Profesionales (PGA, por sus siglas en inglés) mantiene datos sobre desempeño y ganancias de los miembros del PGA Tour. Los 125 mejores jugadores con base en las ganancias totales en los eventos del PGA Tour están exentos para la siguiente temporada. Elaborar la lista de los 125 con más dinero es importante, pues un jugador "exento" califica para ser miembro de tiempo completo del PGA Tour en la siguiente temporada.

Por lo general, se considera que la puntuación promedio es la estadística más importante en términos del éxito en el PGA Tour. Para investigar la relación entre variables como distancia de tiro, precisión de tiro, *greens* en regulación, salidas de búnker y promedio de golpes cortos por ronda que se tienen en una puntuación promedio, en el archivo PGA Tour (sitio web del PGA

TABLA 15.13 Datos de 48 universidades de Estados Unidos

WEB archivo
Alumni

	State	Graduation Rate	% of Classes Under 20	Student-Faculty Ratio	Alumni Giving Rate
Boston College	MA	85	39	13	25
Brandeis University	MA	79	68	8	33
Brown University	RI	93	60	8	40
California Institute of Technology	CA	85	65	3	46
Carnegie Mellon University	PA	75	67	10	28
Case Western Reserve University	OH	72	52	8	31
College of William and Mary	VA	89	45	12	27
Columbia University	NY	90	69	7	31
Cornell University	NY	91	72	13	35
Dartmouth College	NH	94	61	10	53
Duke University	NC	92	68	8	45
Emory University	GA	84	65	7	37
Georgetown University	DC	91	54	10	29
Harvard University	MA	97	73	8	46
Johns Hopkins University	MD	89	64	9	27
Lehigh University	PA	81	55	11	40
Massachusetts Inst. of Technology	MA	92	65	6	44
New York University	NY	72	63	13	13
Northwestern University	IL	90	66	8	30
Pennsylvania State University	PA	80	32	19	21
Princeton University	NJ	95	68	5	67
Rice University	TX	92	62	8	40
Stanford University	CA	92	69	7	34
Tufts University	MA	87	67	9	29
Tulane University	LA	72	56	12	17
U. of California–Berkeley	CA	83	58	17	18
U. of California–Davis	CA	74	32	19	7
U. of California–Irvine	CA	74	42	20	9
U. of California–Los Angeles	CA	78	41	18	13
U. of California–San Diego	CA	80	48	19	8
U. of California–Santa Barbara	CA	70	45	20	12
U. of Chicago	IL	84	65	4	36
U. of Florida	FL	67	31	23	19
U. of Illinois–Urbana Champaign	IL	77	29	15	23
U. of Michigan–Ann Arbor	MI	83	51	15	13
U. of North Carolina–Chapel Hill	NC	82	40	16	26
U. of Notre Dame	IN	94	53	13	49
U. of Pennsylvania	PA	90	65	7	41
U. of Rochester	NY	76	63	10	23
U. of Southern California	CA	70	53	13	22
U. of Texas–Austin	TX	66	39	21	13
U. of Virginia	VA	92	44	13	28
U. of Washington	WA	70	37	12	12
U. of Wisconsin–Madison	WI	73	37	13	13
Vanderbilt University	TN	82	68	9	31
Wake Forest University	NC	82	59	11	38
Washington University–St. Louis	MO	86	73	7	33
Yale University	CT	94	77	7	50

WEB archivo
PGATour

Tour, 2009) se encuentran los datos sobre desempeño al final del año de 125 jugadores con las ganancias totales más altas en los eventos del PGA Tour de 2008. Cada fila del conjunto de datos corresponde a un jugador y están ordenados con base en el total de ganancias. Las descripciones de los datos son las siguientes.

Money (*dinero*). Ganancias totales en eventos del PGA Tour.

Scoring Average (*puntuación promedio*). Número promedio de golpes por ronda completa.

DrDist (*distancia del tiro*). Número promedio de yardas por tiro medido. En el PGA Tour, la distancia de tiro se mide en dos hoyos por ronda. Se tiene cuidado en seleccionar dos hoyos orientados en direcciones opuestas para contrarrestar el efecto del viento. Los golpes se miden hasta el punto donde se detiene la pelota sin importar que esté dentro del recorrido o no.

DrAccu (*precisión del tiro*). Porcentaje de veces en que un tiro de salida se detiene dentro del recorrido (sin importar el club). La precisión del tiro se mide en cada hoyo, excluidos los que sean par 3.

GIR (*greens en regulación*). Porcentaje de veces que un jugador logró alcanzar el *green* en regulación. Se considera un *green* alcanzado en regulación si cualquier parte de la pelota toca la superficie de *putting* luego de que se tomó el golpe GIR. Éste se determina al restar 2 del par (primer golpe en un par 3, segundo en un par 4, tercero en un par 5). En otras palabras, se considera un *green* alcanzado en regulación si el jugador llegó a la superficie de *putting* en par menos dos golpes.

Sand Saves (*salidas de búnker*). Porcentaje de veces que un jugador está en posibilidad de tener un *up and down* en un búnker de arena al lado del *green* (sin importar el marcador). *Up and down* indica que le tomó al jugador 2 tiros o menos introducir la bola en el hoyo desde un búnker de arena al lado del *green*.

PPR (*tiros cortos por ronda*). Número promedio de tiros cortos por ronda.

Scrambling. Porcentaje de veces que un jugador falla en el *green* en regulación, pero aún logra par o mejor que par.

Informe gerencial

1. Para predecir Scoring Average, obtenga una ecuación de regresión estimada usando como variable independiente primero DrDist y después DrAccu. ¿Cuál de estas variables es mejor predictor de Scoring Average? Analice sus hallazgos.
2. Obtenga una ecuación de regresión estimada en la que GIR sea la variable independiente. Compare sus hallazgos con los resultados obtenidos utilizando DrDist y DrAccu.
3. Desarrolle una ecuación de regresión estimada en la que GIR y Sand Saves sean las variables independientes. Analice sus hallazgos.
4. Obtenga una ecuación de regresión estimada en la que GIR y PPR sean las variables independientes. Analice sus hallazgos.
5. Desarrolle una ecuación de regresión estimada con GIR y Scrambling como variables independientes. Analice sus hallazgos.
6. Compare los resultados que se obtuvieron con las ecuaciones de regresión estimada donde se utilizaron como variables independientes GIR y Sand Saves, GIR y PPR, y GIR y Scrambling. Si tuviera que elegir una de esas ecuaciones para predecir Scoring Average, ¿cuál utilizaría? Explique.
7. Obtenga una ecuación de regresión estimada para predecir Scoring Average en la que GIR, Sand Saves y PPR sean las variables independientes. Compare los resultados con una ecuación de regresión estimada en la que GIR, PPR y Scrambling sean las variables independientes.
8. Desarrolle una ecuación de regresión estimada que use GIR, Sand Saves, PPR y Scrambling para predecir Scoring Average. Analice sus hallazgos.

Caso a resolver 4

Predicción del porcentaje de triunfos de la NFL

WEB archivo
NFLStats

La Liga Nacional de Futbol Americano de Estados Unidos (NFL, por sus siglas en inglés) registra diversos datos de desempeño tanto de jugadores en lo individual como de los equipos. En el archivo NLFStats (sitio web de la NFL) se presenta información del desempeño al final del año de la temporada 2005. Cada fila corresponde a un equipo de la NFL, cada uno de los cuales está clasificado por su porcentaje de juegos ganados. A continuación se lista la descripción de los datos.

WinPct	Porcentaje de juegos ganados.
TakeInt	Intercepciones realizadas; número total de intercepciones obtenidas por la defensiva del equipo.
TakeFum	Balones sueltos recuperados; número total de balones sueltos recuperados por la defensiva del equipo.
GiveInt	Intercepciones otorgadas; número total de intercepciones permitidas por la ofensiva del equipo.
GiveFum	Balones sueltos perdidos; número total de balones sueltos perdidos por la ofensiva del equipo.
DefYds/G	Número promedio de yardas permitidas por la defensiva por juego.
RushYds/G	Número promedio de yardas por tierra por juego.
PassYds/G	Número promedio de yardas por aire por juego.
FGPct	Porcentaje de goles de campo.

Informe gerencial

1. Resuma los datos con los métodos de la estadística descriptiva. Comente sus hallazgos.
2. Obtenga una ecuación de regresión estimada para predecir WinPCT usando DefYds/G, RushYds/G, PassYds/G y FGPct. Analice sus hallazgos.
3. En la ecuación de regresión estimada obtenida en el inciso 2), elimine todas las variables independientes que no sean significativas y obtenga una nueva ecuación de regresión estimada para predecir WinPct. Use $\alpha = 0.05$.
4. Algunos analistas de futbol americano consideran que las pérdidas de balón son uno de los factores más importantes para determinar el éxito de un equipo. Si Takeaways = TakeInt + TakeFum, y Giveayaws = GiveInt + GiveFum, sea NetDiff = Takeaways − Giveaways. Obtenga una ecuación de regresión estimada para predecir WinPct empleando NetDiff. Compare estos resultados con la ecuación de regresión estimada obtenida en el inciso 3).
5. Desarrolle una ecuación de regresión estimada para predecir WinPct usando todos los datos proporcionados.

Apéndice 15.1

Regresión múltiple con Minitab

WEB archivo
Butler

En la sección 15.2 se estudió la solución a problemas de regresión múltiple utilizando software con los resultados de Minitab para el caso de Butler Trucking Company. En este apéndice se describen los pasos requeridos para que Minitab genere esos resultados. Primero es necesario ingresar los datos en la hoja de cálculo. Las millas recorridas se ingresan en la columna Cl, el número de entregas en la columna C2 y el tiempo de recorrido (en horas) en la columna C3. Los nombres de las variables, Miles (millas), Deliveries (entregas) y Time (tiempo) se ingresan como encabezados de estas columnas. En los pasos subsecuentes se hará referencia a los datos con los nombres de las variables Miles, Deliveries y Time o los identificadores de las columnas Cl, C2 y C3. A continuación se describen los pasos a seguir con Minitab para obtener los resultados de regresión que se presentan en la figura 15.4.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Seleccione el menú **Regression**.
Paso 3. Elija **Regression**.
Paso 4. Cuando el cuadro de diálogo **Regression** aparezca:
Ingrese Time en la caja **Response**.
Ingrese Miles y Deliveries en el cuadro **Predictors**.
Haga clic en **OK**.

Apéndice 15.2 Regresión múltiple con Excel

WEB archivo
Butler

En la sección 15.2 se trabajó la solución de problemas de regresión múltiple empleando software mediante los resultados proporcionados por Minitab al caso de Butler Trucking Company. En este apéndice se describe el uso de la herramienta de regresión de Excel para obtener la ecuación de regresión múltiple estimada para dicho problema. A medida que se describen los pasos a seguir, consulte la figura 15.14. Primero, en las celdas A1:D1 de la hoja de trabajo se ingresan los rótulos Assignment, Miles, Deliveries y Time, y en las celdas B2:D11 se introducen los datos muestrales. En las celdas A2:A11, los números 1-10 identifican cada observación.

FIGURA 15.14 Resultados de Excel al problema de Butler Trucking con dos variables independientes

	A	B	C	D	E	F	G	H	I	J
1	**Assignment**	**Miles**	**Deliveries**	**Time**						
2	1	100	4	9.3						
3	2	50	3	4.8						
4	3	100	4	8.9						
5	4	100	2	6.5						
6	5	50	2	4.2						
7	6	80	2	6.2						
8	7	75	3	7.4						
9	8	65	4	6						
10	9	90	3	7.6						
11	10	90	2	6.1						
12										
13	SUMMARY OUTPUT									
14										
15	*Regression Statistics*									
16	Multiple R	0.9507								
17	R Square	0.9038								
18	Adjusted R Square	0.8763								
19	Standard Error	0.5731								
20	Observations	10								
21										
22	ANOVA									
23		*df*	*SS*	*MS*	*F*	*Significance F*				
24	Regression	2	21.6006	10.8003	32.8784	0.0003				
25	Residual	7	2.2994	0.3285						
26	Total	9	23.9							
27										
28		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*	
29	Intercept	−0.8687	0.9515	−0.9129	0.3916	−3.1188	1.3813	−4.1986	2.4612	
30	Miles	0.0611	0.0099	6.1824	0.0005	0.0378	0.0845	0.0265	0.0957	
31	Deliveries	0.9234	0.2211	4.1763	0.0042	0.4006	1.4463	0.1496	1.6972	
32										

Los pasos siguientes describen cómo emplear la herramienta de regresión para el análisis de regresión múltiple.

Paso 1. Haga clic en la ficha **Data** en la cinta.
Paso 2. En el grupo **Analysis**, haga clic en **Data Analysis**.
Paso 3. Elija **Regression** en la lista de Analysis Tools.
Paso 4. Cuando el cuadro de diálogo Regression aparezca:
Ingrese D1:D11 en el cuadro **Input Y Range**.
Ingrese B1:C11 en el cuadro **Input X Range**.
Seleccione **Labels**.
Elija **Confidence Level**.
Ingrese 99 en el cuadro **Confidence Level**.
Seleccione **Output Range**.
Ingrese A13 en el cuadro **Output Range** (para indicar la esquina superior izquierda de la sección de la hoja de trabajo donde deberán aparecer los resultados).
Haga clic en **OK**.

En los resultados de Excel que se presentan en la figura 15.14, el rótulo para la variable independiente x_1 es Miles (vea la celda A30) y para la variable independiente x_2 es Deliveries (vea la celda A31). La ecuación de regresión estimada es

$$\hat{y} = -0.8687 + 0.0611x_1 + 0.9234x_2$$

Observe que el uso de la herramienta de regresión de Excel para la regresión múltiple es casi igual que para la regresión lineal simple. La principal diferencia estriba en que en el caso de la regresión múltiple se requiere un rango mayor de celdas para identificar las variables independientes.

Apéndice 15.3 Regresión logística con Minitab

WEB archivo
Simmons

Minitab llama *regresión logística binaria* (*binary logistic regression 0 y 1*) a la regresión logística con una sola variable independiente que sólo puede tomar los valores 0 y 1. En este apéndice se describen los pasos que se requieren en el procedimiento de regresión logística binaria de Minitab para generar los resultados presentados en la figura 15.13 sobre el caso de Simmons. Primero, en una hoja de trabajo de Minitab deben ingresarse los datos. Las cantidades (en miles de $) que gastaron los clientes en las tiendas Simmons se ingresan en la columna C2, los datos sobre la tarjeta de crédito (1 si es de Simmons; 0 si no es el caso) se ingresan en la columna C3 y el dato sobre el uso del cupón (1 si el cliente lo usó; 0 si no lo usó) se introducen en la columna C4. Los nombres de las variables Spending (gasto) Card (tarjeta) y Coupon (cupón) se ingresan en la hoja de cálculo como encabezados de las columnas.

En los pasos subsecuentes se hará referencia a los datos utilizando los nombres de las variables Spending, Card y Coupon o los identificadores de las columnas C2, C3 y C4. Los pasos siguientes generan de los resultados de la regresión logística.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Seleccione el menú **Regression**.
Paso 3. Elija **Binary Logistic Regression**.
Paso 4. Cuando el cuadro de diálogo **Binary Logistic Regression** aparezca:
Ingrese Coupon en el cuadro **Response**.
Ingrese Spending y Card en el cuadro **Model**.
Haga clic en **OK**.

La información presentada en la figura 15.13 aparecerá como parte de los resultados.

Apéndice 15.4 Análisis de regresión múltiple con StatTools

WEB archivo
Butler

En este apéndice se muestra el uso de StatTools para realizar los cálculos del análisis de regresión sobre el problema de Butler Trucking. Se inicia con el uso del Data Set Manager a efecto de crear un conjunto de datos de StatTools para los datos utilizados en el procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes describen el uso de StatTools para probar los resultados de la regresión.

Paso 1. Haga clic en la ficha **StatTools** en la cinta.
Paso 2. En el grupo **Analyses**, haga clic en **Regression and Classification**.
Paso 3. Elija la opción **Regression**.
Paso 4. Cuando el cuadro de diálogo StatTools-Regression aparezca:
Seleccione **Multiple** en el cuadro **Regression Type**.
En la sección **Variables**:
Haga clic en el botón **Format** y seleccione **Unstacked**.
En la columna etiquetada **I** seleccione **Miles**.
En la columna etiquetada **I** seleccione **Deliveries**.
En la columna etiquetada **D** seleccione **Time**.
Haga clic en **OK**.

El resultado del análisis de regresión aparecerá en una nueva hoja de trabajo.

El cuadro de diálogo StatTools-Regression contiene varias de las opciones más avanzadas para desarrollar estimaciones de intervalos de predicción y producir gráficas de residuales. La herramienta StatTools Help proporciona información sobre todas esas opciones.

CAPÍTULO 16

Análisis de regresión: construcción de modelos

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: MONSANTO COMPANY

16.1 MODELO LINEAL GENERAL
- Modelado de relaciones curvilíneas
- Interacción
- Transformaciones que involucran la variable dependiente
- Modelos no lineales que son intrínsecamente lineales

16.2 DETERMINACIÓN DE CUÁNDO AGREGAR O ELIMINAR VARIABLES
- Caso general
- Uso de los valores-*p*

16.3 ANÁLISIS DE UN PROBLEMA MAYOR

16.4 PROCEDIMIENTOS DE SELECCIÓN DE VARIABLES
- Regresión por pasos
- Selección hacia adelante
- Eliminación hacia atrás
- Regresión de los mejores subconjuntos
- La elección final

16.5 MÉTODO DE REGRESIÓN MÚLTIPLE PARA EL DISEÑO DE EXPERIMENTOS

16.6 AUTOCORRELACIÓN Y LA PRUEBA DE DURBIN-WATSON

ESTADÍSTICA *en* LA PRÁCTICA

MONSANTO COMPANY*

SAN LUIS MISSOURI

Monsanto Company funda sus raíces en una inversión de $500 de un empresario y un almacén polvoriento en la orilla del río Mississippi, donde en 1901 John F. Queency comenzó la producción de sacarina. En la actualidad es una de las empresas químicas más grandes del país y produce más de mil productos, que incluyen químicos industriales para fabricar las superficies sintéticas de las canchas deportivas que se usan en los estadios modernos. Monsanto es una empresa multinacional que cuenta con fábricas, laboratorios y centros técnicos, y realiza operaciones de marketing en 65 países.

La División Química de Nutrición de Monsanto produce y comercializa un suplemento de metionina que se utiliza en productos alimenticios para aves de corral, cerdos y ganado. Debido a que los avicultores trabajan con altos volúmenes y bajos márgenes de utilidad, necesitan productos alimenticios rentables para aves de corral con el mayor valor nutricional posible. El alimento con una composición óptima se traducirá en un crecimiento rápido y un alto peso corporal final para un nivel determinado de consumo del alimento. La industria química trabaja en estrecha colaboración con los criadores de aves de corral para optimizar los productos alimenticios. En última instancia, el éxito depende de mantener bajo el costo de las aves en comparación con el costo de la carne y de otros productos cárnicos.

Monsanto utilizó el análisis de regresión para modelar la relación entre el peso corporal y y la cantidad de metionina x adicionada al alimento para aves de corral. Inicialmente se desarrolló la siguiente ecuación estimada de regresión lineal.

$$\hat{y} = 0.21 + 42x$$

Esta ecuación estimada de regresión resultó estadísticamente significativa; sin embargo, el análisis de residuales indicaba que una relación curvilínea sería un mejor modelo para la relación entre el peso corporal y la metionina.



Los investigadores de Monsanto utilizaron el análisis de regresión a efecto de obtener un alimento de composición óptima para los criadores de aves de corral. © Kent Knudson/PhotoLink/Getty Images/PhotoDisc.

Una investigación complementaria llevada a cabo por Monsanto mostró que, aunque pequeñas cantidades de metionina tendían a aumentar el peso corporal, en algún punto éste se estabilizaba y las cantidades adicionales del aminoácido fueron de poco o ningún beneficio. De hecho, cuando la cantidad de metionina aumentaba más allá de los requerimientos nutricionales, el peso corporal tendía a disminuir. La siguiente ecuación estimada de regresión múltiple fue utilizada para modelar la relación curvilínea entre el peso corporal y la metionina.

$$\hat{y} = -1.89 + 1.32x - 0.506x^2$$

Los resultados de la regresión le permitieron a Monsanto determinar el nivel óptimo de metionina a ser utilizado en los productos alimenticios para aves de corral.

En este capítulo se ampliará el estudio sobre el análisis de regresión mostrando de qué forma se pueden desarrollar los modelos curvilíneos como el utilizado por Monsanto. Además, se describirá una variedad de herramientas que servirá para determinar cuál es la variable independiente que lleva a la mejor ecuación de regresión estimada.

* Los autores agradecen a James R. Ryland y Robert M. Schisla, especialistas de la División Química de Nutrición de Monsanto, por proporcionar este artículo para *Estadística en la práctica*.

La construcción de modelos es el proceso de desarrollo de una ecuación de regresión que describe la relación entre una variable dependiente y una o más variables independientes. Los principales problemas en la construcción de modelos consisten en determinar la forma funcional adecuada de la relación y selección de las variables independientes que se incluirán en el modelo. En la sección 16.1 se establece el marco para la construcción de modelos con la introducción del concepto del modelo lineal general. La sección 16.2, la cual proporciona las bases para los procedimientos más complejos basados en software, presenta un modelo general que determina cuándo agregar o eliminar variables independientes. En la sección 16.3 se considera un

problema más amplio de regresión que involucra ocho variables independientes y 25 observaciones. Este problema se utiliza para ilustrar los procedimientos de selección de variables presentados en la sección 16.4, que comprenden la regresión por pasos, el procedimiento de selección hacia adelante, el procedimiento de eliminación hacia atrás y el mejor subconjunto de regresión. En la sección 16.5 se muestra cómo el análisis de regresión múltiple proporciona otro método para resolver problemas de diseño experimental, y en la sección 16.6 se explica cómo utilizar la prueba de Durbin-Watson para detectar la correlación serial o autocorrelación.

16.1 Modelo lineal general

Suponga que se obtienen los datos de una variable independiente y y las k variables independientes $x_1, x_2, \ldots, x_k$. El objetivo es utilizar estos datos para desarrollar una ecuación de regresión estimada que ofrezca la mejor relación entre las variables dependientes e independientes. Como marco general para el desarrollo de relaciones más complejas entre las variables independientes, se introduce el concepto del **modelo lineal general** que involucra p variables independientes.

Si se puede escribir un modelo de regresión en la forma de la ecuación (16.1), entonces son aplicables los procedimientos estándar de regresión múltiple descritos en el capítulo 15.

MODELO LINEAL GENERAL

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p + \epsilon \qquad \textbf{(16.1)}$$

En la ecuación (16.1) cada una de las variables independientes z_j (donde $j = 1, 2, \ldots, p$) es una función de $x_1, x_2, \ldots, x_k$ (las variables para las que se obtuvieron los datos). En algunos casos, cada z_j puede ser una función de una sola variable x. El ejemplo más sencillo es cuando se obtienen los datos de una sola variable x_1 y se desea estimar y utilizando una relación lineal. En este caso, $z_1 = x_1$, y la ecuación (16.1) se convierte en

$$y = \beta_0 + \beta_1 x_1 + \epsilon \qquad \textbf{(16.2)}$$

La ecuación (16.2) es el modelo de regresión lineal simple de la muestra presentado en el capítulo 14, con excepción de que a la variable independiente se le llama x_1 en lugar de x. En la literatura sobre modelos estadísticos, a este modelo se le llama *modelo simple de primer orden con una variable predictora*.

TABLA 16.1

Datos para el ejemplo de Reynolds

Antigüedad en meses	Básculas vendidas
41	275
106	296
76	317
10	376
22	162
12	150
85	367
111	308
40	189
51	235
9	83
12	112
6	67
56	325
19	189

WEB archivo

Reynolds

Modelado de relaciones curvilíneas

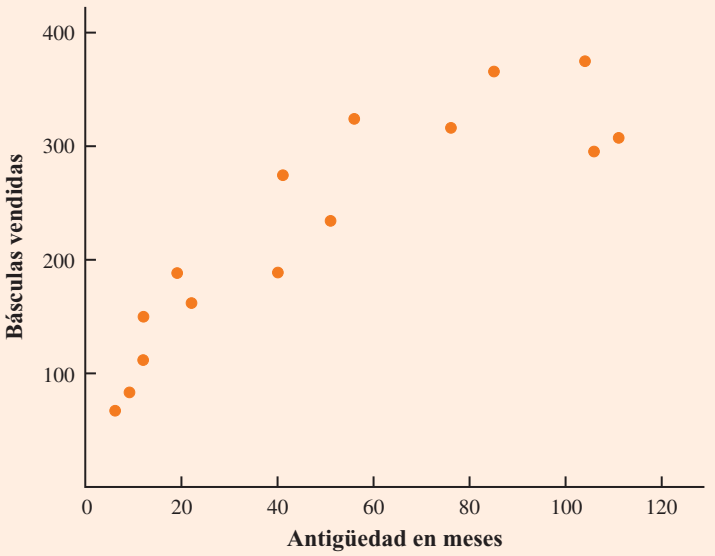
Con la ecuación (16.1) se pueden modelar los tipos de relaciones más complejos. Para ilustrar esto, considere el problema que enfrenta Reynolds, Inc., un fabricante de básculas industriales y de equipo para laboratorio. Los directivos de Reynolds desean investigar la relación que existe entre la antigüedad de su personal de ventas y el número de básculas electrónicas para laboratorio que se venden. En la tabla 16.1 se observa el número de básculas electrónicas vendidas por cada uno de los 15 vendedores elegidos aleatoriamente en el último periodo, el número de meses en los que la empresa contrató a cada uno de ellos, y el número de meses que cada vendedor ha sido empleado por la empresa. La figura 16.1 es el diagrama de dispersión para estos datos e indica una posible relación curvilínea entre el lapso de tiempo empleado y el número de unidades vendidas. Antes de considerar cómo desarrollar una relación curvilínea para Reynolds, se analizarán los resultados de Minitab en la figura 16.2 que corresponden a un modelo simple de primer orden; la regresión estimada es

$$\text{Sales} = 111 + 2.38\ \text{Months}$$

donde

Sales (ventas) = número de básculas electrónicas para laboratorio vendidas
Months (meses) = número de meses de antigüedad del vendedor

FIGURA 16.1 Diagrama de dispersión para el ejemplo de Reynolds



La figura 16.3 es la gráfica correspondiente a los residuales estandarizados. Aunque los resultados de la computadora muestran que la relación es significativa (el valor-p = 0.000) y que una relación lineal explica el alto porcentaje de variabilidad en las ventas (R-sq = 78.1%), la gráfica de residuales estandarizados sugiere que se necesita una relación curvilínea.

Para dar cuenta de la relación curvilínea, en la ecuación (16.1) se establece que $z_1 = x_1$ y $z_2 = x_1^2$ para obtener el siguiente modelo

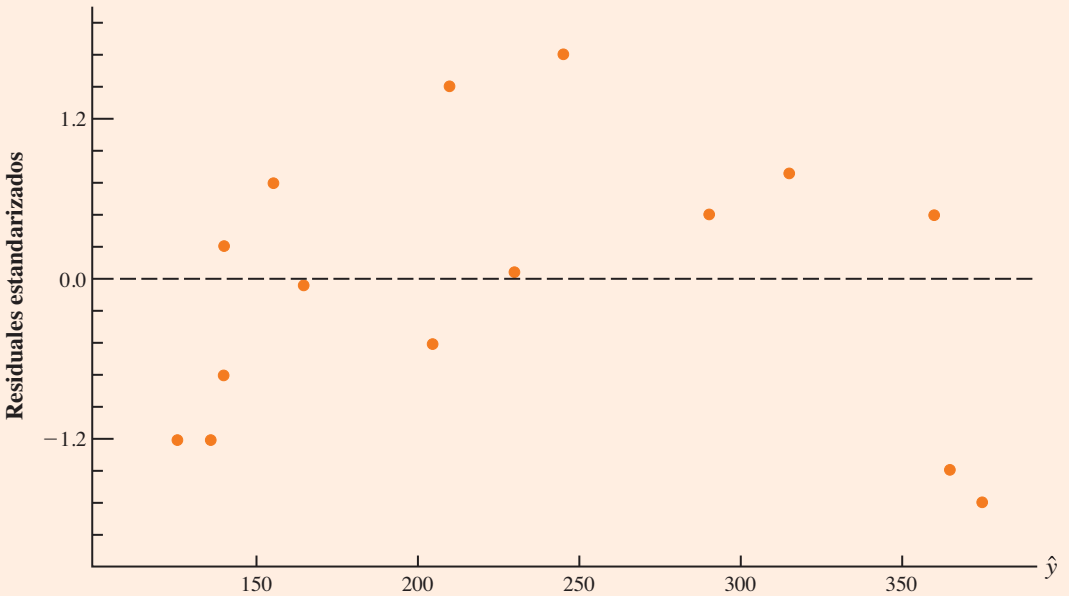
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon \qquad \textbf{(16.3)}$$

A éste se le llama *modelo de segundo orden con una variable predictora*. Para desarrollar la ecuación de regresión estimada correspondiente a este modelo, el software de estadística utilizado necesita los datos originales de la tabla 16.1, así como los datos correspondientes para

FIGURA 16.2 Resultados de Minitab para el ejemplo de Reynolds: modelo de primer orden

```
The regression equation is
Sales = 111 + 2.38 Months

Predictor     Coef  SE Coef     T      p
Constant    111.23    21.63  5.14  0.000
Months      2.3768   0.3489  6.81  0.000

S = 49.5158   R-sq = 78.1%   R-sq(adj) = 76.4%

Analysis of Variance

SOURCE           DF      SS      MS      F      p
Regression        1  113783  113783  46.41  0.000
Residual Error   13   31874    2452
Total            14  145657
```

FIGURA 16.3 Gráfica de residuales estandarizados para el ejemplo de Reynolds: modelo de primer orden



agregar una segunda variable independiente que es el cuadrado del número de meses que el trabajador ha estado con la empresa. En la figura 16.4 se muestran los resultados de Minitab correspondientes al modelo de segundo orden; la ecuación de regresión estimada es

$$\text{Sales} = 45.3 + 6.34\ \text{Months} - 0.0345\ \text{MonthsSq}$$

Los datos de la variable independiente MonthsSq *se obtienen elevando al cuadrado los valores de Months.*

donde

$$\text{MonthsSq (meses al cuadrado)} = \text{cuadrado del número de meses que el vendedor ha trabajado}$$

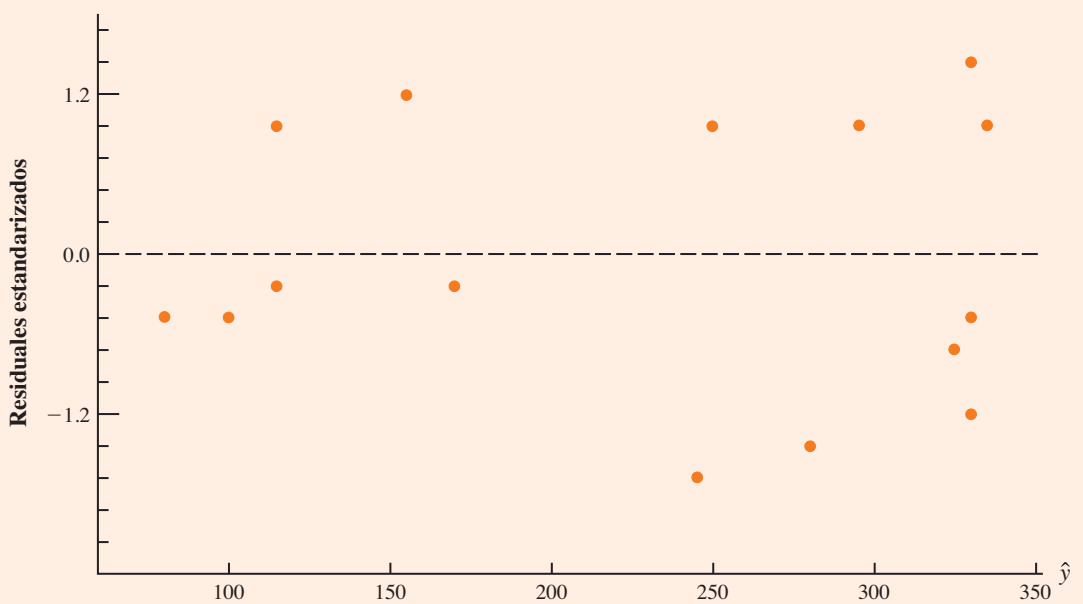
La figura 16.5 es la gráfica de residuales estandarizados correspondiente. Se muestra que el patrón curvilíneo anterior ha desaparecido. En el nivel de significancia 0.05, el resultado de Minitab indica que el modelo general es significativo (el valor-p para la prueba F es 0.000); observe también que el valor-p correspondiente a la relación t para MonthsSq (valor-p = 0.002) es menor que 0.05, y por tanto se puede concluir que la adición de MonthsSq al modelo que involucra los meses es significativa. Como el valor R-sq(adj) es 88.6%, se puede estar satisfecho con el ajuste previsto por esta ecuación de regresión estimada. Lo más importante, sin embargo, es ver lo fácil que es manejar las relaciones curvilíneas en el análisis de regresión.

Muchos tipos de relaciones pueden ser claramente modelados usando la ecuación (16.1). Las técnicas de regresión con las que se ha estado trabajando definitivamente no están limitadas a una relación lineal o en línea recta. En el análisis de regresión múltiple, la palabra *lineal* en el término "modelo lineal general" se refiere únicamente al hecho de que $\beta_0, \beta_1, \ldots, \beta_p$ tienen todos exponentes de 1; esto no implica que la relación entre y y x_1 sea lineal. De hecho, en esta sección se ha visto un ejemplo de cómo la ecuación (16.1) puede utilizarse para modelar una relación curvilínea.

FIGURA 16.4 Resultados de Minitab para el ejemplo de Reynolds: modelo de segundo orden

```
The regression equation is
Sales = 45.3 + 6.34 Months - 0.0345 MonthsSq

Predictor         Coef    SE Coef      T      p
Constant         45.35      22.77   1.99  0.070
Months           6.345      1.058   6.00  0.000
MonthsSq     -0.034486   0.008948  -3.85  0.002

S = 34.4528   R-sq = 90.2%   R-sq(adj) = 88.6%

Analysis of Variance

SOURCE           DF      SS     MS      F      p
Regression        2  131413  65707  55.36  0.000
Residual Error   12   14244   1187
Total            14  145657
```

FIGURA 16.5 Gráfica de residuales estandarizados para el ejemplo de Reynolds: modelo de segundo orden



Interacción

Si el conjunto de datos originales consta de las observaciones para y y las dos variables independientes x_1 y x_2, se puede desarrollar un modelo de segundo orden con dos variables predictoras colocando $z_1 = x_1$, $z_2 = x_2$, $z_3 = x_1^2$, $z_4 = x_2^2$ y $z_5 = x_1x_2$ en el modelo lineal general de la ecuación (16.1). El modelo que se obtiene es el siguiente.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon \qquad \textbf{(16.4)}$$

En este modelo de segundo orden la variable $z_5 = x_1x_2$ se agrega para tomar en cuenta los efectos potenciales de las dos variables en una acción conjunta. A este tipo de efecto se le llama **interacción**.

Para proporcionar un ejemplo de interacción y lo que ésta significa, se revisará el estudio de regresión realizado por Tyler Personal Care para un nuevo shampoo. Los dos factores que se cree que tienen mayor influencia sobre las ventas son el precio de venta por unidad y el gasto de publicidad. Para investigar los efectos de estas dos variables sobre las ventas se parearon los precios de \$2.00, \$2.50 y \$3.00 con los gastos de publicidad de \$50 000 y \$100 000 en 24 mercados de prueba. Las unidades vendidas (en miles) que se observaron se presentan en la tabla 16.2.

La tabla 16.3 es un resumen de estos datos. Observe que las ventas (Sales) medias muestrales correspondientes al precio (Price) de \$2.00 y un gasto en publicidad (Advertising Expenditure) de \$50 000 son de 461 000, y las ventas medias muestrales correspondientes a un precio de \$2.00 y un gasto en publicidad de \$100 000 son de 808 000. Por tanto, cuando el precio se mantiene constante en \$2.00, la diferencia en las ventas medias entre gastos de publicidad de \$50 000 y \$100 000 es de 808 000 − 461 000 = 347 000 unidades. Cuando el precio del producto es de \$2.50, la diferencia es de 646 000 − 364 000 = 282 000 unidades. Por último, cuando el precio es de \$3.00, la diferencia es 375 000 − 332 000 = 43 000 unidades. Claramente, la diferencia en las ventas medias entre los gastos de publicidad de \$50 000 y \$100 000 depende del precio del producto. En otras palabras, a precios de venta más altos, el efecto del aumento en los gastos de publicidad disminuye. Estas observaciones proporcionan la evidencia de la interacción entre las variables precio y gastos de publicidad.

Para ofrecer otro punto de vista de la interacción, la figura 16.6 presenta las ventas medias muestrales de las seis combinaciones precio-gastos de publicidad. Esta gráfica indica también que el efecto en los gastos de publicidad sobre las ventas medias depende del precio del

TABLA 16.2 Datos para el ejemplo de Tyler Personal Care

WEB archivo
Tyler

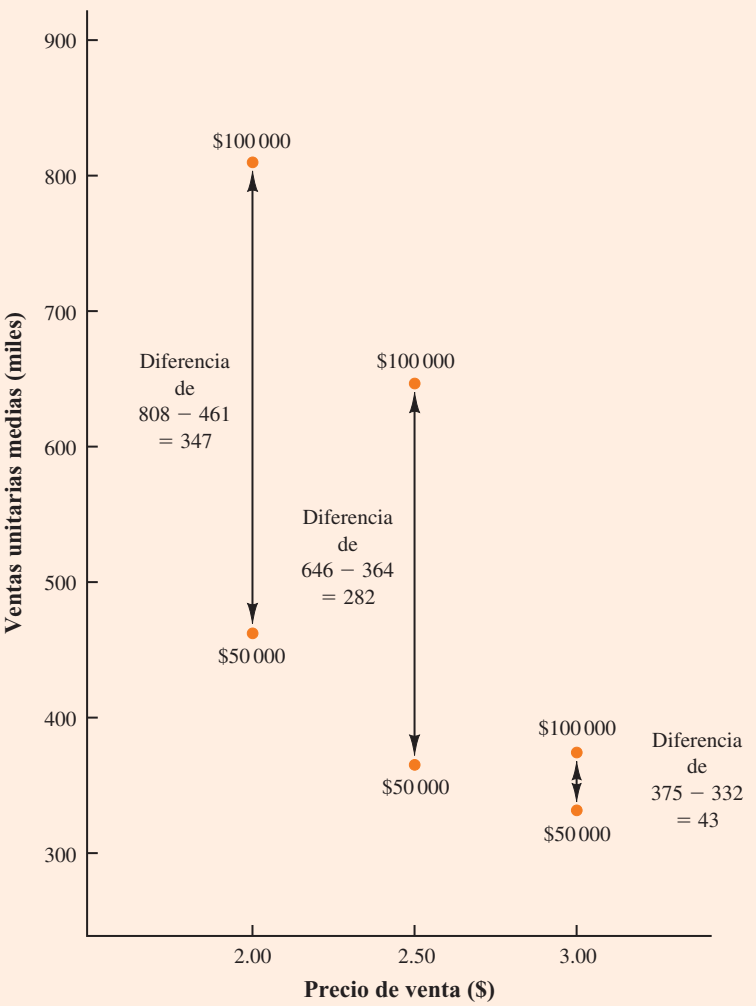
Price	Advertising Expenditure ($1 000s)	Sales (1 000s)	Price	Advertising Expenditure ($1 000s)	Sales (1 000s)
$2.00	50	478	$2.00	100	810
$2.50	50	373	$2.50	100	653
$3.00	50	335	$3.00	100	345
$2.00	50	473	$2.00	100	832
$2.50	50	358	$2.50	100	641
$3.00	50	329	$3.00	100	372
$2.00	50	456	$2.00	100	800
$2.50	50	360	$2.50	100	620
$3.00	50	322	$3.00	100	390
$2.00	50	437	$2.00	100	790
$2.50	50	365	$2.50	100	670
$3.00	50	342	$3.00	100	393

TABLA 16.3 Ventas unitarias medias (miles) para el ejemplo de Tyler Personal Care

		Precio		
		$2.00	**$2.50**	**$3.00**
Gasto en publicidad	**$50 000**	461	364	332
	$100 000	808	646	375

Ventas medias de 808 000 unidades cuando el precio es = $2.00 y el gasto de publicidad es = $100 000

FIGURA 16.6 Ventas unitarias medias (miles) como una función de precio de venta y gasto de publicidad



producto. Observe de nuevo el efecto de la interacción. Cuando existe interacción entre dos variables no se puede estudiar el efecto de una variable sobre la respuesta y en forma independiente de la otra variable. En otras palabras, las conclusiones significativas sólo pueden desarrollarse si se considera el efecto conjunto que ambas variables tienen sobre la respuesta.

Para tomar en cuenta el efecto de interacción se utilizará el siguiente modelo de regresión.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \qquad \textbf{(16.5)}$$

donde,

y = ventas unitarias (miles)
x_1 = precio ($)
x_2 = gasto de publicidad (miles $)

Observe que la ecuación (16.5) refleja la creencia de Tyler de que el número de unidades vendidas depende linealmente del precio de venta y de los gastos de publicidad (cuenta para los términos $\beta_1 x_1$ y $\beta_2 x_2$) y de que existe una interacción entre las dos variables (cuenta para el término $\beta_3 x_1 x_2$).

Para desarrollar una ecuación de regresión estimada se utilizó un modelo lineal general con tres variables independientes (z_1, z_2, z_3).

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \epsilon \qquad \textbf{(16.6)}$$

donde

$z_1 = x_1$
$z_2 = x_2$
$z_3 = x_1 x_2$

La figura 16.7 muestra el resultado de Minitab correspondiente al modelo de interacción para el ejemplo de Tyler Personal Care. El resultado de la ecuación de regresión estimada es

$$\text{Sales} = -276 + 175\ \text{Price} + 19.7\ \text{AdvExp} - 6.08\ \text{PriceAdv}$$

donde

Sales (ventas) = ventas unitarias (miles)
Price (precio) = precio del producto ($)
AdvExp (gastoPubl) = gasto en publicidad ($ miles)
PriceAdv (precioPubl) = término de interacción (precio por AdvExp)

Los datos para la variable independiente PriceAdv *se obtienen multiplicando cada valor del* Precio *por el valor correspondiente de* AdvExp.

Como el modelo es significativo (el valor-p para la prueba F es 0.000) y el valor-p correspondiente a la prueba t para PriceAdv es 0.000, se concluye que la interacción es significativa dado el efecto lineal del precio del producto y los gastos en publicidad. Así, los resultados de la regresión muestran que el efecto de los gastos en publicidad sobre las ventas depende del precio.

Transformaciones que involucran la variable dependiente

Al mostrar cómo el modelo lineal general se puede utilizar para modelar una gran variedad de posibles relaciones entre las variables independientes y la variable dependiente, se ha enfocado la atención en las transformaciones que involucran una o más variables independientes.

FIGURA 16.7 Resultado de Minitab para el ejemplo de Tyler Personal Care

```
The regression equation is
Sales = - 276 + 175 Price + 19.7 AdvExpen - 6.08 PriceAdv

Predictor      Coef  SE Coef       T      p
Constant    -275.8    112.8   -2.44  0.024
Price       175.00    44.55    3.93  0.001
Adver       19.680    1.427   13.79  0.000
PriceAdv   -6.0800   0.5635  -10.79  0.000

S = 28.1739   R-sq = 97.8%   R-sq(adj) = 97.5%

Analysis of Variance

SOURCE          DF      SS      MS       F      p
Regression       3  709316  236439  297.87  0.000
Residual Error  20   15875     794
Total           23  725191
```

TABLA 16.4
Rendimiento en millas por galón y el peso de 12 automóviles

Peso	Millas por gallón
2289	28.7
2113	29.2
2180	34.2
2448	27.9
2026	33.3
2702	26.4
2657	23.9
2106	30.5
3226	18.1
3213	19.5
3607	14.3
2888	20.9

WEB archivo
MPG

A menudo vale la pena tener en cuenta las transformaciones que implican a la variable dependiente y. Como ejemplo de cuando se desea transformar esta variable, considere los datos de la tabla 16.4 en la que se presentan el número de millas por galón y el peso de 12 automóviles. El diagrama de dispersión de la figura 16.8 indica una relación lineal negativa entre estas dos variables. Por tanto, se utilizará un modelo simple de primer orden para relacionarlas. El resultado de Minitab se muestra en la figura 16.9. La ecuación de regresión estimada es:

$$\text{MPG} = 56.1 - 0.0116\ \text{Weight}$$

donde

$$\text{MPG (millas por galón)} = \text{rendimiento de millas por galón}$$
$$\text{Weight (peso)} = \text{peso del automóvil en libras}$$

El modelo es significativo (el valor-p de la prueba F es 0.000) y el ajuste es muy bueno (R-sq = 93.5%). Sin embargo, se aprecia en la figura 16.9 que la observación 3 destaca por tener un residual estandarizado mayor.

La figura 16.10 es la gráfica de los residuales estandarizados correspondientes al modelo de primer orden. La forma que se observa no parece ser la banda horizontal que se espera encontrar si los supuestos sobre el término del error son válidos. En cambio, la variabilidad de los residuales parece aumentar a medida que se incrementa el valor de $\hat{y}$. En otras palabras, se observa que el patrón en forma de cuña al que se refieren los capítulos 14 y 15 es indicativo de una varianza no constante. No se justifica llegar a cualquier conclusión acerca de la importancia estadística del resultado para la ecuación de regresión estimada si los supuestos para la prueba de significancia parecen no estar satisfechos.

A menudo el problema de la varianza no constante se puede corregir transformando la variable dependiente en una escala diferente. Por ejemplo, si se trabaja con el logaritmo de la variable dependiente en lugar de con la variable original, el efecto será comprimir sus valores y por tanto disminuir los efectos de la varianza no constante. La mayoría del software de estadística permite aplicar las transformaciones logarítmicas utilizando ya sea la base 10 (logaritmo común) o la base $e = 2.71828\ldots$ (logaritmo natural). Aquí se aplicó la transformación

FIGURA 16.8 Diagrama de dispersión del ejemplo de millas por galón

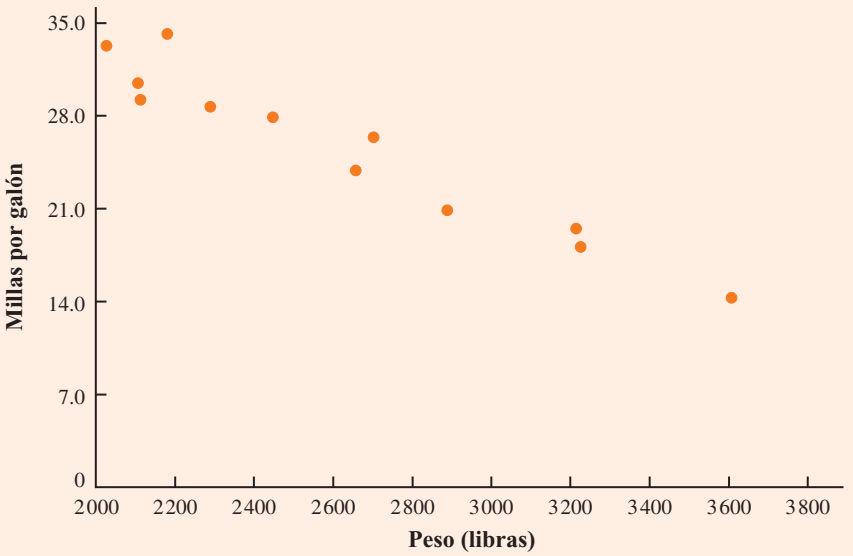
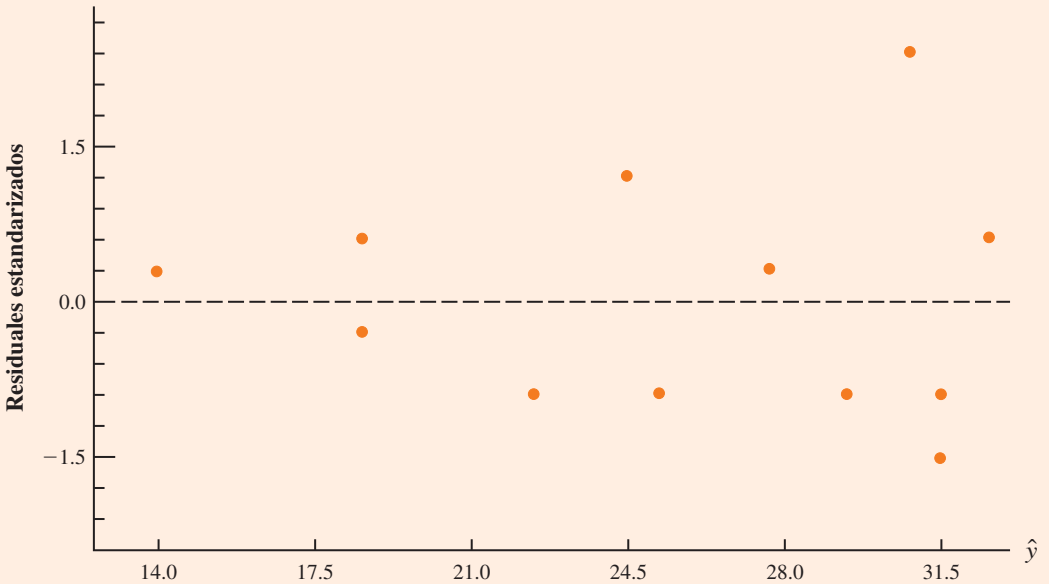


FIGURA 16.9 Resultados de Minitab para el ejemplo de millas por galón

```
The regression equation is
MPG = 56.1 - 0.0116 Weight

Predictor        Coef    SE Coef       T      p
Constant       56.096      2.582   21.72  0.000
Weight     -0.0116436  0.0009677  -12.03  0.000

S = 1.67053   R-sq = 93.5%   R-sq(adj) = 92.9%

Analysis of Variance

SOURCE          DF      SS      MS       F      p
Regression       1  403.98  403.98  144.76  0.000
Residual Error  10   27.91    2.79
Total           11  431.88

Unusual Observations
Obs  Weight     MPG     Fit  SE Fit  Residual  St Resid
  3    2180  34.200  30.713   0.644     3.487      2.26R

R denotes an observation with a large standardized residual.
```

FIGURA 16.10 Gráfica de residuales estandarizados para el ejemplo de millas por galón



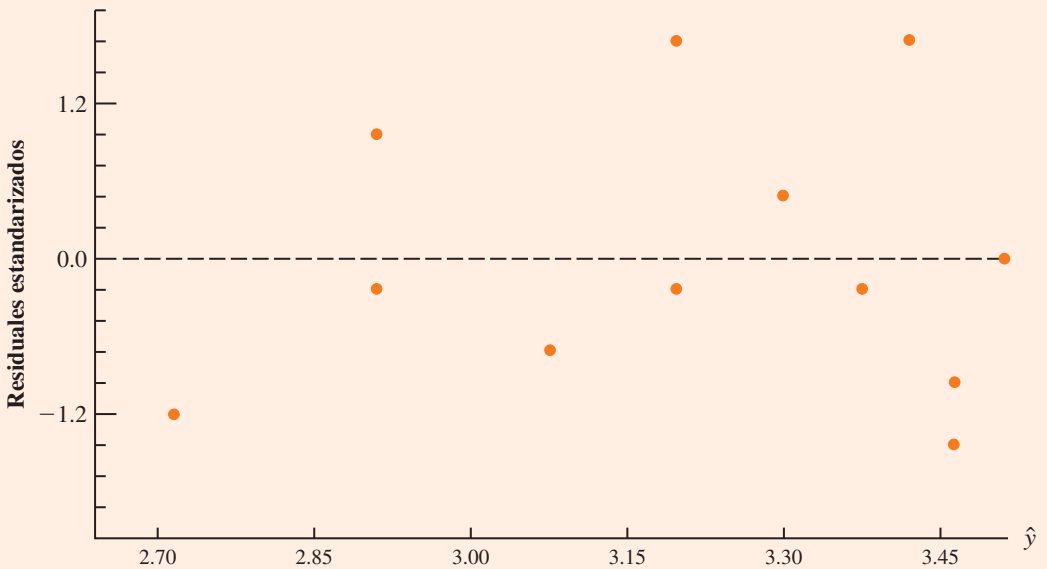
logarítmica natural a los datos en millas por galón y se desarrolló la ecuación de regresión estimada relacionando el peso con dicho logaritmo natural. Los resultados de regresión obtenidos al utilizar el logaritmo natural de millas por galón como variable dependiente, etiquetada como LogeMPG, se muestran en la figura 16.11; la figura 16.12 es la gráfica correspondiente a los residuales estandarizados.

Al observar la gráfica en la figura 16.12 vemos que la forma de cuña ha desaparecido. Por otra parte, ninguna de las observaciones se identifica por tener residuales estandarizados

FIGURA 16.11 Resultado de Minitab para el ejemplo de millas por galón: transformación logarítmica

```
The regression equation is
LogeMPG = 4.52 -0.000501 Weight

Predictor         Coef     SE Coef       T      p
Constant       4.52423     0.09932   45.55  0.000
Weight     -0.00050110  0.00003722  -13.46  0.000

S = 0.0642547   R-sq = 94.8%   R-sq(adj) = 94.2%

Analysis of Variance

SOURCE          DF       SS        MS       F      p
Regression       1  0.74822   0.74822  181.22  0.000
Residual Error  10  0.04129   0.00413
Total           11  0.78950
```

FIGURA 16.12 Gráfica de residuales estandarizados para el ejemplo de millas por galón: transformación logarítmica



mayores. El modelo con el logaritmo de millas por galón como variable dependiente es estadísticamente significativo y proporciona un excelente ajuste a los datos observados. Por tanto, se recomienda utilizar la ecuación de regresión estimada

$$\text{Loge}\text{MPG} = 4.52 - 0.000501 \text{ peso}$$

Para estimar el rendimiento en millas por galón de un automóvil que pesa 25 000 libras, en primer lugar se desarrollará una estimación del logaritmo para el rendimiento de millas por galón.

$$\text{Loge}\text{MPG} = 4.52 - 0.000501(2500) = 3.2675$$

La estimación de las millas por galón se obtiene determinando el número cuyo logaritmo natural es 3.2675. Al utilizar una calculadora con una función exponencial, o elevar e a la potencia 3.2675, se obtienen 26.2 millas por galón.

Otro método para los problemas de la varianza no constante es usar $1/y$ como variable dependiente en vez de y. A este tipo de transformación se le llama *transformación recíproca*. Por ejemplo, si la variable dependiente se mide en millas por galón, la transformación recíproca dará como resultado una nueva variable dependiente cuyas unidades serán 1/(millas por galón) o galones por milla. En general, no hay manera de determinar si una transformación logarítmica o una transformación recíproca funcionará mejor sin tener que tratar cada uno de los términos.

Modelos no lineales que son intrínsecamente lineales

Los modelos en los que los parámetros ($\beta_0, \beta_1, \ldots, \beta_p$) tienen otros exponentes distintos de 1 se denominan *modelos no lineales*. Sin embargo, para el caso del modelo exponencial se puede realizar una transformación de las variables que permita desarrollar un análisis de regresión

con la ecuación (16.1), el modelo lineal general. El modelo exponencial implica la siguiente ecuación de regresión.

$$E(y) = \beta_0\beta_1^x \tag{16.7}$$

Este modelo es adecuado cuando la variable dependiente y aumenta o disminuye en un porcentaje constante en lugar de hacerlo en una cantidad fija a medida que x aumenta.

Como ejemplo, suponga que las ventas de un producto y están relacionadas con los gastos de publicidad x (en miles de dólares) con base en el modelo exponencial siguiente.

$$E(y) = 500(1.2)^x$$

Por tanto, $x = 1$, $E(y) = 500(1.2)^1 = 600$; para $x = 2$, $E(y) = 500(1.2)^2 = 720$, y para $x = 3$, $E(y) = 500(1.2)^3 = 864$. Observe que $E(y)$ no aumenta en una cantidad constante en este caso, sino en un porcentaje constante; el incremento porcentual es de 20%.

Se puede transformar este modelo no lineal en un modelo lineal tomando el logaritmo de am-bos lados de la ecuación (16.7).

$$\log E(y) = \log \beta_0 + x \log \beta_1 \tag{16.8}$$

Ahora bien, si $y' = \log E(y)$, $\beta_0' = \log \beta_0$ y $\beta_1' = \log \beta_1$, se puede expresar la ecuación (16.8) como

$$y' = \beta_0' + \beta_1' x$$

Es evidente que las fórmulas de regresión lineal simple se pueden utilizar ahora para desarrollar estimaciones de β_0' y β_1'. Al denotar las estimaciones como b_0' y b_1', llegamos a la siguiente ecuación de regresión estimada.

$$\hat{y}' = b_0' + b_1' x \tag{16.9}$$

Para obtener las predicciones de la variable dependiente original y dado el valor de x, primero se sustituye el valor de x en la ecuación (16.9) y se calcula $\hat{y}'$. El antilogaritmo de $\hat{y}'$ será la predicción de y, o su valor esperado.

Muchos modelos no lineales pueden transformarse en un modelo lineal equivalente. Sin embargo, estos modelos han tenido un uso limitado en sus aplicaciones en el comercio y la economía. Por lo demás, la base matemática necesaria para su estudio está más allá del alcance de este libro.

Ejercicios

Métodos

AUTOevaluación

1. Considere los siguientes datos para dos variables, x y y.

x	22	24	26	30	35	40
y	12	21	33	35	40	36

   a) Desarrolle una ecuación de regresión estimada para los datos en la forma $\hat{y} = b_0 + b_1 x$.
   b) Utilice los resultados del inciso *a*) para probar una relación significativa entre x y y. Utilice $\alpha = 0.05$.
   c) Desarrolle un diagrama de dispersión de estos datos. ¿El diagrama sugiere una ecuación de regresión estimada de la forma $\hat{y} = b_0 + b_1 x + b_2 x^2$? Explique.

d) Desarrolle una ecuación de regresión estimada con los datos de la forma $\hat{y} = b_0 + b_1x + b_2x^2$.
e) Remítase al inciso *d*). ¿Es significativa la relación entre x, x^2 y y? Utilice $\alpha = 0.05$.
f) Pronostique el valor de y cuando $x = 25$.

2. Considere los siguientes datos para dos variables, x y y.

x	9	32	18	15	26
y	10	20	21	16	22

a) Desarrolle una ecuación de regresión estimada con los datos de la forma $\hat{y} = b_0 + b_1x$. Comente sobre lo apropiado de esta ecuación para predecir y.
b) Desarrolle una ecuación de regresión estimada con los datos de la forma $\hat{y} = b_0 + b_1x + b_2x^2$. Comente sobre lo adecuado de esta ecuación para predecir y.
e) Pronostique el valor de y cuando $x = 20$.

3. Considere los siguientes datos para dos variables, x y y.

x	2	3	4	5	7	7	7	8	9
y	4	5	4	6	4	6	9	5	11

a) ¿Parece haber una relación lineal entre x y y? Explique.
b) Desarrolle la ecuación de regresión estimada relacionando x y y.
c) Trace la gráfica de residuales estandarizados contra $\hat{y}$ para la ecuación de regresión estimada desarrollada en el inciso *b*). ¿Los supuestos del modelo parecen satisfacerse? Explique.
d) Realice una transformación logarítmica de la variable dependiente y. Desarrolle una ecuación de regresión estimada utilizando la variable dependiente transformada. ¿Los supuestos del modelo parecen satisfacerse al utilizar la variable dependiente transformada? ¿Funciona mejor la transformación recíproca en este caso? Explique.

Aplicaciones

4. Un departamento de autopistas estudia la relación entre el flujo del tráfico y la velocidad. El siguiente modelo ha sido admitido como hipótesis

$$y = \beta_0 + \beta_1x + \epsilon$$

donde

y = flujo de tráfico en vehículos por hora
x = velocidad del vehículo en millas por hora

Los siguientes datos fueron obtenidos durante las horas pico en las seis carreteras que salen de la ciudad.

Flujo del tráfico (y)	Velocidad del vehículo (x)
1 256	35
1 329	40
1 226	30
1 335	45
1 349	50
1 124	25

a) Desarrolle una ecuación de regresión estimada con los datos.
b) Utilice $\alpha = 0.01$ para probar una relación significativa.

AUTO evaluación

5. Trabajando con el problema del ejercicio 4, los expertos en estadística sugirieron el uso de la siguiente ecuación de regresión estimada curvilínea.

$$\hat{y} = b_0 + b_1 x + b_2 x^2$$

a) Retome los datos del ejercicio 4 para estimar los parámetros de la ecuación de regresión estimada.
b) Utilice $\alpha = 0.01$ para probar una relación significativa.
c) Estime el flujo del tráfico en vehículos por hora a una velocidad de 38 millas por hora.

6. En un estudio sobre la infraestructura del servicio de emergencia se investigó la relación entre el número de instalaciones y la distancia promedio recorrida para proporcionar este tipo de servicio. La siguiente tabla muestra los datos obtenidos.

Número de instalaciones	Distancia promedio (millas)
9	1.66
11	1.12
16	0.83
21	0.62
27	0.51
30	0.47

a) Desarrolle un diagrama de dispersión con estos datos considerando la distancia promedio recorrida como variable dependiente.
b) ¿El modelo lineal simple parece ser apropiado? Explique.
c) Desarrolle con estos datos la ecuación de regresión estimada que crea que mejor explique la relación entre estas dos variables.

7. Casi todos los sistemas de tren ligero de Estados Unidos usan los vagones eléctricos que circulan sobre rieles construidos al nivel de la calle. Según la Federal Transit Administration, el tren ligero es una de las formas de transporte más seguras, con una tasa de 0.99 accidentes por millón de millas por pasajero en comparación con 2.29 de los autobuses. Los datos siguientes muestran las millas de vías (Miles) y el número de pasajeros (Riders) que usan entre semana los sistemas seleccionados de tren ligero de diversas ciudades (City) (*USA Today*, 7 de enero de 2003).

WEB archivo

LightRail

City	Miles	Riders
Los Angeles	22	70
San Diego	47	75
Portland	38	81
Sacramento	21	31
San Jose	31	30
San Francisco	73	164
Philadelphia	69	84
Boston	51	231
Denver	17	35
Salt Lake City	18	28
Dallas	44	39
New Orleans	16	14
St. Louis	34	42
Pittsburgh	18	25
Buffalo	6	23
Cleveland	15	15
Newark	9	8

a) Desarrolle un diagrama de dispersión con estos datos; utilice la cantidad de millas de vía como variable independiente. ¿El modelo de regresión lineal simple parece apropiado?

b) Utilice el modelo de regresión lineal simple para desarrollar una ecuación de regresión estimada y pronostique el número de pasajeros de un día entre semana dado el número de millas de la vía. Construya una gráfica de residuales estandarizados. Con base en la gráfica, ¿parece ser apropiado el modelo de regresión lineal simple?

c) Realice una transformación logarítmica de la variable dependiente. Desarrolle una ecuación de regresión estimada utilizando la variable dependiente transformada. ¿Los supuestos del modelo se satisfacen usando la variable dependiente transformada?

d) Realice una transformación recíproca en la variable dependiente. Desarrolle una ecuación de regresión estimada utilizando la variable dependiente transformada.

e) ¿Cuál es la ecuación de regresión estimada que recomendaría? Explique.

8. Corvette, Ferrari y Jaguar fabricaron varios automóviles clásicos que continúan aumentando su valor. Los siguientes datos, basados en el Martin Rating System de Automóviles de Colección, muestran la clasificación (Rating) de extravagancia (1-20) y el alto precio (Price/$ miles) de 15 automóviles clásicos con su respectivo fabricante (Make), año (Year) y modelo (Model). (Sitio web de BusinessWeek, febrero de 2006.)

WEB archivo
ClassicCars

Year	Make	Model	Rating	Price ($1 000)
1984	Chevrolet	Corvette	18	1 600
1956	Chevrolet	Corvette 265/225-hp	19	4 000
1963	Chevrolet	Corvette coupe (340-bhp 4-speed)	18	1 000
1978	Chevrolet	Corvette coupe Silver Anniversary	19	1 300
1960–1963	Ferrari	250 GTE 2+2	16	350
1962–1964	Ferrari	250 GTL Lusso	19	2 650
1962	Ferrari	250 GTO	18	375
1967–1968	Ferrari	275 GTB/4 NART Spyder	17	450
1968–1973	Ferrari	365 GTB/4 Daytona	17	140
1962–1967	Jaguar	E-type OTS	15	77.5
1969–1971	Jaguar	E-type Series II OTS	14	62
1971–1974	Jaguar	E-type Series III OTS	16	125
1951–1954	Jaguar	XK 120 roadster (steel)	17	400
1950–1953	Jaguar	XK C-type	16	250
1956–1957	Jaguar	XKSS	13	70

a) Desarrolle un diagrama de dispersión para los datos utilizando la calificación de extravagancia y el precio como variables independientes. ¿El modelo de regresión lineal simple es apropiado?

b) Desarrolle una ecuación de regresión múltiple estimada con $x =$ calificación de extravagancia y x^2 como las dos variables independientes.

c) Considere la relación no lineal mostrada por la ecuación (16.7). Utilice logaritmos para desarrollar una ecuación de regresión estimada para este modelo.

d) ¿Cuál de las ecuaciones de regresión estimada prefiere, la obtenida en el inciso *b*) o la del inciso *c*)? Explique.

9. *Kiplinger's Personal Finance Magazine* clasificó 359 áreas metropolitanas de Estados Unidos a efecto de determinar las mejores ciudades para vivir, trabajar y de esparcimiento. La información en el conjunto de datos llamado MetroAreas muestra los resultados del estudio de Kiplinger para 50 áreas metropolitanas con una población de 1 000 000 o más habitantes (sitio web de Kiplinger, 2 de marzo de 2009). El conjunto de datos incluye las siguientes variables: población, ingresos, índice de costo de vida y creatividad (%). Population es el tamaño de población en miles; Income es el ingreso promedio familiar en miles; el índice del costo de vida se basa en una media nacional 100 y la creatividad (%) es el porcentaje de la fuerza de trabajo en ámbitos creativos como ciencia, ingeniería, arquitectura, educación, arte y entretenimiento. Los trabajadores en los campos creativos son por lo general considerados un factor importante en la vitalidad y habitabilidad de una ciudad, y son claves para su prosperidad económica futura.

WEB archivo
MetroAreas

a) Desarrolle un diagrama de dispersión con los datos del ingreso familiar promedio como variable independiente y el porcentaje de la fuerza de trabajo en el campo creativo como variable dependiente. ¿El modelo de regresión lineal simple parece apropiado?

b) Desarrolle un diagrama de dispersión para estos datos con el índice del costo de vida como variable independiente y el porcentaje de la fuerza de trabajo en el campo creativo como variable dependiente. ¿El modelo lineal simple parece ser apropiado?

c) Utilice los datos proporcionados para desarrollar la ecuación de regresión múltiple estimada a efecto de calcular el porcentaje de la fuerza de trabajo en los campos creativos.

d) Tucson, Arizona, es un área metropolitana con una población de 946 362 habitantes, un ingreso familiar promedio de $42 984, y un índice en el costo de vida de 99. Desarrolle una estimación del porcentaje de la fuerza de trabajo en los ámbitos creativos de Tucson. ¿Existen factores que se deben considerar antes de utilizar esta estimación puntual?

16.2 Determinación de cuándo agregar o eliminar variables

En esta sección se mostrará como usar la prueba *F* para determinar si es conveniente agregar una o más variables independientes a un modelo de regresión múltiple. Esta prueba consiste en determinar la reducción del valor de la suma de cuadrados debido al error como resultado de agregar una o más variables independientes al modelo. En primer lugar se ilustrará cómo utilizar la prueba en el contexto del ejemplo de Butler Trucking.

En el capítulo 15 se presentó el caso de Butler Trucking para ilustrar el uso del análisis de regresión múltiple. Recuerde que los directivos deseaban desarrollar una ecuación de regresión estimada para predecir el tiempo total del recorrido diario de los camiones con dos variables independientes: millas recorridas y número de repartos. Con millas recorridas x_1 como la única variable independiente, el método de mínimos cuadrados proporciona la siguiente ecuación de regresión estimada.

$$\hat{y} = 1.27 + 0.0678x_1$$

En el capítulo 15 se mostró que la suma de cuadrados debido al error con este modelo fue SCE = 8.029. Cuando se le agregó una segunda variable independiente x_2, número de entregas, se obtuvo la siguiente ecuación de regresión estimada.

$$\hat{y} = -0.869 + 0.0611x_1 + 0.923x_2$$

La suma de cuadrados debido al error para este modelo fue SCE = 2.299. Claramente, agregar x_2 dio como resultado una reducción de SCE. La pregunta que se desea responder es: ¿La adición de la variable x_2 lleva a una reducción *significativa* de la SCE?

Utilice la notación SCE(x_1) para denotar la suma de cuadrados debido al error cuando x_1 es la única variable independiente en el modelo, SCE (x_1, x_2) para denotar la suma de cuadrados debido al error cuando tanto x_1 como x_2 están en el modelo, y así sucesivamente. Por tanto, la reducción de SCE que se obtuvo al adicionar x_2 al modelo que sólo tenía x_1 es:

$$\text{SCE}(x_1) - \text{SCE}(x_1, x_2) = 8.029 - 2.299 = 5.730$$

Una prueba *F* fue realizada para determinar si esta reducción es significativa.

El numerador del estadístico *F* es la reducción de SCE dividida entre el número de variables independientes agregadas al modelo original. Aquí ha sido añadida la única variable x_2; por tanto, el numerador del estadístico *F* es

$$\frac{\text{SCE}(x_1) - \text{SCE}(x_1, x_2)}{1} = 5.730$$

El resultado es una medida de la disminución de la SCE por la variable independiente añadida al modelo. El denominador del estadístico F es el cuadrado medio debido al error para el modelo que incluye todas las variables independientes. Para Butler Trucking esto corresponde al modelo que contiene tanto x_1 como x_2, por lo que $p = 2$ y

$$\text{CME} = \frac{\text{SCE}(x_1, x_2)}{n - p - 1} = \frac{2.299}{7} = 0.3284$$

El siguiente estadístico F proporciona las bases para comprobar si la suma de x_2 es estadísticamente significativa.

$$F = \frac{\dfrac{\text{SCE}(x_1) - \text{SCE}(x_1, x_2)}{1}}{\dfrac{\text{SCE}(x_1, x_2)}{n - p - 1}} \qquad \textbf{(16.10)}$$

El número de grados de libertad del numerador para esta prueba F es igual al número de variables añadidas al modelo, y los grados de libertad en el denominador es igual a $n - p - 1$.

Para el problema de Butler Trucking obtenemos

$$F = \frac{\dfrac{5.730}{1}}{\dfrac{2.299}{7}} = \frac{5.730}{0.3284} = 17.45$$

Consulte la tabla 4 del apéndice B. Se determina que para un nivel de significancia de $\alpha = 0.05$, $F_{0.05} = 5.59$. Debido a que $F = 17.45 > F_{0.05} = 5.59$, la hipótesis nula de que x_2 no es estadísticamente significativa puede ser rechazada; en otras palabras, al agregar x_2 al modelo en el que sólo se tiene x_1 se obtiene una reducción significativa en la suma de cuadrados debido al error.

Cuando se desea probar la significancia agregando sólo una variable independiente al modelo, el resultado con la prueba F que se acaba de describir también se puede obtener mediante la prueba t para la significancia de uno de sus parámetros (el cual se describe en la sección 15.4). De hecho, el estadístico F calculado es el cuadrado del estadístico t que se utiliza para probar la significancia de un solo parámetro.

Debido a que la prueba t es equivalente a la prueba F cuando sólo una variable independiente se añade al modelo, ahora se puede aclarar aún más el uso correcto de la prueba t para probar la significancia de un solo parámetro. Si uno de los parámetros no es significativo, la variable correspondiente puede eliminarse del modelo. Sin embargo, si la prueba t muestra que hay dos o más parámetros que no son significativos, nunca se puede eliminar más de una variable independiente del modelo con base en esta prueba; si una variable es eliminada, una segunda variable que inicialmente no era significativa se vuelve significativa.

Ahora cabe considerar si la adición de más de una variable independiente, como un conjunto, da como resultado una reducción significativa de la suma de cuadrados debido al error.

Caso general

Considere el siguiente modelo de regresión múltiple con variables independientes q donde $q < p$.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \epsilon \qquad \textbf{(16.11)}$$

Si a este modelo se le agregan las variables $x_{q+1}, x_{q+2}, \ldots, x_p$, se obtiene un modelo con variables independientes p

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \beta_{q+1} x_{q+1} + \beta_{q+2} x_{q+2} + \cdots + \beta_p x_p + \epsilon \quad \textbf{(16.12)}$$

Para probar si la adición de $x_{q+1}, x_{q+2}, \ldots, x_p$ es estadísticamente significativa, las hipótesis nula y alternativa se pueden plantear de la siguiente manera.

H_0: $\beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0$
H_a: uno o más de los parámetros no es igual a cero.

El siguiente estadístico F proporciona la base para probar si la adición de las variables independientes es estadísticamente significativa.

$$F = \frac{\dfrac{\text{SCE}(x_1, x_2, \ldots, x_q) - \text{SCE}(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{p - q}}{\dfrac{\text{SCE}(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{n - p - 1}} \quad \textbf{(16.13)}$$

Este valor F calculado se compara con F_α, el valor de la tabla con $p - q$ grados de libertad en el numerador y $n - p - 1$ grados de libertad en el denominador. Si $F > F_\alpha$, entonces H_0 es rechazada y se concluye que el conjunto de variables independientes agregadas es estadísticamente significativo. Observe que para el caso especial donde $q = 1$ y $p = 2$, la ecuación (16.13) se reduce a la ecuación (16.10).

Muchos software, como Minitab, proporcionan sumas de cuadrados que corresponden al orden en que cada variable independiente entra al modelo; en estos casos se simplifican los cálculos de la prueba F para determinar si se agrega o elimina un conjunto de variables.

Muchos estudiantes encuentran la ecuación (16.13) un tanto compleja. Para dar una descripción más sencilla de este cociente F, se puede referir el modelo con el menor número de variables independientes como modelo reducido, y el modelo con el mayor número de variables independientes como modelo completo. Si SCE(reducido) denota la suma de cuadrados debido al error del modelo reducido, y SCE(completo) la suma de cuadrados debido al error del modelo completo, se puede escribir el numerador de (16.13) como

$$\frac{\text{SCE(reducido)} - \text{SCE(completo)}}{\text{número de términos extra}} \quad \textbf{(16.14)}$$

Observe que "el número de términos extra" denota la diferencia entre el número de variables independientes en el modelo completo y el número de variables independientes en el modelo reducido. El denominador de la ecuación (16.13) es la suma de cuadrados debido al error en el modelo completo dividida entre los grados de libertad correspondientes; es decir, el denominador es el cuadrado medio debido al error del modelo completo. Al denotar éste como CME(completo), permite escribirlo como

$$F = \frac{\dfrac{\text{SCE(reducido)} - \text{SCE(completo)}}{\text{número de términos extra}}}{\text{CME(completo)}} \quad \textbf{(16.15)}$$

Para ilustrar el uso de este estadístico F, suponga que se tiene un problema de regresión con 30 observaciones. Un modelo con las variables independientes x_1, x_2 y x_3 tiene un error en la suma de cuadrados de 150, y el segundo modelo con las variables independientes x_1, x_2, x_3, x_4 y x_5 tiene un error en la suma de cuadrados de 100. ¿La adición de las dos variables independientes x_4 y x_5 dio como resultado una reducción significativa en la suma de cuadrados debido al error?

Observe primero que los grados de libertad para STC es $30 - 1 = 29$, y que los grados de libertad para la suma de cuadrados debido a la regresión para el modelo completo es 5 (el

número de variables independientes en el modelo completo). Por tanto, los grados de libertad para la suma de cuadrados debido al error en el modelo completo es 29 − 5 = 24 y, por tanto, CME(completo) = 100/24 = 4.17. Así que el estadístico F es

$$F = \frac{\dfrac{150 - 100}{2}}{4.17} = 6.00$$

Este valor F calculado se compara con el valor F que se encuentra en la tabla para 2 grados de libertad en el numerador y 24 grados de libertad en el denominador. En el nivel 0.05 de significancia, la tabla 4 del apéndice B muestra que $F_{0.05} = 3.40$. Debido a que $F = 6.00$ es mayor que 3.40, se concluye que la adición de las variables x_4 y x_5 es estadísticamente significativa.

Uso de los valores-*p*

El criterio del valor-p también puede utilizarse para determinar si tiene ventajas agregar una o más variables independientes al modelo de regresión múltiple. En el ejemplo anterior se mostró cómo llevar a cabo la prueba F para determinar si la adición de las dos variables independientes x_4 y x_5 a un modelo con tres variables independientes, x_1, x_2 y x_3, era estadísticamente significativa. En este ejemplo, el valor que se obtuvo para el estadístico F fue 6.00, y se concluyó (al comparar el valor $F = 6.00$ con el valor crítico $F_{0.05} = 3.40$) que la adición de las variables x_4 y x_5 fue significativa. Al usar Minitab o Excel, el valor-p que corresponde a $F = 6.00$ (2 grados de libertad en el numerador y 24 grados de libertad en el denominador) es 0.008. Como el valor-$p = 0.008 < \alpha = 0.05$, se concluye también que la adición de las dos variables independientes es estadísticamente significativa. Es difícil determinar el valor-p directamente en las tablas de la distribución F, pero el software como Minitab o Excel facilita su cálculo.

NOTAS Y COMENTARIOS

El cálculo del estadístico F también se basa en la diferencia de la suma de cuadrados de la regresión. Para mostrar esta forma del estadístico F, primero se observa que

$$\text{SCE(reducido)} = \text{STC} - \text{SCR(reducido)}$$
$$\text{SCE(completo)} = \text{STC} - \text{SCR(completo)}$$

Por tanto

$$\begin{aligned}\text{SCE(reducido)} - \text{SCE(completo)} &= [\text{STC} - \text{SCR(reducido)}] - [\text{STC} - \text{SCR(completo)}] \\ &= \text{SCR(completo)} - \text{SCR(reducido)}\end{aligned}$$

Así que,

$$F = \frac{\dfrac{\text{SCR(completo)} - \text{SCR(reducido)}}{\text{número de términos extra}}}{\text{CME(completo)}}$$

Ejercicios

Métodos

10. En un análisis de regresión con 27 observaciones, se obtuvo la siguiente ecuación de regresión estimada.

$$\hat{y} = 25.2 + 5.5x_1$$

Para esta ecuación de regresión estimada STC = 1 550 y SCE = 520.

a) Con $\alpha = 0.05$, pruebe si x_1 es significativa.
Suponga que se agregan a este modelo las variables x_2 y x_3 y que se obtiene la siguiente ecuación de regresión.

$$\hat{y} = 16.3 + 2.3x_1 + 12.1x_2 - 5.8x_3$$

Para esta ecuación de regresión estimada, STC = 1 550 y SCE = 100.

b) Utilice una prueba F y un nivel de significancia de 0.05 para determinar si x_2 y x_3 contribuyen significativamente al modelo.

AUTOevaluación

11. En un análisis de regresión con 30 observaciones se obtuvo la siguiente ecuación de regresión estimada.

$$\hat{y} = 17.6 + 3.8x_1 - 2.3x_2 + 7.6x_3 + 2.7x_4$$

Para esta ecuación de regresión estimada, STC = 1 805 y SCR = 1 760.

a) Con $\alpha = 0.05$, pruebe la significancia de la relación entre las variables.
Suponga que las variables x_1 y x_4 se retiran del modelo y se obtiene la siguiente ecuación de regresión estimada.

$$\hat{y} = 11.1 - 3.6x_2 + 8.1x_3$$

Para este modelo, STC = 1 805 y SCR = 1 705.

b) Calcule SCE(x_1, x_2, x_3, x_4).

c) Calcule SCE(x_2, x_3).

d) Utilice la prueba F y 0.05 como nivel de significancia para determinar si x_1 y x_2 contribuyen significativamente al modelo.

Aplicaciones

12. La Ladies Professional Golfers Association (LPGA) mantiene estadísticas sobre el desempeño y las ganancias de los miembros de la LPGA Tour. Las estadísticas de fin de año sobre el desempeño de las 30 jugadoras que obtuvieron las mayores ganancias en los eventos de LPGA Tour de 2005 aparecen en el archivo LPGA Tour (sitio web de LPGA Tour, 2006). Earnings ($1 000) son los ingresos totales en miles de dólares; Scoring Avg. es la puntuación promedio de una jugadora en todos los eventos; Green in Reg. es el porcentaje de las veces en que la jugadora llega al *green* en regulación; Putting Avg. es el promedio de *putts* realizados dentro del *green* en regulación, y Sand Saves es el porcentaje de veces que una jugadora es capaz de conseguir "subir y bajar" cuando se encuentra en la trampa de arena al lado del *green*. Un *green* en regulación es considerado un golpe en cualquier parte de la bola que esté en contacto con la superficie del *putting*, y la diferencia entre el valor del par de hoyos y el número de golpes en el *green* es por lo menos de un par.

WEB archivo
LPGATour

a) Desarrolle una ecuación de regresión estimada para pronosticar la Scoring Avg. de todos los eventos dado el número promedio de *putts* en los golpes dados en Green in Reg.

b) Desarrolle una ecuación de regresión estimada para pronosticar la Scoring Avg. de todos los eventos dado el tiempo promedio en que una jugadora es capaz de golpear el Green in Reg, y el promedio de veces en que consigue "subir y bajar" una vez que se encuentra en la trampa de arena.

c) Con un el nivel de significancia de 0.05, pruebe si las dos variables independientes agregadas en el inciso *b*), el porcentaje de veces en que una jugadora consigue llegar al *green* en regulación y el promedio de veces en que es capaz de "subir y bajar" una vez que se encuentra en la trampa de arena al lado del *green*, contribuyen significativamente el desarrollo de la ecuación de regresión en el inciso *a*). Explique.

13. Vaya al ejercicio 12.

WEB archivo
LPGATour

a) Desarrolle una ecuación de regresión estimada para predecir las ganancias en todos los eventos, dado el número de Putting Avg. en los *green* en regulación.

b) Desarrolle una ecuación de regresión estimada para predecir el total de ganancias en todos los eventos dado el porcentaje de veces en que una jugadora es capaz de llegar al *green* en regulación, la puntuación Putting Avg. lograda en el *green* en regulación y el porcentaje de veces en que la jugadora es capaz de conseguir "subir y bajar" una vez que se encuentra en la trampa de arena al lado del *green*.

c) Con un nivel de 0.05 de significancia, pruebe si las dos variables independientes agregadas en el inciso *b*), el promedio de veces en que la jugadora es capaz de llegar el *green* en regulación y el porcentaje de veces en que consigue "subir y bajar" una vez que está en la trampa de arena al lado del *green*, contribuyen significativamente al desarrollo de la ecuación de regresión estimada en el inciso *a*). Explique.

s) En general, las puntuaciones más bajas llevan a ganancias más grandes. Para investigar esta opción que permita predecir las ganancias totales, desarrolle una ecuación de regresión estimada para pronosticar las ganancias totales en todos los eventos dada la puntuación promedio. ¿Prefiere utilizar esta ecuación o la ecuación desarrollada en el inciso *b*). Explique.

14. Un estudio realizado a lo largo de 10 años por la American Heart Association proporciona datos sobre cómo la edad (Age), la presión arterial (Blood Pressure) y el tabaquismo se relacionan con el riesgo (Risk) de sufrir infartos. Los siguientes son los datos de una parte de este estudio. El riesgo se interpreta como la probabilidad (multiplicada por 100) de que el paciente sufra un infarto en los próximos 10 años. Para la variable fumador (Smoker), 1 indica a un fumador y el 0 a un no fumador.

WEB archivo

Stroke

Risk	Age	Blood Pressure	Smoker
12	57	152	0
24	67	163	0
13	58	155	0
56	86	177	1
28	59	196	0
51	76	189	1
18	56	155	1
31	78	120	0
37	80	135	1
15	78	98	0
22	71	152	0
36	70	173	1
15	67	135	1
48	77	209	1
15	60	199	0
36	82	119	1
8	66	166	0
34	80	125	1
3	62	117	0
37	59	207	1

a) Desarrolle una ecuación de regresión estimada para predecir el riesgo de fumar dada la edad y el nivel de presión sanguínea.

b) Considere la adición de dos variables independientes al modelo desarrollado en el inciso *a*): una para la interacción entre la edad y el nivel de presión arterial y otra que indique si la persona es fumadora. Desarrolle una ecuación de regresión estimada utilizando estas cuatro variables independientes.

c) Con un nivel de 0.05 de significancia, lleve a cabo una prueba para determinar si la adición del término *interacción* y la variable *fumador* contribuyen significativamente a la ecuación de regresión estimada desarrollada en el inciso *a*).

15. En el beisbol, una carrera limpia es cualquiera en la que el lanzador anota al equipo contrario, con excepción de las carreras anotadas como consecuencia de errores. El promedio de carreras anotadas (ERA), la estadística que con mayor frecuencia se utiliza para comparar la efectividad de los lanzadores, se calcula de la siguiente manera.

$$\text{ERA} = \left(\frac{\text{carreras limpias}}{\text{entradas lanzadas}}\right)9$$

WEB archivo
MLBPitching

Observe que el número promedio de carreras limpias por cada entrada se multiplica por 9, el número de entradas en un juego reglamentario. Por tanto, ERA representa el número promedio que el lanzador cede por nueve entradas. Por ejemplo, en 2008 Roy Halladay, un lanzador de los Azulejos de Toronto, lanzó 246 entradas y cedió 76 carreras ganadas; su ERA fue (76/246)9 = 2.78. Para investigar la relación entre el ERA y otras medidas de efectividad en el lanzamiento, los datos de 50 lanzadores en la Major League Baseball durante la temporada 2008 aparecen en el conjunto de datos llamado MLBPitching (página web de la MLB, febrero de 2009). Las descripciones de las variables aparecen en el siguiente conjunto de datos.

W	Número de juegos ganados
L	Número de juegos perdidos
WPCT	Porcentaje de juegos ganados
H/9	Número de golpes cedidos por nueve entradas
HR/9	Número de jonrones cedidos por nueve entradas
BB/9	Número promedio de bases por bola cedidos por nueve entradas.

a) Desarrolle una ecuación de regresión estimada para predecir el promedio de carreras ganadas dado el número promedio de golpes cedidos por nueve entradas.
b) Desarrolle una ecuación de regresión estimada para predecir el promedio de carreras ganadas dados H/9, HR/9 y BB/9.
c) Con un nivel de 0.05 de significancia, pruebe si las dos variables independientes añadidas en el inciso *b*), HR/9 y BB/9, contribuyen significativamente a la ecuación de regresión estimada del inciso *a*).

16.3 Análisis de un problema mayor

Cuando se presentó el análisis de regresión múltiple se utilizó ampliamente el ejemplo de Butler Trucking. El tamaño pequeño de este problema era una ventaja en la exploración de los conceptos introductorios; sin embargo, dificultan ilustrar algunas cuestiones relacionadas con la selección de variables que participan en la construcción de un modelo. Para dar un ejemplo de los procesos de selección de variables que se estudiarán en la próxima sección, se introduce un conjunto de datos que consta de 25 observaciones con ocho variables independientes. El Dr. David W. Cravens, del Departamento de Marketing de Texas Christian University, autorizó el uso de este conjunto de datos. En consecuencia, se le denominará los *datos de Cravens*.[1]

Los datos de Cravens corresponden a una empresa que vende sus productos en varios territorios, cada uno de los cuales está asignado a un representante de ventas. Se realizó el análisis de regresión para determinar si diversas variables (independientes) predictoras podrían explicar las ventas en cada territorio. Los datos de una muestra aleatoria de 25 territorios de ventas se listan en la tabla 16.5; la definición de las variables se presenta en la tabla 16.6.

[1] Para más detalles, véase David W. Cravens, Robert B. Woodruff y Joe C. Stamper, "An analytical approach for evaluating sales territory performance", *Journal of Marketing*, 36 (enero de 1972), pp. 31-37. Copyright © 1972 American Marketing Association.

TABLA 16.5 Datos de Cravens

WEB archivo
Cravens

Sales	Time	Poten	AdvExp	Share	Change	Accounts	Work	Rating
3669.88	43.10	74065.1	4582.9	2.51	0.34	74.86	15.05	4.9
3473.95	108.13	58117.3	5539.8	5.51	0.15	107.32	19.97	5.1
2295.10	13.82	21118.5	2950.4	10.91	−0.72	96.75	17.34	2.9
4675.56	186.18	68521.3	2243.1	8.27	0.17	195.12	13.40	3.4
6125.96	161.79	57805.1	7747.1	9.15	0.50	180.44	17.64	4.6
2134.94	8.94	37806.9	402.4	5.51	0.15	104.88	16.22	4.5
5031.66	365.04	50935.3	3140.6	8.54	0.55	256.10	18.80	4.6
3367.45	220.32	35602.1	2086.2	7.07	−0.49	126.83	19.86	2.3
6519.45	127.64	46176.8	8846.2	12.54	1.24	203.25	17.42	4.9
4876.37	105.69	42053.2	5673.1	8.85	0.31	119.51	21.41	2.8
2468.27	57.72	36829.7	2761.8	5.38	0.37	116.26	16.32	3.1
2533.31	23.58	33612.7	1991.8	5.43	−0.65	142.28	14.51	4.2
2408.11	13.82	21412.8	1971.5	8.48	0.64	89.43	19.35	4.3
2337.38	13.82	20416.9	1737.4	7.80	1.01	84.55	20.02	4.2
4586.95	86.99	36272.0	10694.2	10.34	0.11	119.51	15.26	5.5
2729.24	165.85	23093.3	8618.6	5.15	0.04	80.49	15.87	3.6
3289.40	116.26	26878.6	7747.9	6.64	0.68	136.58	7.81	3.4
2800.78	42.28	39572.0	4565.8	5.45	0.66	78.86	16.00	4.2
3264.20	52.84	51866.1	6022.7	6.31	−0.10	136.58	17.44	3.6
3453.62	165.04	58749.8	3721.1	6.35	−0.03	138.21	17.98	3.1
1741.45	10.57	23990.8	861.0	7.37	−1.63	75.61	20.99	1.6
2035.75	13.82	25694.9	3571.5	8.39	−0.43	102.44	21.66	3.4
1578.00	8.13	23736.3	2845.5	5.15	0.04	76.42	21.46	2.7
4167.44	58.44	34314.3	5060.1	12.88	0.22	136.58	24.78	2.8
2799.97	21.14	22809.5	3552.0	9.14	−0.74	88.62	24.96	3.9

Como paso previo, se considerarán los coeficientes de correlación muestral entre cada par de variables. La figura 16.13 es la matriz de correlación obtenida con Minitab. Observe que el coeficiente de correlación muestral entre Sales y Time es 0.623, entre Sales y Poten es 0.598, y así sucesivamente.

En cuanto a los coeficientes de correlación entre las variables independientes, observará que la correlación entre Time y Accounts es de 0.758, por lo que, si Accounts se utiliza como

TABLA 16.6 Definiciones de las variables para los datos de Cravens

Variable	Definición
(Sales) ventas	Total de ventas acreditadas al representante de ventas
(Time) antigüedad	Antigüedad del empleado en meses
(Poten) potencial	Mercado potencial; ventas industriales totales en unidades en el territorio de ventas*
(AdvExp) gastos en publicidad	Gastos de publicidad en el territorio de ventas
(Share) participación	Participación de mercado; promedio ponderado de los últimos cuatro años
(Change) cambio	Cambio en la participación de mercado durante los últimos cuatro años
(Accounts) cuentas	Cuentas asignadas a los representantes de ventas*
(Work) trabajo	Carga de trabajo; índice ponderado basado en compras anuales y concentraciones de cuentas
(Raiting) clasificación	Evaluación general del representante de ventas en ocho dimensiones de desempeño; evaluación agregada sobre una escala de 1-7

* Estos datos fueron codificados para preservar la confidencialidad.

FIGURA 16.13 Coeficientes de correlación muestral de los datos de Cravens

	Sales	Time	Poten	AdvExp	Share	Change	Accounts	Work
Time	0.623							
Poten	0.598	0.454						
AdvExp	0.596	0.249	0.174					
Share	0.484	0.106	-0.211	0.264				
Change	0.489	0.251	0.268	0.377	0.085			
Accounts	0.754	0.758	0.479	0.200	0.403	0.327		
Work	-0.117	-0.179	-0.259	-0.272	0.349	-0.288	-0.199	
Rating	0.402	0.101	0.359	0.411	-0.024	0.549	0.229	-0.277

variable independiente, Time no agregaría mucho poder explicatorio al modelo. Recuerde la prueba de la regla práctica que se estudió en la sección 15.4, donde se vio que la multicolinealidad puede causar problemas si el valor absoluto del coeficiente de correlación muestral excede 0.7 en cualquiera de las dos variables independientes. Por tanto, debe evitarse incluir tanto Time y Accounts en el mismo modelo de regresión. También el coeficiente de correlación muestral entre Change y Rating, que es 0.549, es alto y puede justificar consideraciones más cuidadosas.

En cuanto a los coeficientes de correlación muestral entre Sales y cada una de las variables independientes, puede dar una rápida indicación de que estas variables son, por sí mismas, buenos predictores. Observe que el mejor predictor de Sales es Accounts, debido a que su coeficiente de correlación muestral es más alto (0.754). Recuerde que para el caso de una variable independiente, el cuadrado del coeficiente de correlación muestral es el coeficiente de determinación. Por tanto, Accounts puede explicar $(0.754)^2(100)$, o 56.85%, de la variabilidad en Sales. Las siguientes variables independientes más importantes son Time, Poten y AdvExp, cada una con un coeficiente de correlación muestral de aproximadamente 0.6.

Aunque existen problemas potenciales de multicolinealidad, se considerará el desarrollo de una ecuación de regresión estimada con las ocho variables independientes. Minitab proporciona los resultados de la figura 16.14. El modelo de regresión múltiple de ocho variables tiene un coeficiente de determinación ajustado de 88.3%. Observe, sin embargo, que los valores-p para la pruebas t de cada uno de los parámetros muestran que sólo Poten, AdvExp y Share son significativos a un nivel de $\alpha = 0.05$, dado el efecto de todas las demás variables. Por tanto, se deseará investigar los resultados que se obtienen si se utilizan sólo estas tres variables. En la figura 16.15 se presentan los resultados proporcionados por Minitab para la ecuación de regresión estimada con estas tres variables. Observe que el coeficiente de determinación ajustado para esta ecuación es 82.7% el cual, aunque no es tan bueno como el de la ecuación de regresión estimada con ocho variables, es alto.

¿Cómo se puede encontrar la ecuación que haga el mejor trabajo dada la información de que se dispone? Un método consiste en calcular todas las regresiones posibles. Es decir, se podrían desarrollar ocho ecuaciones de regresión estimada con una sola variable (cada una de las cuales corresponde a una de las variables independientes), 28 ecuaciones con dos variables (es el número de combinaciones de ocho variables tomadas de dos en dos), y así sucesivamente. En total, para los datos de Cravens, serían 225 diferentes ecuaciones de regresión estimada conteniendo una o más variables independientes que tendrían que ajustarse a los datos.

Con el excelente software disponible en la actualidad es posible calcular todas las regresiones. Sin embargo, implica una gran cantidad de cálculos y revisar abundantes resultados de computadora, muchos de los cuales corresponden a modelos obviamente pobres. Se prefiere un método más sistemático para seleccionar el subconjunto de variables independientes que proporcione la mejor ecuación de regresión estimada. En la siguiente sección se presentan algunos de los métodos más conocidos.

FIGURA 16.14 Resultados de Minitab para el modelo con ocho variables independientes

```
The regression equation is
Sales = - 1508 + 2.01 Time + 0.0372 Poten + 0.151 AdvExp + 199 Share
        + 291 Change + 5.55 Accounts + 19.8 Work + 8 Rating

Predictor      Coef   SE Coef      T      p
Constant    -1507.8     778.6  -1.94  0.071
Time          2.010     1.931   1.04  0.313
Poten      0.037206  0.008202   4.54  0.000
AdvExp      0.15094   0.04711   3.21  0.006
Share        199.08     67.03   2.97  0.009
Change        290.9     186.8   1.56  0.139
Accounts      5.550     4.775   1.16  0.262
Work          19.79     33.68   0.59  0.565
Rating          8.2     128.5   0.06  0.950

S = 449.015   R-sq = 92.2%   R-sq(adj) = 88.3%

Analysis of Variance

SOURCE          DF        SS       MS      F      p
Regression       8  38153712  4769214  23.66  0.000
Residual Error  16   3225837   201615
Total           24  41379549
```

FIGURA 16.15 Resultados de Minitab para el modelo con las variables Poten, AdvExp y Share

```
The regression equation is
Sales = - 1604 + 0.0543 Poten + 0.167 AdvExp + 283 Share

Predictor      Coef   SE Coef      T      p
Constant    -1603.6     505.6  -3.17  0.005
Poten      0.054286  0.007474   7.26  0.000
AdvExp      0.16748   0.04427   3.78  0.001
Share        282.75     48.76   5.80  0.000

S = 545.515   R-sq = 84.9%   R-sq(adj) = 82.7%

Analysis of Variance

SOURCE          DF        SS         MS       F      p
Regression       3  35130228   11710076   39.35  0.000
Residual Error  21   6249321     297587
Total           24  41379549
```

16.4 Procedimientos de selección de variables

Los procedimientos de selección de variables son particularmente útiles en las primeras etapas de construcción de un modelo, pero no pueden sustituir la experiencia y el juicio del analista.

En esta sección se verán cuatro **procedimientos de selección de variables**: la regresión por pasos, la selección hacia adelante, la eliminación hacia atrás y la regresión de los mejores subconjuntos. Dado un conjunto de datos con diversas variables independientes posibles, estos procedimientos se pueden utilizar para identificar cuáles variables proporcionan el mejor modelo. Los tres primeros procedimientos son iterativos; en cada paso se agrega o se elimina una sola variable independiente y se evalúa el nuevo modelo. El procedimiento continúa hasta que un criterio de detención indica que no se puede encontrar un modelo mejor. El último procedimiento (mejores subconjuntos) no evalúa las variables de una en una, sino los modelos de regresión con diferentes subconjuntos de variables independientes.

En los procedimientos de regresión por pasos, de selección hacia adelante y de eliminación hacia atrás, el criterio de selección de una variable independiente para agregarla o eliminarla del modelo, en cada paso, se basa en el estadístico F presentado en la sección 16.2. Suponga que, por ejemplo, se está considerando si agregar x_2 a un modelo en el que interviene x_1, o eliminar x_2 de un modelo en el que participan x_1 y x_2. Para probar si la adición o la eliminación de x_2 es estadísticamente significativa, las hipótesis nula y alternativa se pueden plantear como sigue:

$$H_0: \beta_2 = 0$$
$$H_a: \beta_2 \neq 0$$

En la sección 16.2 [vea la ecuación (16.10)] se mostró que

$$F = \frac{\dfrac{\text{SCE}(x_1) - \text{SCE}(x_1, x_2)}{1}}{\dfrac{\text{SCE}(x_1, x_2)}{n - p - 1}}$$

se puede utilizar como criterio para determinar si la presencia de x_2 en el modelo causa una reducción significativa en la suma de cuadrados debido al error. El valor-p correspondiente a este estadístico F es el criterio utilizado para determinar si una variable independiente debe agregarse o eliminarse del modelo de regresión. La regla de rechazo se aplica siempre: rechazar H_0 si el valor-$p \leq \alpha$.

Regresión por pasos

El procedimiento de regresión por pasos comienza por determinar en cada paso si alguna de las variables *que ya se encuentran en el modelo* debe ser eliminada. Primero se calcula el estadístico F y un valor-p correspondiente para cada una de las variables independientes del modelo. El nivel de significancia a para determinar si una variable independiente debe ser eliminada del modelo es referido en Minitab como *Alpha to remove* (alfa para eliminar). Si el valor-p de cualquier variable independiente es mayor que *Alpha to remove*, la variable independiente que tenga el mayor valor-p se elimina del modelo y el proceso de regresión por pasos inicia un nuevo paso.

Si ninguna de las variables independientes puede ser eliminada del modelo, el procedimiento trata de introducir otra variable independiente. Primero se calcula el estadístico F y el valor-p correspondiente de cada variable que no está en el modelo. Al nivel de significancia a para determinar si una variable independiente debe agregarse al modelo se le conoce como *Alpha to enter* (alfa para ingresar) en Minitab. La variable independiente con el menor valor-p es ingresada al modelo siempre que el valor-p sea menor o igual a *Alpha to enter*. El procedimiento continúa de esta manera hasta que no haya variables independientes que puedan ser eliminadas o agregadas al modelo.

En la figura 16.16 se muestran los resultados obtenidos por Minitab con el procedimiento de regresión por pasos para los datos de Cravens con 0.05 como *Alpha to remove* y 0.05 como

FIGURA 16.16 Resultados de Minitab mediante la regresión por pasos para los datos de Cravens

```
Alpha-to-Enter: 0.05     Alpha-to-Remove: 0.05

Response is Sales on 8 predictors, with N = 25

    Step        1       2        3         4
Constant   709.32   50.29   -327.24   -1441.93

Accounts     21.7    19.0      15.6        9.2
T-Value      5.50    6.41      5.19       3.22
P-Value     0.000   0.000     0.000      0.004

AdvExp               0.227     0.216      0.175
T-Value               4.50      4.77       4.74
P-Value              0.000     0.000      0.000

Poten                        0.0219     0.0382
T-Value                        2.53       4.79
P-Value                       0.019      0.000

Share                                     190
T-Value                                  3.82
P-Value                                 0.001

S             881     650       583        454
R-Sq        56.85   77.51     82.77      90.04
R-Sq(adj)   54.97   75.47     80.31      88.05
C-p          67.6    27.2      18.4        5.4
```

Alpha to enter. Este procedimiento terminó en cuatro pasos. La ecuación de regresión estimada obtenida con el procedimiento de regresión por pasos de Minitab es:

$$\hat{y} = -1\,441.93 + 9.2\,\text{Accounts} + 0.175\,\text{AdvExp} + 0.0382\,\text{Poten} + 190\,\text{Share}$$

Debido a que el procedimiento por pasos no considera cada subconjunto posible de un determinado número de variables independientes, éste no necesariamente selecciona la ecuación de regresión estimada con el mayor valor R-sq.

En la figura 16.16 observe también que $s = \sqrt{\text{CME}}$ se ha reducido de 881 en el mejor modelo con una variable (utilizando Account) a 454. El valor de R-sq se incrementó de 56.85% a 90.04%, y el R-sq(adj) de la ecuación de regresión estimada se recomienda que sea de 88.05%.

En resumen, en cada paso del procedimiento de regresión por pasos lo primero que se considera es si alguna de las variables independientes puede eliminarse del modelo actual. Si ninguna puede descartarse, el procedimiento verifica si cualquier otra que no está actualmente en el modelo puede ser ingresada. Debido a la naturaleza del procedimiento de regresión por pasos, una variable independiente se ingresa en el modelo en un solo paso, es eliminada en un paso subsiguiente, y después se introduce al modelo en un paso posterior. El procedimiento se detiene cuando ninguna variable independiente puede ser eliminada ni agregada al modelo.

Selección hacia adelante

Este procedimiento se inicia sin variables independientes. Se agrega una variable de una en una utilizando el mismo proceso que en la regresión por pasos para determinar si una variable independiente debe ser ingresada al modelo. Sin embargo, el procedimiento de selección hacia

adelante no permite que una variable sea eliminada del modelo una vez que se ha ingresado, y se detiene cuando el valor-*p* de cada una de las variables independientes que no está en el modelo es mayor que *Alpha to enter*.

La ecuación de regresión estimada obtenida con Minitab mediante el procedimiento de selección hacia adelante es

$$\hat{y} = -1441.93 + 9.2\,\text{Accounts} + 0.175\,\text{AdvExp} + 0.0382\,\text{Poten} + 190\,\text{Share}$$

Por tanto, para los datos de Cravens, el procedimiento de selección hacia adelante (con 0.05 de *Alpha to enter*) lleva a la misma ecuación de regresión estimada que el procedimiento por pasos.

Eliminación hacia atrás

Este procedimiento comienza con un modelo que incluye todas las variables independientes. A continuación elimina cada variable de una en una utilizando el mismo procedimiento de regresión por pasos. Sin embargo, la eliminación hacia atrás no permite que una variable independiente sea introducida nuevamente una vez que ha sido eliminada, y se detiene cuando ninguna de las variables independientes del modelo tiene un valor-*p* superior a *Alpha to remove*.

La ecuación de regresión estimada obtenida con el procedimiento de eliminación hacia atrás de Minitab para los datos de Cravens (con 0.05 para *Alpha to remove*) es

$$\hat{y} = -1\,312 + 3.8\,\text{Time} + 0.0444\,\text{Poten} + 0.152\,\text{AdvExp} + 259\,\text{Share}$$

Al comparar las ecuaciones de regresión estimada obtenidas mediante los procedimientos de eliminación hacia atrás y de selección hacia adelante, se observa que hay tres variables independientes comunes a ambos: AdvExp, Poten y Share. Sin embargo, en la eliminación hacia atrás se incluyó Time en lugar de Accounts.

La selección hacia adelante y la eliminación hacia atrás pueden llevar a modelos diferentes.

La selección hacia adelante y la eliminación hacia atrás son los dos extremos en la construcción de modelos; el primero comienza sin ninguna variable independiente en el modelo y añade una a una las variables independientes, mientras que el segundo comienza con todas las variables independientes en el modelo y las elimina una por una. Los dos procedimientos llevan a la ecuación de regresión estimada. Es posible, sin embargo, que puedan dar lugar a dos diferentes ecuaciones de regresión estimada, como se observó con los datos de Cravens. Decidir cuál utilizar sigue siendo tema de discusión. En última instancia, se debe aplicar el juicio del analista. El procedimiento de los mejores subconjuntos para la construcción de modelos que se estudia a continuación proporciona información adicional para la construcción de modelos a considerar antes de tomar una decisión definitiva.

Regresión de los mejores subconjuntos

La regresión por pasos, la selección hacia adelante y la eliminación hacia atrás son los métodos para elegir el modelo de regresión mediante la adición o eliminación, una por una, de las variables independientes. Ninguno de ellos garantiza que se encuentre el mejor modelo para un número determinado de variables. Por tanto, estos métodos de variables de una por una son bien vistos como heurística para una buena selección del modelo de regresión.

Algunos software utilizan un procedimiento llamado *regresión de los mejores subconjuntos*, que permite al usuario encontrar el mejor modelo de regresión dado un número determinado de variables independientes. Minitab posee dicho procedimiento. La figura 16.17 es parte de los resultados de computadora obtenidos con el procedimiento de los mejores subconjuntos para el conjunto de datos de Craven.

En estos resultados se identifican las dos mejores ecuaciones de regresión estimada con una sola variable, las dos mejores ecuaciones con dos variables, las dos mejores ecuaciones de tres variables, y así sucesivamente. El criterio para determinar cuáles son las mejores ecuaciones

FIGURA 16.17 Parte de los resultados de Minitab obtenidos mediante la regresión de los mejores subconjuntos

Vars	R-sq	R-sq(adj)	s	Time	Poten	AdvExp	Shares	Change	Accounts	Work	Rating
1	56.8	55.0	881.09						X		
1	38.8	36.1	1049.3	X							
2	77.5	75.5	650.39			X			X		
2	74.6	72.3	691.11		X			X			
3	84.9	82.7	545.52		X	X	X				
3	82.8	80.3	582.64		X	X			X		
4	90.0	88.1	453.84		X	X	X		X		
4	89.6	87.5	463.93	X	X	X	X				
5	91.5	89.3	430.21	X	X	X	X	X			
5	91.2	88.9	436.75		X	X	X	X	X		
6	92.0	89.4	427.99	X	X	X	X	X	X		
6	91.6	88.9	438.20		X	X	X	X	X	X	
7	92.2	89.0	435.66	X	X	X	X	X	X	X	
7	92.0	88.8	440.29	X	X	X	X	X	X		X
8	92.2	88.3	449.02	X	X	X	X	X	X	X	X

de regresión estimada para cualquier número de predictores es el valor del coeficiente de determinación (R-sq). Por ejemplo, la mejor ecuación de regresión estimada con una sola variable independiente es Accounts con un R-sq = 56.8%; AdvExp y Accounts, con R-sq = 77.5%, proporcionan la mejor ecuación con dos variables independientes, y Poten, AdvExp y Shares, con R-sq = 84.9%, aportan la mejor ecuación con tres variables independientes. Para los datos de Cravens, el mayor coeficiente de determinación ajustado (Adj. R-sq = 89.4%) es el del modelo con seis variables independientes: Time, Poten, AdvExp, Shares, Change y Accounts. Sin embargo, el mejor modelo con cuatro variables (Poten, AdvExp, Shares y Accounts) tiene un coeficiente de determinación ajustado casi igual de alto (88.1%). Se prefiere, por lo general, el modelo más sencillo con el menor número de variables.

La elección final

El análisis realizado hasta ahora sobre los datos de Cravens es una buena preparación para la elección de un modelo final, pero debe efectuarse un análisis precio a la elección definitiva. Como se señaló en los capítulos 14 y 15, se debe realizar un análisis cuidadoso de los residuales. Se desea que la gráfica de residuales del modelo elegido parezca una banda horizontal. Suponga que en los residuales no se encuentre problema alguno y que se desea usar los resultados del procedimiento de los mejores subconjuntos que ayuden a elegir un modelo.

Este procedimiento muestra que el mejor modelo de cuatro variables es el que contiene las variables independientes Poten, AdvExp, Share y Accounts. Este resulta ser también el modelo con cuatro variables identificadas con el procedimiento de regresión por pasos. La tabla 16.7 es útil en la toma de la decisión final. Ésta muestra varios posibles modelos que contienen todas o algunas de estas cuatro variables independientes.

TABLA 16.7 Modelos seleccionados con Accounts, Poten, AdvExp y Share

Modelo	Variables independientes	Adj. R-sq
1	Accounts	55.0
2	AdvExp, Accounts	75.5
3	Poten, Share	72.3
4	Poten, AdvExp, Accounts	80.3
5	Poten, AdvExp, Share	82.7
6	Poten, AdvExp, Share, Accounts	88.1

En la tabla 16.7 se observa que el modelo que sólo tiene AdvExp y Accounts es bueno. Su coeficiente de determinación ajustado es 75.5%, mientras que el modelo con las cuatro variables sólo proporciona un aumento de 12.6 puntos porcentuales. El modelo más sencillo de dos variables podría ser el que se prefiera si, por ejemplo, es difícil medir el potencial de mercado (Poten). Sin embargo, si se cuenta fácilmente con los datos y se necesita una gran precisión en las predicciones de ventas, es claro que se preferirá el modelo constructor de las cuatro variables.

NOTAS Y COMENTARIOS

1. El procedimiento por pasos requiere que *Alpha to remove* sea mayor o igual que *Alpha to enter*. Este requisito impide que una misma variable sea eliminada y luego introducida en un mismo paso.
2. Las funciones de las variables independientes se pueden utilizar a efecto de crear nuevas variables independientes para usarse con cualquiera de los procedimientos de esta sección. Por ejemplo, si se desea tener x_1x_2 en el modelo para considerar la interacción, se utilizarían los datos de x_1 y x_2 con el fin de crear los datos para $z = x_1x_2$.
3. Ninguno de los procedimientos que agregan o eliminan variables de una en una garantiza la identificación del mejor modelo de regresión, pero estos métodos son excelentes para encontrar buenos modelos, sobre todo cuando poca multicolinealidad está presente.

Ejercicios

Métodos

16. Un estudio proporcionó datos de las variables que pueden estar relacionadas con el número de semanas en que un trabajador de manufactura ha estado desempleado. La variable dependiente del estudio (semanas) se define como el número de éstas en que un trabajador ha estado sin empleo debido a un despido. Las siguientes variables independientes se utilizaron en el análisis.

WEB archivo
Layoffs

Age (*edad*)	Edad del trabajador
Educ (*educación*)	Número de años de estudio
Married (*casado*)	Una variable ficticia; 1 si está casado, 0 si no es el caso
Head (*cabeza*)	Una variable ficticia; 1 si es cabeza de familia, 0 si no es el caso
Tenure (*ocupación*)	Número de años en el trabajo anterior
Manager (*gerente*)	Una variable ficticia; 1 si la ocupación es gerente, 0 si no es el caso
Sales (*ventas*)	Una variable ficticia; 1 si la ocupación es en ventas, 0 si no es el caso

Los datos están disponibles en el archivo llamado *Layoffs*.

a) Desarrolle la mejor ecuación de regresión estimada con una variable.

b) Utilice el procedimiento por pasos para desarrollar la mejor ecuación de regresión estimada. Utilice los valores de 0.05 para *Alpha to enter* y *Alpha to remove*.

c) Con el procedimiento de selección hacia adelante desarrolle la mejor ecuación de regresión estimada. Utilice el valor de 0.05 para *Alpha to enter*.

d) Con el procedimiento de eliminación hacia atrás desarrolle la mejor ecuación de regresión estimada. Utilice el valor de 0.05 para *Alpha to remove*.

e) Utilice el procedimiento de regresión de los mejores subconjuntos para obtener la mejor ecuación de regresión estimada.

WEB archivo
LPGATour2

17. The Ladies Professional Golfers Association (LPGA) lleva estadísticas sobre el desempeño y las ganancias de los miembros del LPGA Tour. Las estadísticas de fin de año sobre el papel de las 30 jugadoras que obtuvieron las mejores ganancias totales en la LPGA Tour de 2005 aparecen en el archivo titulado LPGATour2 (sitio web de LPGATour, 2006). Earnings (ganancias) constituyen el resultado total en miles de dólares en todos los eventos de la gira; Scoring Avg. es la puntuación promedio para todos los eventos; Drive Average es la distancia promedio en yardas alcanzada en el *drive* por la jugadora; Greens in Reg. es el porcentaje de veces que la golfista llega al *green* en regulación; Putting Avg. es el promedio de *putts* en el *green* en regulación, y Sand Saves es el porcentaje de veces que una jugadora es capaz de logra "subir y bajar" (*up and down*) cuando se encuentra en la trampa de arena al lado del *green*. Éste se considera un golpe en la regulación si alguna parte de la bola toca la superficie del *putting* y la diferencia entre el valor del par de hoyos y el número de golpes que lleva a golpear el *green* es por lo menos de 2. DriveGreens denota una nueva variable independiente que representa la interacción entre la distancia media alcanzada en el *drive* por la jugadora y el porcentaje de veces que es capaz de alcanzar el *green* en regulación. Utilice los métodos de esta sección a efecto de desarrollar la mejor ecuación de regresión múltiple estimada para calcular el Scoring Avg. de una jugadora en todos los eventos.

18. Jeff Sagarin ha proporcionando, desde 1985, evaluaciones deportivas en *USA Today*. En el beisbol sus pronósticos estadísticos RPG (runs/game) toman en cuenta las estadísticas de ofensiva del jugador, y se afirma que es la mejor medida del verdadero valor de este indicador. Los datos siguientes muestran el RPG y una gran variedad de estadísticas de ofensiva para la temporada 2005 de la Major League Baseball para 20 miembros de los Yankees de Nueva York (sitio web de USA Today, 3 de marzo de 2006). Los encabezados de columna se definen como sigue: RPG, estadístico que predice el número de carreras por juego; H, batazos; 2B, dobles; 3B, triples; HR, jonrones; RBI, carreras impulsadas; BB, bases por bolas; SO, ponches; SB, bases robadas; CS, atrapado en robo de base; OBP, en bases porcentuales; SLG, porcentaje de potencia, y AVG, promedio de bateo.

WEB archivo
Yankees

Player	RPG	H	2B	3B	HR	RBI	BB	SO	SB	CS	OBP	SLG	AVG
D Jeter	6.51	202	25	5	19	70	77	117	14	5	0.389	0.45	0.309
H Matsui	6.32	192	45	3	23	116	63	78	2	2	0.367	0.496	0.305
A Rodriguez	9.06	194	29	1	48	130	91	139	21	6	0.421	0.61	0.321
G Sheffield	6.93	170	27	0	34	123	78	76	10	2	0.379	0.512	0.291
R Cano	5.01	155	34	4	14	62	16	68	1	3	0.32	0.458	0.297
B Williams	4.14	121	19	1	12	64	53	75	1	2	0.321	0.367	0.249
J Posada	5.36	124	23	0	19	71	66	94	1	0	0.352	0.43	0.262
J Giambi	9.11	113	14	0	32	87	108	109	0	0	0.44	0.535	0.271
T Womack	2.91	82	8	1	0	15	12	49	27	5	0.276	0.28	0.249
T Martinez	5.08	73	9	0	17	49	38	54	2	0	0.328	0.439	0.241
M Bellhorn	4.07	63	20	0	8	30	52	112	3	0	0.324	0.357	0.21
R Sierra	3.27	39	12	0	4	29	9	41	0	0	0.265	0.371	0.229
J Flaherty	1.83	21	5	0	2	11	6	26	0	0	0.206	0.252	0.165
B Crosby	3.48	27	0	1	1	6	4	14	4	1	0.304	0.327	0.276
M Lawton	5.15	6	0	0	2	4	7	8	1	0	0.263	0.25	0.125
R Sanchez	3.36	12	1	0	0	2	2	3	0	1	0.326	0.302	0.279
A Phillips	2.13	6	4	0	1	4	1	13	0	0	0.171	0.325	0.15
M Cabrera	1.19	4	0	0	0	0	0	2	0	0	0.211	0.211	0.211
R Johnson	3.44	4	2	0	0	0	1	4	0	0	0.3	0.333	0.222
F Escalona	5.31	4	1	0	0	2	1	4	0	0	0.375	0.357	0.286

Considere el estadístico RPG como la variable dependiente.

a) Desarrolle la mejor ecuación de regresión estimada de una variable.

b) Utilice los métodos de esta sección para desarrollar la mejor ecuación de regresión múltiple estimada para calcular el RPG de un jugador.

WEB archivo
Stroke

19. Consulte el ejercicio 14. Utilice edad, presión sanguínea, si la persona es o no fumadora, y cualquier interacción que involucre estas variables; obtenga una ecuación de regresión estimada que sirva para predecir el riesgo. Describa brevemente el proceso que elija para obtener una ecuación de regresión estimada para estos datos.

16.5 Método de regresión múltiple para el diseño de experimentos

En la sección 15.7 se abordó el uso de variables ficticias en el análisis de regresión múltiple. En esta sección se muestra cómo en una ecuación de regresión múltiple estas variables pueden proporcionar otro método para resolver problemas de diseño experimental o diseño de experimentos. Se demostrará el método de regresión múltiple para el diseño experimental aplicado al ejemplo de Chemitech Inc., el diseño completamente aleatorizado presentado en el capítulo 13.

Recuerde que Chemitech desarrolló un nuevo sistema de filtro para el suministro público de agua. La empresa comprará los componentes para este sistema a varios proveedores y los ensamblará en su fábrica de Colombia, Carolina del Sur. Se tienen tres métodos de ensamble, identificados como A, B y C. Los gerentes de Chemitech desean determinar cuál puede producir el mayor número de sistemas de filtro por semana.

Se seleccionó una muestra aleatoria de 15 empleados, y cada uno de los tres métodos fue asignado al azar a 5 sujetos. El número de unidades ensambladas por empleado se muestra en la tabla 16.8. Las medias muestrales del número de unidades ensambladas con cada uno de los tres métodos es el siguiente.

Métodos de ensamble	Número medio producido
A	62
B	66
C	52

Aunque B parece tener una tasa de producción superior a la de los otros métodos, la cuestión es determinar si las tres medias muestrales observadas son lo suficientemente distintas como para concluir que las medias poblacionales correspondientes a los tres métodos de ensamble son diferentes.

Al aplicar el método de regresión a este problema empezamos definiendo las variables ficticias que se utilizan para indicar cuál de los métodos de ensamble fue utilizado. Debido a que

TABLA 16.8 Número de unidades producidas por los 15 trabajadores

Método		
A	**B**	**C**
58	58	48
64	69	57
55	71	59
66	64	47
67	68	49

TABLA 16.9 Variables ficticias para el experimento de Chemitech

A	B	
1	0	Observación relacionada con el método de ensamble A
0	1	Observación relacionada con el método de ensamble B
0	0	Observación relacionada con el método de ensamble C

en el caso de Chemitech existen tres métodos o tratamientos, se necesitan dos variables ficticias. En general, si el factor que se investigará implica distintos niveles o tratamientos k, es necesario definir $k - 1$ variables ficticias. Para el experimento de Chemitech se determinan las variables ficticias A y B como se muestran en la tabla 16.9.

Se pueden manejar las variables ficticias para relacionar el número de unidades y producidas por semana con el método de ensamble utilizado por el empleado.

$$\begin{aligned} E(y) &= \text{valor esperado del número de unidades producidas por semana} \\ &= \beta_0 + \beta_1 \text{A} + \beta_2 \text{B} \end{aligned}$$

Por tanto, si nos interesa el valor esperado del número de unidades ensambladas a la semana por un empleado que utiliza el método C, el procedimiento de asignación de valores numéricos a las variables ficticias se traduciría en A = B = 0. La ecuación de regresión múltiple se reduce entonces a

$$E(y) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$$

Se interpreta β_0 como el valor esperado del número de unidades ensambladas a la semana por el empleado que utiliza el método C. En otras palabras, β_0 es la media del número de unidades generadas por semana utilizando el método C.

A continuación se considera la forma de la ecuación de regresión múltiple para cada uno de estos métodos. Mediante A, los valores de las variables ficticias son A = 1 y B = 0, y

$$E(y) = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$$

Para el método B se establece A = 0 y B = 1, y

$$E(y) = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$$

Observamos que $\beta_0 + \beta_1$ representa la media del número de unidades ensambladas por semana utilizando el método A, y $\beta_0 + \beta_2$ representa la media del número de unidades ensambladas por semana utilizando el método B.

Ahora se desea obtener las estimaciones del coeficiente de β_0, β_1 y β_2 para desarrollar una estimación del número medio de unidades generadas por semana con cada uno de los métodos. En la tabla 16.10 se presentan los datos muestrales, que consisten de 15 observaciones de A, B y y. La figura 16.18 presenta los resultados correspondientes a la regresión múltiple de Minitab. Se observa que las estimaciones de β_0, β_1 y β_2 son $b_0 = 52$, $b_1 = 10$ y $b_2 = 14$. Por tanto, la mejor estimación de la media del número de unidades ensambladas por semana con cada uno de los métodos es la siguiente.

Método de ensamble	Estimación de $E(y)$
A	$b_0 + b_1 = 52 + 10 = 62$
B	$b_0 = 52 + 14 = 66$
C	$b_0 = 52$

TABLA 16.10 Datos de entrada para el diseño completamente aleatorizado de Chemitech

WEB archivo
Chemitech

A	B	y
1	0	58
1	0	64
1	0	55
1	0	66
1	0	67
0	1	58
0	1	69
0	1	71
0	1	64
0	1	68
0	0	48
0	0	57
0	0	59
0	0	47
0	0	49

Observe que las estimaciones de los números medios de las unidades producidas con cada uno de los tres métodos de ensamble obtenidas mediante el análisis de regresión son las mismas que las medias muestrales presentadas previamente.

Note cómo ahora se pueden utilizar los resultados del análisis de regresión múltiple para realizar la prueba del ANOVA sobre la diferencia entre las medias de las tres plantas. En primer lugar, observe que si las medias no difieren

$$E(y) \text{ para el método A} - E(y) \text{ para el método C} = 0$$
$$E(y) \text{ para el método B} - E(y) \text{ para el método C} = 0$$

FIGURA 16.18 Resultados de la regresión múltiple para el diseño completamente aleatorizado de Chemitech

```
The regression equation is
y = 52.0 + 10.0 A + 14.0 B

Predictor      Coef  SE Coef       T      P
Constant     52.000    2.380   21.84  0.000
A            10.000    3.367    2.97  0.012
B            14.000    3.367    4.16  0.001

S = 5.32291   R-Sq 60.5%   R-Sq(adj) = 53.9%

Analysis of Variance

SOURCE           DF       SS      MS      F      P
Regression        2   520.00  260.00   9.18  0.004
Residual Error   12   340.00   28.33
Total            14   860.00
```

Debido a que β_0 es igual a $E(y)$ para el método C, y $\beta_0 + \beta_1$ es igual a $E(y)$ para el método A, la primera diferencia es igual a $(\beta_0 + \beta_1) - \beta_0 = \beta_1$. Por otra parte, como $\beta_0 + \beta_2$ es igual a $E(y)$ al emplear el método B, la segunda diferencia es igual a $(\beta_0 + \beta_2) - \beta_0 = \beta_2$. Se concluye que los tres métodos no difieren si $\beta_1 = 0$ y $\beta_2 = 0$. Por tanto, la hipótesis nula para una prueba sobre la diferencia entre las medias se puede establecer como

$$H_0 : \beta_1 = \beta_2 = 0$$

Suponga que el nivel de significancia es $\alpha = 0.05$. Recuerde que para probar este tipo de hipótesis nula acerca de la significancia de la relación de regresión se utiliza la prueba F de significancia general. El resultado de Minitab en la figura 16.18 muestra que el valor-p correspondiente a $F = 9.18$ es 0.004. Debido a que el valor-$p = 0.004 < \alpha = 0.05$, se rechaza $H_0 : \beta_1 = \beta_2 = 0$, y se concluye que las medias de los tres métodos de ensamble no son las mismas. Ya que la prueba F muestra que la relación de regresión múltiple es significativa, se puede realizar una prueba t para determinar la significancia de cada uno de los parámetros β_1 y β_2. Con $\alpha = 0.05$, el valor-p de 0.012 y 0.001 en los resultados de Minitab indica que $H_0 : \beta_1 = 0$ y $H_0 : \beta_2 = 0$ pueden ser rechazadas. Por tanto, ambos parámetros son estadísticamente significativos. Así que se puede concluir que tanto las medias de los parámetros A y C como las medias de los parámetros B y C son diferentes.

Ejercicios

Métodos

AUTO evaluación

20. Considere un diseño completamente aleatorizado que involucre cuatro tratamientos: A, B, C y D. Escriba la ecuación de regresión múltiple para analizar estos datos. Defina todas las variables.

21. Escriba una ecuación de regresión múltiple para analizar los datos de un bloque de diseño aleatorizado que involucre tres tratamientos y dos bloques. Defina todas las variables.

22. Escriba una ecuación de regresión múltiple para analizar los datos de un diseño bifactorial de dos niveles para el factor A y tres niveles para el factor B. Defina todas las variables.

Aplicaciones

AUTO evaluación

23. Jacobs Chemical Company desea estimar el tiempo promedio (en minutos) requerido para mezclar un lote de material en las máquinas de tres fabricantes. Para limitar los costos de las pruebas, se mezclaron cuatro lotes del material en las máquinas de cada uno de los tres fabricantes. El tiempo necesario para mezclar el material es el siguiente.

Fabricante 1	Fabricante 2	Fabricante 3
20	28	20
26	26	19
24	31	23
22	27	22

a) Escriba una ecuación de regresión múltiple para analizar los datos.
b) ¿Cuáles son las mejores estimaciones de los coeficientes en su ecuación de regresión?

c) En términos de los coeficientes de las ecuaciones de regresión, ¿qué hipótesis se tienen que probar a efecto de determinar si el tiempo promedio para mezclar un lote de material es el mismo para los tres fabricantes?

d) Con $\alpha = 0.05$ como nivel de significancia, ¿qué conclusión se obtiene?

24. En la publicidad de cuatro pinturas, todas ofrecen el mismo tiempo de secado. Para comprobar las afirmaciones del fabricante se realizaron pruebas a cinco muestras de cada pintura. El tiempo en minutos en que la primera mano se seca lo suficiente para aplicar una segunda capa fue registrado en cada muestra. Los datos obtenidos son los siguientes.

Pintura 1	Pintura 2	Pintura 3	Pintura 4
128	144	133	150
137	133	143	142
135	142	137	135
124	146	136	140
141	130	131	153

a) Use $\alpha = 0.05$ para probar si existe alguna diferencia significativa en el tiempo promedio de secado entre las pinturas.

b) ¿Cuál es la estimación del tiempo promedio de secado de la pintura 2? ¿Cómo se obtiene el resultado por medio de software?

25. Un concesionario de vehículos lleva a cabo una prueba para determinar si el tiempo necesario para ajustar un motor pequeño depende de si se utiliza un analizador de motor computarizado o un analizador electrónico. Dado que el tiempo de ajuste varía entre los vehículos de tamaño compacto, mediano o grande, se utilizaron estos tres tipos de automóviles como bloques del experimento. Los datos que se obtuvieron (en minutos) son los siguientes.

		Automóvil		
		Pequeño	**Mediano**	**Grande**
Analizador	**Computarizado**	50	55	63
	Electrónico	42	44	46

Utilice $\alpha = 0.05$ para probar si hay diferencias significativas.

26. Una empresa de ventas por catálogo diseñó un experimento factorial para probar el efecto del tamaño y diseño de un anuncio publicitario sobre el número (en miles) de catálogos solicitados. Se consideraron tres diseños y dos tamaños distintos de anuncios. Los datos que se obtuvieron son los siguientes. Pruebe si existen efectos significativos debido a tipo de diseño y tamaño del anuncio, o si existe interacción. Utilice $\alpha = 0.05$.

		Tamaño del anuncio de publicidad	
		Pequeño	**Grande**
Diseño	**A**	8	12
		12	8
	B	22	26
		14	30
	C	10	18
		18	14

16.6 Autocorrelación y la prueba de Durbin-Watson

A menudo, los datos utilizados para los estudios de regresión en la economía y los negocios están correlacionados en el tiempo. No es raro que el valor de y en el periodo t, denotado por y_t, esté relacionado con el valor de y en un periodo anterior. En tales casos, se dice que existe una **autocorrelación** (también llamada **correlación serial**) en los datos. Si el valor de y en el periodo t se relaciona con su valor en el periodo $t - 1$, existe una autocorrelación de primer orden. Si el valor de y en el periodo t se relaciona con el valor de y en el periodo $t - 2$, existe una autocorrelación de segundo orden, y así sucesivamente.

Uno de los supuestos del modelo de regresión indica que los términos del error son independientes. Sin embargo, cuando existe la autocorrelación, este supuesto es violado. En el caso de la autocorrelación de primer orden, el error en el periodo t, denotado ϵ_t, estará relacionada con el error en el periodo $t - 1$, denotado ϵ_{t-1}. Dos casos de autocorrelación de primer orden se ilustran en la figura 16.19. El grupo A muestra una autocorrelación positiva y el grupo B una autocorrelación negativa. En la autocorrelación positiva se espera un residual positivo en un periodo, seguido por un residual negativo en el periodo siguiente, y así sucesivamente. En la autocorrelación negativa se espera un residual positivo en un periodo, seguido de un residual negativo en el periodo siguiente, después un residual positivo, y así sucesivamente.

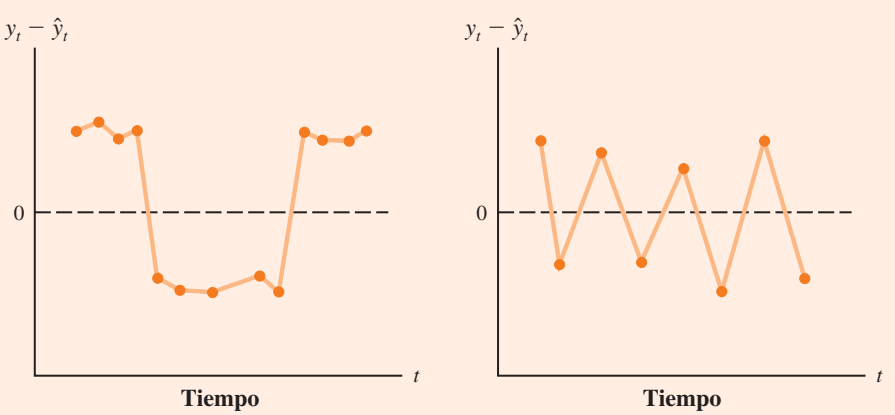
Cuando la autocorrelación está presente, se pueden cometer errores serios al realizar la prueba de significancia estadística basada en el modelo de regresión supuesto. Por tanto, es importante detectar la autocorrelación y emprender acciones correctivas. A continuación se mostrará cómo el estadístico de Durbin-Watson se utiliza para detectar la autocorrelación de primer orden.

Suponga que los valores de ϵ no son independientes sino que están relacionadas de la siguiente manera.

$$\epsilon_t = \rho\epsilon_{t-1} + z_t \qquad \textbf{(16.16)}$$

donde ρ es un parámetro con un valor absoluto menor que 1, y z_t es una variable aleatoria distribuida normal e independientemente con una media de 0 y una varianza de σ^2. En la ecuación (16.16) se observa que si $\rho = 0$, los términos del error no están relacionados, y cada uno tiene una media de 0 y una varianza de σ^2. En este caso no hay autocorrelación y se satisfacen los

FIGURA 16.19 Dos conjuntos de datos con autocorrelación de primer orden



Grupo A. Autocorrelación positiva **Grupo B. Autocorrelación negativa**

supuestos de regresión. Si $\rho > 0$, se tiene una autocorrelación positiva; si $\rho < 0$, la autocorrelación es negativa. En cualquiera de estos casos, los supuestos de regresión sobre el término del error son violados.

En la prueba de **Durbin-Watson** para la autocorrelación se utilizan los residuales a efecto de determinar si $\rho = 0$. Para simplificar la notación del estadístico de Durbin-Watson, el residual *i*ésimo se denota como $e_i = y_i - \hat{y}_i$. El estadístico de prueba de Durbin-Watson se calcula de la siguiente manera.

ESTADÍSTICO DE PRUEBA DE DURBIN-WATSON

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \qquad \textbf{(16.17)}$$

Si los valores sucesivos de los residuales están muy cercanos (autocorrelación positiva), el valor del estadístico de prueba de Durbin-Watson será pequeño. Si los valores sucesivos de los residuales están muy separados (autocorrelación negativa), el valor del estadístico de Durbin-Watson será grande.

Los rangos en el valor del estadístico de prueba Durbin-Watson son de 0 a 4, con un valor de 2 indicando que no existe autocorrelación. Durbin y Watson desarrollaron tablas para determinar cuándo su estadístico de prueba indica la presencia de una autocorrelación. La tabla 16.11 muestra los límites inferiores y superiores (d_L y d_U) para las pruebas de hipótesis con $\alpha = 0.05$; n denota el número de observaciones. La hipótesis nula a probar siempre es que no hay autocorrelación.

$$H_0: \rho = 0$$

La hipótesis alternativa para probar la autocorrelación positiva es

$$H_a: \rho > 0$$

TABLA 16.11 Valores críticos en la prueba de Durbin-Watson para la autocorrelación

Nota. Las entradas en la tabla son los valores críticos en una prueba de Durbing-Watson de una cola para la autocorrelación. En las pruebas de dos colas, el nivel de significancia se duplica.

Puntos de significancia de d_L y d_U: $\alpha = 0.05$
Número de variables independientes

	1		2		3		4		5	
n*	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

* Interpolar linealmente los valores n intermedios.

FIGURA 16.20 Prueba de hipótesis de autocorrelación utilizando la prueba de Durbin-Watson

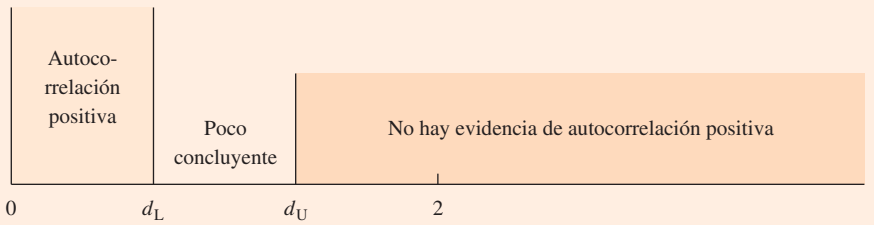


Gráfico A. Prueba de autocorrelación positiva

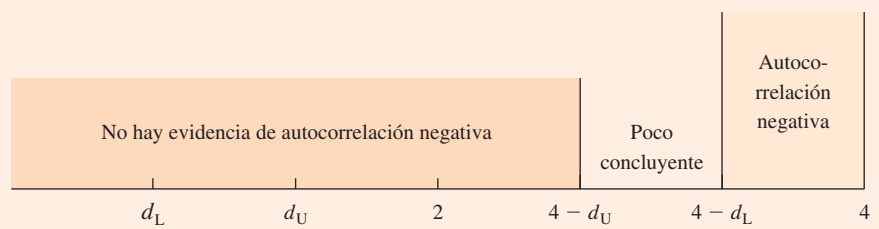


Gráfico B. Prueba de autocorrelación negativa

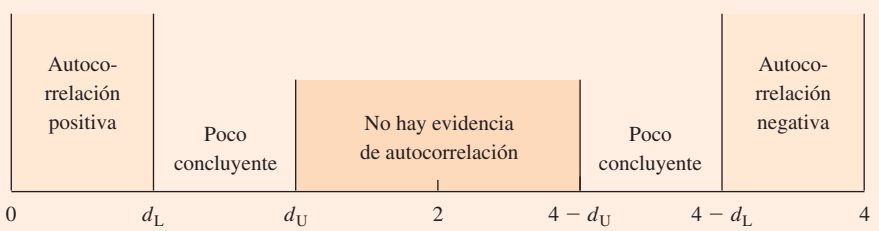


Gráfico C. Prueba de autocorrelación de dos colas

La hipótesis alternativa para probar la autocorrelación negativa es

$$H_a: \rho < 0$$

También es posible una prueba de dos colas. En este caso la hipótesis alternativa es

$$H_a: \rho \neq 0$$

La figura 16.20 muestra cómo se utilizan los valores de d_L y d_U de la tabla 16.11 para probar si existe autocorrelación. En el gráfico A se presenta la prueba para una autocorrelación positiva. Si $d < d_L$, se concluye que existe una autocorrelación positiva. Si $d_L \leq d \leq d_U$, se dice que la prueba no es concluyente. Si $d > d_U$, se concluye que no hay evidencia de una autocorrelación positiva.

En el gráfico B se muestra la prueba de autocorrelación negativa. Si $d > 4 - d_L$, se concluye que existe una autocorrelación negativa. Si $4 - d_U \leq d \leq 4 - d_L$, se dice que la prueba no es concluyente. Si $d < 4 - d_U$, se concluye que no hay evidencia de una autocorrelación negativa.

En el gráfico C se ilustra la prueba de dos colas. Si $d < d_L$ o $d > 4 - d_L$, entonces rechazamos H_0 y se concluye que existe la autocorrelación. Si $d_L \leq d \leq d_U$, o $4 - d_U \leq d \leq 4 - d_L$, la prueba no es concluyente. Si $d_U < d < 4 - d_U$, se concluye que no hay evidencia de autocorrelación.

Si se identifica una autocorrelación significativa, se debe verificar si se omitieron una o varias variables independientes que tengan un efecto de orden temporal sobre la variable dependiente. Si no existen variables que puedan ser identificadas, incluir una variable independiente que mida el tiempo en el que se realiza la observación (por ejemplo, el valor de esta variable puede ser 1 para la primera observación, 2 para la segunda observación, etc.) a veces eliminará o reducirá la autocorrelación. Cuando estos intentos de reducir o eliminar la autocorrelación no funcionan, la transformación de las variables dependientes o independientes suele ser útil; un estudio de estas transformaciones se puede encontrar en libros más avanzados sobre análisis de regresión.

Observe que las tablas de Durbin-Watson listan el menor tamaño de la muestra como 15. Esto se debe a que la prueba es generalmente inconclusa para las muestras de menor tamaño; de hecho, muchos expertos en estadística creen que el tamaño de la muestra debe ser por lo menos de 50 para que se obtengan resultados que valgan la pena.

Ejercicios

Aplicaciones

27. Los datos siguientes muestran los precios (Price) diarios de cierre (en dólares por acción) de IBM considerando las fechas (Date) del 3 de noviembre de 2005 hasta el 1 de diciembre 2005 (*Compustat*, 26 de febrero de 2006).

WEB archivo
IBM

Date	Price ($)
Nov. 3	82.87
Nov. 4	83.00
Nov. 7	83.61
Nov. 8	83.15
Nov. 9	82.84
Nov. 10	83.99
Nov. 11	84.55
Nov. 14	84.36
Nov. 15	85.53
Nov. 16	86.54
Nov. 17	86.89
Nov. 18	87.77
Nov. 21	87.29
Nov. 22	87.99
Nov. 23	88.80
Nov. 25	88.80
Nov. 28	89.11
Nov. 29	89.10
Nov. 30	88.90
Dic. 1	89.21

a) Defina la variable independiente Period, donde Period = 1 corresponda al dato del 3 de noviembre, Period = 2 corresponda al dato del 4 de noviembre, y así sucesivamente. Desarrolle la ecuación de regresión estimada para predecir el precio del cierre dado el valor de Period.

b) A nivel de 0.05 de significancia, pruebe si existe autocorrelación positiva en los datos.

28. Consulte el conjunto de datos de Cravens de la tabla 16.5. En la sección 16.3 se mostró que el coeficiente de determinación ajustado de la ecuación de regresión estimada que contenía

Accounts (cuentas), AdvExp (gastPubl), Poten (potencial) y Share (participación) era de 88.1%. Utilice 0.05 como nivel de significancia y aplique la prueba de Durbin-Watson para determinar si existe autocorrelación positiva.

Resumen

En este capítulo se estudiaron varios conceptos utilizados en la construcción de modelos para ayudar a identificar la mejor ecuación de regresión estimada. En primer lugar, se presentó el concepto de un modelo lineal general para mostrar cómo pueden extenderse los métodos estudiados en los capítulos 14 y 15 para manejar la relación curvilínea y los efectos de interacción. Después se estudió cómo utilizar las transformaciones de la variable dependiente para ser tomadas en cuenta en problemas como el de una varianza no constante en términos del error.

En muchas aplicaciones del análisis de regresión se emplea un gran número de variables independientes. Para agregar o eliminar variables a un modelo de regresión se estudió un método general basado en el estadístico F. Después se presentó un problema grande en el que se tenían 25 observaciones y ocho variables independientes. También se vio que cuando se tienen problemas más grandes, uno de los asuntos a resolver es encontrar el mejor subconjunto de variables independientes. Para esta tarea existen varios procedimientos de selección de variables: regresión por pasos, selección hacia adelante, eliminación hacia atrás y la regresión de los mejores subconjuntos.

En la sección 16.5 se amplió el estudio acerca de cómo los modelos de regresión múltiple pueden ser desarrollados para ofrecer otro método en la solución de problemas del análisis de varianza y de diseño de experimentos. El capítulo concluye con una aplicación del análisis de residuales a efecto de mostrar la prueba de Durbin-Watson para la autocorrelación.

Glosario

Autocorrelación Correlación en los errores que se presenta cuando los términos del error en puntos sucesivos de tiempo están relacionados.

Correlación serial Es lo mismo que la autocorrelación.

Interacción El efecto de dos variables independientes cuando actúan juntas.

Modelo lineal general Un modelo de la forma $y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_q + \epsilon$, en el que cada una de las variables independientes $z_j (j = 1, 2, \ldots, p)$ es una función de x_1, $x_2 \ldots, x_k$, las variables para las que se han reunido los datos.

Procedimientos de selección de variables Métodos de selección de un subconjunto de variables independientes para un modelo de regresión.

Prueba de Durbin-Watson Prueba para determinar si la autocorrelación de primer orden existe.

Fórmulas clave

Modelo lineal general

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p + \epsilon \qquad \textbf{(16.1)}$$

Estadístico de prueba F para agregar o eliminar variables $p - q$

$$F = \frac{\dfrac{\text{SCE}(x_1, x_2, \ldots, x_q) - \text{SCE}(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{p - q}}{\dfrac{\text{SCE}(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{n - p - 1}} \qquad \textbf{(16.13)}$$

Autocorrelación de primer orden

$$\epsilon_t = \rho\epsilon_{t-1} + z_t \tag{16.16}$$

Estadístico de prueba de Durbin-Watson

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{16.17}$$

Ejercicios complementarios

29. Los precios más bajos de las impresoras láser a color las convierten en una gran alternativa frente a las impresoras de inyección de tinta. *PC World* revisó y evaluó la capacidad de 10 impresoras láser a color. Los siguientes datos muestran la velocidad de impresión en páginas por minuto (Speed ppm) de gráficos a color y la calificación (Rating) general de *PC World* para cada impresora examinada (*PC World*, diciembre de 2005). Se incluyen fabricantes y modelo (Make and Model).

WEB archivo
ColorPrinter

Make and Model	Speed (ppm)	Rating
Dell 3000cn	3.4	83
Oki Data C5200n	5.2	81
Konica Minolta MagiColor 2430DL	2.7	79
Brother HL-2700CN	3.1	78
Lexmark C522n	3.8	77
HP Color LaserJet 3600n	5.6	74
Xerox Phaser 6120n	1.6	73
Konica Minolta MagiColor 2450	1.6	71
HP Color LaserJet 2600n	2.6	70
HP Color LaserJet 2550L	1.1	61

a) Desarrolle un diagrama de dispersión de los datos utilizando la velocidad de impresión (Speed) como variable independiente. ¿El modelo de regresión lineal simple parece apropiado?

b) Desarrolle una ecuación de regresión múltiple estimada con x = velocidad y x_2 como las dos variables independientes.

c) Considere el modelo no lineal mostrado en la ecuación (16.7). Utilice logaritmos para transformar este modelo no lineal en un modelo lineal equivalente y desarrolle la ecuación de regresión estimada correspondiente. ¿Esta ecuación proporciona un mejor ajuste que la obtenida en el inciso *b*)?

30. *Consumer Reports* probó 19 marcas y modelos (Make and Model) de bicicletas de ruta, fitness y de confort. Las de ruta están diseñadas para viajes largos por caminos (Road); las fitness para realizar ejercicios o viajes diarios al trabajo, y las de confort (Comfort) para viajes de placer en caminos generalmente planos. Los datos siguientes muestran el tipo (Type), peso (Weight), en libras y el precio (Price, $) para las 19 bicicletas probadas (sitio web de Consumer Reports, febrero de 2009).

WEB archivo
Bikes

Brand and Model	Type	Weight	Price($)
Klein RÃªve v	Road	20	1 800
Giant OCR Composite 3	Road	22	1 800
Giant OCR 1	Road	22	1 000
Specialized Roubaix	Road	21	1 300
Trek Pilot 2.1	Road	21	1 320

(*Continúa*)

Brand and Model	Type	Weight	Price($)
Cannondale Synapse 4	Road	21	1 050
LeMond Poprad	Road	22	1 350
Raleigh Cadent 1.0	Road	24	650
Giant FCR3	Fitness	23	630
Schwinn Super Sport GS	Fitness	23	700
Fuji Absolute 2.0	Fitness	24	700
Jamis Coda Comp	Fitness	26	830
Cannondale Road Warrior 400	Fitness	25	700
Schwinn Sierra GS	Comfort	31	340
Mongoose Switchback SX	Comfort	32	280
Giant Sedona DX	Comfort	32	360
Jamis Explorer 4.0	Comfort	35	600
Diamondback Wildwood Deluxe	Comfort	34	350
Specialized Crossroads Sport	Comfort	31	330

a) Desarrolle un diagrama de dispersión con el peso (Weight) como variable independientes, y el precio (Price) como variable dependiente. ¿El modelo de regresión lineal simple parece apropiado?

b) Desarrolle una ecuación de regresión estimada múltiple con x = peso y x^2 como las dos variables independientes.

c) Utilice las siguientes variables ficticias para desarrollar una ecuación de regresión estimada a efecto de predecir el precio con base en el tipo de bicicleta: Type_fitness = 1 si es de fitness, 0 en caso contrario; y Type_Comfort = 1 si la bicicleta es de confort, 0 en caso contrario. Compare los resultados con los obtenidos en el inciso *b*).

d) Tomando en cuenta la posible interacción entre el tipo de bicicleta y su peso, desarrolle una nueva ecuación de regresión estimada para predecir el precio de la bicicleta dado el tipo, el peso y cualquier interacción entre el peso y cada una de las variables ficticias definidas en el inciso *c*). ¿Qué ecuación parece ser el mejor predictor del precio? Explique.

31. Un estudio investigó la relación entre el retraso en la auditoría (Delay), el periodo desde el final del año fiscal de una empresa hasta la fecha del dictamen del auditor, y las variables que describen al cliente y al auditor. Algunas variables independientes incluidas en el estudio son las siguientes.

Industry	Variable ficticia con valor 1 si se trata de una industria y 0 si se trata de un banco, institución de ahorro y préstamo o compañía de seguros.
Public	Variable ficticia con valor 1 si la empresa fue negociada en la bolsa organizada o por venta libre; de lo contrario, codificar en 0.
Quality	Medida de la calidad general de los controles internos, a juicio del auditor, en una escala de cinco puntos que van desde "prácticamente ninguna" (1) hasta "excelente" (5).
Finished	Medida que va de 1 a 4, a juicio del auditor, donde 1 indica "todo el trabajo realizado después del fin de año" y 4 "la mayor parte del trabajo realizado antes del fin de año".

En una muestra de 40 empresas se obtuvieron los siguientes datos.

WEB archivo
Audit

Delay	Industry	Public	Quality	Finished
62	0	0	3	1
45	0	1	3	3
54	0	0	2	2
71	0	1	1	2
91	0	0	1	1

Delay	Industry	Public	Quality	Finished
62	0	0	4	4
61	0	0	3	2
69	0	1	5	2
80	0	0	1	1
52	0	0	5	3
47	0	0	3	2
65	0	1	2	3
60	0	0	1	3
81	1	0	1	2
73	1	0	2	2
89	1	0	2	1
71	1	0	5	4
76	1	0	2	2
68	1	0	1	2
68	1	0	5	2
86	1	0	2	2
76	1	1	3	1
67	1	0	2	3
57	1	0	4	2
55	1	1	3	2
54	1	0	5	2
69	1	0	3	3
82	1	0	5	1
94	1	0	1	1
74	1	1	5	2
75	1	1	4	3
69	1	0	2	2
71	1	0	4	4
79	1	0	5	2
80	1	0	1	4
91	1	0	4	1
92	1	0	1	4
46	1	1	4	3
72	1	0	5	2
85	1	0	5	1

a) Desarrolle la ecuación de regresión estimada con todas las variables independientes.
b) ¿La ecuación obtenida en el inciso *a*) proporciona un buen ajuste? Explique.
c) Desarrolle un diagrama de dispersión que muestre Delay como una función de Finished. ¿Qué indica este diagrama sobre la relación entre ambas variables?
d) Con base en sus observaciones acerca de la relación entre Delay y Finished, desarrolle una ecuación de regresión estimada distinta a la desarrollada en el inciso *a*) que explique la mayor proporción posible de variabilidad de Delay.

32. Consulte los datos del ejercicio 31. Considere un modelo en el que para predecir Delay se utilice sólo Industry. Con un nivel de significancia de 0.01, pruebe cualquier autocorrelación positiva en los datos.

33. Consulte los datos del ejercicio 31.
a) Obtenga una ecuación de regresión estimada para predecir el retraso utilizando Industry y Quality.
b) Represente gráficamente los residuales obtenidos de la ecuación de regresión estimada desarrollada en el inciso *a*) en función del orden en que se presentan los datos. ¿Parece existir alguna autocorrelación en los datos? Explique.
c) Con un nivel 0.05 de significancia, pruebe si existe una autocorrelación positiva en los datos.

34. Se realizó un estudio para investigar la actividad de los compradores, los cuales fueron clasificados como inactivos (Nonbrowsers), poco activos (Light Browsers) y muy activos (Heavy

Browsers). Por cada comprador en el estudio se obtuvo una medida para determinar qué tan cómodo se sentía en determinada tienda. Las puntaciones más altas indican una mayor comodidad. Suponga que los siguientes datos provienen de este estudio. Utilice un nivel de significancia de 0.05 para probar las diferencias en los niveles de comodidad entre los tres tipos de compradores.

WEB archivo
Browsing

Nonbrowser	Light Browser	Heavy Browser
4	5	5
5	6	7
6	5	5
3	4	7
3	7	4
4	4	6
5	6	5
4	5	7

35. La revista *Money* publicó los precios y los datos correspondientes a los 418 automóviles modelo 2003 más populares. Una de las variables fue el valor de reventa de los mismos, expresado como un porcentaje del precio de reventa sugerido por el fabricante. Los datos se clasificaron con base en el tamaño y el tipo de vehículo. La siguiente tabla muestra el valor de reventa de 10 automóviles compactos (Small), 10 medianos (Midsize), 10 de lujo (Luxury) y 10 deportivos (Sports) todos elegidos aleatoriamente (*Money*, marzo de 2003).

WEB archivo
Resale

Small	Midsize	Luxury	Sports
26	26	36	41
31	29	38	39
41	41	38	30
32	27	39	34
27	26	35	40
34	33	26	43
31	27	40	42
38	29	47	39
27	35	41	44
42	39	32	50

Utilice $\alpha = 0.05$ y determine si existe alguna diferencia significativa entre los valores medios de reventa de los cuatro tipos de vehículos.

Caso a resolver 1 Análisis de las estadísticas de la PGA Tour

La Asociación de Golfistas Profesional (PGA) lleva un registro sobre el desempeño y las ganancias de los miembros del PGA Tour. Con base en las ganancias totales de los eventos del PGA Tour, los 125 mejores jugadores están exentos para la siguiente temporada. Integrar la lista de las ganancias de los primeros 125 es importante, ya que un jugador que está "exento" ha calificado para ser miembro de tiempo completo de la PGA Tour para la siguiente temporada.

La puntuación promedio general se considera la estadística más importante en términos de éxito en el PGA Tour. Para investigar la relación entre el promedio de puntuación y variables como la distancia del *drive*, precisión de conducción, *greens* en regulación, trampa de arena y promedio de *putts* por ronda, se presentan los datos del final del año sobre el desempeño de los 125 jugadores que registraron las mejores ganancias en los eventos de la PGA Tour 2008 obtenidos del archivo PGATour (sitio web de la PGA Tour, 2009). Cada fila del conjunto de datos corresponde a un jugador, y los datos se han clasificado con base en las ganancias totales. A continuación se presenta la descripción de los datos.

WEB archivo
PGATour

Money	Ganancias totales en los eventos de la PGA Tour.
Scoring Average	Número promedio de golpes por ronda completa.
DrDist (*distancia de salto*)	Número promedio de yardas por *drive* medido. En la PGA Tour se mide la distancia del *drive* en dos hoyos por ronda. Hay que tener cuidado al seleccionar los dos agujeros que se colocan en direcciones opuestas para contrarrestar el efecto del viento. Los *drives* son medidos en el punto en el que se detiene, independientemente de si llega a la calle o no.
DrAccu (*precisión de conducción*)	Porcentaje de veces que un golpe de *tee* llega a la calle (independientemente de su club). La precisión de conducción se mide en cada hoyo, excluido el par 3.
GIR (green *en regulación*)	Porcentaje de veces que un jugador alcanza el *green* en regulación. El *green* se considera un golpe en regulación si cualquier parte de la bola toca la superficie del *green* después de que la carrera del GIR se ha tomado. La carrera se determina restando 2 del par (primera carrera en un par 3; 2 en un par 4; 3 en un par 5). En otras palabras, un *green* se considera un golpe en regulación si el jugador ha llegado a la superficie del *green* en el par en menos de dos golpes.
Sand Saves (*trampa de arena*)	Porcentaje de veces que un jugador es capaz de "subir y bajar" una vez en un búnker de arena al lado del *green* (independientemente de la puntuación). "Subir y bajar" indica que el jugador tuvo dos tiros o menos para poner la bola en el agujero de un búnker de arena al lado del *green*.
PPR (putts *por vuelta*)	Número promedio de *putts* por ronda.
Scrambling (*resolución*)	Porcentaje de veces que un jugador falla el *green* en regulación pero todavía realiza un par o lo supera.
Bounce Back (*recuperación*)	Porcentaje de veces que un jugador tiene más de un par de hoyos y luego menos de un par en el siguiente hoyo. En otras palabras, es el porcentaje de hoyos con un mal *bogey* y en el siguiente hoyo con un mejor *birdie*.

Informe gerencial

Suponga que un representante de la PGA Tour lo contrata a efecto de analizar los datos para una presentación que se llevará a cabo en la reunión anual de la asociación. El representante le pregunta si sería posible utilizar estos datos para determinar las medidas del desempeño que sean el mejor predictor de la puntuación promedio de un jugador. Utilice los métodos presentados en éste y en los anteriores capítulos para analizar los datos. Prepare un informe que resuma su análisis para el representante de la PGA Tour, en el que incluya los principales resultados estadísticos, sus conclusiones y recomendaciones. Prepare todo el material técnico adecuado en un apéndice.

Caso a resolver 2 Rendimiento de combustible en los automóviles

WEB archivo
Cars

Todos los automóviles nuevos que se venden en Estados Unidos incluyen una etiqueta sobre consumo de combustible que muestra el rendimiento en millas por galón que se espera que el vehículo pueda alcanzar en ciudad y en carretera. En la *Fuel Economy Guide* del Departamento de Energía de Estados Unidos se encuentra esta información para cualquier automóvil o camión. Una parte de los datos para 230 vehículos se encuentra en el archivo Cars (sitio web de U.S. Department of Energy, 21 de marzo de 2003). Las descripciones de los datos se listan a continuación.

Class	Clase de automóvil (compacto, mediano, grande)
Manufacturer	Fabricante del automóvil
Carline name	Nombre del vehículo
Displ	Desplazamiento del motor en litros
Cyl	Número de cilindros del motor (4, 6, 8)
Trans	Tipo de transmisión (automática, manual)
Cty	Consumo de combustible en la ciudad en millas por galón
Hwy	Consumo de combustible en carretera en millas por galón

Informe gerencial

Utilice los métodos presentados en éste y los capítulos anteriores para analizar el conjunto de datos proporcionado. El objetivo del estudio es desarrollar sendas ecuaciones de regresión estimada para estimar el consumo de combustible en la ciudad y el consumo de combustible en carretera. Presente un resumen y análisis en el que se incluyan los resultados estadísticos más importantes, conclusiones y recomendaciones dentro de un informe gerencial. Incorpore cualquier material técnico adecuado en el apéndice (el resultado de la computadora, las gráficas de residuales, etcétera).

Apéndice 16.1 Procedimientos de selección de variables con Minitab

WEB archivo
Cravens

En la sección 16.4 se estudió el uso de los procedimientos de selección de variables para la solución de problemas de regresión múltiple. En la figura 16.16 se mostraron los resultados obtenidos con la regresión por pasos de Minitab aplicados para los datos de Cravens, y en la figura 16.17 los resultados de los mejores subconjuntos de Minitab. En este apéndice se describen los pasos necesarios para obtener los resultados en ambas figuras, así como en los procedimientos de selección hacia adelante y eliminación hacia atrás. En primer lugar, los datos de la tabla 16.5 se deben introducir en una hoja de cálculo de Minitab. Los valores de Sales, Time, Poten, AdvExp, Share, Change, Accounts, Work y Rating se ingresan en las columnas C1-C9 de la hoja.

Procedimiento por pasos usando Minitab

Las siguientes tareas se pueden utilizar para obtener con Minitab los resultados de regresión por pasos de los datos de Cravens.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Seleccione el menú **Regression**.
Paso 3. Elija **Stepwise**.
Paso 4. Cuando el cuadro de diálogo **Stepwise Regression** aparezca:
Ingrese Sales en el cuadro de diálogo **Response**.
Ingrese Time, Poten, AdvExp, Share, Change, Accounts, Work y Rating en el cuadro **Predictors**.
Seleccione el botón **Methods**.
Paso 5. Cuando el cuadro de diálogo **Stepwise-Method** aparezca:
Seleccione **Stepwise** (**forward and backward**).
Ingrese 0.05 en cuadro de diálogo **Alpha to enter**.
Introduzca 0.05 en el cuadro de diálogo **Alpha to remove**.
Haga clic en **OK**.
Paso 6. Cuando el cuadro de diálogo **Stepwise Regression** aparezca:
Haga clic en **OK**.

Procedimiento de selección hacia adelante utilizando Minitab

Para el procedimiento de selección hacia adelante de Minitab, simplemente hay que modificar el paso 5 del procedimiento de regresión por pasos como se muestra a continuación.

Paso 5. Cuando el cuadro de diálogo **Stepwise-Methods** aparezca:
Seleccione **Forward Selection**.
Ingrese 0.05 en el cuadro de diálogo **Alpha to enter**.
Haga clic en **OK**.

Procedimiento de eliminación hacia atrás utilizando Minitab

Para utilizar el procedimiento de eliminación hacia atrás de Minitab, simplemente hay que modificar el paso 5 del procedimiento de regresión por pasos como se muestra a continuación.

Paso 5. Cuando el cuadro de diálogo **Stepwise-Methods** aparezca:
Seleccione **Backward elimination**.
Ingrese 0.05 en el cuadro de diálogo **Alpha to remove**.
Haga clic en **OK**.

Procedimiento de los mejores subconjuntos usando Minitab

Con los siguientes pasos se obtienen los resultados de Minitab de la regresión de los mejores subconjuntos para los datos de Cravens.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Seleccione el menú **Regression**.
Paso 3. Elija **Best Subsets**.
Paso 4. Cuando el cuadro de diálogo **Best Subsets Regression** aparezca.
Ingrese Sales en el cuadro **Response**.
Ingrese Time, Poten, AdvExp, Share, Change, Accounts, Work y Rating en el cuadro **Predictors**.
Haga clic en **OK**.

Apéndice 16.2 Procedimientos de selección de variables con StatTools

WEB archivo
Cravens

En este apéndice se verá cómo se puede utilizar StatTools para realizar tres procedimientos de selección de variables: regresión por pasos, selección hacia adelante y eliminación hacia atrás. En primer lugar, se muestra cómo se obtienen los resultados de regresión por pasos para el problema de Cravens.

Comience por utilizar el Data Set Manager para crear un conjunto de datos StatTools mediante el procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes describen cómo se puede utilizar StatTools para obtener los resultados de la regresión por pasos.

Paso 1. Haga clic en la barra de herramientas **StatTools**.
Paso 2. En el grupo **Analyses**, haga clic en **Regression and Classification**.
Paso 3. Elija la opción **Regression**.
Paso 4. Cuando el cuadro de diálogo StatTools-Regression aparezca:
Seleccione **Stepwise** en el cuadro **Regression Type**.
En la sección **Variables**:
Haga clic en el botón de **Format** y seleccione **Unstacked**.
En la columna **D** selecciones **Sales**.
En la columna **I** elija **Time, Poten, AdvExp, Share, Change, Accounts, Work** y **Rating**.

En la sección **Parameters**:
Seleccione **Use p-Values**.
Ingrese 0.05 en el cuadro **p-Values to Enter**.
Ingrese 0.05 en el cuadro **p-Value to Leave**.
En la sección **Advance Options**, seleccione **Include Detailed Step Information**.
Haga clic en **OK**.

El resultado de la regresión por pasos para el problema de Cravens aparece en una hoja de cálculo nueva.

El cuadro de diálogo Regression de StatTools contiene opciones más avanzadas para el desarrollo de estimaciones del intervalo de predicción y la realización de gráficas de residuales. La opción Help proporciona información sobre todas estas opciones. StatTools también puede utilizarse para realizar los procedimientos de selección hacia adelante y eliminación hacia atrás. Los pasos necesarios son muy similares a los del procedimiento de selección hacia adelante. La principal diferencia estriba en que en el paso 4 se debe seleccionar ya sea Forward o Backward del cuadro de herramientas Regression Type. Si opta por Forward, tiene que escribir un valor en el cuadro p-Value to Enter, y si elige Backward debe introducir el valor en la caja de herramientas p-Value to Leave.

CAPÍTULO 17

Números índice

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: OFICINA DE ESTADÍSTICAS LABORALES, DEPARTAMENTO DEL TRABAJO DE ESTADOS UNIDOS

17.1 PRECIOS RELATIVOS

17.2 ÍNDICES DE PRECIOS AGREGADOS

17.3 CÁLCULO DEL ÍNDICE DE PRECIOS AGREGADO A PARTIR DE LOS PRECIOS RELATIVOS

17.4 ALGUNOS ÍNDICES DE PRECIOS IMPORTANTES
Índice de precios al consumidor
Índice de precios al productor
Promedios Down Jones

17.5 DEFLACTACIÓN DE UNA SERIE MEDIANTE ÍNDICES DE PRECIOS

17.6 ÍNDICES DE PRECIOS: OTRAS CONSIDERACIONES
Selección de artículos
Selección de un periodo base
Variaciones en la calidad

17.7 ÍNDICES DE CANTIDAD

ESTADÍSTICA *en* LA PRÁCTICA

OFICINA DE ESTADÍSTICAS LABORALES, DEPARTAMENTO DEL TRABAJO DE ESTADOS UNIDOS

WASHINGTON, D.C.

El Departamento del Trabajo de Estados Unidos, a través de su Oficina de Estadísticas Laborales, recaba y distribuye los índices y estadísticas que son indicadores de la actividad comercial y económica del país. Por ejemplo, el departamento elabora y publica el índice de precios al consumidor, el índice de precios al productor y estadísticas sobre las horas y los ingresos promedio de diversos grupos de trabajadores. Quizás el índice más ampliamente citado es el índice de precios al consumidor, elaborado por la Oficina de Estadísticas Laborales. A menudo se utiliza como una medida de la inflación.

En marzo de 2009 la Oficina de Estadísticas Laborales informó que el índice de precios al consumidor (IPC) había aumentado 0.5% en febrero. El nivel de 212.2 en este mes fue 0.3% más alto que el de febrero de 2008. Sobre una base ajustada estacionalmente, el IPC aumentó 0.4% en febrero después de incrementarse 0.3% en enero. El aumento de 8.3% en el índice de precios de la gasolina parece que fue el causante principal, mientras que el índice de alimentos disminuyó 0.1%. Algunos economistas consideran favorable la noticia del incremento del IPC porque reduce la probabilidad de un periodo deflacionario.

Un día antes, la Oficina de Estadísticas Laborales había informado que el índice de precios al productor (IPP) se incrementó 0.1% en febrero con el ajuste estacional. Este aumento siguió al de 0.8% en enero, y a una disminución de 1.9% en diciembre. El IPP mide los cambios en los precios de los mercados al mayoreo y al menudeo, y se le considera el principal indicador en las variaciones del índice de precios al consumidor. El ritmo mínimo en el aumento en febrero fue fuertemente influido, por la disminución de la tasa de incremento en los bienes de energía. Este índice subió 1.3% en febrero después de aumentar 3.7% en enero.



Los precios de la gasolina son un componente del índice de precios al consumidor. © Jeff Chiu/AP Photo.

En este capítulo se verá como se calculan diversos índices, como el índice de precios al consumidor y el índice de precios al productor, y cómo deben interpretarse.

Cada mes el gobierno de Estados Unidos publica una variedad de índices diseñados para ayudar a las personas a entender las condiciones económicas y de negocios actuales de este país. El más conocido y citado es probablemente el índice de precios al consumidor (IPC). Como su nombre lo indica, el IPC es un indicador de lo que sucede con los precios que pagan los consumidores por los artículos que compran. En concreto, mide las variaciones de precios durante un periodo. A partir de un punto determinado o *periodo base* y su índice que corresponde a 100, el IPC puede utilizarse para comparar los precios al consumidor del periodo actual con los del periodo base. Por ejemplo, un IPC de 125 refleja la condición de que los precios al consumidor en su conjunto son alrededor de 25% más altos que los precios del periodo base. Aunque relativamente pocas personas comprenden exactamente lo que este número significa, se sabe lo suficiente sobre el IPC para entender que su aumento se traduce en precios más altos.

A pesar de que el IPC es tal vez el índice más conocido, existen otros índices de los sectores privado y gubernamental que ayudan a medir y comprender las condiciones económicas de un periodo en comparación con las condiciones económicas de otros periodos. El propósito de este capítulo es describir los tipos de índices más utilizados para empezar a construir algunos números índice sencillos y obtener una mejor comprensión de cómo se calculan.

17.1

Precios relativos

TABLA 17.1

Costo de la gasolina regular (todas las fórmulas)

Año	Precio por galón ($)
1990	1.30
1991	1.10
1992	1.09
1993	1.07
1994	1.08
1995	1.11
1996	1.22
1997	1.20
1998	1.03
1999	1.14
2000	1.48
2001	1.42
2002	1.34
2003	1.56
2004	1.85
2005	2.27
2006	2.57
2007	2.80
2008	3.25

Fuente. U.S. Energy Information Administration.

La forma más sencilla de un índice de precios muestra cómo comparar el precio actual por unidad de un producto determinado con el precio por unidad del mismo artículo en el periodo base. Por ejemplo, en la tabla 17.1 se presentan los precios de un galón de gasolina normal desde 1990 hasta 2008. Para facilitar las comparaciones con otros años, la cifra real del costo por galón se puede convertir a un **precio relativo**, que expresa el precio unitario en cada periodo como un porcentaje del precio de la unidad en un periodo base.

$$\text{Precio relativo en el periodo } t = \frac{\text{precio en el periodo } t}{\text{precio en el periodo base}}(100) \qquad \textbf{(17.1)}$$

Respecto de los precios de la gasolina en la tabla 17.1, con el año 1990 como base, se pueden calcular los precios relativos de un galón de gasolina regular desde 1990 hasta 2008. Estas relaciones de precios se muestran en la tabla 17.2. Observe la facilidad con que el precio de un año se compara con el precio del año base y el resultado es el precio relativo. Por ejemplo, el precio relativo de 85.4 en 1995 indica que en este año la gasolina costaba 14.6% menos que en el año base de 1990. Del mismo modo, en 2002 el precio relativo fue de 103.1, lo que indica un incremento de 3.1% en el precio de la gasolina en comparación con el año base. Y el precio relativo de 2008 era de 250.0, que indica un incremento de 150% en el precio del combustible regular a partir de 1990 como año base. Los precios relativos como el de la gasolina regular, son muy útiles en términos de comprensión e interpretación de los cambios en las condiciones económicas y de negocios en el tiempo.

17.2

Índices de precios agregados

TABLA 17.2

Precios relativos para un galón de gasolina regular

Año	Precio relativo (Base 1990)
1990	(1.30/1.30)100 = 100.0
1991	(1.10/1.30)100 = 84.6
1992	(1.09/1.30)100 = 83.8
1993	(1.07/1.30)100 = 82.3
1994	(1.08/1.30)100 = 83.1
1995	(1.11/1.30)100 = 85.4
1996	(1.22/1.30)100 = 93.8
1997	(1.20/1.30)100 = 92.3
1998	(1.03/1.30)100 = 79.2
1999	(1.14/1.30)100 = 87.7
2000	(1.48/1.30)100 = 113.8
2001	(1.42/1.30)100 = 109.2
2002	(1.34/1.30)100 = 103.1
2003	(1.56/1.30)100 = 120.0
2004	(1.85/1.30)100 = 142.3
2005	(2.27/1.30)100 = 174.6
2006	(2.57/1.30)100 = 197.7
2007	(2.80/1.30)100 = 215.4
2008	(3.25/1.30)100 = 250.0

Aunque los precios relativos pueden utilizarse para identificar la variación de los precios de artículos individuales en el tiempo, a menudo se está más interesado en la variación del precio de un grupo de artículos como un todo. Por ejemplo, si se desea un índice que mida la variación del costo general de vida en el tiempo, sí se deseará que el índice se base en la variación de los precios de diversos artículos como alimentación, vivienda, vestido, transporte, atención médica, etc. Un **índice de precios agregado** se desarrolla con el propósito específico de medir la variación combinada de un grupo de artículos.

Considere el desarrollo de un índice de precios agregado para un grupo de artículos clasificados como gastos de operación normales por el uso del automóvil. Para ejemplificar, se limitan los artículos incluidos en el grupo a gasolina, aceite, neumáticos y póliza de seguro.

En la tabla 17.3 se presentan los datos de los cuatro componentes del índice de gastos de operación de un automóvil para los años 1990 y 2008. Con 1990 como periodo base, un índice de precios agregado para los cuatro componentes dará una medida de la variación en los gastos normales por el uso del automóvil en el periodo 1990-2008.

Un índice agregado no ponderado se obtiene simplemente sumando los precios unitarios en el año de interés (por ejemplo, 2008) y dividiendo el resultado entre la suma de los precios unitarios en el año base (1990). Sean

$$P_{it} = \text{precio unitario del artículo } i \text{ en el periodo } t$$
$$P_{i0} = \text{precio unitario del artículo } i \text{ en el periodo base}$$

Un índice agregado no ponderado del periodo t se denota como I_t y está dado por

$$I_t = \frac{\Sigma P_{it}}{\Sigma P_{i0}}(100) \qquad \textbf{(17.2)}$$

donde las sumas incluyen todos los artículos del grupo.

TABLA 17.3 Datos para el índice de gastos de operación de un automóvil

Artículo	Precio unitario ($) 1990	2008
Galón de gasolina	1.30	3.25
Cuarto de galón de aceite	2.10	8.00
Neumáticos	130.00	140.00
Póliza de seguro	820.00	1 030.00

Un índice agregado no ponderado de gastos por el uso normal de un automóvil en 2008 ($t = 2008$) está dado por

$$I_{2008} = \frac{3.25 + 8.00 + 140.00 + 1\,030.00}{1.30 + 2.10 + 130.00 + 820.00}(100)$$
$$= \frac{1\,181.25}{953.4}(100) = 124$$

Del índice de precios agregado no ponderado se concluye que el precio de los gastos normales por el funcionamiento del automóvil sólo se ha incrementado 24% durante el periodo comprendido entre 1990 y 2008. Pero observe que el método agregado no ponderado para establecer un índice de precios compuesto para los gastos del automóvil está fuertemente influido por los artículos cuyos precios por unidad son más altos. En consecuencia, los artículos con precios unitarios relativamente bajos, como la gasolina y el petróleo, están dominados por los artículos con un alto precio unitario, como los neumáticos y el seguro. El índice agregado no ponderado para los gastos del automóvil también está fuertemente influido por las variaciones en los precios de los neumáticos y el seguro.

Si la cantidad utilizada de cada uno de los artículos es la misma, con un índice no ponderado se obtienen los mismos resultados que con un índice ponderado. Sin embargo, en la práctica las cantidades utilizadas rara vez son las mismas.

Debido a la sensibilidad que muestran los índices no ponderados entre uno o varios artículos de precio más alto, este tipo de índice no es muy utilizado. Con un índice de precios agregado ponderado se obtiene una mejor comparación en el uso de distintas cantidades.

La filosofía detrás del **índice de precios agregado ponderado** indica que cada artículo del grupo debe ser ponderado con base en su importancia. En la mayoría de los casos el uso es la mejor medida de importancia. Por tanto, se debe obtener una medida de la cantidad de uso para los diferentes artículos en el grupo. La tabla 17.4 proporciona la información del uso anual de cada artículo que se debe tener en cuenta en los gastos del automóvil con base en el uso estándar de un vehículo de tamaño mediano que recorre aproximadamente 15 000 millas por año. Los ponderadores de las cantidades que se listan muestran el uso anual esperado para este tipo de situación.

TABLA 17.4
Información del uso anual para el índice de gastos de operación de un automóvil

Artículo	Ponderador de la cantidad*
Galones de gasolina	1 000
Cuarto de galón de aceite	15
Neumáticos	2
Póliza de seguro	1

* Basado en 15 000 millas por año. La vida de uso de los neumáticos es de 30 000 millas.

Sea Q_i = la cantidad de uso para el artículo i. El índice de precios agregado ponderado en el periodo t está dado por

$$I_t = \frac{\Sigma P_{it} Q_i}{\Sigma P_{i0} Q_i}(100) \qquad \textbf{(17.3)}$$

donde las sumas son de todos los artículos del grupo. Aplicado a los gastos de operación del automóvil, el índice de precios agregado ponderado se obtiene al dividir los costos totales de operación del uso del automóvil de 2008 entre los costos operativos totales de 1990.

Sea $t = 2008$, y la cantidad de uso se pondera en la tabla 17.4. Se obtiene el siguiente índice de precios agregado ponderado de los gastos por el uso del automóvil en ese año.

$$I_{2008} = \frac{3.25(1\,000) + 8.00(15) + 140.00(2) + 1\,030.00(1)}{1.30(1\,000) + 2.10(15) + 130.00(2) + 820.00(1)}(100)$$
$$= \frac{4\,680}{2\,411.5}(100) = 194$$

A partir de este índice de precios agregado ponderado se concluye que el precio de los gastos de operación de un automóvil se ha incrementado 94% durante el periodo 1990-2008.

Es evidente que, en comparación con el índice agregado no ponderado, el índice ponderado proporciona un indicador más preciso de la variación del precio en el uso del automóvil durante el periodo de referencia. Al tomar en cuenta la cantidad utilizada de gasolina, se contrarresta el pequeño incremento porcentual de los costos del seguro. El índice ponderado muestra un mayor aumento en los gastos del automóvil que el índice no ponderado. En general, el índice ponderado, con cantidades de uso como ponderación, es el método preferido para establecer un índice de precios de un grupo de elementos.

En el índice de precios agregado ponderado de la fórmula (17.3) observe que la cantidad, el término Q_i, no tiene un segundo subíndice que indique el tiempo. Esto se debe a que las cantidades Q_i se consideran fijas y no varían con el tiempo como los precios. Las ponderaciones fijas o cantidades son especificadas por el diseñador del índice al emplear las que considere representativas del uso estándar. Una vez establecidas, se mantienen constantes o fijas para todos los periodos en que se utilice el índice. Los índices de otros años que no sean 2008 exigen la recolección de nuevos datos de precios P_{it}, pero las cantidades siguen siendo la ponderación Q_i.

En un caso especial del índice agregado de ponderación fija, las cantidades se determinan con el uso en el año base. En este caso escribimos $Q_i = Q_{i0}$ con cero como subíndice, que indica el ponderador de la cantidad del año base; la fórmula (17.3) se convierte en

$$I_t = \frac{\Sigma P_{it} Q_{i0}}{\Sigma P_{i0} Q_{i0}} (100) \qquad \textbf{(17.4)}$$

Cuando los ponderadores de las cantidades fijas se determinan a partir del año base, el índice agregado ponderado se conoce con el nombre de **índice de Laspeyres**.

Otra opción para determinar los ponderadores de las cantidades consiste en revisar las cantidades en cada periodo. En este caso se determina Q_{it} por cada año en que el índice es calculado. El índice agregado ponderado en el periodo t con estos ponderadores de las cantidades está dado por

$$I_t = \frac{\Sigma P_{it} Q_{it}}{\Sigma P_{i0} Q_{it}} (100) \qquad \textbf{(17.5)}$$

Observe que los mismos ponderadores de las cantidades son utilizados para el periodo base (periodo 0) y para el periodo t. Sin embargo, los ponderadores se basan en el uso en el periodo t, no en el periodo base. A este índice agregado ponderado se le conoce como **índice de Paasche**. Éste tiene la ventaja de basarse en los estándares de uso actuales. Sin embargo, este método de cálculo de un índice agregado ponderado tiene dos desventajas: las cantidades normales de uso Q_{it} deben ser determinadas cada año, aumentando así el tiempo y el costo de la recolección de datos, y cada año se deben volver a calcular los números índice de los años anteriores para reflejar el efecto de los nuevos ponderadores de cantidades. Debido a estas desventajas, el índice de Laspeyres es el más ampliamente utilizado. El índice de los gastos por el uso del automóvil se calculó con las cantidades del periodo base, por lo que se trata de un índice de Laspeyres. Si se hubieran utilizado las cifras del año 2008, representaría un índice de Paasche. De hecho, debido a que los automóviles han registrado un consumo de gasolina más eficiente, éste disminuyó, y con el índice de Paasche se obtiene una cifra distinta que con el índice de Laspeyres.

Ejercicios

Métodos

AUTOevaluación

1. En la tabla siguiente se presentan los precios y cantidades utilizadas de dos productos correspondientes a 2007 y 2009.

	Cantidad		Precio unitario ($)	
Artículo	2007	2009	2007	2009
A	1 500	1 800	7.50	7.75
B	2	1	630.00	1 500.00

a) Calcule los precios relativos para cada artículo en 2009 usando 2007 como periodo base.
b) Estime un índice de precios agregado no ponderado de los dos artículos en 2009 usando 2007 como periodo base.
c) Calcule un índice de precios agregado ponderado para los dos artículos con el método Laspeyres.
d) Obtenga un índice de precios agregado ponderado de los dos artículos mediante el método de Paasche.

2. Un artículo cuyo precio relativo es 132 cuesta $10.75 en 2009. El año base es 1992.
a) ¿En qué porcentaje aumentó o disminuyó el costo del artículo en el periodo de 17 años?
b) ¿Cuánto costaba el artículo en 1992?

Aplicaciones

AUTOevaluación

3. Un fabricante compra un componente idéntico de tres proveedores independientes que difieren en el precio unitario y la cantidad ofrecida. Los datos correspondientes a 2007 y 2009 son los siguientes.

Proveedor	Cantidad (2007)	Precio unitario ($) 2007	2009
A	150	5.45	6.00
B	200	5.60	5.95
C	120	5.50	6.20

a) Calcule por separado los precios relativos de los componentes de cada uno de los proveedores. Compare el incremento de precios de los proveedores durante el periodo de dos años.
b) Obtenga un índice de precios agregado ponderado para el componente en 2009.
c) Calcule un índice de precios agregado no ponderado de los componentes en 2009. ¿Cuál es la interpretación de este índice para la empresa fabricante?

4. R&B Beverages, Inc. ofrece una línea completa de cervezas, vino y bebidas refrescantes que distribuye en puntos de venta minoristas en el centro de Illinois. Los datos de los precios unitarios para 2006 y 2009, y las cantidades vendidas (cajas) en 2006 son los siguientes.

Artículo	Cantidad 2006 (cajas)	Precio unitario ($) 2006	2009
Cerveza	35 000	17.50	20.15
Vino	5 000	100.00	118.00
Bebidas refrescantes	60 000	8.00	8.80

Calcule el índice agregado ponderado de las ventas de bebidas de R&B Beverage en 2009, con 2006 como periodo base.

5. En el método UEPS de valuación de inventarios debe establecerse un índice de precios para el inventario con fines de impuestos. Los ponderadores de cantidades se basan en los niveles de inventario de final de año. Utilice el precio unitario de principios de año como precio del periodo base y desarrolle un índice agregado ponderado del valor total del inventario al final del año. ¿Qué tipo de índice de precios agregado ponderado se debe desarrollar para la valuación de inventarios UEPS?

Producto	Inventario final	Precio unitario ($) Principio	Fin
A	500	0.15	0.19
B	50	1.60	1.80
C	100	4.50	4.20
D	40	12.00	13.20

17.3 Cálculo del índice de precios agregado a partir de los precios relativos

En la sección 17.1 se definió el concepto de precios relativos y se mostró cómo se pueden calcular a partir del precio unitario en el periodo actual y del precio unitario del periodo base. Ahora se desea mostrar cómo se calculan directamente los índices de precios agregados como los desarrollados en la sección 17.2 a partir de la información sobre el precio relativo de cada artículo del grupo. Debido al uso limitado de los índices no ponderados, se restringirá la atención a los índices de precios agregados ponderados. Se vuelve a los índices de gastos relacionados con el uso del automóvil de la sección anterior. La información necesaria para los cuatro artículos se encuentra en la tabla 17.5

Debemos comprobar que los precios y las cantidades se manejen en las mismas unidades. Por ejemplo, si los precios son por caja, la cantidad debe ser el número de cajas y no, por ejemplo, el número de unidades.

Sea w_1 la ponderación correspondiente al precio relativo para el artículo i. La expresión general del promedio ponderado de los precios relativos está dada por

$$I_t = \frac{\sum \frac{P_{it}}{P_{i0}}(100)w_i}{\Sigma w_i} \tag{17.6}$$

La elección adecuada de las ponderaciones en la ecuación (17.6) permitirá calcular un índice de precios agregado ponderado a partir de los precios relativos. Esta elección resulta de multiplicar el precio del periodo base por la cantidad de uso.

$$w_i = P_{i0}Q_i \tag{17.7}$$

La sustitución de $w_i = P_{i0}Q_i$ en la ecuación (17.6) proporciona la siguiente expresión para un índice de precios relativo ponderado.

$$I_t = \frac{\sum \frac{P_{it}}{P_{i0}}(100)(P_{i0}Q_i)}{\Sigma P_{i0}Q_i} \tag{17.8}$$

Como en el numerador se cancelan los términos P_{i0}, una expresión equivalente para el índice de precios relativo ponderado es

$$I_t = \frac{\Sigma P_{it}Q_i}{\Sigma P_{i0}Q_i}(100)$$

Por tanto, se observa que el índice de precios relativo ponderado $w_i = P_{i0}Q_i$ proporciona un índice de precios idéntico al índice agregado ponderado presentado en la sección 17.2 por la

TABLA 17.5 Precios relativos para el índice de gastos de operación de un automóvil

	Precio unitario ($)			
Artículo	1990 (P_0)	2008 (P_t)	Precio relativo $(P_t/P_0)100$	Uso Anual
Galón de gasolina	1.30	3.25	250.0	1 000
Cuarto de galón de aceite	2.10	8.00	381.0	15
Neumáticos	130.00	140.00	107.7	2
Póliza de seguro	820.00	1 030.00	125.6	1

TABLA 17.6 Índice de gastos de operación de un automóvil (1990-2008) basado en los precios relativos ponderados

Artículo	Precios relativos $(P_{it}/P_{i0})(100)$	Precio base ($) P_{i0}	Cantidad Q_i	Carga $w_i = P_{i0}Q_i$	Precios relativos ponderados $(P_{it}/P_{i0})(100)w_i$
Gasolina	250.0	1.30	1 000	1 300.00	325 000.00
Aceite	381.0	2.10	15	31.50	12 001.50
Neumáticos	107.7	130.00	2	260.00	28 002.00
Seguros	125.6	820.00	1	820.00	102 992.00
			Totales	2 411.50	467 995.50

$$I_{2008} = \frac{467\,995.50}{2\,411.50} = 194$$

ecuación (17.3). Al utilizar las cantidades del periodo base (es decir, $Q_i = Q_{i0}$) en la ecuación (17.7) se obtiene el índice de Laspeyres. Con el uso de cantidades del periodo actual (es decir, $Q_i = Q_{it}$) en la ecuación (17.7) se obtiene un índice de Paasche.

De vuelta a los datos de gastos por el uso del automóvil, se pueden utilizar los precios relativos de la tabla 17.5 y la ecuación (17.6) para calcular el promedio ponderado de los precios relativos. Los resultados obtenidos mediante las ponderaciones especificadas por la ecuación (17.7) se encuentran en la tabla 17.6. El número índice 194 representa un aumento de 94% en los gastos de operación del automóvil, que es el mismo identificado por el cálculo del índice agregado ponderado en la sección 17.2.

Ejercicios

Métodos

AUTOevaluación

6. Los precios relativos de tres artículos, así como sus precios y uso en el periodo base se muestran en la tabla siguiente. Calcule un índice de precios agregado ponderado para el periodo actual.

		Periodo base	
Artículo	Precio relativo	Precio	Uso
A	150	22.00	20
B	90	5.00	50
C	120	14.00	40

Aplicaciones

AUTOevaluación

7. Mitchell Chemical Company produce una sustancia química para la industria que es una mezcla de tres componentes químicos. El costo por libra a principios de año, el costo por libra al final del año y la proporción de la mezcla son los siguientes.

	Costo por libra ($)		Cantidad (libras) por cada 100 libras del producto
Ingrediente	Inicio	Final	
A	2.50	3.95	25
B	8.75	9.90	15
C	0.99	0.95	60

a) Calcule los precios relativos de cada uno de los tres ingredientes.
b) Calcule un promedio ponderado de los precios relativos para desarrollar un índice del costo anual de las materias primas utilizadas en el producto. ¿Cuál es su interpretación sobre este valor del índice?

8. Un portafolio de inversiones consta de cuatro acciones. El precio de compra, el precio actual y el número de acciones se presentan en la tabla siguiente.

Acción	Precio de compra/acción ($)	Precio actual/acción ($)	Número de acciones
Holiday Trans	15.50	17.00	500
NY Electric	18.50	20.25	200
KY Gas	26.75	26.00	500
PQ Soaps	42.25	45.50	300

Construya una media ponderada de los precios relativos como un índice del desempeño del portafolio hasta la fecha. Interprete este índice de precios.

9. Calcule los precios relativos de los productos de R&B Beverages del ejercicio 4. Utilice un promedio ponderado de los precios relativos para demostrar que este método proporciona el mismo índice que el método agregado ponderado.

17.4 Algunos índices de precios importantes

Se han identificado los procedimientos para calcular los índices de precios de un artículo o de grupos de artículos. Ahora veremos algunos índices de precios que son indicadores importantes de las condiciones económicas y de negocios. Se considerarán, en específico, el índice de precios al consumidor, el índice de precios al productor y los promedios Dow Jones.

Índice de precios al consumidor

El IPC incluye los gastos en servicios (por ejemplo, pagos médicos y a dentistas) y todos los impuestos directamente relacionados con la compra y el uso de un artículo.

El **índice de precios al consumidor (IPC)** es publicado mensualmente por la Oficina de Estadísticas Laborales de Estados Unidos, y es la principal medida del costo de vida en este país. El conjunto de artículos empleado para elaborar el índice está conformado por una *canasta de mercado* de 400 productos, que incluye alimentación, vivienda, vestido, transporte y medicamentos. El IPC es un índice de precios agregado ponderado con precios fijos.[1] La ponderación aplicada a cada artículo de la canasta de mercado se obtiene de un estudio de uso de las familias de todo Estados Unidos.

En febrero de 2009, el IPC, calculado con un índice base de 100 de 1982-1984, fue de 212.2. Esta cifra significa que el costo de la canasta del mercado de bienes y servicios aumentó 112.2% desde el periodo base de 1982-1984. Las series de tiempo de 45 años del IPC desde 1960 hasta 2005 se muestran en la figura 17.1. Observe cómo el IPC refleja el comportamiento de una fuerte inflación en la economía a finales de 1970 y principios de 1980.

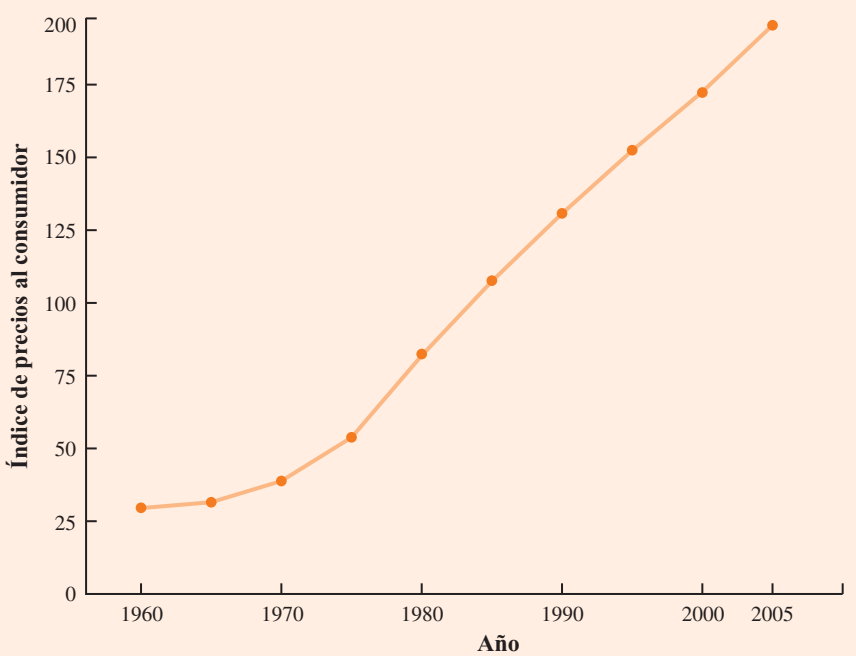
Índice de precios al productor

El IPP ha sido diseñado como una medida de la variación en los precios de los productos nacionales; las importaciones no se incluyen.

El **índice de precios al productor (IPP)** también es publicado mensualmente por la Oficina de Estadísticas Laborales de Estados Unidos y mide las variaciones mensuales de los precios en los mercados principales del país. El IPP se basa en los precios de la primera transacción de

[1] La Oficina de Estadísticas Laborales publica de hecho dos índices de precios al consumidor: uno para todos los consumidores urbanos (IPC-U) y una versión ajustada para trabajadores asalariados y personal administrativo (IPC-W). El IPC-U es el más citado, y se publica periódicamente en *The Wall Street Journal*.

FIGURA 17.1 Índice de precios al consumidor, 1960-2005 (base 1982-1984 = 100)



cada artículo en los mercados no minoristas. Todas las mercancías que se venden en las transacciones comerciales de estos mercados están representadas. La encuesta abarca materias primas y productos manufacturados y procesados en cualquier nivel de elaboración, incluyendo la producción de las industrias clasificadas como manufactura, agricultura, silvicultura, pesca, minería, gas y electricidad, y servicios públicos. Uno de los usos comunes de este índice es llevar un indicador principal de la tendencia futura en los precios al consumidor y en el costo de vida. El aumento del IPP refleja que hay incrementos en los precios al productor que con el tiempo pasarán a los consumidores mediante mayores precios al menudeo.

Las ponderaciones para los diferentes artículos en el IPP se basan en el valor de los embarques. El promedio ponderado de los precios relativos se calcula utilizando el método de Laspeyres. En febrero de 2009 el IPP fue calculado con un índice base de 100 para 1982, y fue de 171.3.

Promedios Dow Jones

Los **promedios Dow Jones** son índices que permiten conocer la evolución de los precios y los movimientos de acciones ordinarias. El más conocido es el índice industrial Dow Jones (DJIA), que se basa en los precios de las acciones comunes de 30 grandes empresas. Consiste en la suma de los precios de estas acciones dividida por un número, que se corrige de vez en cuando para ajustarse a las divisiones de las acciones o fusiones de las empresas que participan en el índice. A diferencia de los otros índices de precios estudiados, no se expresa como porcentaje de los precios del año base. Las empresas específicas utilizadas en julio de 2009 para calcular el DJIA se listan en la tabla 17.7.

Charles Henry Dow publicó su primer índice accionario el 3 de julio de 1884 en el Customer's Afternoon Letter. *Once acciones, de las cuales nueve eran ferroviarias, fueron incluidas en el primer índice. Un promedio comparable para el DJIA fue publicado por primera vez el 1 de octubre de 1928.*

Otros promedios Dow Jones se calculan con 20 acciones de transporte y 15 acciones de empresas de servicios públicos. Los promedios Dow Jones se calculan y difunden diariamente en *The Wall Street Journal* y otras publicaciones financieras.

TABLA 17.7 Las 30 empresas incluidas en el índice industrial Dow Jones (julio de 2009)

3m	Disney	Kraft Foods
Alcoa	DuPont	McDonald's
American Express	ExxonMobil	Merck
AT&T	General Electric	Microsoft
Bank of America	Hewlett-Packard	Pfizer
Boeing	Home Depot	Procter & Gamble
Caterpillar	IBM	Travelers
Chevron Corp.	Intel	United Technologies
Coca-Cola	Johnson & Johnson	Verizon
Cisco Systems	J. P. Morgan Chase	Wal-Mart Stores

Source. Barron's, 13 de julio de 2009.

17.5 Deflactación de una serie mediante índices de precios

Las series de tiempo se deflactan para eliminar los efectos de la inflación.

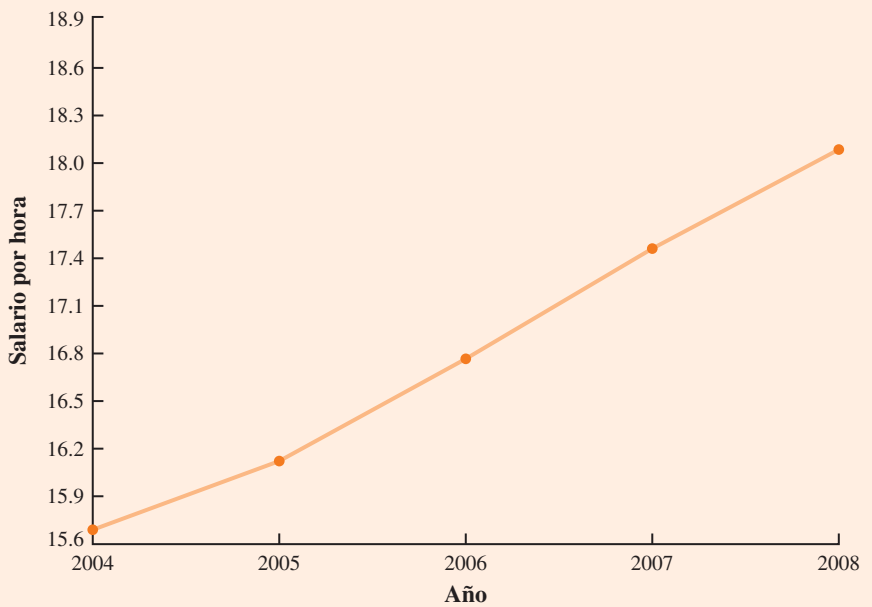
Muchas series de negocios y económicas medidas en el tiempo, como ventas de empresas, ventas industriales e inventarios, se miden en cantidad de dinero. Estas series de tiempo a menudo muestran un patrón de crecimiento cada vez mayor, que es generalmente interpretado como una indicación del aumento en el volumen físico relacionado con estas actividades. Por ejemplo, un aumento de 10% en el monto de dinero de un inventario podría interpretarse como que el inventario físico es 10% más grande. Tales percepciones pueden ser engañosas cuando una serie de tiempo se mide en términos de dinero y el monto total es una combinación de cambios tanto en los precios como en las cantidades. Por tanto, en el periodo en que las variaciones en los precios son significativas, los cambios en el monto de dinero pueden no corresponder a variaciones en la cantidad, a menos que la serie de tiempo sea ajustada para eliminar el efecto de los cambios en los precios.

Por ejemplo, de 1976 a 1980, el importe total del gasto en la industria de la construcción aumentó aproximadamente 75%. Esta cifra indica un crecimiento excelente en la actividad del sector. Sin embargo, los precios en la construcción aumentaban muy rápido, en ocasiones incluso más rápido que la tasa de 75%. De hecho, mientras que el gasto total en construcción fue en aumento, la actividad en el sector permanecía relativamente constante o, en casos como la edificación de casas nuevas, disminuía. Para interpretar correctamente la actividad en el sector de la construcción del periodo 1976-1980 se debe ajustar la serie de los gastos totales mediante un índice de precios para eliminar el efecto del incremento de los precios. Cada vez que se elimina este efecto en una serie de tiempo, se dice que estamos *deflactando la serie*.

En relación con el ingreso de las personas y los salarios, se suele escuchar discusiones acerca de "salarios reales" o del "poder de compra" de los salarios. Estos conceptos se relacionan con la noción de deflactar un índice de salario por hora. Por ejemplo, la figura 17.2 muestra el patrón de los salarios por hora de los trabajadores de la industria en el periodo 2004-2008. Se observa una tendencia de aumentos salariales de $15.69 a $18.07 por hora. ¿Los trabajadores de la industria estarán contentos con este aumento? La respuesta depende de lo que ocurra con el poder de compra de los salarios. Si se puede comparar este poder de compra de $15.69 por hora de 2004 con el de $18.07 por hora de 2008, se estará en mejores condiciones de juzgar la mejora relativa de los salarios.

La tabla 17.8 incluye informes tanto de la tasa salarial por hora y el IPC (calculado con un índice base de 100 de 1982-1984) para el periodo 2004-2008. Con estos datos se mostrará cómo utilizar el IPC para deflactar el índice de los salarios por hora. La serie deflactada se obtiene al

FIGURA 17.2 Salario por hora actual de los trabajadores de la industria



dividir el salario por hora de cada año entre el valor correspondiente del IPC y multiplicarlo por 100. El índice deflactado de los salarios por hora de los trabajadores de la industria se incluye en la tabla 17.9; en la figura 17.3 se presenta una gráfica que muestra los salarios deflactados o reales.

Los salarios reales son una mejor medida del poder de compra de los salarios actuales. De hecho, muchos contratos sindicales piden que se ajusten con base en los cambios en el costo de vida.

¿Qué indica la serie deflactada acerca de los salarios reales o el poder de compra de los trabajadores de la industria durante el periodo 2004-2008? En términos de dinero del periodo base (1982-1984 = 100), el salario por hora no aumentó durante ese lapso. Después de eliminar el efecto inflacionario vemos que el poder de compra de los trabajadores sólo aumentó $0.08 durante el periodo de cuatro años. Este efecto se ve en la figura 17.3. Por tanto, la ventaja de utilizar índices de precios para deflactar una serie estriba en que se obtiene una imagen más clara de los cambios reales ocurridos en términos monetarios.

Este proceso de deflactar una serie de medidas en el tiempo tiene una aplicación importante en el cálculo del producto interno bruto (PIB), que es el valor total de todos los bienes y servicios producidos en un país. Obviamente, el PIB mostrará, con el tiempo, las ganancias que

TABLA 17.8 Salarios por hora para los trabajadores de la industria e índices de precios al consumidor, 2004-2008.

Año	Salario por hora ($)	IPC
2004	15.69	188.9
2005	16.12	195.3
2006	16.76	201.6
2007	17.45	207.3
2008	18.07	215.3

Fuente. Oficina de Estadísticas Laborales. El IPC se calcula con un índice base 100 de 1982-1984.

TABLA 17.9 Serie deflactada de salarios por hora de los trabajadores de la industria. 2004-2008

Año	Salario por hora deflactado
2004	($15.69/188.9)(100) = $8.31
2005	($16.12/195.3)(100) = $8.25
2006	($16.76/201.6)(100) = $8.31
2007	($17.45/207.3)(100) = $8.42
2008	($18.07/215.3)(100) = $8.39

FIGURA 17.3 Salarios reales por hora de los trabajadores de la industria, 2004-2008



se deben en parte a los aumentos de precios si el PIB no es deflactado por un índice de precios. Por tanto, para ajustar el valor total de los bienes y servicios de manera que reflejen los verdaderos cambios en el volumen de bienes y servicios producidos y vendidos, el PIB debe calcularse con un índice de precios como deflactor. El proceso es similar al analizado para el cálculo de los salarios reales.

Ejercicios

Aplicaciones

AUTOevaluación

10. El salario promedio por hora de los trabajadores de la industria en febrero de 1996 fue de $11.86, y en febrero de 2009 subió a $18.55. El IPC en febrero de 1996 fue de 154.9 y en febrero de 2009 aumentó a 212.2.
    a) Deflacte los salarios por hora de 1996 y de 2009 para determinar las tasas de los salarios reales.
    b) ¿Cuál es la variación porcentual en los salarios por hora actuales de 1996 a 2009?
    c) ¿Cuál es la variación porcentual de los salarios reales en el mismo periodo?

11. Los salarios promedio por hora de los trabajadores de la industria de servicios en cuatro años, desde 2002 hasta 2005, se muestran enseguida. Utilice la información proporcionada por el índice de precios al consumidor para deflactar la serie de salarios. Calcule el aumento o disminución porcentual de los salarios reales desde 2003 hasta 2005.

Año	Salarios por hora	IPC (base 1982-1984)
2002	18.52	179.9
2003	18.95	184.0
2004	19.23	188.9
2005	19.46	195.3

Fuente. Oficina de Estadísticas Laborales.

12. La Oficina del Censo de Estados Unidos reportó las siguientes cantidades de pedidos de la industria para los años de 2005 a 2007.

Año	Pedidos de la industria ($ miles de millones)
2005	4 742
2006	5 020
2007	5 081

a) El IPC para el periodo 2005-2007 fue proporcionado en la tabla 17.8. Utilice esta información para deflactar la serie de pedidos de la industria y comente acerca de los patrones que siguen estos pedidos en términos de dólares constantes.

b) Los siguientes índices de precios al productor (bienes de consumo terminados) corresponden al periodo 2005-2007, con 1982 como año base. Utilice el IPP para deflactar la serie.

Año	IPP (1982 = 100)
2005	155.8
2006	160.3
2007	166.6

c) ¿Cree usted que el IPC o el IPP es el más apropiado para deflactor los pedidos de la industria?

13. El total de los volúmenes de ventas al detalle de Dooley Retail Outlets' para los años seleccionados desde 1982 se muestra en la tabla siguiente. También se lista el IPC con el índice base de 1982-1984. Deflacte las cifras de volumen de ventas sobre la base de dólares constantes de 1982-1984, y comente sobre el volumen de las ventas de la empresa en términos de dólares deflactados.

Año	Ventas al detalle ($)	IPC (1982-1984 base)
1982	380 000	96.5
1987	520 000	113.6
1992	700 000	140.3
1997	870 000	160.5
2002	940 000	179.9
2007	990 000	207.3

17.6 Índices de precios: otras consideraciones

En las secciones anteriores se describieron varios métodos para calcular los índices de precios, se estudió el uso de algunos de los índices más importantes y se presentó un procedimiento con índices de precios para deflactar series de tiempo. Diversos temas deben ser considerados para comprender mejor cómo se construyen y usan los índices de precios. Algunos se verán en esta sección.

Selección de artículos

El propósito principal de un índice de precios es medir la variación en el tiempo para un conjunto específico de artículos, productos, y así sucesivamente. Cuando este conjunto es muy grande, el índice no puede basarse en todos los artículos que lo integran. Por el contrario, se debe utilizar una muestra de artículos representativos. Mediante la recolección de información de precios y cantidad de los elementos de la muestra se espera obtener una buena idea del comportamiento de los precios de todos los artículos que el índice representa. Por ejemplo, para el índice de precios al consumidor, la cantidad de productos que podría considerarse en las poblaciones de los artículos que normalmente compra un consumidores es de 2 000 o más. Sin embargo, el índice sólo se basa en las características de precio y cantidad de 400. La selección de los artículos específicos para el índice no es una tarea trivial. Estudios de los patrones de compra del consumidor, así como el buen criterio se incluyen en el proceso de selección. Una muestra aleatoria simple no se utiliza para elegir los 400 artículos.

Después del proceso de selección inicial, el grupo de artículos considerado para calcular el índice debe ser revisado periódicamente y modificado cada vez que cambian los patrones de compra. Por tanto, la decisión de cuáles artículos incluir en un índice debe ser resuelto antes de que éste pueda desarrollarse y otra vez antes de que sea revisado.

Selección de un periodo base

La mayoría de los índices se establece sobre un periodo base con valor de 100 en algún momento específico. Todos los valores futuros del índice se relacionan con el valor del periodo base. Qué periodo base es apropiado para un índice no es una pregunta fácil de responder. Debe basarse en el criterio de la persona que desarrolla el índice.

Muchos índices establecidos por el gobierno de Estados Unidos hasta 2009 utilizan el periodo base de 1982. Como pauta general, el periodo base no debe estar muy alejado de la época actual. Por ejemplo, un índice de precios al consumidor con un periodo base de 1945 sería difícil de entender para la mayoría de las personas, debido a que no están familiarizadas con las condiciones de ese año. Por tanto, el periodo base para la mayoría de los índices se ajusta a una época más reciente. El periodo base del IPC cambió de 1967 al promedio de 1982-1984 en 1988. El IPP utiliza actualmente 1982 con el mismo propósito (es decir, 1982 = 100).

Variaciones en la calidad

El objetivo de un índice de precios es medir los cambios en los precios en el tiempo. Lo ideal es que estos datos sean recabados para el mismo conjunto de artículos en diferentes épocas, y después calcular el índice. Un supuesto básico es que el precio sea identificado para los mismos artículos en cada periodo. Un problema se presenta cuando un producto varía en calidad de un periodo a otro. Por ejemplo, un fabricante puede modificar la calidad usando materiales menos costosos, o con menos funciones, y así sucesivamente, de un año a otro. El precio aumenta en los años siguientes, pero es el precio de un producto de menor calidad. En consecuencia, aumenta en realidad más de lo que indica el precio de lista del artículo. Es difícil, si no es que imposible, ajustar un índice a la disminución en la calidad de un artículo.

Una mejora significativa en la calidad también puede aumentar el precio de un producto. La parte del precio que está relacionada con el mejoramiento de la calidad debe ser excluida del cálculo del índice. Sin embargo, ajustar un índice para un aumento de precio relacionado con la mayor calidad de un artículo es extremadamente difícil, si no es que imposible.

Aunque en la práctica sea común ignorar variaciones menores en la calidad en el desarrollo de un índice de precios, las variaciones importantes deben ser tomadas en cuenta, ya que se puede modificar la descripción del producto de un periodo a otro. Si la descripción cambia, el índice debe ser modificado para reflejarlo; en algunos casos el producto puede ser eliminado del índice.

Sin embargo, en diversas situaciones una mejora sustancial de la calidad va acompañada de una disminución en el precio. Esta situación menos común ha sido el caso de las computadoras personales durante la década de 1990 y principios de la primera década de los años 2000.

17.7 Índices de cantidad

Además de los índices de precios descritos en las secciones anteriores, otros tipos de números índice son útiles, en particular los que se aplican para medir los cambios en los niveles de cantidad en el tiempo. A este tipo de índice se le denomina **índice de cantidad**.

Recuerde que en el desarrollo del índice de precios agregado ponderado en la sección 17.2, para calcular un número índice para el periodo t se requirieron datos sobre los precios unitarios de un periodo base (P_0) y del periodo t (P_t). La ecuación (17.3) proporcionó el índice de precios agregado ponderado como

$$I_t = \frac{\Sigma P_{it} Q_i}{\Sigma P_{i0} Q_i} (100)$$

El numerador, $\Sigma P_{it} Q_i$, representa el valor total de las cantidades fijas de los artículos del índice en el periodo t. El denominador $\Sigma P_{i0} Q_i$ representa el valor total de las mismas cantidades fijas de los artículos del índice en el año 0.

El cálculo de un índice de cantidad agregado ponderado es similar al de un índice de precios agregado ponderado. Las cantidades de cada artículo se miden en el periodo base y en el periodo t, con Q_{i0} y Q_{it}, respectivamente, representando aquellas cantidades para el artículo i. Las cantidades son después ponderadas por un precio fijo, el valor agregado, o algún otro factor. El "valor agregado" de un producto es el valor de venta menos el costo de los insumos adquiridos. La fórmula para calcular un índice de cantidad agregado ponderado para el periodo t es

$$I_t = \frac{\Sigma Q_{it} w_i}{\Sigma Q_{i0} w_i} (100) \qquad \textbf{(17.9)}$$

En algunos índices de cantidad, la ponderación para el artículo i se toma como el precio del periodo base (P_{i0}), en cuyo caso el índice de cantidad agregado ponderado es

$$I_t = \frac{\Sigma Q_{it} P_{i0}}{\Sigma Q_{i0} P_{i0}} (100) \qquad \textbf{(17.10)}$$

Los índices de cantidad también se pueden calcular sobre la base de cantidades relativas ponderadas. Una fórmula para esta versión de un índice de cantidad es la siguiente.

$$I_t = \frac{\sum \frac{Q_{it}}{Q_{i0}} (Q_{i0} P_i)}{\Sigma Q_{i0} P_i} (100) \qquad \textbf{(17.11)}$$

Esta fórmula es la versión para las cantidades de la fórmula de precios relativos ponderados de la ecuación (17.8) desarrollada en la sección 17.3.

El **índice de producción industrial**, elaborado por el consejo de la Reserva Federal de Estados Unidos, es probablemente el índice de cantidad más conocido. Se publica mensualmente sobre un periodo base de 2002. Está diseñado para medir las variaciones en el volumen de los niveles de producción de una variedad de artículos clasificados, además de la minería y los servicios públicos. En febrero de 2009 el índice fue de 99.7.

Ejercicios

Métodos

AUTO evaluación

14. A continuación se listan los datos de las cantidades de tres artículos vendidos en 1995 y 2009, junto con sus precios de venta en 1995. Calcule el índice de cantidad agregado ponderado para 2009.

	Cantidad vendida		
Artículo	1995	2009	Precio unitario 1995 ($)
A	350	300	18.00
B	220	400	4.90
C	730	850	15.00

Aplicaciones

AUTO evaluación

15. Una compañía de transporte traslada cuatro productos básicos para un distribuidor en particular. Los embarques totales de mercancías en 1994 y 2009, así como los precios de 1994, se reportan en la tabla siguiente.

	Embarques		Precio/embarque
Mercancía	1994	2009	1994
A	120	95	$1 200
B	86	75	1 800
C	35	50	2 000
D	60	70	1 500

Desarrolle un índice de cantidad agregado ponderado con 1994 como año base. Opine sobre el aumento o la disminución de las cantidades durante el periodo 1994-2009.

16. Un concesionario de automóviles presentó un informe de ventas de 1992 y 2009 para los tres modelos listados en la tabla siguiente. Calcule las cantidades relativas y utilícelas para desarrollar un índice de cantidad agregado ponderado para 2009 con los datos de los dos años.

	Ventas		Precio medio por venta
Modelo	1992	2009	(1992)
Sedán	200	170	$15 200
Deportivo	100	80	17 000
Vagoneta	75	60	16 800

Resumen

Los índices de precio y cantidad son medidas importantes de los cambios en los niveles de precio y cantidad en el entorno de negocios y económico. Los precios relativos son simplemente la razón entre el precio unitario actual de un artículo respecto de un precio unitario del periodo base multiplicado por 100, con un valor de 100 que indica que no hay ninguna diferencia en el precio actual y el precio del periodo base. Los índices de precios agregados se crean como una medida compuesta del cambio general en los precios de determinado grupo de artículos o productos. Por lo general, los artículos en un índice de precios agregado son ponderados por la cantidad de uso. Un índice de precios agregado ponderado también se puede calcular al ponderar los precios relativos mediante las cantidades de uso de los artículos en el índice.

El índice de precios al consumidor y el índice de precios al productor son ampliamente citados y sus años base son 1982-1984 y 1982, respectivamente. El promedio industrial Dow Jones es otro índice de precios ampliamente referenciado. Es una suma ponderada de los precios de 30 acciones comunes de las grandes empresas. A diferencia de muchos otros índices, no se establece como porcentaje de un valor del periodo base.

A menudo, los índices de precios se utilizan para deflactar algunas series económicas que se miden en el tiempo. Se vio que el IPC puede ser utilizado para deflactar los salarios por hora a efecto de obtener un índice de salarios reales. La elección de los artículos que se incluyen en el índice, la selección de un periodo base y el ajuste por cambios en la calidad son importantes consideraciones adicionales en el desarrollo de un número índice. El índice de cantidad se estudió brevemente, y el de producción industrial se mencionó como un índice de cantidad importante.

Glosario

Índice de cantidad Índice diseñado para medir las variaciones en las cantidades en el tiempo.
Índice de Laspeyres Índice de precios agregado ponderado en el cual la ponderación de cada artículo es su cantidad en el periodo base.
Índice de Paasche Índice de precios agregado ponderado en el que la ponderación de cada artículo es la cantidad en el periodo actual.
Índice de precios agregado Índice de precios compuesto basado en los precios de un grupo de artículos.
Índice de precios agregado ponderado Índice de precios compuesto en el que los precios de los artículos son ponderados por su importancia relativa.
Índice de precios al consumidor (IPC) Índice de precios mensual que utiliza las variaciones del precio en la canasta de mercado de los bienes y servicios de consumo para medir los cambios en los precios al consumidor en el tiempo.
Índice de precios al productor (IPP) Índice de precios mensual diseñado para medir los cambios en los precios de los bienes vendidos en los mercados primarios (es decir, la primera compra de un producto en los mercados no minoristas).
Índice de producción industrial Índice de cantidad diseñado para medir las variaciones en el volumen físico o en los niveles de producción de los bienes industriales en el tiempo.
Precio relativo Índice de precios para un determinado artículo que se calcula dividiendo un precio unitario actual entre un precio unitario base y multiplicando el resultado por 100.
Promedios Down Jones Índice de precios agregado que permite conocer la tendencia de los precios y los movimientos en la Bolsa de Valores.

Fórmulas clave

Precio relativo en el periodo *t*

$$\frac{\text{Precio en el periodo } t}{\text{Precio en el periodo base}}(100) \qquad \textbf{(17.1)}$$

Índice de precios agregado no ponderado del periodo *t*

$$I_t = \frac{\Sigma P_{it}}{\Sigma P_{i0}}(100) \qquad \textbf{(17.2)}$$

Índice de precios agregado ponderado en el periodo *t*

$$I_t = \frac{\Sigma P_{it} Q_i}{\Sigma P_{i0} Q_i}(100) \qquad \textbf{(17.3)}$$

Promedio ponderado de los precios relativos

$$I_t = \frac{\sum \frac{P_{it}}{P_{i0}}(100) w_i}{\Sigma w_i} \qquad \textbf{(17.6)}$$

Factor de ponderación de la ecuación (17.6)

$$w_i = P_{i0} Q_i \qquad \textbf{(17.7)}$$

Índice de cantidad agregado ponderado

$$I_t = \frac{\Sigma Q_{it} w_i}{\Sigma Q_{i0} w_i}(100) \qquad \textbf{(17.9)}$$

Ejercicios complementarios

17. Los precios de venta promedio de viviendas unifamiliares nuevas para el periodo 2004-2007 son los siguientes (sitio web de Census Bureau, 19 de marzo de 2009).

Año	Precio ($ miles)
2004	221.0
2005	240.9
2006	246.5
2007	247.9

a) Utilice 2004 como año base y desarrolle un índice de precios para la vivienda nueva en este periodo de cuatro años.

b) Utilice 2005 como año base y desarrolle un índice de precios para la vivienda nueva en este periodo de cuatro años.

18. Nickerson Manufacturing Company reporta los siguientes datos de cantidades enviadas y costos unitarios para cada uno de cuatro productos.

Productos	Cantidades en el periodo base (2003)	Costo unitario medio del envío ($)	
		2003	2009
A	2000	10.50	15.90
B	5000	16.25	32.00
C	6500	12.20	17.40
D	2500	20.00	35.50

a) Calcule el precio relativo de cada producto.
b) Obtenga un índice de precios agregado ponderado que refleje la variación en el costo de los pedidos durante un periodo de cuatro años.

19. Con los datos del ejercicio 18 calcule un índice de Paasche para el costo de los envíos si las cantidades en 2009 son 4 000, 3 000, 7 500 y 3 000 para cada uno de los cuatro productos.

20. Boran Stockbrokers, Inc. selecciona cuatro acciones con el fin de desarrollar su propio índice de comportamiento del mercado accionario. A continuación se presentan los precios por acción de enero y marzo de 2009 para un periodo base de 2007. Las cantidades del año base se fijan sobre los volúmenes históricos de las cuatro acciones.

			Precio por acción ($)		
Acción	Industria	Cantidad en 2007	2007 Base	Enero 2009	Marzo 2009
A	Petrolera	100	31.50	22.75	22.50
B	De la computación	150	65.00	49.00	47.50
C	Siderúrgica	75	40.00	32.00	29.50
D	Bienes raíces	50	18.00	6.50	3.75

Utilice el periodo base de 2007 para calcular el índice Boran de enero y marzo de 2009. Opine acerca de qué señala el índice respecto de lo que está sucediendo en el mercado accionario.

21. Calcule los precios relativos de las cuatro acciones que componen el índice Boran del ejercicio 20. Utilice los agregados ponderados de los precios relativos para calcular los índices Boran de enero y marzo de 2009.

22. Considere la información siguiente de los precios relativos y cantidades de producción de granos en Iowa (sitio web de Census Bureau, 19 de marzo de 2009).

Producto	Cantidades en 1991 (millones de fanegas)	Precio base por fanega ($)	1991-2007 Precios relativos
Maíz	1 427	2.30	173.9
Soya	350	5.51	197.8

¿Cuál es el índice de precios agregado ponderado de 2007 para los granos en Iowa?

23. La lista siguiente reporta los precios y datos de cantidad de la fruta fresca para los años 1988 y 2007 (sitio web de Census Bureau, 19 de marzo de 2009). Los datos de cantidad reflejan el consumo per cápita en libras y los precios se indican por libra.

Fruta	Consumo per cápita 1988 (libras)	Precio 1988 ($/libras)	Precio 2007 ($/libras)
Plátano	24.3	0.41	0.53
Manzanas	19.9	0.71	1.12
Naranjas	13.9	0.56	0.91
Peras	3.2	0.64	1.27

a) Calcule el precio relativo de cada producto.
b) Obtenga un índice de precios agregado ponderado de estos productos. Opine acerca de la variación de precios en las frutas durante un periodo de 19 años.

24. Los salarios iniciales (base a nueve meses) de los profesores de administración de empresas en una de las principales universidades del medio oeste de Estados Unidos se presentan a continuación. Utilice el IPC para deflactar los datos de los sueldos a dólares constantes. Opine acerca de la tendencia de los sueldos en la educación superior indicada por estos datos.

Año	Sueldo inicial ($)	IPC (1982-1984 base)
1970	14 000	38.8
1975	17 500	53.8
1980	23 000	82.4
1985	37 000	107.6
1990	53 000	130.7
1995	65 000	152.4
2000	80 000	172.2
2005	110 000	195.3

25. Los cinco precios históricos de una determinada acción y el índice de precios al consumidor con un periodo base de 1982-1984 son los siguientes.

Año	Precio por acción ($)	IPC (1982-1984 base)
2004	51.00	188.9
2005	54.00	195.3
2006	58.00	201.6
2007	59.50	207.3
2008	59.00	215.3

Deflacte el precio de las acciones y comente sobre los aspectos de inversión de estas acciones.

26. Una importante empresa de manufactura presentó un informe de la cantidad y valor del producto para 2005 y 2009 en la tabla siguiente. Calcule un índice de cantidades agregado ponderado con los datos. Opine acerca de lo que significa este índice de cantidad.

	Cantidades		
Producto	2005	2009	Valores ($)
A	800	1200	30.00
B	600	500	20.00
C	200	500	25.00

CAPÍTULO 18

Análisis de series de tiempo y elaboración de pronósticos

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: OCCUPATIONAL HEALTH CLINIC DE NEVADA

18.1 PATRONES DE UNA SERIE DE TIEMPO
- Patrón horizontal
- Patrón de tendencia
- Patrón estacional
- Patrones estacional y de tendencia
- Componente cíclico
- Selección de un método de elaboración de pronósticos

18.2 EXACTITUD DEL PRONÓSTICO

18.3 PROMEDIOS MÓVILES Y SUAVIZAMIENTO EXPONENCIAL
- Promedios móviles
- Promedios móviles ponderados
- Suavizamiento exponencial

18.4 PROYECCIÓN DE LA TENDENCIA
- Regresión de tendencia lineal
- Suavizamiento exponencial lineal de Holt
- Regresión de tendencia no lineal

18.5 ESTACIONALIDAD Y TENDENCIA
- Estacionalidad sin tendencia
- Estacionalidad y tendencia
- Modelos basados en datos mensuales

18.6 DESCOMPOSICIÓN DE SERIES DE TIEMPO
- Cálculo de los índices estacionales
- Desestacionalización de una serie de tiempo
- Uso de una serie de tiempo desestacionalizada para identificar tendencias
- Ajustes estacionales
- Modelos basados en datos mensuales
- Patrón cíclico

ESTADÍSTICA *en* LA PRÁCTICA

OCCUPATIONAL HEALTH CLINIC*

SPARKS, NEVADA

La Occupational Health Clinic de Nevada es un centro médico de propiedad privada que se encuentra en Sparks, Nevada, y se especializa en medicina del trabajo. Ha operado en el mismo lugar por más de 20 años, y en el último bienio había registrado una fase de rápido crecimiento. La facturación mensual creció de $57 000 a más de $300 000 en 26 meses, cuando el edificio principal de la clínica se incendió.

La póliza de seguro de la unidad médica cubría la propiedad física y el equipo, así como la pérdida de ingresos totales debido a la interrupción de su funcionamiento normal. La reclamación del seguro de propiedad fue un asunto relativamente sencillo, ya que consistió en determinar el valor de la propiedad física y del equipo que se perdió durante el incendio. Sin embargo, determinar el valor de la pérdida de ingresos durante los siete meses que se tardó en reconstruir el edificio era un tema complejo, que requirió negociaciones entre los propietarios y la compañía de seguros. No hubo reglas prestablecidas que pudieran ayudar a calcular "lo que hubiera sucedido" con la facturación de la clínica si el incendio no se hubiera producido.

Para estimar la pérdida de ingresos, la clínica utilizó un método de elaboración de pronósticos para proyectar el crecimiento que habrían registrado los ingresos durante el periodo de siete meses de pérdida de negocio. La historia real de la facturación antes del incendio sirvió como base para un modelo de elaboración de pronósticos de tendencia lineal y patrones estacionales como los que se discuten en el presente capítulo. Este modelo de elaboración de pronósticos permitió a la clínica establecer una estimación precisa de la pérdida, que fue aceptada finalmente por la compañía de seguros.



Una médico de la Occupational Health Clinic de Nevada checa la presión arterial de una paciente. © Bob Pardue–Medical Lifestyle/Alamy.

* Agradecemos a los autores Bard Betz, director de Operaciones, y a Curtis Brauer, asistente ejecutivo administrativo, de Occupational Health Clinic de Nevada, por proporcionar este artículo para *Estadística en la práctica.*

Un pronóstico no es más que una predicción de lo que sucederá en el futuro. Los gerentes deben aprender a aceptar que, independientemente de la técnica que se utilice, no podrán tener pronósticos perfectos.

El propósito de este capítulo es presentar el análisis de series de tiempo y de elaboración de pronósticos. Suponga que se le ha solicitado preparar los pronósticos trimestrales de ventas de cada uno de los productos de la empresa para el próximo año. Los programas de producción, compra de materias primas, las políticas de inventarios y el monto de las ventas se verán afectados por el pronóstico trimestral que proporcione. En consecuencia, un pronóstico deficiente puede dar lugar a una mala planeación y a incrementar los costos para la empresa. ¿Cómo se debe proceder para obtener un pronóstico trimestral del volumen de ventas? Un buen criterio, intuición y estar concientes de la situación de la economía pueden dar una idea aproximada o "una sensación" de lo que es probable que suceda en el futuro, pero convertir esa sensación en un número que sea utilizado como el prónostico de ventas para el próximo año es difícil.

Los métodos de elaboración de pronósticos se pueden clasificar como cualitativos o cuantitativos. Los primeros implican la necesidad del criterio de expertos para obtener los pronósticos. Dichos métodos son apropiados cuando los datos históricos de la variable a pronosticar no apliquen o no estén disponibles. Los métodos cuantitativos se pueden utilizar cuando 1) la información del pasado acerca de la variable que se desea pronosticar esté disponible; 2) la información pueda cuantificarse, y 3) sea razonable suponer que el patrón del pasado continúe en

el futuro. En estos casos, los pronósticos se pueden obtener mediante un método de series de tiempo o un método causal. Este capítulo se centrará exclusivamente en los métodos de elaboración de pronósticos cuantitativos.

Si los datos históricos se limitan a los valores pasados de la variable que se pronostica, al procedimiento de elaboración de pronósticos se le llama *método de series de tiempo*, y los datos históricos se refieren como una serie de tiempo. El objetivo del análisis de la serie de tiempo es descubrir un patrón en los datos históricos o de series de tiempo para después extrapolar el modelo al futuro. El pronóstico se basa únicamente en los valores pasados de la variable o en los errores de pronóstico del pasado.

Los métodos de elaboración de pronósticos causales se basan en el supuesto de que la variable a pronosticar tiene una relación de causa y efecto con una o más variables. En el estudio del análisis de regresión de los capítulos 14, 15 y 16 se mostró cómo una o más variables independientes podrían ser utilizadas para pronosticar el valor de una sola variable dependiente. En cuanto a los análisis de regresión como una herramienta de elaboración de pronósticos, se puede observar el valor de la serie de tiempo que se desea pronosticar como la variable dependiente. Por tanto, si se identifica un buen conjunto de variables independientes relacionadas o explicativas, podemos desarrollar una ecuación de regresión y predecir la serie de tiempo. Por ejemplo, las ventas de muchos productos están influidas por los gastos de publicidad, por lo que el análisis de regresión sirve para desarrollar una ecuación que muestre cómo las ventas y la publicidad están relacionadas. Una vez que se determina el presupuesto de publicidad para el siguiente periodo, se podría sustituir este valor en la ecuación y obtener una predicción para el volumen de ventas de ese periodo. Observe que si se utiliza un método de series de tiempo para obtener el pronóstico, los gastos de publicidad no serían considerados, es decir, en este método el pronóstico se basa únicamente en las ventas del pasado.

Al tratar el tiempo como variable independiente y la serie de tiempo como una variable dependiente, el análisis de regresión también puede utilizarse como un método de series de tiempo. Para diferenciar la aplicación del análisis de regresión en estos dos casos, se utilizan los términos *regresión de corte transversal* y *regresión de series de tiempo*. Por tanto, la regresión de series de tiempo se refiere al uso del análisis de regresión cuando la variable independiente es el tiempo. Debido a que este capítulo se enfoca en los métodos de series de tiempo, se deja la discusión acerca de la aplicación del análisis de regresión como un método de elaboración de pronósticos causal a textos más avanzados sobre la materia.

18.1

Patrones de una serie de tiempo

WEB archivo

Gasoline

TABLA 18.1

Serie de tiempo de las ventas de gasolina

Semana	Ventas (miles de galones)
1	17
2	21
3	19
4	23
5	18
6	16
7	20
8	18
9	22
10	20
11	15
12	22

Una **serie de tiempo** es una secuencia de observaciones en una variable que se mide en puntos sucesivos en el tiempo o sobre un periodo sucesivo. Las medidas pueden ser tomadas cada hora, día, semana, mes o año, o en cualquier otro intervalo regular.[1] El patrón de datos es un factor importante en la comprensión de cómo las series de tiempo se han comportado en el pasado. Si se espera que tal comportamiento continúe en el futuro, se puede utilizar el patrón anterior como guía en la selección de un método de elaboración de pronósticos adecuado.

Para identificar los datos del patrón subyacente, un primer paso útil es construir una **gráfica de series de tiempo**. Ésta es una representación gráfica de la relación entre el tiempo y las variables de serie de tiempo: el tiempo está en el eje horizontal y los valores de la serie de tiempo en el eje vertical. Se revisarán algunos tipos comunes de patrones de datos que se pueden identificar al examinar una gráfica de series de tiempo.

Patrón horizontal

Un **patrón horizontal** se presenta cuando los datos fluctúan alrededor de una media constante. Para ilustrar una serie de tiempo con un patrón horizontal, observe los datos de la ta-

[1] El estudio se limitará a las series de tiempo en las que los valores de las series sean medidos en intervalos iguales. Los casos en los que las observaciones se realizan en intervalos desiguales quedan fuera del alcance de este libro.

FIGURA 18.1 Gráfica de la serie de tiempo de las ventas de gasolina



WEB archivo
GasolineRevised

bla 18.1. Estos datos muestran el número de galones de gasolina vendidos por un distribuidor en Bennington, Vermont, en las últimas 12 semanas. El valor medio o promedio para esta serie de tiempo es 19.25 o 19 250 galones por semana. La figura 18.1 muestra una gráfica de serie de tiempo para estos datos. Observe cómo éstos fluctúan alrededor de una media muestral de 19 250 galones. Aunque la variabilidad aleatoria está presente, se diría que estos datos siguen un patrón horizontal.

El concepto de **series de tiempo estacionarias**[2] designa una serie de tiempo cuyas propiedades estadísticas son independientes del tiempo. Esto significa, en particular, que

1. El proceso de generación de los datos tiene una media constante.
2. La variabilidad de la serie de tiempo es constante en el tiempo.

Una gráfica para una serie de tiempo estacionaria exhibe siempre un patrón horizontal. Pero la sola observación de un patrón horizontal no es evidencia suficiente para concluir que la serie de tiempo sea estacionaria. Los libros más avanzados sobre elaboración de pronósticos estudian los procedimientos para determinar si una serie de tiempo es estacionaria y proporcionan métodos para transformarla de no estacionaria en estacionaria.

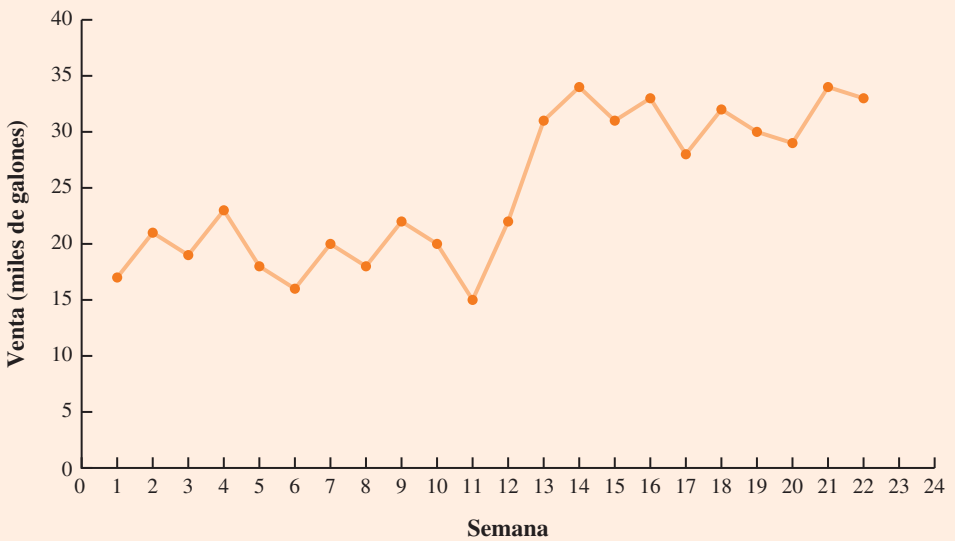
Los cambios en las condiciones de negocios a menudo pueden dar lugar a que una serie de tiempo que tiene un patrón horizontal cambie a un nuevo nivel. Por ejemplo, suponga que un distribuidor firma un contrato con el Departamento de Policía de Vermont para proveer de gasolina a los automóviles de la policía local ubicados al sur del estado. Con este nuevo contrato el distribuidor espera tener un gran incremento en las ventas semanales a partir de la semana 13. La tabla 18.2 muestra el número de galones de gasolina que se venden para la serie de tiempo original y para las 10 semanas después de firmar el nuevo contrato. La figura 18.2 muestra la gráfica correspondiente de la serie de tiempo. Observe el aumento en el nivel de la serie de tiempo a partir de la semana 13. Este cambio hace más difícil elegir un método de elaboración de pronósticos adecuado. La selección de un método que se adapte bien a los cambios en el nivel de una serie de tiempo es una consideración importante en muchas aplicaciones prácticas.

TABLA 18.2
Serie de tiempo de las ventas de gasolina después de obtener el contrato con la policía de Vermont

Semana	Ventas (miles de galones)
1	17
2	21
3	19
4	23
5	18
6	16
7	20
8	18
9	22
10	20
11	15
12	22
13	31
14	34
15	31
16	33
17	28
18	32
19	30
20	29
21	34
22	33

[2] Para una definición formal de series de tiempo estacionarias, remítase a G. E. P, Box, G. M. Jenkins y G. C. Reinsell, *Time series analysis: forecasting and control* (*Análisis de series de tiempo: pronóstico y control*), 3a. ed., Englewood Cliffs, NJ, Prentice Hall, 1994, p. 23.

FIGURA 18.2 Gráfica de series de tiempo de las ventas de gasolina después de obtener el contrato con la policía de Vermont



Patrón de tendencia

WEB archivo
Bicycle

Aunque los datos de las series de tiempo presentan fluctuaciones aleatorias, estas series también pueden mostrar cambios o movimientos graduales hacia valores relativamente mayores o menores durante un periodo. Si una gráfica de series de tiempo muestra este tipo de comportamiento, se dice que existe un **patrón de tendencia**. La tendencia, por lo general, es el resultado de factores a largo plazo, como el aumento o disminución de la población o la variación de sus características demográficas, la tecnología y/o preferencias de los consumidores, etcétera.

Para ilustrar una serie de tiempo con un patrón de tendencia, considere las series de tiempo de ventas de bicicletas de un fabricante en particular en los últimos 10 años, como se muestra en la tabla 18.3 y la figura 18.3. Observe que en el primer año se vendieron 21 600 bicicletas, en el segundo 22 900, y así sucesivamente. En el año 10, el último año, se han vendido 31 400 bicicletas. La inspección visual de la gráfica de las serires de tiempo permite apreciar algunos movimientos ascendentes y descendentes en los últimos 10 años, pero la serie de tiempo también parece tener una tendencia sistemática de aumento o disminución.

TABLA 18.3
Serie de tiempo de las ventas de bicicletas

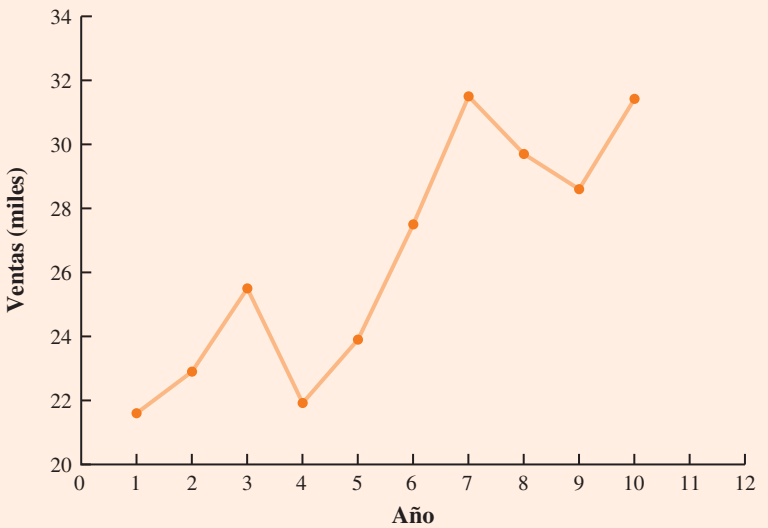
Año	Ventas (miles)
1	21.6
2	22.9
3	25.5
4	21.9
5	23.9
6	27.5
7	31.5
8	29.7
9	28.6
10	31.4

La tendencia en la serie de tiempo de las ventas de bicicletas parece ser lineal y creciente con el tiempo, pero a veces una tendencia se puede describir mejor por otros tipos de patrones. Por ejemplo, los datos en la tabla 18.4 y la gráfica correspondiente a la serie de tiempo de la figura 18.4 muestran las ventas de un medicamento contra el colesterol, dado que la empresa obtuvo la aprobación de la FDA hace 10 años. La serie de tiempo se incrementó de una manera no lineal, es decir, la tasa de variación de los ingresos no aumentó en una cantidad constante de un año a otro. De hecho, los ingresos parecen estar creciendo de manera exponencial. Las relaciones exponenciales de este tipo son apropiadas cuando la variación porcentual de un periodo a otro es relativamente constante.

Patrón estacional

La tendencia de una serie de tiempo se puede identificar con el análisis de las variaciones multianuales en los datos históricos. Los patrones estacionales son reconocidos al identificarse los mismos patrones de repetición en periodos sucesivos. Por ejemplo, un fabricante de albercas espera tener pocas ventas en los meses de otoño e invierno, y aumentarlas en los meses de primavera y verano. Los fabricantes de equipos de remoción de nieve y de ropa de invierno,

FIGURA 18.3 Gráfica de la serie de tiempo de las ventas de bicicletas



WEB archivo
Cholesterol

sin embargo, prevén exactamente lo contrario. Como era de esperar, el patrón de una gráfica de series de tiempo que tiene un comportamiento repetitivo en un periodo de un año debido a la influencia estacional se llama **patrón estacional**. Aunque por lo general se considera que las variaciones estacionales son aquellas que se representan en un lapso de un año, los datos de series de tiempo también pueden presentar patrones estacionales de menos de un año. Por ejemplo, el volumen de tráfico diario muestra en un día un comportamiento "estacional", donde los valores máximos se presentan en las horas pico, un flujo moderado el resto del día y al comienzo de la noche, y un flujo ligero desde la medianoche hasta la madrugada.

TABLA 18.4
Serie de tiempo de ingresos por medicamentos contra el colesterol ($ millones)

Año	Ingresos
1	23.1
2	21.3
3	27.4
4	34.6
5	33.8
6	43.2
7	59.5
8	64.4
9	74.2
10	99.3

Como ejemplo de un patrón estacional, considere el número de sombrillas vendidas en una tienda de ropa en los últimos cinco años. La tabla 18.5 muestra la serie de tiempo con los datos de año (Year), trimestre (Quarter) y ventas (Sales), y la figura 18.5 ilustra la gráfica correspondiente. La gráfica de una serie de tiempo no indica ninguna tendencia a largo plazo en las ventas. De hecho, a menos que observe cuidadosamente los datos, es posible concluir que éstos siguen un patrón horizontal. Pero una inspección más cercana revela un patrón regular en los datos. Es decir, el primer y tercer trimestre presentan ventas moderadas, el segundo trimestre tiene ventas más altas, y el cuarto trimestre tiende a tener el menor volumen de ventas. Por tanto, se concluye que existe un patrón estacional trimestral.

Patrones de tendencia y estacional

Algunas series de tiempo son una combinación de un patrón de tendencia y estacional. Por ejemplo, los datos de la tabla 18.6 y la gráfica correspondiente de las series de tiempo en la figura 18.6 muestran las ventas (Sales) de televisores por trimestre (Quarter) y año (Year) de un fabricante en particular en los últimos cuatro años. Claramente se presenta una tendencia creciente. Sin embargo, la figura 18.6 indica también que las ventas son menores en el segundo trimestre de cada año y que aumentan a partir de los trimestres 3 y 4. Por tanto, se llega a la conclusión de que un patrón estacional también está presente en las ventas de televisores. En estos casos se utiliza un método de elaboración de pronósticos que tiene la capacidad para tratar la tendencia y la estacionalidad.

Patrón cíclico

El **patrón cíclico** existe si la gráfica de la serie de tiempo muestra una secuencia de puntos que caen de manera alterna por arriba y debajo de la línea de tendencia por más de un año. Muchas

FIGURA 18.4 Gráfica de la serie de tiempo de las ventas de medicamentos contra el colesterol ($ millones)

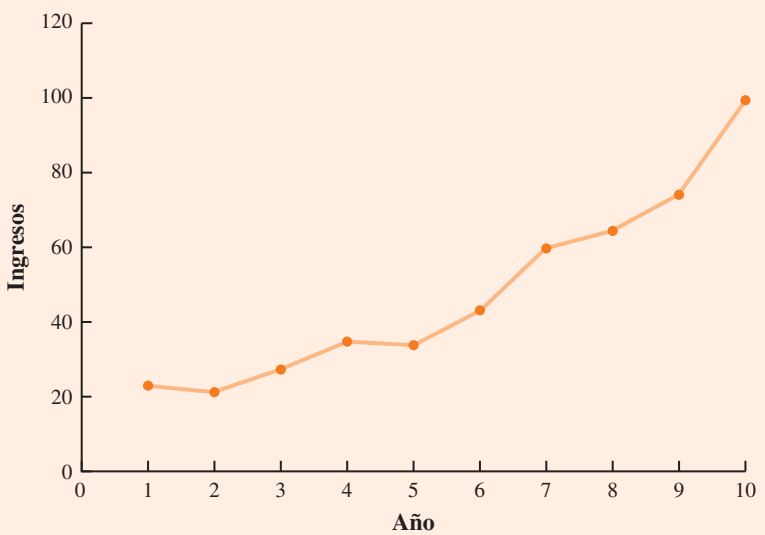


TABLA 18.5 Serie de tiempo de las ventas de sombrillas

WEB archivo
Umbrella

Year	Quarter	Sales
1	1	125
	2	153
	3	106
	4	88
2	1	118
	2	161
	3	133
	4	102
3	1	138
	2	144
	3	113
	4	80
4	1	109
	2	137
	3	125
	4	109
5	1	130
	2	165
	3	128
	4	96

series de tiempo económicas suelen mostrar un comportamiento cíclico con observaciones regulares que caen por debajo y por encima de la línea de tendencia. A menudo, el patrón cíclico se debe a ciclos multianuales de la economía. Por ejemplo, periodos de inflación moderada seguidos por periodos de inflación rápida pueden dar lugar a que la serie de tiempo alterne hacia arriba y hacia abajo de la línea general de tendencia creciente (por ejemplo, una serie de tiempo sobre el costo de vivienda). Los ciclos económicos son extremadamente difíciles, si no es que

FIGURA 18.5 Gráfica de series de tiempo de las ventas de sombrillas

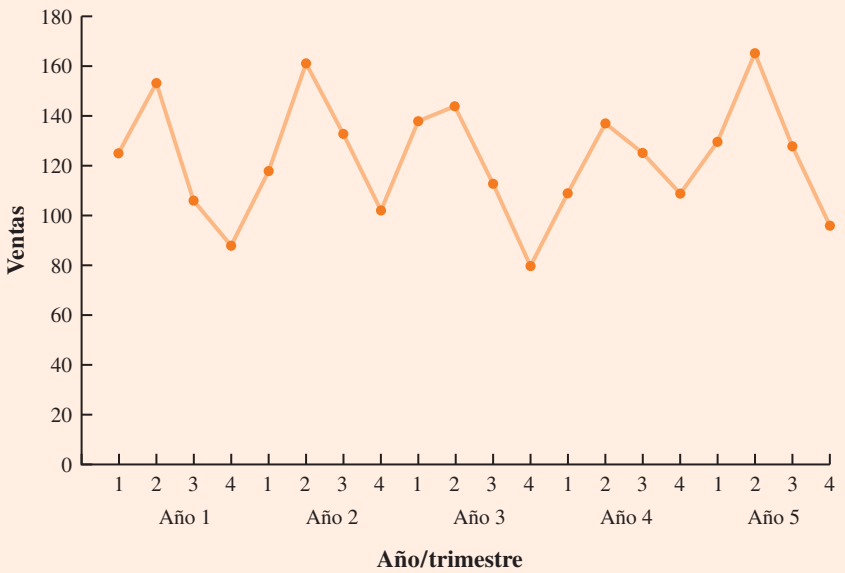


TABLA 18.6 Serie de tiempo de las ventas de televisores

WEB archivo
TVSales

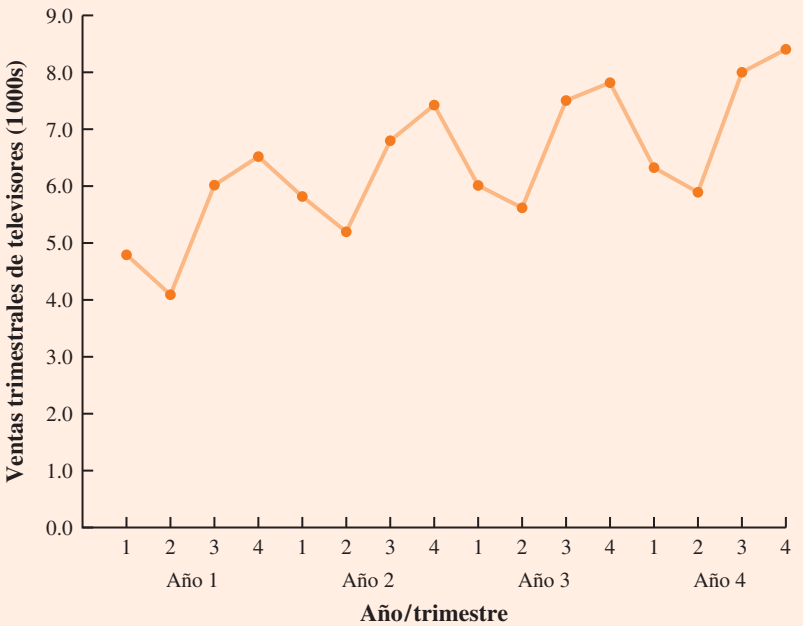
Year	Quarter	Sales (1 000s)
1	1	4.8
	2	4.1
	3	6.0
	4	6.5
2	1	5.8
	2	5.2
	3	6.8
	4	7.4
3	1	6.0
	2	5.6
	3	7.5
	4	7.8
4	1	6.3
	2	5.9
	3	8.0
	4	8.4

imposibles de predecir. Como resultado, los efectos cíclicos a menudo se combinan con efectos de tendencia a largo plazo y se conocen como *efecto de tendencia-cíclico*. Este capítulo no trata de los efectos cíclicos que puedan presentarse en las series de tiempo.

Selección de un método de elaboración de pronósticos

El patrón subyacente en la serie de tiempo es un factor importante en la selección de un método de elaboración de pronósticos. Por tanto, la gráfica correspondiente debe ser una de las primeras tareas a desarrollar cuando se trate de determinar qué método de elaboración de pronósticos utilizar. Si se observa un patrón horizontal, entonces tenemos que seleccionar un método apropiado para este tipo de patrón. Del mismo modo, si se observa una tendencia en los datos, en-

FIGURA 18.6 Gráfica de la serie de tiempo del conjunto de ventas trimestrales de televisores



tonces se debe utilizar un método de elaboración de pronósticos con la capacidad para manejar la tendencia con efectividad. Las siguientes dos secciones ilustran los métodos que se pueden utilizar en situaciones donde el patrón subyacente es horizontal, es decir, no están presentes los efectos de tendencia o estacionales. Después se considerarán los métodos apropiados cuando la tendencia y/o la estacionalidad están presentes en los datos.

18.2 Exactitud del pronóstico

Esta sección se inicia con la elaboración de pronósticos de la serie de tiempo de las ventas de gasolina mostradas en la tabla 18.1, utilizando el más simple de todos los métodos de elaboración de pronósticos: uno que utiliza el volumen de la última semana de ventas como predictor de la siguiente semana. Por ejemplo, un distribuidor vendió 17 mil galones de gasolina en la semana 1; este valor se utiliza como el pronóstico para la semana 2. Después, se utiliza 21, el valor real de las ventas en la semana 2, como el pronóstico para la semana 3, y así sucesivamente. Las predicciones obtenidas para los datos históricos con este método se muestran en la columna Pronóstico en la tabla 18.7. Debido a su sencillez, a este enfoque se le denomina a menudo método de elaboración de pronósticos *ingenuo*.

¿Qué tan exactos son los pronósticos al utilizar el método *ingenuo*? Para responder esta pregunta se presentan varias medidas de exactitud para los pronósticos. Estas medidas se utilizan para determinar qué tan bien un método particular es capaz de reproducir los datos de las series de tiempo que están disponibles. Al seleccionar el enfoque que tiene la mejor exactitud de los datos ya conocidos, se espera que aumente la probabilidad de obtener un mejor pronóstico para periodos futuros.

El concepto clave relacionado con la medida de exactitud del pronóstico es el **error de pronóstico**, definido como

$$\text{Error de pronóstico} = \text{valor real} - \text{pronóstico}$$

TABLA 18.7 Cálculos y medidas de exactitud de pronósticos utilizando el valor más reciente como pronóstico para el próximo periodo

Semana	Valor de la serie de tiempo	Pronóstico	Error de pronóstico	Valor absoluto del error de pronóstico	Error de pronóstico cuadrado	Error porcentual	Valor absoluto del error porcentual
1	17						
2	21	17	4	4	16	19.05	19.05
3	19	21	−2	2	4	−10.53	10.53
4	23	19	4	4	16	17.39	17.39
5	18	23	−5	5	25	−27.78	27.78
6	16	18	−2	2	4	−12.50	12.50
7	20	16	4	4	16	20.00	20.00
8	18	20	−2	2	4	−11.11	11.11
9	22	18	4	4	16	18.18	18.18
10	20	22	−2	2	4	−10.00	10.00
11	15	20	−5	5	25	−33.33	33.33
12	22	15	7	7	49	31.82	31.82
		Totales	5	41	179	1.19	211.69

Por ejemplo, debido a que el distribuidor vendió en realidad 21 mil galones de gasolina en la semana 2 y el pronóstico al utilizar el volumen de ventas en la semana 1 fue de 17 mil galones, el error de pronóstico en la semana 2 es

$$\text{Error de pronóstico en la semana 2} = 21 - 17 = 4$$

El hecho de que el error de pronóstico sea positivo, indica que en la semana 2 el método de elaboración de pronósticos subestimó el valor real de las ventas. A continuación utilice 21, el valor real de las ventas en la semana 2, como pronóstico para la semana 3. Ya que el valor real de las ventas en la semana 3 es 19, el error de pronóstico para esta semana es $19 - 21 = -2$. En este caso, el error negativo indica que en la semana 3 el pronóstico sobrestimó el valor real. Así, el error de pronóstico puede ser positivo o negativo dependiendo de si es demasiado bajo o demasiado alto. Un resumen completo de los errores de pronóstico para este método ingenuo se muestra en la tabla 18.7, en la columna Error de pronóstico.

En el análisis de regresión, un residual se define como la diferencia entre el valor observado y el valor estimado de la variable dependiente. Los errores de pronóstico son análogos a los residuales en el análisis de regresión.

Una medida sencilla de exactitud de los pronósticos es la media o promedio de errores de pronóstico. La tabla 18.7 muestra que la suma de estos errores para la serie de tiempo de las ventas de gasolina es 5, por lo que la media o promedio del error de pronóstico es $5/11 = 0.45$. Observe que aunque la serie de tiempo de gasolina se compone de 12 valores, al calcular la media del error se divide la suma de los errores entre 11, ya que existen solamente 11 errores de pronóstico. Debido a que la media del error de pronóstico es positiva, el método arroja pronósticos bajos; es decir, los valores observados tienden a ser mayores que los pronosticados. Debido a que los errores de pronóstico positivos y negativos tienden a compensarse entre sí, es probable que la media del error sea pequeña, así que ésta no es una medida muy útil para la exactitud del pronóstico.

El **error absoluto medio**, que se denota EAM, es una medida de exactitud del pronóstico que evita el problema de los errores positivos y negativos que se compensan entre sí. Como es de esperar, dado su nombre, EAM es el promedio de los valores absolutos de los errores de pronóstico. La tabla 18.7 muestra que la suma de los valores absolutos de los errores de pronóstico es 41; por tanto

$$\text{EAM} = \text{promedio del valor absoluto de los errores de pronóstico} = \frac{41}{11} = 3.73$$

En el análisis de regresión el cuadrado medio debido al error (CME) o error cuadrático medio es la suma de los residuales cuadrados dividida entre sus grados de libertad. En el pronóstico, el CME es el promedio de la suma de los errores de pronóstico cuadrados.

Otra medida que evita el problema de los errores de pronóstico positivos y negativos que se compensan entre sí se obtiene al calcular el promedio de los errores de pronóstico cuadrados. Esta medida de exactitud en los pronósticos se llama **cuadrado medio debido al error**, denotado como CME, o error cuadrático medio. En la tabla 18.7, la suma de los cuadrados debido al error es 179: por tanto,

$$\text{CME} = \text{promedio de la suma de los errores de pronóstico cuadrados } \frac{179}{11} = 16.27$$

El tamaño del EAM y del CME depende de la escala de los datos. Como resultado, es difícil hacer comparaciones de los distintos intervalos de tiempo, como la de un método de pronósticos de ventas mensuales de gasolina con un método de elaboración de pronósticos de ventas semanal, o hacer comparaciones de las distintas series de tiempo. Para hacer comparaciones como éstas se debe trabajar con las medidas relativas o porcentuales de los errores. El **error porcentual absoluto medio**, denotado como EPAM, es una medida de este estilo. Para calcular el EPAM, en primer lugar se debe determinar el error porcentual de cada pronóstico. Por ejemplo, el error porcentual que corresponde al pronóstico de 17 en la semana 2 se calcula dividiendo el error de pronóstico en la semana 2 entre el valor real en la semana 2 y multiplicando el resultado por 100. Para esta semana, el error porcentual se calcula de la siguiente manera.

$$\text{Error porcentual para la semana 2} = \frac{4}{21}(100) = 19.05\%$$

Por tanto, el error de pronóstico para la semana 2 es 19.05% del valor observado en tal semana. Un resumen completo de los errores porcentuales se muestra en la tabla 18.7, en la columna Error porcentual. En la siguiente columna se muestran los valores absolutos de este porcentaje.

La tabla 18.7 indica que la suma de los valores absolutos de los errores porcentuales es 211.69, por lo que

$$\text{EPAM} = \text{promedio del valor absoluto de los errores porcentuales de pronóstico} = \frac{211.69}{11} = 19.24\%$$

En resumen, al utilizar el método de elaboración de pronósticos ingenuo (la más reciente observación), se obtuvieron las siguientes medidas de exactitud del pronóstico.

$$\begin{aligned}\text{EAM} &= 3.73\\ \text{CME} &= 16.27\\ \text{EPAM} &= 19.24\%\end{aligned}$$

Estas medidas de exactitud miden simplemente qué tan bien el método de elaboración de pronósticos es capaz de predecir los valores históricos de las series de tiempo. Ahora, suponga que se desea predecir las ventas para un periodo futuro, como la semana 13. En este caso, el pronóstico es 22, el valor real de las series de tiempo en la semana 12. ¿Es ésta una estimación exacta de ventas para la semana 13? Desafortunadamente no hay manera de abordar el tema de la exactitud relacionada con el pronóstico para periodos futuros. Pero si se elige un método de elaboración de pronósticos que funcione bien para los datos históricos, y se piensa que el patrón histórico continuará en el futuro, se deben obtener resultados que, en última instancia, probaron ser buenos.

Antes de concluir esta sección, se considerará otro método de elaboración de pronósticos para las series de tiempo en las ventas de gasolina de la tabla 18.1 Suponga que se utiliza el promedio de todos los datos históricos disponibles como pronóstico para el próximo periodo. Comience por elaborar un pronóstico para la semana 2. Ya que existe sólo un valor histórico disponible antes de la semana 2, el pronóstico para ésta es sólo el valor de la serie de tiempo para la semana 1, por lo que el pronóstico es de 17 mil galones de gasolina. Para calcular el pronóstico de la semana 3 se toma el promedio de los valores de ventas en las semanas 1 y 2. Por tanto, el resultado que se obtiene es el que se indica a continuación.

TABLA 18.8 Cálculo y medidas de exactitud del pronóstico al utilizar el promedio de todos los datos históricos como pronóstico del próximo periodo

Semana	Valor de la serie de tiempo	Pronóstico	Error de pronóstico	Valor absoluto del error de pronóstico	Error de pronóstico cuadrado	Error porcentual	Valor absoluto del error porcentual
1	17						
2	21	17.00	4.00	4.00	16.00	19.05	19.05
3	19	19.00	0.00	0.00	0.00	0.00	0.00
4	23	19.00	4.00	4.00	16.00	17.39	17.39
5	18	20.00	−2.00	2.00	4.00	−11.11	11.11
6	16	19.60	−3.60	3.60	12.96	−22.50	22.50
7	20	19.00	1.00	1.00	1.00	5.00	5.00
8	18	19.14	−1.14	1.14	1.31	−6.35	6.35
9	22	19.00	3.00	3.00	9.00	13.64	13.64
10	20	19.33	0.67	0.67	0.44	3.33	3.33
11	15	19.40	−4.40	4.40	19.36	−29.33	29.33
12	22	19.00	3.00	3.00	9.00	13.64	13.64
		Totales	4.53	26.81	89.07	2.76	141.34

$$\text{Pronóstico para la semana 3} = \frac{17 + 21}{2} = 19$$

De la misma forma, el pronóstico para la semana 4 indica,

$$\text{Pronóstico para la semana 4} = \frac{17 + 21 + 19}{3} = 19$$

Los pronósticos obtenidos al utilizar este método para las series de tiempo de las ventas de gasolina se muestran en la tabla 18.8, en la columna Pronóstico. Con estos resultados se obtuvieron los siguientes valores de EAM, CME y EPAM.

$$\text{EAM} = \frac{26.81}{11} = 2.44$$

$$\text{CME} = \frac{89.07}{11} = 8.10$$

$$\text{EPAM} = \frac{141.34}{11} = 12.85\%$$

Ahora se puede determinar la exactitud de los dos métodos de elaboración de pronósticos que se han considerado en esta sección mediante la comparación de los valores de EAM, CME y EPAM.

	Método ingenuo	Promedio de los valores pasados
EAM	3.73	2.44
CME	16.27	8.10
EPAM	19.24%	12.85%

Para cada medida, el promedio de los valores pasados proporciona pronósticos más precisos que al utilizar la observación más reciente como pronóstico para el próximo periodo. En general, si la serie de tiempo subyacente es estacionaria, el promedio de todos los datos históricos siempre proporcionará mejores resultados.

Pero suponga que la serie de tiempo subyacente no es estacionaria. En la sección 18.1 se menciona que las variaciones en las condiciones de negocios suelen dar lugar a una serie de tiempo con un patrón horizontal que cambia a un nuevo nivel. Se estudió una situación en la que el distribuidor de gasolina firmó un contrato con la policía del estado de Vermont para proveer de combustible a las patrullas de policías del sur del estado. La tabla 18.2 muestra el número de galones de gasolina que se vendieron para la serie de tiempo original y para las 10 semanas después de firmado el nuevo contrato, y la figura 18.2 presenta la gráfica que corresponde a las series de tiempo. Observe el cambio en el nivel de la semana 13 para la serie de tiempo resultante. Cuando ocurre este cambio, le toma tiempo al método de elaboración de pronósticos que utiliza el promedio de todos los datos históricos ajustarse a un nuevo nivel de series de tiempo. Pero en este caso el método ingenuo simple se ajusta muy rápidamente a los cambios en el nivel, debido a que utiliza la observación más reciente como pronóstico.

Las medidas de exactitud de los pronósticos son factores importantes en la comparación de distintos métodos de elaboración de pronósticos, pero se debe tener cuidado de no depender demasiado de ellas. El buen criterio y el conocimiento sobre las condiciones de negocios que puedan afectar el pronóstico también deben tomarse en cuenta cuidadosamente en la elección de un método. La exactitud de los pronósticos históricos no es la única consideración, sobre todo si es probable que la serie de tiempo cambie en el futuro.

En la siguiente sección se presentarán métodos más sofisticados para el desarrollo de los pronósticos de una serie de tiempo que muestren un patrón horizontal. Al utilizar las medidas de exactitud de los pronósticos desarrolladas aquí, se logrará determinar si dichos métodos proporcionan más exactitud a los pronósticos que la obtenida utilizando los enfoques sencillos que se ilustran en esta sección. Los métodos que se presentarán también tienen la ventaja de adaptarse a situaciones donde las series de tiempo cambian a un nuevo nivel. La capacidad de un método de pronósticos para adaptarse rápidamente a estos cambios es una consideración importante, especialmente en situaciones de elaboración de pronósticos a corto plazo.

Ejercicios

Métodos

AUTOevaluación

1. Considere los datos de las siguientes series de tiempo.

Semana	1	2	3	4	5	6
Valor	18	13	16	11	17	14

Utilizando el método ingenuo (el valor más reciente) como pronóstico para la semana próxima, calcule las siguientes medidas de exactitud de los pronósticos.

a) Error absoluto medio.
b) Cuadrado medio debido al error o error cuadrático medio.
c) Error porcentual absoluto medio.
d) ¿Cuál es el pronóstico para la semana 7?

AUTOevaluación

2. Consulte los datos de las series de tiempo del ejercicio 1. Utilice el promedio de todos los datos históricos como pronóstico para el próximo periodo y calcule las siguientes medidas de exactitud de los pronósticos.

a) Error absoluto medio.
b) Cuadrado medio debido al error o error cuadrático medio.
c) Error porcentual absoluto medio.
d) ¿Cuál es el pronóstico para la semana 7?

AUTO evaluación

3. En los ejercicios 1 y 2 se utilizaron distintos métodos de elaboración de pronósticos. ¿Cuál parece dar la mejor exactitud del pronóstico para los datos históricos? Explique.
4. Considere los datos siguientes de series de tiempo.

Mes	1	2	3	4	5	6	7
Valor	24	13	20	12	19	23	15

a) Calcule el valor del CME utilizando el valor más reciente como pronóstico para el periodo próximo. ¿Cuál es el pronóstico para el mes 8?
b) Calcule el valor del CME al utilizar el promedio de todos los datos disponibles como pronóstico para el siguiente periodo. ¿Cuál es el pronóstico para el mes 8?
c) ¿Qué método parece proveer el mejor pronóstico?

18.3 Promedios móviles y suavizamiento exponencial

En esta sección se estudiarán tres métodos de elaboración de pronósticos que son apropiados para una serie de tiempo de patrón horizontal: promedios móviles, promedios móviles ponderados y suavizamiento exponencial. Estos métodos también se adaptan bien a los cambios de nivel de un patrón horizontal como se observó en las series de tiempo de las ventas prolongadas de gasolina (tabla 18.2 y figura 18.2). Sin embargo, no funcionan muy bien sin alguna modificación cuando existen efectos importantes de tendencia, cíclicos o estacionales. Debido a que el objetivo de cada uno de estos métodos es "suavizar" las fluctuaciones aleatorias en las series de tiempo, se les conoce como *métodos de suavizamiento*. Son fáciles de utilizar y en general proporcionan un alto nivel de exactitud para pronósticos a corto plazo, como el del periodo siguiente.

Promedios móviles

El método de **promedios móviles** utiliza el promedio de los valores de los k datos más recientes de la serie de tiempo como pronóstico para el próximo periodo. En términos matemáticos, un pronóstico de promedio móvil de orden k es el siguiente.

PRONÓSTICO DE PROMEDIO MÓVIL DE ORDEN k

$$F_{t+1} = \frac{\sum(\text{los } k \text{ valores más recientes de los datos})}{k} = \frac{Y_t + Y_{t-1} + \ldots + Y_{t-k+1}}{k} \qquad (18.1)$$

donde

F_{t+1} = pronóstico de la serie de tiempo para el periodo $t + 1$
Y_t = valor real de la serie de tiempo en el periodo t

El término *móvil* se utiliza porque cada vez que en la serie de tiempo hay una nueva observación, ésta sustituye a la observación más antigua de la ecuación y se calcula un nuevo promedio. Como resultado, el promedio se modifica, o se mueve, conforme se disponga de una nueva observación.

Para ilustrar el método de los promedios móviles, regrese a los datos de las ventas de gasolina de la tabla 18.1 y de la figura 18.1. La gráfica de la figura 18.1 indica que la serie de tiempo de las ventas de gasolina tiene un patrón horizontal. Por tanto, se pueden aplicar los métodos de suavizamiento de esta sección.

Para utilizar los promedios móviles a efecto de pronosticar las series de tiempo, primero se debe seleccionar el orden, o el número de los valores de las series de tiempo que se incluirán en el promedio móvil. Si sólo los valores más recientes se consideran relevantes, es preferible utilizar un valor pequeño de k. Si existen valores más antiguos que se consideren relevantes, entonces es mejor un valor grande de k. Como se mencionó antes, una serie de tiempo con un patrón horizontal puede cambiar con el tiempo a un nuevo nivel. Un promedio móvil se adaptará al nuevo nivel y seguirá brindando pronósticos adecuados después de k periodos. Así, un valor menor de k hará un seguimiento más rápido en el cambio en una serie de tiempo, pero los valores mayores serán más eficaces para el suavizamiento de las fluctuaciones aleatorias en el tiempo. Así que el criterio de negocios basado en el entendimiento del comportamiento de una serie de tiempo es de gran ayuda en la elección de un buen valor de k.

Para ilustrar cómo los promedios móviles pueden utilizarse para pronosticar las ventas de gasolina, se utilizará un promedio móvil de tres semanas ($k = 3$). Se comienza por calcular el pronóstico de ventas en la semana 4 con la media de los valores de la serie de tiempo en las semanas 1 a 3.

$$F_4 = \text{promedio de las semanas 1 a 3} = \frac{17 + 21 + 19}{3} = 19$$

Por tanto, el pronóstico del promedio móvil de ventas en la semana 4 es 19 o 19 mil galones de gasolina. Debido a que el valor real observado en esta semana es 23, el error de pronóstico en la semana 4 es $23 - 19 = 4$.

A continuación se calcula el pronóstico de ventas en la semana 5 al promediar los valores de la serie de tiempo de las semanas 2 a 4.

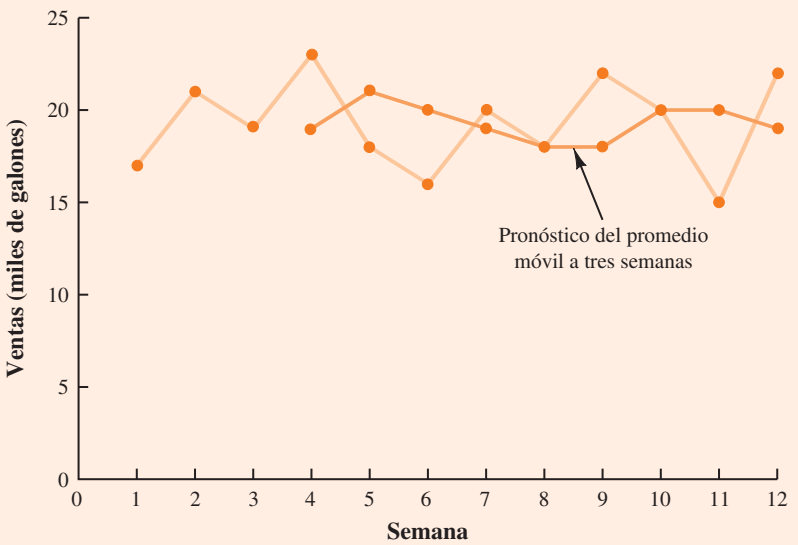
$$F_5 = \text{promedio de las semanas 2 a 4} = \frac{21 + 19 + 23}{3} = 21$$

Por tanto, el pronóstico de las ventas en la semana 5 es 21 y el error relacionado con este indicador es $18 - 21 = -3$. Un resumen completo del pronóstico del promedio móvil para las series de tiempo en las tres semanas de ventas de gasolina se proporciona en la tabla 18.9. La figura 18.7 muestra la gráfica de la serie de tiempo original y el pronóstico del promedio móvil de tres semanas. Observe cómo la gráfica de los pronósticos por promedio móvil ha tendido a suavizar las fluctuaciones aleatorias en la serie de tiempo.

TABLA 18.9 Resumen de los cálculos del promedio móvil para tres semanas

Semana	Valor de la serie de tiempo	Pronóstico	Error de pronóstico	Valor absoluto del error de pronóstico	Error de pronóstico cuadrado	Error porcentual	Valor absoluto del error porcentual
1	17						
2	21						
3	19						
4	23	19	4	4	16	17.39	17.39
5	18	21	−3	3	9	−16.67	16.67
6	16	20	−4	4	16	−25.00	25.00
7	20	19	1	1	1	5.00	5.00
8	18	18	0	0	0	0.00	0.00
9	22	18	4	4	16	18.18	18.18
10	20	20	0	0	0	0.00	0.00
11	15	20	−5	5	25	−33.33	33.33
12	22	19	3	3	9	13.64	13.64
		Totales	0	24	92	−20.79	129.21

FIGURA 18.7 Gráfica de series de tiempo de las ventas de gasolina y pronósticos del promedio móvil a tres semanas



Para pronosticar las ventas en la semana 13, el siguiente periodo en el futuro, se calcula simplemente el promedio de los valores de la serie de tiempo en las semanas 10, 11 y 12.

$$F_{13} = \text{promedio de las semanas 10 a 12} = \frac{20 + 15 + 22}{3} = 19$$

Por tanto, el pronóstico para la semana 13 es 19 o 19 mil galones de gasolina.

Exactitud del pronóstico En la sección 18.2 se estudiaron tres medidas de exactitud del pronóstico: EAM, CME y EPAM. Al utilizar los cálculos del promedio móvil de tres semanas de la tabla 18.9, los valores para estas tres medidas de exactitud del pronóstico son

$$\text{EAM} = \frac{24}{9} = 2.67$$

$$\text{CME} = \frac{92}{9} = 10.22$$

$$\text{EPAM} = \frac{129.21}{9} = 14.36\%$$

En situaciones donde es necesario comparar los métodos de elaboración de pronósticos para distintos periodos, son preferibles las medidas relativas como EPAM para comparar un pronóstico de ventas semanales con un pronóstico de ventas mensuales.

En la sección 18.2 también se mostró que al utilizar las observaciones más recientes como pronóstico para la siguiente semana (un promedio móvil de orden $k = 1$) dio como resultado los valores de EAM = 3.73, CME = 16.27 y EPAM = 19.24%. Así, en cada caso el método de promedio móvil para las tres semanas proporcionó pronósticos más exactos que el simple uso de la observación más reciente como pronóstico.

Para determinar si con un orden distinto de k se pueden obtener pronósticos más precisos con el promedio móvil, se recomienda el uso del método de prueba y error para determinar el valor de k que minimiza el CME. Para la serie de tiempo de ventas de gasolina se puede mostrar que el valor mínimo del CME corresponde a un promedio móvil de orden $k = 6$ con CME = 6.79. Si se está dispuesto a asumir que el orden del promedio móvil que es mejor para los datos

históricos también será mejor para los valores futuros de la serie de tiempo, el pronóstico para el promedio móvil más preciso en las ventas de gasolina se puede obtener utilizando un promedio móvil de orden $k = 6$.

Promedios móviles ponderados

Un pronóstico de promedio móvil de orden k = 3 es un caso especial del método de promedios móviles ponderados en el que cada peso es igual a 1/3.

En el método del promedio móvil, cada observación en los cálculos recibe el mismo peso. Una variante, conocida como **promedios móviles ponderados**, consiste en seleccionar un peso diferente para cada uno de los valores y después calcular el promedio ponderado de los k valores más recientes como pronóstico. En la mayoría de los casos la observación más reciente recibe el mayor peso, y los pesos se reducen para los datos más antiguos. Utilice la serie de tiempo de las ventas de gasolina para ilustrar el cálculo de un promedio móvil ponderado de tres semanas. Asigne un peso de 3/6 a la observación más reciente, un peso de 2/6 a la segunda observación más reciente, y un peso de 1/6 a la tercera observación más reciente. Utilizando este promedio ponderado, el pronóstico para la semana 4 se calcula como sigue.

$$\text{Pronóstico para la semana 4} = {}^{1}\!/_{6}(17) + {}^{2}\!/_{6}(21) + {}^{3}\!/_{6}(19) = 19.33$$

Observe que en el método del promedio móvil ponderado la suma de los pesos es igual a 1.

Exactitud del pronóstico Para utilizar el método de promedios móviles ponderados, primero debe seleccionar la cantidad de valores que se incluirán en el promedio móvil ponderado y después elegir los pesos para cada uno de los valores. En general, si se cree que el pasado reciente es un mejor predictor del futuro que el pasado distante, habrá que asignar pesos mayores a las observaciones más recientes. Sin embargo, si la serie de tiempo es muy variable, puede ser mejor elegir pesos aproximadamente iguales para todos los datos. El único requisito en la selección de los pesos es que su suma debe ser igual a 1. Para estimar si con una determinada combinación de cantidad de datos y de pesos se obtiene un pronóstico más preciso que con otra combinación, se recomienda utilizar el CME como medida de exactitud del pronóstico. Es decir, si se supone que la combinación que es mejor para el pasado también será mejor para el futuro, se utilizará la combinación del número de valores y pesos que minimice el CME de la serie de tiempo histórica para pronosticar el siguiente valor en la serie de tiempo.

Suavizamiento exponencial

Existen varios procedimientos de suavizamiento exponencial. El método que aquí se presenta se refiere a menudo como suavizamiento exponencial sencillo. En la siguiente sección se muestra cómo un suvizamiento exponencial que utiliza dos constantes de suavizamiento puede ser utilizado para pronosticar una serie de tiempo con tendencia lineal.

El **suavizamiento exponencial** también utiliza un promedio ponderado de los valores pasados de la serie de tiempo como pronóstico; es un caso especial del método de promedio móvil ponderado en el que se elige sólo un peso, aquel para la observación más reciente. Los pesos de los valores para los demás datos se calculan automáticamente y son más pequeños conforme las observaciones se vuelven más antiguas. La ecuación de suavizamiento exponencial es la siguiente.

PRONÓSTICO DE SUAVIZAMIENTO EXPONENCIAL

$$F_{t+1} = \alpha Y_t + (1 - \alpha)F_t \qquad \textbf{(18.2)}$$

donde

F_{t+1} = pronóstico para el periodo $t + 1$ de la serie de tiempo
Y_t = valor real de la serie de tiempo en el periodo t
F_t = pronóstico para el periodo t de la serie de tiempo
α = constante de suavizamiento $(0 \leq \alpha \leq 1)$

La ecuación (18.2) muestra que el pronóstico para el periodo $t + 1$ es un promedio ponderado del valor real en el periodo t y del valor pronosticado para el periodo t. El peso dado al valor real en el periodo t es la **constante de suavizamiento** α, y el peso dado al pronóstico para el periodo t es $1 - \alpha$. Resulta que el pronóstico exponencial para cualquier periodo es en realidad un promedio ponderado de *todos los valores reales anteriores* de la serie de tiempo. Se ilustra lo anterior con una serie de tiempo que sólo implica los datos de tres periodos: Y_1, Y_2 y Y_3.

Para empezar los cálculos, sea F_1 el valor real de la serie de tiempo en el periodo 1, es decir, $F_1 = Y_1$. Por tanto, el pronóstico para el periodo 2 es

$$\begin{aligned} F_2 &= \alpha Y_1 + (1 - \alpha)F_1 \\ &= \alpha Y_1 + (1 - \alpha)Y_1 \\ &= Y_1 \end{aligned}$$

Observe que el pronóstico de suavizamiento exponencial para el periodo 2 es igual al valor real de la serie de tiempo en el periodo 1.

El pronóstico para el periodo 3 es

$$F_3 = \alpha Y_2 + (1 - \alpha)F_2 = \alpha Y_2 + (1 - \alpha)Y_1$$

Por último, al sustituir esta expresión para F_3 en la expresión para F_4 obtenemos

$$\begin{aligned} F_4 &= \alpha Y_3 + (1 - \alpha)F_3 \\ &= \alpha Y_3 + (1 - \alpha)[\alpha Y_2 + (1 - \alpha)Y_1] \\ &= \alpha Y_3 + \alpha(1 - \alpha)Y_2 + (1 - \alpha)^2 Y_1 \end{aligned}$$

El término suavizamiento exponencial *proviene del carácter exponencial del sistema de ponderación de los valores históricos.*

Observe ahora que F_4 es un promedio ponderado de los tres primeros valores de la serie de tiempo. La suma de los coeficientes o pesos de Y_1, Y_2 y Y_3 es igual a 1. Con un argumento similar se puede demostrar que, en general, cualquier pronóstico F_{t+1} es un promedio ponderado de todos los valores anteriores de la serie de tiempo.

A pesar de que con el suavizamiento exponencial se obtiene un pronóstico que es el promedio ponderado de todas las observaciones anteriores, no deben conservarse todos los datos del pasado para calcular el pronóstico del periodo siguiente. De hecho, la ecuación (18.2) muestra que una vez que el valor de la constante de suavizamiento α es elegida, sólo se necesitan dos informaciones para calcular el pronóstico: Y_t, el valor real de la serie de tiempo para el periodo t, y F_t, el pronóstico para el periodo t.

Para ilustrar el método de suavizamiento exponencial, considere de nuevo la serie de tiempo de los precios de la gasolina presentada en la tabla 18.1 y en la figura 18.1. Como ya se explicó, para iniciar los cálculos se establece un pronóstico de suavizamiento exponencial para el periodo 2 igual al valor real de la serie de tiempo en el periodo 1. Por tanto, como $Y_1 = 17$, para empezar con los cálculos del suavizamiento exponencial se pone $F_2 = 17$. Referente a los datos de la serie de tiempo en la tabla 18.1, se encuentra que el valor real de la serie de tiempo en el periodo 2 es $Y_2 = 21$. Por tanto, el error de pronóstico del periodo 2 es $21 - 17 = 4$.

Al continuar con los cálculos del suavizamiento mediante una constante de suavización $\alpha = 0.2$, se obtiene el siguiente pronóstico para el periodo 3:

$$F_3 = 0.2Y_2 + 0.8F_2 = 0.2(21) + 0.8(17) = 17.8$$

Una vez que se conoce el valor real de la serie de tiempo en el periodo 3, $Y_3 = 19$, se puede generar un pronóstico para el periodo 4 de la siguiente manera.

$$F_4 = 0.2Y_3 + 0.8F_3 = 0.2(19) + 0.8(17.8) = 18.04$$

Al continuar con los cálculos para el suavizamiento exponencial se determinan los valores de los pronósticos semanales que se muestran en la tabla 18.10. Observe que no se ha mostrado

TABLA 18.10 Resumen de los pronósticos de suavizamiento exponencial y los errores de pronóstico de la serie de tiempo para las ventas de gasolina con $\alpha = 0.2$ como constante de suavizamiento

Semana	Valores de la serie de tiempo	Pronóstico	Error de pronóstico	Error de pronóstico cuadrado
1	17			
2	21	17.00	4.00	16.00
3	19	17.80	1.20	1.44
4	23	18.04	4.96	24.60
5	18	19.03	−1.03	1.06
6	16	18.83	−2.83	8.01
7	20	18.26	1.74	3.03
8	18	18.61	−0.61	0.37
9	22	18.49	3.51	12.32
10	20	19.19	0.81	0.66
11	15	19.35	−4.35	18.92
12	22	18.48	3.52	12.39
		Totales	10.92	98.80

un pronóstico de suavizamiento exponencial o un error de pronóstico para la semana 1, ya que no se obtuvo ningún pronóstico. Para la semana 12, se tiene que $Y_{12} = 22$ y $F_{12} = 18.48$. Se puede utilizar esta información para generar un pronóstico sobre la semana 13.

$$F_{13} = 0.2Y_{12} + 0.8F_{12} = 0.2(22) + 0.8(18.48) = 19.18$$

Por tanto, el pronóstico de suavizamiento exponencial de la cantidad vendida en la semana 13 es 19.18, o 19 180 galones de gasolina. Con este pronóstico, la empresa, como consecuencia, puede hacer planes y tomar decisiones.

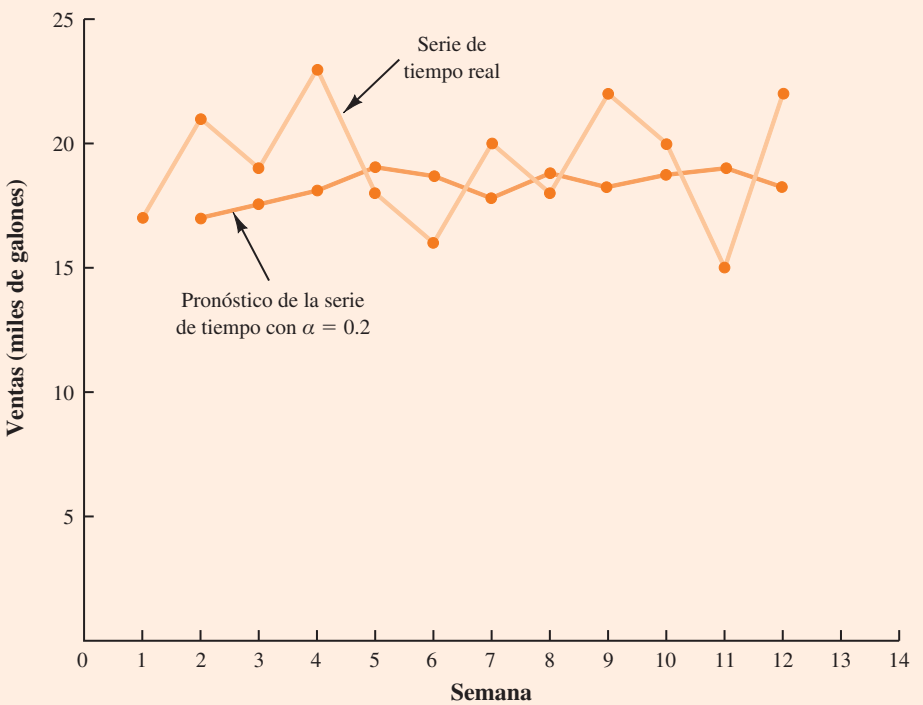
La figura 18.8 muestra la gráfica de los valores reales y pronosticados de la serie de tiempo. Observe en especial cómo los pronósticos "suavizan" la irregularidad de las fluctuaciones de la serie de tiempo.

Exactitud del pronóstico En los cálculos anteriores para el suavizamiento exponencial se utilizó una constante de suavizamiento de $\alpha = 0.2$. Aunque cualquier valor para α entre 0 y 1 es aceptable, algunos darán mejores pronósticos que otros. Una idea de cómo elegir el mejor valor para α se obtiene al revisar el modelo básico de suavizamiento exponencial de la siguiente manera.

$$\begin{aligned} F_{t+1} &= \alpha Y_t + (1 - \alpha)F_t \\ F_{t+1} &= \alpha Y_t + F_t - \alpha F_t \\ F_{t+1} &= F_t + \alpha(Y_t - F_t) \end{aligned} \qquad \textbf{(18.3)}$$

Así, el nuevo pronóstico F_{t+1} es igual al anterior F_t más un ajuste, el cual es la constante de suavizamiento α multiplicada por el error de pronóstico más reciente $Y_t - F_t$. Es decir, el pronóstico para el periodo $t + 1$ se obtiene al ajustar el pronóstico para el periodo t mediante una fracción del error de pronóstico. Si en la serie de tiempo existe una variabilidad aleatoria considerable, se prefiere un valor pequeño para la constante de suavizamiento. La razón de esta elección estriba en que gran parte del error de pronóstico se debe a la variabilidad aleatoria, y no se quiere reaccionar de forma exagerada y ajustar los pronósticos muy rápidamente. Para una serie de tiempo con una variabilidad aleatoria relativamente pequeña, los errores de pronóstico tienden más a representar un cambio en el nivel de la serie. Por tanto, los valores mayores para

FIGURA 18.8 Series de tiempo real y pronosticada de las ventas de gasolina con constante de suavizamiento $\alpha = 0.2$



la constante de suavizamiento tienen la ventaja de ajustar rápidamente los pronósticos, lo que permite adaptarlos más pronto a las condiciones cambiantes.

El criterio que se utilizará a efecto de determinar el valor adecuado para la constante de suavizamiento α es el mismo que el propuesto para determinar el número de periodos a incluir en el cálculo de los promedios móviles. Es decir, se elige el valor de α que minimice el cuadrado medio debido al error (CME) o error cuadrático medio. Un resumen de los cálculos del CME para el pronóstico de suavizamiento exponencial de las ventas de gasolina con $\alpha = 0.2$ se muestra en la tabla 18.10. Observe que hay un error cuadrado menos que el número de periodos, porque no se tenía el valor anterior con el que se pudiera obtener un pronóstico para el periodo 1. El valor de la suma de los errores de pronóstico cuadrados es 98.80, por lo que CME $= 98.80/11 = 8.98$. ¿Habrá un valor de α distinto que proporcione mejores resultados en términos de un valor menor del CME? La forma más sencilla de responder esta pregunta es simplemente probar otros valores para α. Después se comparan los cuadrados medios del error con el valor de 8.98 del CME obtenido mediante una constante de suavizamiento $\alpha = 0.2$.

Los resultados del suavizamiento exponencial con $\alpha = 0.3$ se muestran en la tabla 18.11. El valor de la suma de los errores de pronóstico cuadrados es 102.83, por lo que CME $= 102.83/11 = 9.35$. Observe que con CME $= 9.35$ para este conjunto de datos reales, una constante de suavizamiento de $\alpha = 0.3$ resulta en pronósticos menos exactos que si se emplea una constante de suavizamiento de $\alpha = 0.2$. Por tanto, se preferirá esta constante original de $\alpha = 0.2$. Al utilizar otros valores de α se puede hallar un "buen" valor para la constante de suavizamiento. Este valor puede ser utilizado en el modelo de suavizamiento exponencial a efecto de obtener pronósticos para el futuro. En un momento posterior, después de obtener nuevas observaciones de la serie de tiempo, se analizan nuevamente los datos recabados para determinar si la constante de suavizamiento debe ser modificada para obtener mejores resultados.

TABLA 18.11 Resumen de los pronósticos de suavizamiento exponencial y de los errores de pronóstico para las ventas de gasolina con constante de suavizamiento $\alpha = 0.3$

Semana	Valores de la serie de tiempo	Pronóstico	Error de pronóstico	Error de pronóstico cuadrado
1	17			
2	21	17.00	4.00	16.00
3	19	18.20	0.80	0.64
4	23	18.44	4.56	20.79
5	18	19.81	−1.81	3.28
6	16	19.27	−3.27	10.69
7	20	18.29	1.71	2.92
8	18	18.80	−0.80	0.64
9	22	18.56	3.44	11.83
10	20	19.59	0.41	0.17
11	15	19.71	−4.71	22.18
12	22	18.30	3.70	13.69
		Totales	8.03	102.83

NOTAS Y COMENTARIOS

1. Los paquetes de hoja de cálculo son una ayuda eficaz en la elección de un valor adecuado para α en el suavizamiento exponencial. Con los datos de las series de tiempo y las fórmulas de elaboración de pronósticos, en una hoja de cálculo se pueden probar diferentes valores de α y elegir el que proporciona el error de pronóstico más pequeño utilizando una o más medidas de exactitud de pronóstico (EAM, CME o EPAM).
2. Presentamos el promedio móvil y los métodos de suavizamiento exponencial en el contexto de una serie de tiempo estacionaria. Estos métodos también pueden utilizarse para pronosticar una serie de tiempo no estacionaria que cambia de nivel pero no muestra una tendencia o estacionalidad. Los promedios móviles con valores pequeños de k se pueden adaptar más rápidamente que los promedios móviles con valores mayores de k. Los modelos de suavizamiento exponencial con constantes de suavizamiento más cercanas a 1 se adaptan más rápidamente que los modelos con valores más pequeños de la constante de suavizamiento.

Ejercicios

Métodos

AUTOevaluación

5. Considere los datos siguientes de serie de tiempo.

Semana	1	2	3	4	5	6
Valor	18	13	16	11	17	14

a) Construya una gráfica de serie de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Obtenga los pronósticos con un promedio móvil a tres semanas para esta serie de tiempo. Calcule el CME y un pronóstico para la semana 7.
c) Utilice $\alpha = 0.2$ para calcular los pronósticos de suavizamiento exponencial de la serie de tiempo. Calcule el CME y dé un pronóstico para la semana 7.

d) Compare el método del promedio móvil a tres semanas con el método de suavizamiento exponencial utilizando $\alpha = 0.2$. ¿Cuál parece dar un pronóstico más preciso basado en el CME? Explique.

e) Utilice una constante de suavizamiento de $\alpha = 0.4$ para calcular el pronóstico de suavizamiento exponencial. ¿Una constante de suavizamiento de 0.2 o de 0.4 parece proporcionar pronósticos más precisos basados en el CME? Explique.

6. Considere los datos siguientes de serie de tiempo.

Semana	1	2	3	4	5	6	7
Valor	24	13	20	12	19	23	15

Construya una gráfica de serie de tiempo. ¿Qué tipo de patrón existe en los datos?

a) Desarrolle el pronóstico de promedio móvil para tres semanas en esta serie de tiempo. Calcule el CME y proporcione un pronóstico para la semana 8.

b) Utilice $\alpha = 0.2$ para calcular los pronósticos de suavizamiento exponencial de la serie de tiempo. Calcule el CME y proporcione un pronóstico para la semana 8.

c) Compare el método del promedio móvil a tres semanas con el método de suavizamiento exponencial utilizando $\alpha = 0.2$. ¿Cuál parece dar pronósticos más precisos basados en el CME?

d) Utilice una constante de suavizamiento de $\alpha = 0.4$ para calcular los pronósticos de suavizamiento exponencial. ¿Una constante de suavizamiento de 0.2 o de 0.4 parece ofrecer pronósticos más precisos basados en el CME? Explique.

WEB archivo
Gasoline

7. Regrese a los datos de la serie de tiempo para las ventas de gasolina de la tabla 18.1.

a) Calcule los promedios móviles de la serie de tiempo a 4 y 5 semanas.

b) Calcule el CME de los pronósticos obtenidos con los promedios móviles de 4 y 5 semanas.

c) ¿Cuál parece ser de los datos pasados el mejor número de semanas a utilizar (3, 4 o 5) para el cálculo del promedio móvil? Recuerde que el CME para el promedio móvil de tres semanas es 10.22.

WEB archivo
Gasoline

8. Consulte de nuevo los datos de la serie de tiempo de las ventas de gasolina de la tabla 18.1.

a) Utilice 1/2 como el peso de la observación más reciente, 1/3 para la segunda observación más reciente y 1/6 para la tercera observación más reciente. Calcule un promedio móvil ponderado de las tres semanas para la serie de tiempo.

b) Determine el CME del promedio móvil ponderado del inciso *a*). ¿Prefiere éste que el promedio móvil no ponderado? Recuerde que el CME del promedio móvil ponderado es 10.22.

c) Suponga que se le permite elegir cualesquiera pesos, siempre y cuando su suma sea 1. ¿Siempre será posible elegir un conjunto de pesos que hagan que el CME sea menor para el promedio móvil ponderado que para un promedio móvil no ponderado? ¿Por qué?

WEB archivo
Gasoline

9. Con los datos de la serie de tiempo de las ventas de gasolina de la tabla 18.1, muestre el pronóstico de suavizamiento exponencial utilizando $\alpha = 0.1$.

a) Al aplicar la medida de exactitud del CME, ¿preferiría una constante de suavizamiento de $\alpha = 0.1$ o $\alpha = 0.2$ para la serie de tiempo de las ventas de gasolina?

b) ¿Los resultados son los mismos si se aplica EAM como medida de exactitud?

c) ¿Cuáles son los resultados si se utiliza EPAM?

10. Con una constante de suavizamiento de $\alpha = 0.2$, la ecuación (18.2) muestra que el pronóstico para la semana 13 de las ventas de gasolina listadas en la tabla 18.1 está dado por $F_{13} = 0.2Y_{12} + 0.8F_{12}$. Sin embargo, el pronóstico para la semana 12 está dado por $F_{12} = 0.2Y_{11} + 0.8F_{11}$. Por tanto, se podrían combinar estos dos resultados para mostrar que el pronóstico sobre la semana 13 se puede escribir como

$$F_{13} = 0.2Y_{12} + 0.8(0.2Y_{11} + 0.8F_{11}) = 0.2Y_{12} + 0.16Y_{11} + 0.64Y_{11} + 0.64F_{11}$$

a) Aplique el hecho de que $F_{11} = 0.2Y_{10} + 0.8F_{10}$ (y de manera similar para F_{10} y F_9) y continúe expandiendo la expresión para F_{13} hasta que ésta se escriba en términos de los datos de los valores pasados Y_{12}, Y_{11}, Y_{10}, Y_9 y Y_8, y del pronóstico para el periodo 8.

b) Remítase a los coeficientes o pesos de los valores del pasado Y_{12}, Y_{11}, Y_{10}, Y_9 y Y_8. ¿Qué puede decir acerca de los pesos que el suavizamiento exponencial proporciona a los valores pasados al obtener un nuevo pronóstico? Compare estos pesos con los del método del promedio móvil.

Aplicaciones

11. Para Hawkins Company, los porcentajes de los embarques mensuales recibidos en los últimos 12 meses son 80, 82, 84, 83, 83, 84, 85, 84, 82, 83, 84 y 83.
    a) Construya una gráfica para la serie de tiempo. ¿Qué tipo de patrón existe en los datos?
    b) Compare el pronóstico obtenido mediante promedios móviles de tres meses con el pronóstico obtenido por el método de suavizamiento exponencial con $\alpha = 0.2$. ¿Con cuál se obtienen pronósticos más precisos al utilizar el CME como medida de exactitud?
    c) ¿Cuál es el pronóstico para el próximo mes?

12. A continuación se proporcionan las tasas de interés de bonos corporativos triple A de 12 meses consecutivos.

    9.5 9.3 9.4 9.6 9.8 9.7 9.8 10.5 9.9 9.7 9.6 9.6

    a) Construya una gráfica de serie de tiempo. ¿Qué tipo de patrón existe en los datos?
    b) Desarrolle promedios móviles de tres y cuatro meses para esta serie de tiempo. ¿Con cuál de estos promedios móviles se obtiene un pronóstico más exacto basado en el CME? Explique.
    c) ¿Cuál es el pronóstico para el promedio móvil del próximo mes?

AUTOevaluación

13. Los valores de los contratos de construcción en Alabama (en millones de dólares) para un periodo de 12 meses son los siguientes.

    240 350 230 260 280 320 220 310 240 310 240 230

    a) Construya una gráfica de serie de tiempo. ¿Qué tipo de patrón existe en los datos?
    b) Compare el método del promedio móvil a tres meses con el pronóstico de suavizamiento exponencial utilizando $\alpha = 0.2$. ¿Con cuál se obtienen pronósticos más precisos basados en el CME?
    c) ¿Cuál es el pronóstico para el próximo mes?

14. En la siguiente serie de tiempo se muestran las ventas de un producto en particular en los últimos 12 meses.

Mes	Ventas	Mes	Ventas
1	105	7	145
2	135	8	140
3	120	9	100
4	105	10	80
5	90	11	100
6	120	12	110

    a) Construya una gráfica de serie de tiempo. ¿Qué tipo de patrón existe en los datos?
    b) Utilice $\alpha = 0.3$ para calcular los pronósticos de suavizamiento exponencial de la serie de tiempo.
    c) Utilice una constante de suavizamiento de $\alpha = 0.5$ para calcular los pronósticos de suavizamiento exponencial. ¿Cuál de las constantes de suavizamiento, 0.3 o 0.5, parece proporcionar pronósticos más precisos basados en el CME?

15. Los datos siguientes son los valores del Commodity Futures Index de 10 semanas: 7.35, 7.40, 7.55, 7.56, 7.60, 7.52, 7.52, 7.70, 7.62 y 7.55.
    a) Construya una gráfica de serie de tiempo. ¿Qué tipo de patrón existe en los datos?
    b) Calcule los pronósticos de suavizamiento exponencial para $\alpha = 0.2$.
    c) Calcule los pronósticos de suavizamiento exponencial para $\alpha = 0.3$.
    d) ¿Cuál de las constantes de suavizamiento exponencial proporciona pronósticos más precisos basados en el CME? Elabore el pronóstico para la semana 11.

16. Las calificaciones (rating) Nielsen (porcentajes de audiencia televisiva de hogares en Estados Unidos) del Torneo Masters Golf de 1997 a 2008 son las siguientes (*Golf Magazine*, enero de 2009).

WEB archivo
Masters

Year	Rating
1997	11.2
1998	8.6
1999	7.9
2000	7.6
2001	10.7
2002	8.1
2003	6.9
2004	6.7
2005	8.0
2006	6.9
2007	7.6
2008	7.3

El rating de 11.2 puntos en 1997 indica que 11.2% de los hogares estadounidenses se sintonizó para ver a Tiger Woods triunfar en su primer torneo de golf más importante y convertirse en el primer afroestadounidense en ganar el Masters. Tiger Woods lo ganó en 2001 y 2005.

a) Construya una gráfica de serie de tiempo. ¿Qué tipo de patrón existe en los datos? Opine sobre algunos factores que pueden haber influido en el modelo mostrado en la gráfica de series de tiempo para este periodo.

b) Dado el patrón de la gráfica de series de tiempo desarrollado en el inciso *a*), ¿cree que los métodos de pronóstico estudiados en esta sección son adecuados para obtener los pronósticos para esta serie de tiempo? Explique.

c) ¿Recomendaría utilizar sólo los ratings de Nielsen de 2002-2008 para pronosticar el rating de 2009, o debería usarse toda la serie de tiempo desde 1997 hasta 2008? Explique.

18.4 Proyección de la tendencia

En esta sección se presentan tres métodos de elaboración de pronósticos que son apropiados para las series de tiempo que exhiben un patrón de tendencia. En primer lugar, se muestra qué tan sencillo es el uso de la regresión lineal para pronosticar la serie de tiempo con tendencia lineal. Después se ilustra cómo obtener los pronósticos mediante el suavizamiento exponencial lineal de Holt, una extensión del único suavizamiento exponencial que utiliza dos constantes de suavizamiento: la primera se toma en cuenta para el nivel de la serie de tiempo y la segunda para la tendencia lineal de los datos. Por último, se muestra cómo la capacidad de ajuste a una curva en el análisis de regresión permite pronosticar series de tiempo con una tendencia curvilínea o no lineal.

WEB archivo
Bicycle

TABLA 18.12
Serie de tiempo de las ventas de bicicletas

Año	Ventas (en miles)
1	21.6
2	22.9
3	25.5
4	21.9
5	23.9
6	27.5
7	31.5
8	29.7
9	28.6
10	31.4

Regresión de tendencia lineal

En la sección 18.1 se utilizaron las ventas de bicicletas manejadas en la tabla 18.3 y en la figura 18.3 para ilustrar una serie de tiempo con un patrón de tendencia. Esta serie de tiempos se utilizará para ilustrar cómo la regresión lineal simple permite pronosticar una serie de tiempo con tendencia lineal. Los datos de la serie de tiempo de las ventas de bicicletas se repiten en la tabla 18.12 y en la figura 18.9.

Aunque la gráfica de serie de tiempo de la figura 18.9 muestra algún movimiento ascendente y descendente en los últimos 10 años, se podría estar de acuerdo en que la línea de tendencia mostrada en la figura 18.10 proporciona una aproximación razonable del movimiento de la serie a largo plazo. Se pueden utilizar los métodos de regresión lineal simple (vea el capítulo 14) a efecto de obtener una tendencia lineal para la serie de tiempo de las ventas de bicicletas.

FIGURA 18.9 Gráfica de serie de tiempo para las ventas de bicicletas

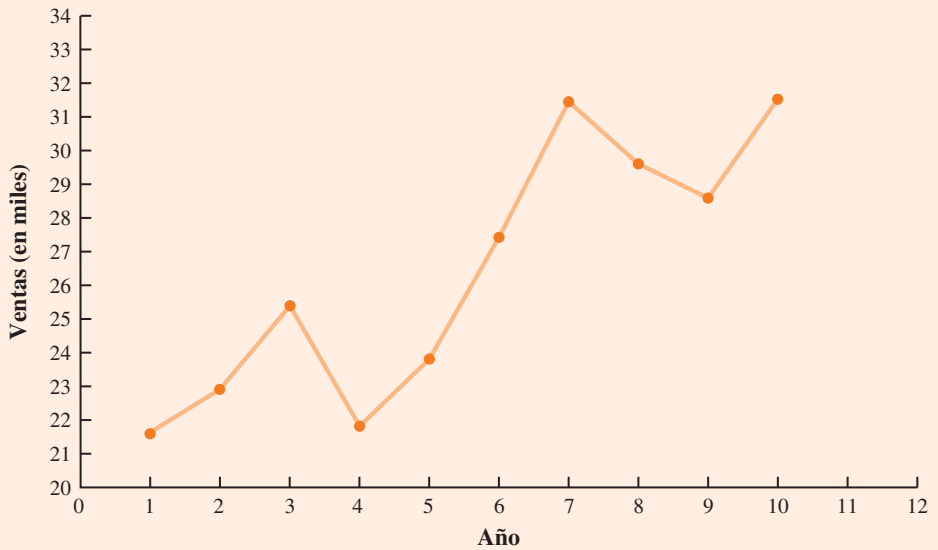
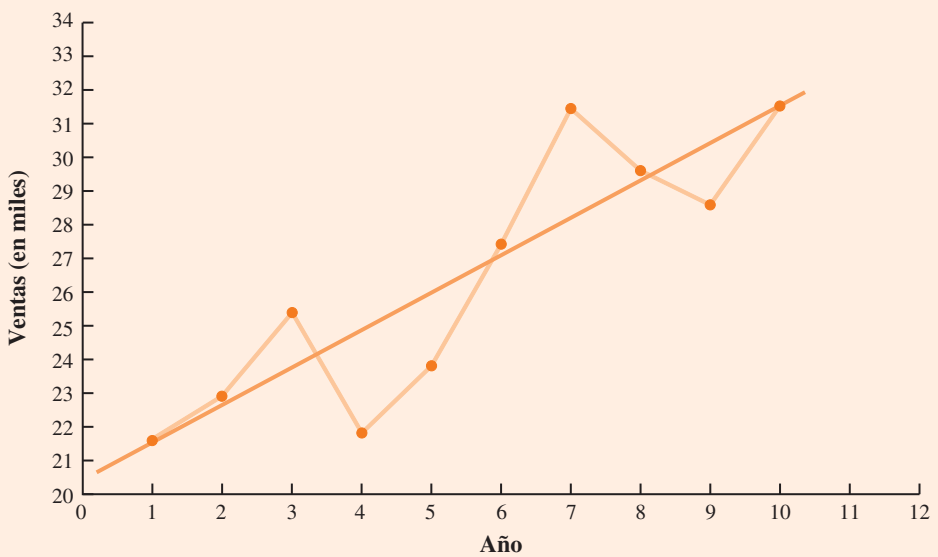


FIGURA 18.10 Tendencia representada por una función lineal de la serie de tiempo de ventas de bicicletas



En el capítulo 14, la ecuación de regresión estimada que describe una relación lineal entre una variable independiente x y una variable dependiente y se expresó como:

$$\hat{y} = b_0 + b_1 x$$

donde $\hat{y}$ es el valor estimado o predicho de y. Para enfatizar el hecho de que en un pronóstico la variable independiente es el tiempo, se sustituirá x por t, así como $\hat{y}$ por T_t para hacer patente que se está estimando la tendencia de una serie de tiempo. Por tanto, para calcular la tendencia lineal en una serie de tiempo se utilizará la siguiente ecuación de regresión estimada.

ECUACIÓN DE TENDENCIA LINEAL

$$T_t = b_0 + b_1 t \tag{18.4}$$

donde

T_t = pronóstico de tendencia lineal en el periodo t
b_0 = intersección de la recta de tendencia lineal
b_1 = pendiente de la recta de tendencia lineal
t = periodo

En la ecuación (18.4) la variable tiempo comienza en $t = 1$ correspondiente a la primera observación de la serie de tiempo (año 1 para la serie de tiempo de las ventas de bicicletas) y continúa hasta que $t = n$ correspondiente a la observación más reciente (año 10 para la serie de tiempo de las ventas de bicicletas). Así, $t = 1$ corresponde al valor de la serie más antigua y $t = 10$ corresponde al año más reciente.

A continuación se presentan las fórmulas para calcular los coeficientes de regresión estimados b_0 y b_1 de la ecuación (18.4).

CÁLCULO DE LA PENDIENTE Y DE LA INTERSECCIÓN PARA UNA TENDENCIA LINEAL*

$$b_1 = \frac{\sum_{t=1}^{n}(t - \bar{t})(Y_t - \bar{Y})}{\sum_{t=1}^{n}(t - \bar{t})^2} \tag{18.5}$$

$$b_0 = \bar{Y} - b_1\bar{t} \tag{18.6}$$

donde

Y_t = valor de la serie de tiempo en el periodo t
n = número de periodos (número de observaciones)
$\bar{Y}$ = valor promedio de la serie de tiempo
$\bar{t}$ = valor promedio de t

*Una fórmula alternativa para b_1 es

$$b_1 = \frac{\sum_{t=1}^{n} tY_t - \left(\sum_{t=1}^{n} t \sum_{t=1}^{n} Y_t\right)/n}{\sum_{t=1}^{n} t^2 - \left(\sum_{t=1}^{n} t\right)^2/n}$$

Esta forma de la ecuación (18.5) se recomienda a menudo cuando se utiliza una calculadora para obtener b_1.

A efecto de calcular la ecuación de tendencia lineal para la serie de tiempo de las ventas de bicicletas, se comienza por determinar $\bar{t}$ y $\bar{Y}$ utilizando la información de la tabla 18.12.

$$\bar{t} = \frac{\sum_{t=1}^{n} t}{n} = \frac{55}{10} = 5.5$$

$$\bar{Y} = \frac{\sum_{t=1}^{n} Y_t}{n} = \frac{264.5}{10} = 26.45$$

TABLA 18.13 Resumen de los cálculos de la tendencia lineal para la serie de tiempo de las ventas de bicicletas

	t	Y_t	$t - \bar{t}$	$Y_t - \bar{Y}$	$(t - \bar{t})(Y_t - \bar{Y})$	$(t - \bar{t})^2$
	1	21.6	−4.5	−4.85	21.825	20.25
	2	22.9	−3.5	−3.55	12.425	12.25
	3	25.5	−2.5	−0.95	2.375	6.25
	4	21.9	−1.5	−4.55	6.825	2.25
	5	23.9	−0.5	−2.55	1.275	0.25
	6	27.5	0.5	1.05	0.525	0.25
	7	31.5	1.5	5.05	7.575	2.25
	8	29.7	2.5	3.25	8.125	6.25
	9	28.6	3.5	2.15	7.525	12.25
	10	31.4	4.5	4.95	22.275	20.25
Totales	55	264.5			90.750	82.50

Con estos valores y la información de la tabla 18.13 se puede calcular la pendiente y la intersección de la tendencia lineal para la serie de tiempo de las ventas de bicicletas.

$$b_1 = \frac{\sum_{t=1}^{n}(t - \bar{t})(Y_t - \bar{Y})}{\sum_{t=1}^{n}(t - \bar{t})^2} = \frac{90.75}{82.5} = 1.1$$

$$b_0 = \bar{Y} - b_1\bar{t} = 26.45 - 1.1(5.5) = 20.4$$

Por tanto, la ecuación de tendencia lineal es

$$T_t = 20.4 + 1.1t$$

La pendiente de 1.1 indica que en los últimos 10 años la empresa experimentó un promedio de crecimiento en las ventas de cerca de 1 100 unidades por año. Si se supone que la tendencia en las ventas de la última década es un buen indicador del futuro, esta ecuación de tendencia puede utilizarse para obtener los pronósticos sobre periodos futuros. Por ejemplo, al sustituir en la ecuación $t = 11$ se obtiene la proyección de tendencia o el pronóstico para el próximo año T_{11}.

$$T_{11} = 20.4 + 1.1(11) = 32.5$$

Por tanto, al utilizar la proyección de tendencia se podrá pronosticar un valor para las ventas de 32 500 bicicletas para el próximo año.

Para calcular la exactitud del método de elaboración de pronósticos de la proyección de tendencia se utilizará el CME. La tabla 18.14 muestra el cálculo de la suma de los errores cuadrados para la serie de tiempo de las ventas de bicicletas. Así que para esta serie,

$$\text{CME} = \frac{\sum_{t=1}^{n}(Y_t - F_t)^2}{n} = \frac{30.7}{10} = 3.07$$

Debido a que el pronóstico de regresión de tendencia lineal utiliza el mismo procedimiento de análisis de regresión que se presentó en el capítulo 14, se pueden utilizar los procedimientos estándar para realizar los cálculos del análisis de regresión con Minitab o Excel. La figura 18.11 muestra el resultado obtenido al utilizar el módulo de Minitab del análisis de regresión de la serie de tiempo de las ventas de bicicletas.

TABLA 18.14 Resumen de los pronósticos de tendencia lineal y de errores de pronóstico para la serie de tiempo de ventas de bicicletas

Año	Ventas (en miles) Y_t	Pronóstico T_t	Error de pronóstico	Error de pronóstico cuadrado
1	21.6	21.5	0.1	0.01
2	22.9	22.6	0.3	0.09
3	25.5	23.7	1.8	3.24
4	21.9	24.8	−2.9	8.41
5	23.9	25.9	−2.0	4.00
6	27.5	27.0	0.5	0.25
7	31.5	28.1	3.4	11.56
8	29.7	29.2	0.5	0.25
9	28.6	30.3	−1.7	2.89
10	31.4	31.4	0.0	0.00
			Total	30.70

En la figura 18.11 el valor del CME en la tabla de ANOVA es

$$\text{CME} = \frac{\text{suma de cuadrados debido al error}}{\text{grados de libertad}} = \frac{30.7}{8} = 3.837$$

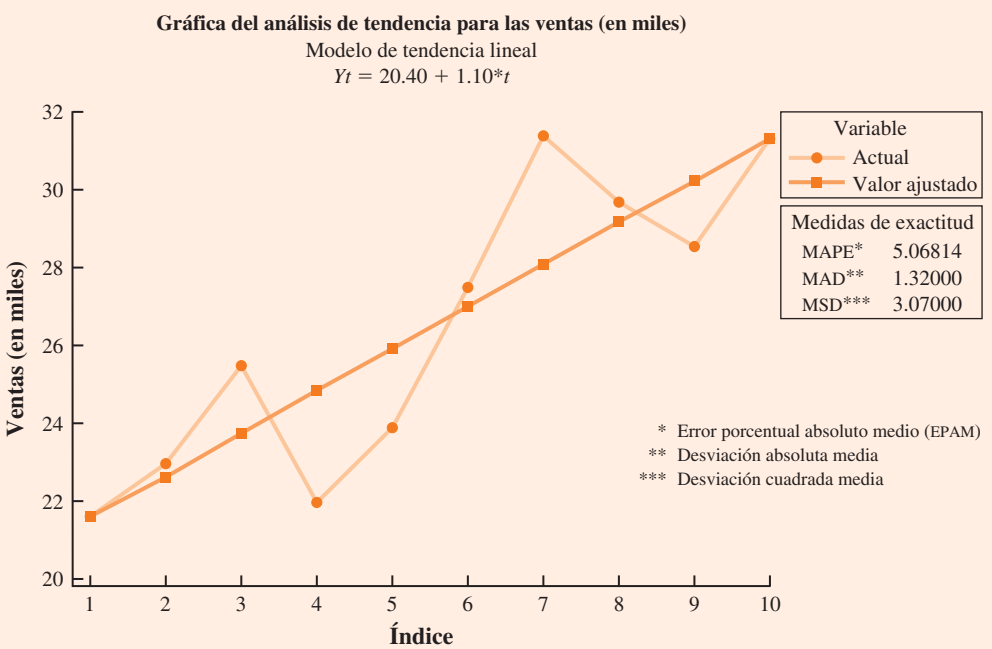
En los resultados de Minitab del análisis de tendencia MSD *es la desviación cuadrada media, es decir, el promedio de los errores de pronóstico cuadrados.*

Este valor del CME difiere del valor calculado antes porque la suma de los errores cuadrados se divide entre 8 en vez de 10; por tanto, en el CME el resultado de regresión no es la media de los errores de pronóstico cuadrados. Sin embargo, la mayoría de los paquetes de pronóstico calculan el valor del CME tomando la media de los errores cuadrados. Por tanto, al utilizar los paquetes de series de tiempo para desarrollar una ecuación de tendencia, el resultado del valor del CME puede diferir ligeramente del que se obtendría con un método de regresión general. Por ejemplo, en la figura 18.12 se muestra la parte gráfica del resultado obtenido al utilizar el procedimiento de análisis de tendencia de series de tiempo de Minitab. Observe que MSD = 3.07 es el promedio de los errores de pronóstico cuadrados.

FIGURA 18.11 Resultado de regresión en Minitab para la serie de tiempo de las ventas de bicicletas

```
The regression equation is
Y = 20.4 + 1.10 t

Predictor     Coef   SE Coef       T       p
Constant   20.400     1.338   15.24   0.000
t          1.1000    0.2157    5.10   0.001

S = 1.95895   R-sq = 76.5%   R-sq(adj) = 73.5%

Analysis of Variance

SOURCE          DF       SS      MS      F      p
Regression       1   99.825  99.825  26.01  0.001
Residual Error   8   30.700   3.837
Total            9  130.525
```

FIGURA 18.12 Resultado del análisis de tendencia lineal en Minitab de la serie de tiempo de las ventas de bicicletas



Suavizamiento exponencial lineal de Holt

El suavizamiento exponencial lineal de Holt suele llamarse suavizamiento exponencial doble.

Charles Holt desarrolló una versión de suavizamiento exponencial para pronosticar una serie de tiempo con tendencia lineal. Recuerde que el procedimiento de suavizamiento exponencial estudiado en la sección 18.3 utiliza α como constante de suavizamiento para "suavizar" la aleatoriedad o las fluctuaciones irregulares en una serie de tiempo, y que los pronósticos para el periodo $t + 1$ se obtienen mediante la ecuación

$$F_{t+1} = \alpha Y_t + (1 - \alpha)F_t$$

Los pronósticos por el método de **suavizamiento exponencial lineal de Holt** se obtienen mediante dos constantes de suavizamiento α y β, y tres ecuaciones.

ECUACIONES DE SUAVIZAMIENTO EXPONENCIAL LINEAL DE HOLT

$$L_t = \alpha Y_t + (1 - \alpha)(L_{t-1} + b_{t-1}) \quad \textbf{(18.7)}$$

$$b_t = \beta(L_t - L_{t-1}) + (1 - \beta)\, b_{t-1} \quad \textbf{(18.8)}$$

$$F_{t+k} = L_t + b_t\, k \quad \textbf{(18.9)}$$

donde

L_t = estimación del nivel de la serie de tiempo para el periodo t

b_t = estimación de la pendiente de la serie de tiempo para el periodo t

α = constante de suavizamiento para el nivel de la serie de tiempo

β = constante de suavizamiento para la pendiente de la serie de tiempo
F_{t+k} = pronóstico para el periodo k en el futuro
k = número de periodos próximos a pronosticar

Aplique el método de Holt a la serie de tiempo de las ventas de bicicletas listadas en la tabla 18.12 y utilice $\alpha = 0.1$ y $\beta = 0.2$. Para empezar con el método se necesitan los valores de L_1, la estimación del nivel de la serie de tiempo en el año 1, y b_1 la estimación de la pendiente de la serie de tiempo en el año 1. El método utilizado comúnmente consiste en determinar $L_1 = Y_1$ y $b_1 = Y_2 - Y_1$. Al utilizar este procedimiento inicial obtenemos

$$L_1 = Y_1 = 21.6$$
$$b_1 = Y_2 - Y_1 = 22.9 - 21.6 = 1.3$$

Mediante la ecuación (18.9) con $k = 1$, el pronóstico de las ventas en el año 2 es $F_2 = L_1 + b_1 = 21.6 + 1.3(1) = 22.9$. Después se continúa utilizando las ecuaciones (18.7) a (18.9) para calcular las estimaciones del nivel y la tendencia para el año 2, así como un pronóstico para el año 3.

Primero utilizamos la ecuación (18.7) y la constante de suavizamiento $\alpha = 0.1$ para calcular la estimación del nivel de la serie de tiempo para el año 2.

$$L_2 = 0.1(22.9) + 0.9(21.6 + 1.3) = 22.9$$

Observe que $21.6 + 1.3$ es el pronóstico de ventas para el año 2. Por tanto, la estimación del nivel de la serie de tiempo en este año obtenida mediante la ecuación (18.7) es simplemente un promedio ponderado del valor observado en el año 2 (con un peso de $\alpha = 0.1$) y el pronóstico para el año 2 (con un peso de $1 - \alpha = 1 - 0.1 = 0.9$). En general, los valores mayores de α dan más peso al valor observado (Y_t), mientras que valores menores dan más peso al valor pronosticado ($L_{t-1} + b_{t-1}$).

A continuación utilice la ecuación (18.8) y la constante de suavizamiento $\beta = 0.2$ para calcular la pendiente de la serie de tiempo en el año 2.

$$b_2 = 0.2(22.9 - 21.6) + (1 - 0.2)(1.3) = 1.3$$

La estimación de tal pendiente es un promedio ponderado de la diferencia en el nivel estimado de la serie de tiempo entre el año 2 y el año 1 (con un peso de $\beta = 0.2$) y la estimación de la pendiente del año 1 (con un peso de $1 - \beta = 1 - 0.2 = 0.8$). En general, los valores mayores de β dan más peso a la diferencia de los niveles estimados, mientras que los valores menores se lo conceden a la pendiente estimada del último periodo.

Al utilizar las estimaciones de L_2 y b_2 recién obtenidas, el pronóstico de ventas para el año 3 se calcula mediante la ecuación (18.9):

$$F_3 = L_2 + b_2 = 22.9 + 1.3(1) = 24.2$$

Se realizan otros cálculos en forma similar, los cuales se muestran en la tabla 18.15. La suma de los errores de pronóstico cuadrados es 39.678; por ende, CME = 39.678/9 = 4.41.

¿Con valores diferentes para la constante de suavizamiento α y β se obtienen pronósticos más precisos? Para responder esta pregunta habría que probar diferentes combinaciones de α y β para determinar si se puede encontrar una combinación con la que se obtenga un valor menor del CME de 4.41, el valor obtenido utilizando las constantes de suavizamiento $\alpha = 0.1$ y $\beta = 0.2$. Se puede realizar la búsqueda de buenos valores α y β por ensayo y error o mediante software de estadística más avanzado que tenga la opción de seleccionar un conjunto óptimo de constantes de suavizamiento.

TABLA 18.15 Resumen de los cálculos de Holt para el suavizamiento exponencial lineal de la serie de tiempo de las ventas de bicicletas utilizando $\alpha = 0.1$ y $\beta = 0.2$

Año	Ventas (miles) Y_t	Nivel estimado L_t	Tendencia estimada b_t	Pronóstico F_t	Error de pronóstico	Error de pronóstico cuadrado
1	21.6	21.600	1.300			
2	22.9	22.900	1.300	22.900	0.000	0.000
3	25.5	24.330	1.326	24.200	1.300	1.690
4	21.9	25.280	1.251	25.656	−3.756	14.108
5	23.9	26.268	1.198	26.531	−2.631	6.924
6	27.5	27.470	1.199	27.466	0.034	0.001
7	31.5	28.952	1.256	28.669	2.831	8.016
8	29.7	30.157	1.245	30.207	−0.507	0.257
9	28.6	31.122	1.189	31.402	−2.802	7.851
10	31.4	32.220	1.171	32.311	−0.911	0.830
					Total	39.678

Observe que las estimaciones, ambas en el año 10, del nivel de la serie de tiempo es $L_1 = 32.220$ y de la pendiente es $b_1 = 1.171$. Si se asume que la tendencia de los últimos 10 años en las ventas es un buen indicador del futuro, la ecuación (18.9) puede utilizarse para desarrollar pronósticos para periodos futuros. Por ejemplo, al sustituir $t = 11$ en la ecuación (18.9) se obtiene la proyección de tendencia para el próximo año o el pronóstico, F_{11}.

$$F_{11} = L_{10} + b_{10}(1) = 32.220 + 1.171 = 33.391$$

Por tanto, al utilizar el suavizamiento exponencial lineal de Holt se pronostica que el año próximo las ventas serán de 33 391 bicicletas.

Regresión de tendencia no lineal

WEB archivo
Cholesterol

Para modelar una tendencia suele utilizarse el modelo de función lineal. Sin embargo, como ya se vio, algunas veces las series de tiempo tienen tendencias curvilíneas o no lineales. Como ejemplo, considere los ingresos anuales en millones de dólares para las ventas de un medicamento contra el colesterol en los primeros 10 años. La tabla 18.16 muestra la serie de tiempo y la figura 18.13 la gráfica correspondiente. Por ejemplo, los ingresos del año 1 fueron de $23.1 millones; los del año 2 de $21.3 millones, y así sucesivamente. La gráfica de serie de tiempo indica una tendencia general creciente o a la alza. Pero a diferencia de la serie de tiempo de las ventas de bicicletas, la tendencia lineal no parece ser apropiada. En cambio, parece requerirse una función curvilínea para modelar la tendencia a largo plazo.

TABLA 18.16 Serie de tiempo de las ventas de medicamentos contra el colesterol (millones de $)

Año (t)	Ingresos (millones de $)
1	23.1
2	21.3
3	27.4
4	34.6
5	33.8
6	43.2
7	59.5
8	64.4
9	74.2
10	99.3

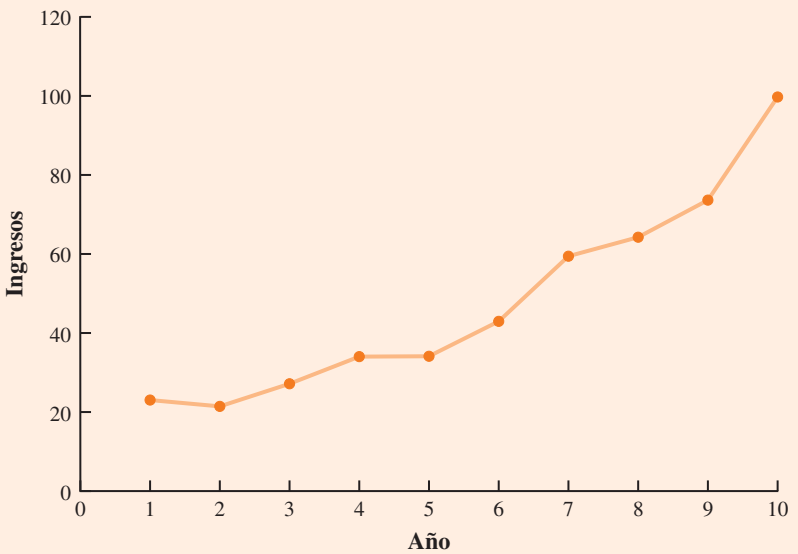
Ecuación de tendencia cuadrática Una variedad de funciones no lineales puede utilizarse para obtener una estimación de la tendencia de la serie de tiempo del colesterol. Por ejemplo, considere la siguiente ecuación de tendencia cuadrática:

$$T_t = b_0 + b_1 t + b_2 t^2 \qquad \textbf{(18.10)}$$

Para la serie de tiempo del colesterol, $t = 1$ corresponde al año 1, $t = 2$ corresponde al año 2, y así sucesivamente.

El modelo lineal general estudiado en la sección 16.1 puede utilizarse para calcular los valores de b_0, b_1 y b_2. Existen dos variables independientes, el año y el año al cuadrado, mientras que la variable dependiente son las ventas en millones de dólares. Así que la primera observa-

FIGURA 18.13 Gráfica de la serie de tiempo de ventas del medicamento contra el colesterol (millones de $)



ción es 1, 1, 23.1; la segunda es 2, 4, 21.3; la tercera es 3, 9, 27.4, y así sucesivamente. La figura 18.14 muestra el resultado de Miniab sobre la regresión múltiple para el modelo de tendencia cuadrática; la ecuación de regresión estimada es

$$\text{Ingresos (\$ millones)} = 24.2 - 2.11\ \text{Year} + 0.922\ \text{YearSq}.$$

donde

$$\text{Year (año)} = 1, 2, 3, \ldots, 10$$
$$\text{YearSq (año cuadrado)} = 1, 4, 9, \ldots, 100$$

FIGURA 18.14 Resultado de Minitab de la regresión de tendencia cuadrática de la serie de tiempo de ventas de bicicletas

```
The regression equation is
Revenue = 24.2 - 2.11 Year + 0.922 YearSq

Predictor      Coef   SE Coef       T       p
Constant     24.182     4.676    5.17   0.001
Year         -2.106     1.953   -1.08   0.317
YearSq       0.9216    0.1730    5.33   0.001

S = 3.97578   R-Sq = 98.1%   R-Sq(adj) = 97.6%

Analysis of Variance

SOURCE           DF      SS      MS       F      p
Regression        2  5770.1  2885.1  182.52  0.000
Residual Error    7   110.6    15.8
Total             9  5880.8
```

Al utilizar el procedimiento estándar de regresión múltiple se requiere calcular los valores para el año al cuadrado como segunda variable independiente. Se puede utilizar el análisis de tendencia para series de tiempo de Minitab alternativamente para obtener los mismos resultados. No se requiere obtener valores para el año al cuadrado y es más fácil de usar. Se recomienda este método en la solución de los ejercicios que involucren tendencias cuadráticas.

Ecuación de tendencia exponencial Otra alternativa que podemos utilizar para modelar el patrón no lineal mostrado por la serie de tiempo del colesterol es ajustar un modelo exponencial a los datos. Por ejemplo, considere la siguiente ecuación de tendencia exponencial.

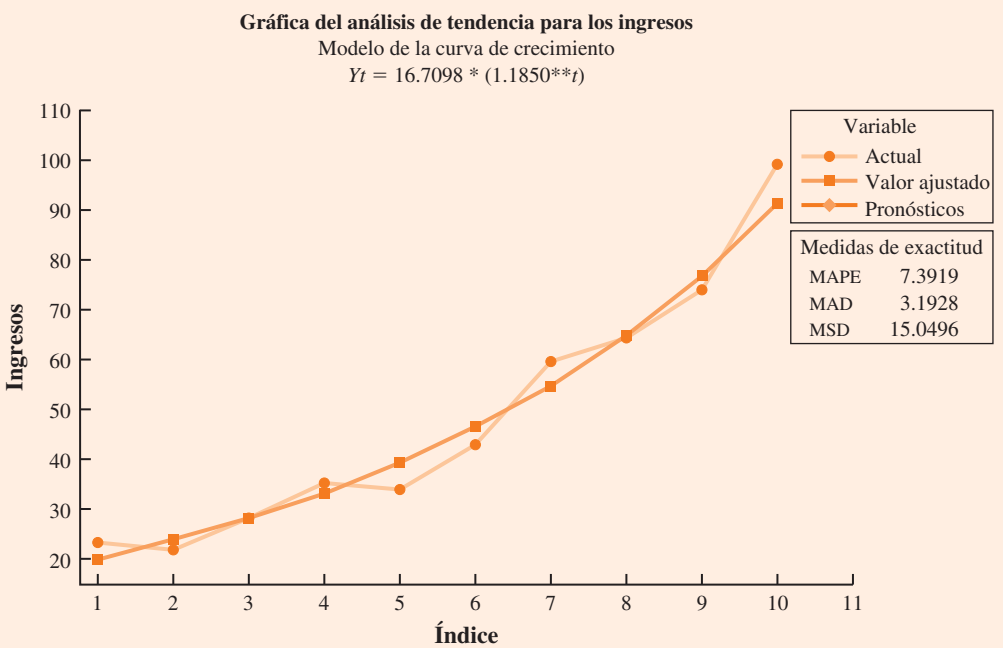
$$T_t = b_0(b_1)^t \tag{18.11}$$

Para entender mejor esta ecuación, suponga que $b_0 = 20$ y $b_1 = 1.2$. Después, para $t = 1$, $T_1 = 20(1.2)^1 = 24$; para $t = 2$, $T_2 = 20(1.2)^2 = 28.8$; y para $t = 3$, $T_3 = 20(1.2)^3 = 34.56$. Observe que T_t no está aumentando por un monto constante como en el caso del modelo de tendencia lineal, sino en un porcentaje constante; el incremento porcentual es de 20%.

En su módulo de serie de tiempo, Minitab tiene la capacidad para calcular una ecuación de tendencia exponencial y puede utilizarse entonces para el pronóstico. Desafortunadamente, Excel no tiene esta capacidad. No obstante, en la sección 16.1 se describe cómo, al tomar logaritmos de los términos de la ecuación (18.11), la metodología del modelo general lineal puede utilizarse para calcular la ecuación de tendencia exponencial.

El módulo de la serie de tiempo de Minitab es muy fácil de utilizar para desarrollar una ecuación de tendencia exponencial. No hay necesidad de trabajar con los logaritmos y usar el análisis de regresión para calcularla. En la figura 18.15 se muestra la gráfica obtenida de la computadora con el procedimiento del análisis de tendencia de la serie de tiempo de Minitab que se adapta a una ecuación de tendencia exponencial.

FIGURA 18.15 Resultado del análisis de tendencia de crecimiento exponencial de Minitab de la serie de tiempo de las ventas de medicamentos contra el colesterol



NOTAS Y COMENTARIOS

La regresión de tendencia lineal se basa en determinar la ecuación de regresión estimada que minimiza la suma de los errores de pronóstico cuadrados y, por consiguiente, del CME. Por tanto, se esperaría que la regresión de tendencia lineal sea mejor que el suavizamiento exponencial lineal de Holt en términos del CME. Por ejemplo, para la serie de tiempo de las ventas de bicicletas, el valor del CME al utilizar la regresión lineal es 3.07, comparado con el valor de 3.97 que utiliza el suavizamiento exponencial lineal de Holt. La regresión de tendencia lineal también proporciona un pronóstico más exacto con la medida EAM de exactitud del pronóstico. Para la serie de tiempo de las ventas de bicicletas, la regresión de tendencia lineal resulta con un valor de EAM de 1.32 en comparación con un valor de 1.67 según el método lineal de Holt. Sin embargo, al basarse en el EPAM, el suavizamiento exponencial lineal de Holt (EPAM = 5.07%) es mejor que la regresión de tendencia lineal (6.42%). Por tanto, para la serie de tiempo de ventas de bicicletas, decidir cuáles son los métodos con los que se obtienen pronósticos más exactos depende de qué medida de la exactitud del pronóstico se utilice.

Ejercicios

Métodos

AUTOevaluación

17. Considere los datos siguientes de serie de tiempo.

t	1	2	3	4	5
Y_t	6	11	9	14	15

a) Construya una gráfica de series de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Desarrolle la ecuación de tendencia lineal para esta serie de tiempo.
c) ¿Cuál es el pronóstico para $t = 6$?

18. Consulte la serie de tiempo del ejercicio 17. Utilice el método de suavizamiento lineal exponencial de Holt con $\alpha = 0.3$ y $\beta = 0.5$ y obtenga un pronóstico para $t = 6$.

19. Considere la siguiente serie de tiempo.

t	1	2	3	4	5	6	7
Y_t	120	110	100	96	94	92	88

a) Construya una gráfica de series de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Desarrolle la ecuación de tendencia lineal para esta serie de tiempo.
c) ¿Cuál es el pronóstico para $t = 8$?

20. Considere la siguiente serie de tiempo.

t	1	2	3	4	5	6	7
Y_t	82	60	44	35	30	29	35

a) Construya una gráfica de series de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Utilice Minitab o Excel para desarrollar la ecuación de tendencia cuadrática para la serie de tiempo.
c) ¿Cuál es el pronóstico para $t = 8$?

Aplicaciones

AUTOevaluación

21. Debido a los altos costos de inscripción en las universidades estatales y privadas, las matrículas en los colegios de educación profesional técnica (community colleges) se han incrementado drásticamente en los últimos años. Los siguientes datos muestran la inscripción (en miles) en el Jefferson Community College de 2001 a 2009.

Año	Periodo (*t*)	Inscripción (en miles)
2001	1	6.5
2002	2	8.1
2003	3	8.4
2004	4	10.2
2005	5	12.5
2006	6	13.3
2007	7	13.7
2008	8	17.2
2009	9	18.1

a) Construya una gráfica de series de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Desarrolle la ecuación de tendencia lineal para esta serie de tiempo.
c) ¿Cuál es el pronóstico para 2010?

22. El Seneca Children's Fund (SCF) es una organización de caridad local que dirige un campamento de verano para niños desprotegidos. El consejo de administración ha trabajado muy duro en los últimos años para reducir la cantidad de gastos generales, un factor importante en la forma en que las organizaciones de caridad son recomendadas por los organismos independientes. Los siguientes datos muestran el porcentaje del total de dinero recaudado que SCF ha invertido en gastos administrativos y en campañas de recaudación de fondos para 2003-2009.

Año	Periodo (*t*)	Gasto (%)
2003	1	13.9
2004	2	12.2
2005	3	10.5
2006	4	10.4
2007	5	11.5
2008	6	10.0
2009	7	8.5

a) Construya una gráfica de series de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Desarrolle la ecuación de tendencia lineal para esta serie de tiempo.
c) Pronostique el porcentaje de gastos administrativos para 2010.
d) Si SCF puede mantener su actual tendencia en la reducción de gastos administrativos, ¿cuánto tiempo le llevará alcanzar un nivel de 5% o menos?

23. El presidente de una pequeña empresa de manufactura está preocupado por el continuo aumento en los costos de fabricación de los últimos años. Las cifras siguientes presentan una serie de tiempo del costo por unidad del producto principal de la empresa en los últimos ocho años.

Año	Costo unitario ($)	Año	Costo unitario ($)
1	20.00	5	26.60
2	24.50	6	30.00
3	28.20	7	31.00
4	27.50	8	36.00

a) Construya una gráfica de series de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Desarrolle la ecuación de tendencia lineal para esta serie de tiempo.
c) ¿En qué porcentaje han aumentado los costos de la empresa cada año?
d) Proporcione un cálculo estimado del costo unitario para el próximo año.

24. FRED® (Datos económicos de la Reserva Federal), una base de datos con más de 3 000 series de tiempo económicas de Estados Unidos, contiene datos históricos sobre los tipos de cambio. Los

datos siguientes muestran el tipo de cambio (Rate) por año (Year) y mes (Month) para Estados Unidos y China (página web del Banco de la Reserva Federal de St. Louis). Las unidades para el tipo de cambio son el número de yuanes chinos por un dólar estadounidense.

WEB archivo
ExchangeRate

Year	Month	Rate
2007	October	7.5019
2007	November	7.4210
2007	December	7.3682
2008	January	7.2405
2008	February	7.1644
2008	March	7.0722
2008	April	6.9997
2008	May	6.9725
2008	June	6.8993
2008	July	6.8355

a) Construya una gráfica de series de tiempo. ¿Existe una tendencia lineal?
b) Desarrolle con Minitab o Excel una ecuación de tendencia lineal para esta serie de tiempo.
c) Utilice la ecuación de tendencia para pronosticar el tipo de cambio en agosto de 2008.
d) ¿Se sentiría usted cómodo utilizando la ecuación de tendencia para pronosticar el tipo de cambio de diciembre de 2008?

25. En la siguiente serie de tiempo se presentan las ventas de automóviles de B.J. Scott Motors, Inc. de 10 años.

Año	Ventas	Año	Ventas
1	400	6	260
2	390	7	300
3	320	8	320
4	340	9	340
5	270	10	370

a) Construya una gráfica de series de tiempo. Comente acerca de la conveniencia de utilizar una tendencia lineal.
b) Utilice Minitab o Excel para desarrollar una ecuación de tendencia cuadrática para pronosticar las ventas.
c) Con la ecuación de tendencia del inciso *b*), pronostique las ventas para el año 11.
d) Sugiera una alternativa de una ecuación de tendencia cuadrática para pronosticar las ventas. Explique.

26. Giovanni Food Products prepara y vende pizzas congeladas en las escuelas públicas del este de Estados Unidos. Al utilizar una estrategia de marketing muy agresiva, ha sido capaz de incrementar sus ingresos anuales en $10 millones en los últimos 10 años. Pero el incremento de la competencia ha disminuido su tasa de crecimiento. A continuación se muestran los ingresos (Revenue) por año (Year) en millones de dólares de la última década.

WEB archivo
Pasta

Year	Revenue
1	8.53
2	10.84
3	12.98
4	14.11
5	16.31
6	17.21
7	18.37
8	18.45
9	18.40
10	18.43

a) Construya una gráfica de series de tiempo. Comente acerca de la conveniencia de utilizar una tendencia lineal.
b) Utilice Minitab o Excel para desarrollar una ecuación de tendencia cuadrática para pronosticar los ingresos.
c) Utilizando la ecuación de tendencia desarrollada en el inciso *b*), pronostique los ingresos en el año 11.

27. La revista *Forbes* clasifica los equipos de la NFL por su valor anual. Los datos siguientes son los valores de Indianapolis Colts de 1998 a 2008 (sitio web de Forbes). Se proporcionan año (Year), periodo (Period) y valor (Value) en millones de $.

WEB archivo
NFLValue

Year	Period	Value ($millions)
1998	1	227
1999	2	305
2000	3	332
2001	4	367
2002	5	419
2003	6	547
2004	7	609
2005	8	715
2006	9	837
2007	10	911
2008	11	1076

a) Construya una gráfica de series de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Utilice Minitab o Excel y obtenga una ecuación de tendencia cuadrática para pronosticar el valor del equipo.
c) Al utilizar Minitab o Excel, obtenga la ecuación de tendencia exponencial para pronosticar el valor del equipo.
d) Con Minitab o Excel obtenga la ecuación de tendencia lineal para pronosticar el valor del equipo.
e) ¿Qué ecuación recomendaría utilizar para estimar el valor del equipo en 2009?
f) Utilice el modelo del inciso *e*) para pronosticar el valor de los Colts en 2009.

18.5 Estacionalidad y tendencia

En esta sección se muestra cómo desarrollar pronósticos para una serie de tiempo que tiene un patrón estacional. En la medida en que la estacionalidad existe, se debe incorporar a los modelos de elaboración de pronósticos para garantizar un pronóstico exacto. Se comenzará por considerar una serie de tiempo estacional sin tendencia y después se estudiará cómo modelar la estacionalidad con la tendencia.

Estacionalidad sin tendencia

Como ejemplo, considere el número de paraguas vendidos en una tienda de ropa en los últimos cinco años. La tabla 18.17 muestra la serie de tiempo y la figura 18.16 la gráfica correspondiente. Esta última no indica una tendencia a largo plazo en las ventas. De hecho, a menos que se observen cuidadosamente los datos, es posible concluir que éstos siguen un patrón horizontal y que un suavizamiento exponencial sencillo podría utilizarse para pronosticar las ventas. Pero una inspección más cercana a la gráfica de serie de tiempo revela un patrón en los datos. Es decir, el primer y tercer trimestres registran ventas moderadas, el segundo tiene las ventas más altas, y el cuarto tiende a ser el trimestre más bajo en volumen de ventas. Por tanto, se podría concluir que un patrón estacional trimestral está presente.

WEB archivo
Umbrella

En el capítulo 15 se mostró cómo utilizar las variables ficticias o binarias (dummy) para tratar las variables cualitativas independientes en un modelo de regresión múltiple. Se puede utilizar el mismo método para modelar una serie de tiempo con un patrón estacional al tratar la estación

como variable categórica. Recuerde que cuando esta variable tiene k niveles, se necesitan $k - 1$ variables ficticias. Por tanto, si hay cuatro estaciones, se requieren tres variables ficticias. Por ejemplo, la serie de tiempo de la temporada de ventas de sombrillas es una variable cualitativa con cuatro niveles: trimestre 1, trimestre 2, trimestre 3 y trimestre 4. Por tanto, para modelar los efectos estacionales en la serie de tiempo de las sombrillas se necesitan $4 - 1 = 3$ variables ficticias. Éstas pueden ser codificadas de la siguiente manera.

$$\text{Qtr 1} = \begin{cases} 1 \text{ si el trimestre es 1} \\ 0 = \text{en otro caso} \end{cases} \quad \text{Qtr 2} = \begin{cases} 1 \text{ si el trimestre es 2} \\ 0 = \text{en otro caso} \end{cases} \quad \text{Qtr 3} = \begin{cases} 1 \text{ si el trimestre es 3} \\ 0 = \text{en otro caso} \end{cases}$$

Usando $\hat{Y}$ para denotar el valor estimado o pronosticado de las ventas, la forma general de la ecuación de regresión estimada relacionada con el número de sombrillas vendidas con base en el trimestre (Qtr) en que las ventas ocurren es la siguiente:

$$\hat{Y} = b_0 + b_1 \text{ Qtr1} + b_2 \text{ Qtr2} + b_3 \text{ Qtr3}$$

La tabla 18.18 es la serie de tiempo de las ventas de sombrillas que muestra los valores codificados de las variables ficticias. Al utilizar estos datos y el procedimiento de regresión de Minitab, se obtuvo el resultado de computadora mostrado en la figura 18.17. La ecuación de regresión múltiple estimada obtenida es

$$\text{Ventas} = 95.0 + 29.0 \text{ Qtr1} + 57.0 \text{ Qtr2} + 26.0 \text{ Qtr3}$$

Esta ecuación se puede utilizar para pronosticar las ventas trimestrales para el próximo año.

Trimestre 1. Ventas $= 95.0 + 29.0(1) + 57.0(0) + 26.0(0) = 124$
Trimestre 2. Ventas $= 95.0 + 29.0(0) + 57.0(1) + 26.0(0) = 152$
Trimestre 3. Ventas $= 95.0 + 29.0(0) + 57.0(0) + 26.0(1) = 121$
Trimestre 4. Ventas $= 95.0 + 29.0(0) + 57.0(1) + 26.0(0) = 95$

TABLA 18.17
Serie de tiempo de las ventas de sombrillas

Año	Trimestre	Ventas
1	1	125
	2	153
	3	106
	4	88
2	1	118
	2	161
	3	133
	4	102
3	1	138
	2	144
	3	113
	4	80
4	1	109
	2	137
	3	125
	4	109
5	1	130
	2	165
	3	128
	4	96

FIGURA 18.16 Gráfica de la serie de tiempo de ventas de sombrillas

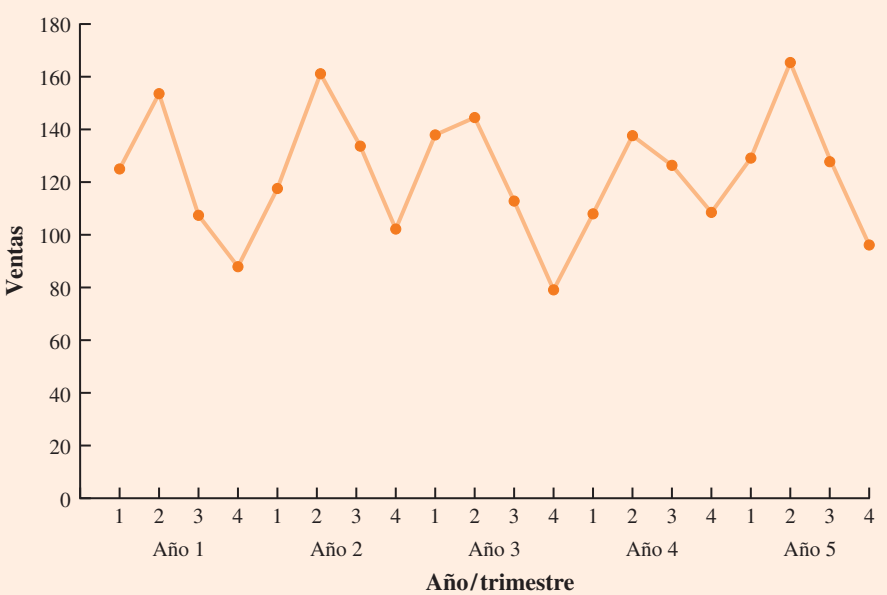


TABLA 18.18 Serie de tiempo de ventas de sombrillas con variables ficticias

Año	Trimestre	Trim 1	Trim 2	Trim 3	Ventas
1	1	1	0	0	125
	2	0	1	0	153
	3	0	0	1	106
	4	0	0	0	88
2	1	1	0	0	118
	2	0	1	0	161
	3	0	0	1	133
	4	0	0	0	102
3	1	1	0	0	138
	2	0	1	0	144
	3	0	0	1	113
	4	0	0	0	80
4	1	1	0	0	109
	2	0	1	0	137
	3	0	0	1	125
	4	0	0	0	109
5	1	1	0	0	130
	2	0	1	0	165
	3	0	0	1	128
	4	0	0	0	96

Es interesante señalar que se habrían podido obtener los pronósticos trimestrales para el próximo año simplemente calculando el número promedio de sombrillas que se venden cada trimestre, como se muestra en la siguiente tabla.

Año	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
1	125	153	106	88
2	118	161	133	102
3	138	144	113	80
4	109	137	125	109
5	130	165	128	96
Promedio	124	152	121	95

Con todo, el resultado de regresión mostrado en la figura 18.17 proporciona información adicional que puede utilizarse para evaluar la exactitud del pronóstico y determinar la signifi-

FIGURA 18.17 Resultado de la regresión de Minitab para la serie de tiempo de ventas de sombrillas

```
The regression equation is
Sales = 95.0 + 29.0 Qtr1 + 57.0 Qtr2 + 26.0 Qtr3

Predictor     Coef  SE Coef      T      P
Constant    95.000    5.065  18.76  0.000
Qtr1        29.000    7.162   4.05  0.001
Qtr2        57.000    7.162   7.96  0.000
Qtr3        26.000    7.162   3.63  0.002
```

cancia de los resultados. Y para los tipos más complejos de situaciones problemáticas, como tratar con una serie de tiempo que tiene tanto los efectos de tendencia como estacionales, el método de un promedio simple no funcionará.

Estacionalidad y tendencia

WEB archivo

TVSales

TABLA 18.19

Serie de tiempo de las ventas de televisores

Año	Trimestre	Ventas (miles $)
1	1	4.8
	2	4.1
	3	6.0
	4	6.5
2	1	5.8
	2	5.2
	3	6.8
	4	7.4
3	1	6.0
	2	5.6
	3	7.5
	4	7.8
4	1	6.3
	2	5.9
	3	8.0
	4	8.4

Ahora ampliaremos el método de regresión para incluir situaciones en las que la serie de tiempo contiene tanto el efecto estacional como una tendencia lineal, y mostraremos cómo pronosticar la serie de tiempo de las ventas trimestrales de televisores presentadas en la sección 18.1. Los datos respectivos se muestran en la tabla 18.19. La gráfica de series de tiempo de la figura 18.18 indica que las ventas son muy bajas en el segundo trimestre de cada año y que aumentan en los trimestres 3 y 4. Por tanto, se concluye que existe un patrón estacional para las ventas de televisores. Pero la serie de tiempo tiene también una tendencia lineal ascendente que tendrá que tomarse en cuenta para obtener pronósticos exactos de las ventas trimestrales. Es fácil manejar y combinar el método de las variables ficticias por estacionalidad con el método de regresión de la serie de tiempo que se estudió en la sección 18.3 para el manejo de la tendencia lineal.

La forma general de la ecuación de regresión múltiple estimada para modelar tanto los efectos estacionales trimestrales como la tendencia lineal en la serie de tiempo de los televisores es la siguiente.

$$\hat{Y}_t = b_0 + b_1\,\text{Qtr1} + b_2\,\text{Qtr2} + b_3\,\text{Qtr3} + b_4 t$$

donde

$\hat{Y}_t$ = estimación o pronóstico de ventas en el periodo t
Qtr1 = 1 si el periodo t corresponde al primer trimestre del año; 0 en caso contrario
Qtr2 = 1 si el periodo t corresponde al segundo trimestre del año; 0 en caso contrario
Qtr3 = 1 si el periodo t corresponde al tercer trimestre del año; 0 en caso contrario
t = periodo

FIGURA 18.18 Gráfica de serie de tiempo para el conjunto de ventas de televisores

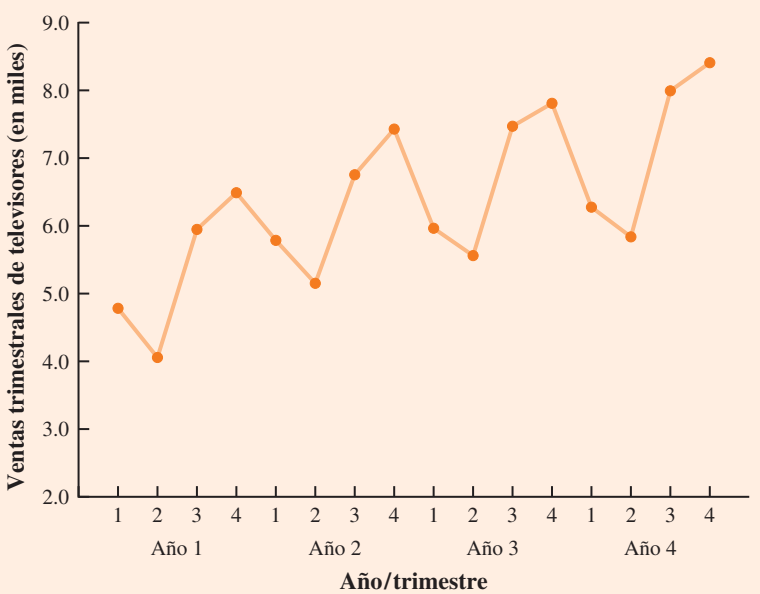


TABLA 18.20 Serie de tiempo de las ventas de televisores con variables ficticias y periodo

Año	Trimestre	Trim 1	Trim 2	Trim 3	Periodo	Ventas (miles)
1	1	1	0	0	1	4.8
	2	0	1	0	2	4.1
	3	0	0	1	3	6.0
	4	0	0	0	4	6.5
2	1	1	0	0	5	5.8
	2	0	1	0	6	5.2
	3	0	0	1	7	6.8
	4	0	0	0	8	7.4
3	1	1	0	0	9	6.0
	2	0	1	0	10	5.6
	3	0	0	1	11	7.5
	4	0	0	0	12	7.8
4	1	1	0	0	13	6.3
	2	0	1	0	14	5.9
	3	0	0	1	15	8.0
	4	0	0	0	16	8.4

La tabla 18.20 es la serie de tiempo revisada de las ventas de televisores que incluye los valores codificados de las variables ficticias y el periodo t. Al utilizar los datos de la tabla 18.20 y el procedimiento de regresión de Minitab se obtuvo el resultado de computadora que se muestra en la figura 18.19. La ecuación de regresión múltiple estimada es

$$\text{Ventas} = 6.07 - 1.36\,\text{Qtr1} - 2.03\,\text{Qtr2} - 0.304\,\text{Qtr3} + 0.146t \qquad \textbf{(18.12)}$$

Ahora se puede utilizar la ecuación (18.12) a efecto de pronosticar las ventas trimestrales para el próximo año, el cual es el año 5 para la serie de tiempo de ventas de televisores; es decir, los periodos 17, 18, 19 y 20.

Pronóstico para el periodo 17 (trimestre 1 en el año 5)

$$\text{Ventas} = 6.07 - 1.36(1) - 2.03(0) - 0.304(0) + 0.146(17) = 7.19$$

Pronóstico para el periodo 18 (trimestre 2 en el año 5)

$$\text{Ventas} = 6.07 - 1.36(0) - 2.03(1) - 0.304(0) + 0.146(18) = 6.67$$

FIGURA 18.19 Resultado de regresión de Minitab de la serie de tiempo de ventas de sombrillas

```
The regression equation is
Sales (1000s) = 6.07 - 1.36 Qtr1 - 2.03 Qtr2 - 0.304
                Qtr3 + 0.146 Period

Predictor      Coef  SE Coef       T      P
Constant     6.0688   0.1625   37.35  0.000
Qtr1        -1.3631   0.1575   -8.66  0.000
Qtr2        -2.0337   0.1551  -13.11  0.000
Qtr3        -0.3044   0.1537   -1.98  0.073
Period      0.14562  0.01211   12.02  0.000
```

Pronóstico para el periodo 19 (trimestre 3 en el año 5)

$$\text{Ventas} = 6.07 - 1.36(0) - 2.03(0) - 0.304(1) + 0.146(19) = 8.54$$

Pronóstico para el periodo 20 (trimestre 4 en el año 5)

$$\text{Ventas} = 6.07 - 1.36(0) - 2.03(0) - 0.304(0) + 0.146(20) = 8.99$$

Por tanto, tomando en cuenta los efectos tanto estacionales como de tendencia lineal en las ventas de televisores, las estimaciones de las ventas trimestrales en el año 5 son 7 190, 6 670, 8 540 y 8 990.

Las variables ficticias en la ecuación de regresión múltiple estimada realmente ofrecen cuatro ecuaciones de regresión múltiple estimadas, una para cada trimestre. Por ejemplo, si el periodo t corresponde al trimestre 1, la ecuación estimada para las ventas es

$$\textit{Trimestre}\ 1.\ \text{Ventas} = 6.07 - 1.36(1) - 2.03(0) - 0.304(0) + 0.146t = 4.71 + 0.146t$$

Del mismo modo, si el periodo t corresponde a los trimestres 2, 3 y 4, las estimaciones para las ventas trimestrales son:

$$\textit{Trimestre}\ 2.\ \text{Ventas} = 6.07 - 1.36(0) - 2.03(1) - 0.304(0) + 0.146t = 4.04 + 0.146t$$
$$\textit{Trimestre}\ 3.\ \text{Ventas} = 6.07 - 1.36(0) - 2.03(0) - 0.304(1) + 0.146t = 5.77 + 0.146t$$
$$\textit{Trimestre}\ 4.\ \text{Ventas} = 6.07 - 1.36(0) - 2.03(0) - 0.304(0) + 0.146t = 6.07 + 0.146t$$

La pendiente de la tendencia lineal para cada ecuación de pronóstico trimestral es 0.146, lo que indica un crecimiento en las ventas de alrededor de 146 televisores por trimestre. La única diferencia en las cuatro ecuaciones estriba en que tienen diferentes intersecciones. Por ejemplo, la intersección en la ecuación del trimestre 1 es 4.71 y para el trimestre 4 es 6.07. Por tanto, las ventas en el trimestre 1 son $4.71 - 6.07 = -1.36$ o 1 360 televisores menos que en el trimestre 4. En otras palabras, el coeficiente estimada en la regresión para Qtr1 en la ecuación (18.12) proporciona una estimación de la diferencia en las ventas entre los trimestres 1 y 4. Interpretaciones similares pueden darse para -2.03, el coeficiente estimado para la variable ficticia Qtr2, y para -0.304, el coeficiente estimado para la variable ficticia Qtr3.

Modelos basados en datos mensuales

Cuando una variable categórica como la estación tiene k niveles, son necesarias $k - 1$ variables ficticias.

En el ejemplo anterior de las ventas de televisores se mostró cómo utilizar las variables ficticias para tomar en cuenta los efectos estacionales trimestrales de la serie de tiempo. Debido a que hubo cuatro niveles para la variable cualitativa estacional, se requirieron tres variables ficticias. Sin embargo, numerosas empresas utilizan los pronósticos mensuales más que los trimestrales. Para los datos mensuales, la estación es una variable cualitativa con 12 niveles y, por tanto, son obligatorias $12 - 1 = 11$ variables ficticias. Por ejemplo, las 11 variables ficticias pueden codificarse de la siguiente manera:

$$\text{Mes 1} = \begin{cases} 1 \text{ si es enero} \\ 0 \text{ en otro caso} \end{cases}$$
$$\text{Mes 2} = \begin{cases} 1 \text{ si es febrero} \\ 0 \text{ en otro caso} \end{cases}$$
$$\vdots$$
$$\text{Mes 11} = \begin{cases} 1 \text{ si es noviembre} \\ 0 \text{ en otro caso} \end{cases}$$

Fuera de este cambio, el método de regresión múltiple para el manejo de la estacionalidad es el mismo.

Ejercicios

Métodos

AUTOevaluación

28. Considere la siguiente serie de tiempo.

Trimestre	Año 1	Año 2	Año 3
1	71	68	62
2	49	41	51
3	58	60	53
4	78	81	72

a) Construya una gráfica de series de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Utilice las siguientes variables ficticias para desarrollar una ecuación de regresión estimada que tome en cuenta los efectos estacionales en los datos: Qtr1 = 1 si es el trimestre 1, 0 en caso contrario; Qtr2 = 1 si es el trimestre 2, 0 en caso contrario; Qtr 3 = 1 si es el trimestre 3, 0 en caso contrario.
c) Calcule los pronósticos trimestrales para el año siguiente.

29. Considere los datos siguientes de series de tiempo.

Trimestre	Año 1	Año 2	Año 3
1	4	6	7
2	2	3	6
3	3	5	6
4	5	7	8

a) Construya una gráfica de serie de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Utilice las siguientes variables ficticias para desarrollar una ecuación de regresión estimada que tome en cuenta cualquier efecto estacional y de tendencia lineal en los datos: Qtr1 = 1 si el trimestre es 1, 0 en caso contrario; Qtr 2 = 1 si el trimestre es 2, 0 en caso contrario; Qtr3 = 1 si el trimestre es 3, 0 en caso contrario.
c) Calcule los pronósticos trimestrales para el próximo año.

Aplicaciones

30. Los datos de las ventas trimestrales (número de ejemplares vendidos) para un libro de texto universitario en los últimos tres años son los siguientes.

Trimestre	Año 1	Año 2	Año 3
1	1690	1800	1850
2	940	900	1100
3	2625	2900	2930
4	2500	2360	2615

a) Construya una gráfica de series de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Utilice las siguientes variables ficticias para desarrollar una ecuación de regresión estimada que tome en cuenta los efectos estacionales de los datos: Qrt = 1 si el trimestre es 1, 0 en caso contrario; Qtr2 = 1 si el trimestre es 2, 0 en caso contrario; Qtr3 = 1 si el trimestre es 3, 0 en caso contrario.

c) Calcule los pronósticos trimestrales para el próximo año.

d) Suponga que $t = 1$ se refiere a la observación del trimestre 1 del año 1; $t = 2$ a la observación del trimestre 2 del año 1; . . . y $t = 12$ a la observación del trimestre 4 del año 3. Utilice las variables ficticias definidas en el inciso *b*) así como *t*, para desarrollar una ecuación de regresión estimada que tome en cuenta los efectos estacionales y cualquier tendencia lineal en la serie de tiempo. A partir de los efectos estacionales en los datos y la tendencia lineal, calcule los pronósticos trimestrales para el próximo año.

31. Especialistas en el control de la contaminación del aire en el sur de California monitorean cada hora la cantidad de ozono, dióxido de carbono y dióxido de nitrógeno contenida en el aire. Los datos de serie de tiempo de cada hora presentan estacionalidad, con niveles de contaminantes que muestran patrones que varían cada hora (Hour) durante el día. El 15, 16 y 17 de julio se observaron los siguientes niveles para las 12 horas a partir de las 6:00 a.m. hasta las 6:00 p.m.

WEB archivo
Pollution

15 de julio	25	28	35	50	60	60	40	35	30	25	25	20
16 de julio	28	30	35	48	60	65	50	40	35	25	20	20
17 de julio	35	42	45	70	72	75	60	45	40	25	25	25

a) Construya una gráfica de series de tiempo. ¿Qué tipo de patrón existe en los datos?

b) Utilice las siguientes variables ficticias para obtener una ecuación de regresión estimada que tome en cuenta los efectos estacionales de los datos.

Hour1 = 1 si la lectura se realizó entre las 6:00 a.m. y las 7:00 a.m.; 0 de otra forma
Hour2 = 1 si la lectura se realizó entre las 7:00 a.m. y las 8:00 a.m.; 0 de otra forma
.
.
.
Hour11 = 1 si la lectura se realizó entre las 4:00 p.m. y las 5:00 p.m.; 0 de otra forma

Note que cuando los valores de las 11 variables ficticias son iguales a 0, la observación corresponde a la hora entre las 5:00 p.m. y las 6:00 p.m.

c) Utilizando la ecuación de regresión estimada obtenida en el inciso *a*), calcule estimaciones de los niveles de dióxido de nitrógeno para el 18 de julio.

d) Suponga que $t = 1$ se refiere a la observación en la hora 1 del 15 de julio; $t = 2$ a la observación en la hora 2 del 15 de julio, . . . y $t = 36$ a la observación en la hora 12 del 17 de julio. Utilice las variables ficticias definidas en el inciso *b*) y *t*, para desarrollar una ecuación de regresión estimada que tome en cuenta los efectos estacionales y de tendencia lineal de la serie de tiempo. Con base en los efectos estacionales de los datos y la tendencia lineal, calcule las estimaciones de los niveles de dióxido de nitrógeno para el 18 de julio.

32. South Shore Construction edifica muelles y diques permanentes a lo largo de la costa sur de Long Island, en Nueva York. Aunque la empresa ha estado en el negocio sólo cinco años, sus ingresos han aumentado de $308 000 en el primer año de operación hasta $1 084 000 en el año más reciente. Los siguientes datos muestran los ingresos por trimestre (Quarter) en miles de dólares para cada año (Year).

WEB archivo
SouthShore

Quarter	Year 1	Year 2	Year 3	Year 4	Year 5
1	20	37	75	92	176
2	100	136	155	202	282
3	175	245	326	384	445
4	13	26	48	82	181

a) Construya una gráfica de series de tiempo. ¿Qué tipo de patrón existe en los datos?

b) Utilice las siguientes variables ficticias para desarrollar una ecuación de regresión estimada que tome en cuenta los efectos estacionales de los datos. Qtr1 = 1 si es el trimestre 1, 0 en caso contrario; Qtr2 = 1 si es el trimestre 2, 0 en caso contrario; Qtr3 = 1 si es

el trimestre 3, 0 en caso contrario. Con base sólo en los efectos estacionales de los datos, calcule las estimaciones de las ventas trimestrales para el año 6.

c) Suponga que el periodo = 1 se refiere a la observación en el trimestre 1 del año 1; periodo = 2 a la observación del trimestre 2 del año 1; . . . y periodo = 20 a la observación en el trimestre 4 del año 5. Utilice las variables ficticias que se definen en el inciso *b*) y el periodo para desarrollar una ecuación de regresión estimada que tome en cuenta los efectos estacionales y de cualquier tendencia lineal de la serie de tiempo. Con base en ambos efectos, calcule las estimaciones de las ventas trimestrales para el año 6.

33. El consumo de energía eléctrica se mide en kilowatts-hora (kWh). La compañía de servicios local ofrece un programa de interrupción por el cual los clientes comerciales participantes reciben tarifas favorables, pero deberán reducir el consumo si la empresa se los pide. Timko Products ha acordado reducir el consumo los jueves desde las 8:00 p.m. Para determinar los ahorros respectivos, la empresa debe calcular el uso de energía normal de Timko durante este periodo. Los datos de su gasto de energía eléctrica para las 72 horas anteriores se muestran a continuación. Se incluyen los consumos de los días lunes (Monday), martes (Tuesday), miércoles (Wednesday) y jueves (Thursday) con su respectivo periodo (Time Period).

WEB archivo
Power

Time Period	Monday	Tuesday	Wednesday	Thursday
12–4 a.m.	—	19 281	31 209	27 330
4–8 a.m.	—	33 195	37 014	32 715
8–12 noon	—	99 516	119 968	152 465
12–4 p.m.	124 299	123 666	156 033	
4–8 p.m.	113 545	111 717	128 889	
8–12 midnight	41 300	48 112	73 923	

a) Construya una gráfica de series de tiempo. ¿Qué tipo de patrón existe en los datos?

b) Utilice las siguientes variables ficticias para desarrollar una ecuación de regresión estimada que tome en cuenta los efectos estacionales.

Time1 = 1 para el periodo de 12 a 4 a.m.; 0 en caso contrario
Time2 = 1 para el periodo de 4 a 8 a.m.; 0 en caso contrario
Time3 = 1 para el periodo de 8 a 12 del mediodía; 0 en caso contrario
Time4 = 1 para el periodo de 12 a 4 p.m.; 0 en caso contrario
Time5 = 1 para el periodo de 4 a 8 p.m.; 0 en caso contrario

c) Utilice la ecuación de regresión estimada obtenida en el inciso *b*) para calcular el uso normal de Timko a lo largo del periodo en que el servicio será interrumpido.

d) Suponga que el periodo = 1 se refiere a la observación del lunes en el periodo de 12 a 4 p.m.; el periodo = 2 a la observación del lunes del periodo de 4 a 8 p.m.; . . . y el periodo = 18 a la observación del jueves en el periodo 8 a 12 del mediodía. Utilice las variables ficticias definidas en el inciso *b*), así como el periodo para desarrollar una ecuación de regresión estimada que tome en cuenta los efectos estacionales y de cualquier tendencia lineal de la serie de tiempo.

e) Utilice la ecuación obtenida en el inciso *d*) para estimar el uso normal de Timko a lo largo del periodo en que el servicio estará interrumpido.

34. Los gastos del mantenimiento del césped ($) por mes (Month) para un edificio de apartamentos de seis unidades en el sur de la Florida durante tres años (Year) son los siguientes.

WEB archivo
AptExp

Month	Year 1	Year 2	Year 3
January	170	180	195
February	180	205	210
March	205	215	230
April	230	245	280
May	240	265	290

Month	Year 1	Year 2	Year 3
June	315	330	390
July	360	400	420
August	290	335	330
September	240	260	290
October	240	270	295
November	230	255	280
December	195	220	250

a) Construya una gráfica de series de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Desarrolle una ecuación de regresión estimada que tome en cuenta cualquier efecto de estacionalidad o de tendencia lineal en los datos. Utilice las siguientes variables ficticias que tomen en cuenta los efectos estacionales: Jan = 1 si es enero, 0 en caso contrario; Feb = 1 si es febrero, 0 en caso contrario; Mar = 1 si es marzo, 0 en caso contrario; . . . Nov = 1 si es noviembre, 0 en caso contrario. Observe que al utilizar este método de codificación, cuando las 11 variables ficticias son 0, la observación corresponde a un gasto en diciembre.
c) Calcule los pronósticos mensuales para el siguiente año con base en los efectos tanto de tendencia como estacionales.

18.6 Descomposición de series de tiempo

En esta sección se enfoca la atención en lo que se conoce como **descomposición de series de tiempo**, la cual se puede utilizar para separar o descomponer una serie de tiempo en su parte de tendencia y estacional y en su componente irregular, aunque este método puede utilizarse para el pronóstico, y su aplicación principal es conseguir una mejor comprensión de la serie de tiempo. Muchas series de tiempo para las empresas y la economía de Estados Unidos son sostenidas y publicadas por agencias gubernamentales como la Oficina del Censo y la Oficina de Estadísticas Laborales. Estas agencias utilizan la descomposición de series de tiempo para crear series desestacionalizadas.

Entender qué sucede en realidad con una serie de tiempo a menudo depende del uso de los datos desestacionalizados. Por ejemplo, podríamos estar interesados en saber si el consumo de energía eléctrica está aumentando en nuestra área. Suponga que se entera de que éste se redujo 3% en septiembre con respecto al mes anterior. Se debe ser cuidadoso al ejercitar el uso de esa información, porque cada vez que una influencia estacional está presente, esas comparaciones pueden ser engañosas si los datos no han sido desestacionalizados. El hecho de que el consumo de energía eléctrica se haya reducido 3% entre agosto y septiembre podría ser por el efecto estacional que se relaciona con una disminución en el uso del aire acondicionado y no por una disminución del uso de la energía eléctrica a largo plazo. En efecto, después de ajustar el efecto estacional, se podría incluso determinar que el consumo de energía eléctrica aumentó. Muchas series de tiempo, como las estadísticas de desempleo y las ventas de casas y al detalle (o al menudeo), están sujetas a fuertes influencias estacionales. Es importante desestacionalizar dichos datos, antes de emitir un juicio acerca de cualquier tendencia a largo plazo.

Los métodos de descomposición de series tiempo asumen que Y_t, el valor real de la serie de tiempo en el periodo t, es una función de tres componentes: un componente de tendencia, un componente estacional y un componente irregular o de error. El cómo estos tres componentes se combinan para generar los valores observados de la serie de tiempo depende de si se asume que la relación entre ellos se describe mejor por un modelo aditivo o un modelo multiplicativo.

Un **modelo de descomposición aditiva** tiene la siguiente forma:

$$Y_t = \text{Tendencia}_t + \text{Estacional}_t + \text{Irregular}_t \qquad \textbf{(18.13)}$$

donde

$$\text{Tendencia}_t = \text{valor de la tendencia en el periodo } t$$
$$\text{Estacional}_t = \text{valor estacional en el periodo } t$$
$$\text{Irregular}_t = \text{valor irregular en el periodo } t$$

El componente irregular corresponde al término de error ε en el modelo de regresión lineal simple estudiado en el capítulo 14.

En el modelo aditivo, los valores de los tres componentes simplemente se suman para obtener el valor real de las series de tiempo Y_t. El componente irregular o de error toma en cuenta la variabilidad de la serie de tiempo que no puede ser explicada por los componentes de tendencia y estacional.

Un modelo aditivo es apropiado en situaciones en las que las fluctuaciones estacionales no dependen del nivel de la serie de tiempo. El modelo de regresión que incorpora efectos estacionales y de tendencia en la sección 18.5 es un modelo aditivo. Éste es apropiado si las fluctuaciones estacionales en el periodo anterior son casi del mismo tamaño que las fluctuaciones estacionales en periodos posteriores. Sin embargo, si las fluctuaciones estacionales cambian en el tiempo y son cada vez mayores a medida que aumenta el volumen de ventas debido a una tendencia lineal a largo plazo, entonces se debe utilizar el modelo multiplicativo. Muchas series de tiempo para las empresas y para la economía siguen este patrón.

Un **modelo de descomposición multiplicativa** toma la siguiente forma:

$$Y_t = \text{Tendencia}_t \times \text{Estacional}_t \times \text{Irregular}_t \qquad \textbf{(18.14)}$$

donde

$$\text{Tendencia}_t = \text{valor de la tendencia en el periodo } t$$
$$\text{Estacional}_t = \text{índice estacional en el periodo } t$$
$$\text{Irregular}_t = \text{índice irregular en el periodo } t$$

La Oficina del Censo utiliza un modelo multiplicativo en conjunción con su metodología para desestacionalizar las series de tiempo.

En este modelo los componentes de tendencia, estacional e irregular se multiplican para dar el valor de la serie de tiempo. La tendencia se mide en las unidades de producto de la serie que se pronostica. Sin embargo, los componentes estacional e irregular se miden en términos relativos, con valores superiores a 1.00 indicando los efectos por arriba de la tendencia y con valores menores a 1.00 indicando los efectos por debajo de la tendencia.

Debido a que este es el método más utilizado en la práctica, nuestro análisis de descomposición de las series de tiempo se limitará a mostrar cómo se desarrollan las estimaciones de los componentes de tendencia y estacional de un modelo multiplicativo. A modo de ejemplo, se trabajará con la serie de tiempo de las ventas trimestrales de televisores presentada en la sección 18.5; los datos de las ventas trimestrales se muestran en la tabla 18.19 y la gráfica correspondiente de la serie de tiempo se presenta en la figura 18.18. Después de demostrar cómo se descompone una serie de tiempo con el modelo multiplicativo, se estudiará cómo los índices estacionales y el componente de tendencia pueden ser recombinados para elaborar un pronóstico.

Cálculo de los índices estacionales

La figura 18.18 indica que las ventas son más bajas en el segundo trimestre de cada año y aumentan en los trimestres 3 y 4. Por tanto, se concluye que existe un comportamiento estacional para la serie de tiempo de las ventas de televisores. El procedimiento utilizado para determinar la influencia estacional de cada trimestre empieza por calcular un promedio móvil para separar los componentes estacional e irregular de los datos, lo que deja una serie de tiempo que contiene sólo la tendencia y cualquier variación aleatoria restante que no fue eliminada por los cálculos del promedio móvil.

Como se trabaja con series trimestrales, se utilizarán cuatro valores en cada promedio móvil. El cálculo del promedio móvil de los primeros cuatro trimestres de ventas de televisores es

$$\text{Primer promedio móvil} = \frac{4.8 + 4.1 + 6.0 + 6.5}{4} = \frac{21.4}{4} = 5.35$$

Observe que el cálculo del promedio móvil de los primeros cuatro trimestres da el promedio trimestral de las ventas durante el año 1 de la serie de tiempo. Para continuar con este cálculo se agrega el valor 5.8 correspondiente al primer trimestre del año 2 y se elimina el 4.8 del primer trimestre del año 1. Por tanto, el segundo promedio móvil es

$$\text{Segundo promedio móvil} = \frac{4.1 + 6.0 + 6.5 + 5.8}{4} = \frac{22.4}{4} = 5.60$$

De manera similar, el cálculo del tercer promedio móvil es $(6.0 + 6.5 + 5.8 + 5.2)/4 = 5.875$.

Antes de continuar con el cálculo de los promedios móviles de toda la serie de tiempo, regrese al primero que resultó en un valor de 5.35. Éste es el promedio trimestral del volumen de ventas para el año 1. Al retroceder en su cálculo, parece razonable asociar el valor 5.35 con el "central" del grupo del promedio móvil. Sin embargo, observe que como en cada promedio móvil intervienen cuatro trimestres, no hay trimestre central. El valor 5.35 corresponde en realidad al periodo 2.5, la segunda mitad del trimestre 2 y la primera mitad del trimestre 3. De manera similar, al pasar al valor del siguiente promedio móvil, que es 5.60, el trimestre central corresponderá al periodo 3.5, la última mitad del trimestre 3 y la primera mitad del 4.

Los dos valores del promedio móvil que se calculan no corresponden directamente a los trimestres originales de la serie de tiempo. Esta dificultad se resuelve calculando el promedio de los dos promedios móviles. Ya que el centro del primero es el periodo 2.5 (la mitad de un periodo o trimestre más temprano) y el centro del segundo es el periodo 3.5 (la mitad del periodo o trimestre más tarde), el promedio de los dos promedios móviles se centra en el trimestre 3, exactamente donde debe estar. Este promedio se conoce como *promedio móvil centrado*, y para el periodo 3 es $(5.35 + 5.60)/2 = 5.475$, mientras que para el periodo 4 es $(5.60 + 5.875)/2 = 5.738$. La tabla 18.21 muestra un resumen completo de los cálculos del promedio móvil y del promedio móvil centrado para los datos de las ventas de televisores.

¿Qué información se obtiene de los promedios móviles centrados de la tabla 18.21 de esta serie de tiempo? La figura 18.20 muestra una gráfica de los valores reales de la serie de tiempo y de los valores de los promedios móviles centrados. Observe sobre todo cómo estos últimos tienden a "suavizar" tanto las fluctuaciones estacionales como las irregulares de la serie de tiempo. Los promedios móviles centrados representan la tendencia en los datos y cualquier variación aleatoria que no se ha eliminado con el uso de los promedios móviles para suavizar los datos.

Antes se demostró que el modelo de descomposición multiplicativa es

$$Y_t = \text{Tendencia}_t \times \text{Estacional}_t \times \text{Irreglar}_t$$

Al dividir cada lado de esta ecuación entre el componente de tendencia T_1, se puede identificar el efecto estacional irregular en la serie de tiempo.

Los valores del componente irregular estacional son llamados a menudo valores de la serie de tiempo sin tendencia.

$$\frac{Y_t}{\text{Tendencia}_t} = \frac{\text{Tendencia}_t \times \text{Estacional}_t \times \text{Irregular}_t}{\text{Tendencia}_t} = \text{Estacional}_t \times \text{Irregular}_t$$

Por ejemplo, el tercer trimestre del año 1 muestra un valor de tendencia de 5.475 (el promedio móvil centrado). Así $6.0/5.475 = 1.096$ es el valor combinado estacional-irregular del componente irregular. En la tabla 18.22 se resumen los valores del componente estacional-irregular de toda la serie de tiempo.

Considere los valores del componente estacional-irregular para el tercer trimestre: 1.096, 1.075 y 1.109. Los valores de la parte estacional-irregular mayores de 1.00 indican efectos por encima de la tendencia estimada, y los valores menores de 1.00 indican efectos por debajo de la tendencia estimada. Así, los tres valores del componente estacional-irregular para el trimestre 3 muestran un efecto por encima del promedio en el tercer trimestre. Ya que año con año las

TABLA 18.21 Cálculos de los promedios móviles centrados de la serie de tiempo de las ventas de televisores

Año	Trimestre	Ventas (en miles)	Promedio móvil de cuatro trimestres	Promedio móvil centrado
1	1	4.8		
1	2	4.1		
			5.350	
1	3	6.0		5.475
			5.600	
1	4	6.5		5.738
			5.875	
2	1	5.8		5.975
			6.075	
2	2	5.2		6.188
			6.300	
2	3	6.8		6.325
			6.350	
2	4	7.4		6.400
			6.450	
3	1	6.0		6.538
			6.625	
3	2	5.6		6.675
			6.725	
3	3	7.5		6.763
			6.800	
3	4	7.8		6.838
			6.875	
4	1	6.3		6.938
			7.000	
4	2	5.9		7.075
			7.150	
4	3	8.0		
4	4	8.4		

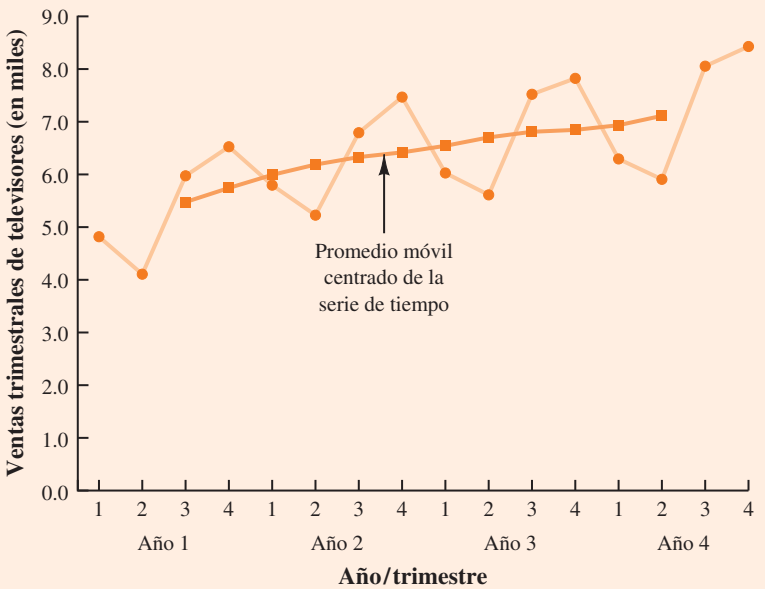
fluctuaciones en los valores estacional-irregulares se deben principalmente al error aleatorio, se pueden promediar los valores calculados para eliminar la influencia irregular y obtener una estimación de la influencia estacional del tercer trimestre.

$$\text{Efecto estacional del trimestre 3} = \frac{1.096 + 1.075 + 1.109}{3} = 1.09$$

Al número 1.09 se le conoce como *índice estacional* para el tercer trimestre. La tabla 18.23 resume los cálculos necesarios para obtener los índices estacionales de la serie de tiempo de las ventas de televisores. Los índices estacionales de los cuatro trimestres son 0.93, 0.84, 1.09 y 1.14.

La interpretación de los índices estacionales en la tabla ofrece una idea sobre el componente estacional de las ventas de televisores. El mejor trimestre de ventas es el cuarto, con ventas promedio de 14% por encima de la tendencia estimada. El peor, o más bajo, es el segundo trimestre; su índice estacional de 0.84 indica que el promedio de ventas está 16% por debajo de la tendencia estimada. El componente estacional se corresponde claramente con la expectativa intuitiva de que el interés por ver televisión y, por tanto, los patrones de compra de televisores

FIGURA 18.20 Serie de tiempo de las ventas trimestrales de televisores y su promedio móvil centrado



tienden a alcanzar el punto máximo en el cuarto trimestre debido a la próxima temporada de invierno y a la reducción de las actividades al aire libre. Las bajas ventas del segundo trimestre reflejan un menor interés por ver televisión debido a las actividades en primavera y antes del verano de los clientes potenciales.

TABLA 18.22 Valores del componente estacional-irregular de la serie de tiempo de las ventas de televisores

Año	Trimestre	Ventas (en miles)	Promedio móvil centrado	Valor estacional-irregular
1	1	4.8		
1	2	4.1		
1	3	6.0	5.475	1.096
1	4	6.5	5.738	1.133
2	1	5.8	5.975	0.971
2	2	5.2	6.188	0.840
2	3	6.8	6.325	1.075
2	4	7.4	6.400	1.156
3	1	6.0	6.538	0.918
3	2	5.6	6.675	0.839
3	3	7.5	6.763	1.109
3	4	7.8	6.838	1.141
4	1	6.3	6.938	0.908
4	2	5.9	7.075	0.834
4	3	8.0		
4	4	8.4		

TABLA 18.23 Cálculo de los índices estacionales de la serie de tiempo de ventas de televisores

Trimestre	Valores estacional-irregular			Índice estacional
1	0.971	0.918	0.908	0.93
2	0.840	0.839	0.834	0.84
3	1.096	1.075	1.109	1.09
4	1.133	1.156	1.141	1.14

Algunas veces es necesario un último ajuste para obtener los índices estacionales. Debido a que el modelo multiplicativo requiere que el índice estacional promedio sea igual a 1.00, la suma de los cuatro índices de la tabla 18.23 debe ser igual a 4.00. En otras palabras, los efectos estacionales incluso deben nivelarse a lo largo del año. En el ejemplo visto aquí, el promedio de los índices estacionales es igual a 1.00, y por tanto no es necesario ningún tipo de ajuste. En otros casos puede requerirse un ligero ajuste. Para realizarlo, se multiplica cada índice estacional por el número de estaciones, dividido entre la suma de los índices estacionales sin ajustar. Por ejemplo, cuando se tienen datos trimestrales se multiplica cada índice estacional por 4/(suma de los índices estacionales no ajustados). En algunos ejercicios se requerirá hacer este ajuste para obtener el índice estacional adecuado.

Desestacionalización de una serie de tiempo

Una serie de tiempo a la que se le han eliminado los efectos estacionales se conoce como **serie de tiempo desestacionalizada**, y al proceso de uso de los índices estacionales para eliminar los efectos estacionales de una serie de tiempo se le conoce como *desestacionalizar la serie de tiempo*. Al utilizar un modelo de descomposición multiplicativa se desestacionaliza una serie de tiempo dividiendo cada observación entre el índice estacional correspondiente. El modelo de descomposición multiplicativa es

Las series de tiempo económicas ajustadas por variaciones estacionales son generalmente reportadas en publicaciones como el Survey of Current Business, The Wall Street Journal *y* BusinessWeek.

$$Y_t = \text{Tendencia}_t \times \text{Estacional}_t \times \text{Irregular}_t$$

Así, cuando se divide cada una de las observaciones de la serie de tiempo (Y_t) entre su índice estacional correspondiente, los datos resultantes muestran únicamente la tendencia y la variabilidad aleatoria (el componente irregular). La serie de tiempo desestacionalizada para las ventas de televisores se resume en la tabla 18.24. Una gráfica de la serie de tiempo desestacionalizada se muestra en la figura 18.21.

Uso de una serie de tiempo desestacionalizada para identificar tendencias

La gráfica de la serie de tiempo desestacionalizada para las ventas de televisores que se muestra en la figura 18.21 parece tener una tendencia lineal ascendente. Para identificar esta tendencia, se ajustará una ecuación de tendencia lineal para la serie de tiempo desestacionalizada utilizando el mismo método que se muestra en la sección 18.4. La única diferencia estriba en que se ajustará la línea de tendencia a los datos desestacionalizados en lugar de a los datos originales.

Recuerde que para una tendencia lineal la ecuación de regresión estimada puede expresarse como

$$T_t = b_0 + b_1 t$$

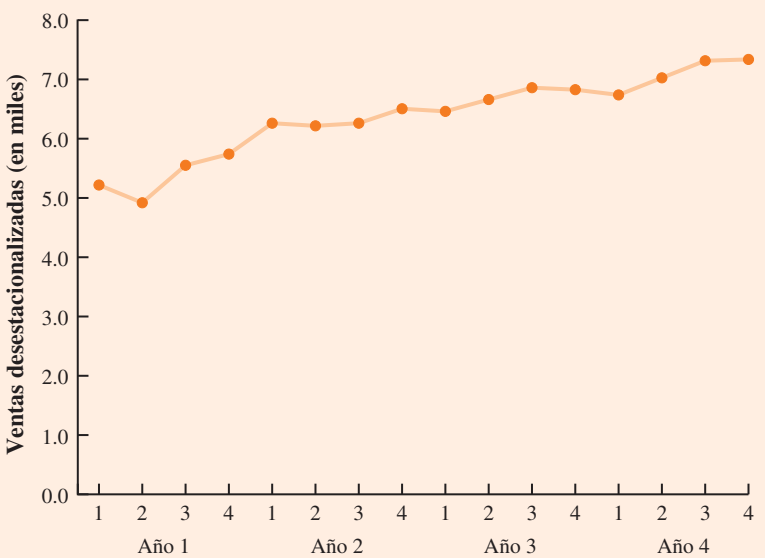
donde

T_t = pronóstico de tendencia lineal en el periodo t
b_0 = intersección de la recta de tendencia lineal
b_1 = pendiente de la línea de tendencia
t = periodo

TABLA 18.24 Valores desestacionalizados de la serie de tiempo de ventas de televisores

Año	Trimestre	Periodo	Ventas (en miles)	Índice estacional	Ventas desestacionalizadas
1	1	1	4.8	0.93	5.16
	2	2	4.1	0.84	4.88
	3	3	6.0	1.09	5.50
	4	4	6.5	1.14	5.70
2	1	5	5.8	0.93	6.24
	2	6	5.2	0.84	6.19
	3	7	6.8	1.09	6.24
	4	8	7.4	1.14	6.49
3	1	9	6.0	0.93	6.45
	2	10	5.6	0.84	6.67
	3	11	7.5	1.09	6.88
	4	12	7.8	1.14	6.84
4	1	13	6.3	0.93	6.77
	2	14	5.9	0.84	7.02
	3	15	8.0	1.09	7.34
	4	16	8.4	1.14	7.37

FIGURA 18.21 Valores desestacionalizados de la serie de tiempo de ventas de televisores



En la sección 18.4 se obtuvo la fórmula para calcular los valores de b_0 y b_1. Para ajustar una recta de tendencia lineal a los datos desestacionalizados de la tabla 18.24, el único cambio estriba en que al calcular b_0 y b_1 se utilizan los valores de la serie de tiempo desestacionalizada en lugar de los valores observados Y_t.

La figura 18.22 muestra los resultados de computadora obtenidos con el procedimiento de análisis de regresión de Minitab para estimar la línea de tendencia de la serie de tiempo desestacionalizada de los televisores. La ecuación de tendencia lineal estimada es

$$\text{Ventas desestacionalizadas} = 5.10 + 0.148\,t$$

FIGURA 18.22 Resultados de regresión de Minitab para la serie desestacionalizada de ventas de televisores

```
The regression equation is
Deseasonalized Sales = 5.10 + 0.148 Period

Predictor      Coef  SE Coef      T      P
Constant     5.1050   0.1133  45.07  0.000
Period      0.14760  0.01171  12.60  0.000

S = 0.215985   R-Sq = 91.9%   R-Sq(adj) = 91.3%

Analysis of Variance

Source           DF      SS      MS       F      P
Regression        1  7.4068  7.4068  158.78  0.000
Residual Error   14  0.6531  0.0466
Total            15  8.0599
```

La pendiente de 0.148 indica que en los últimos 16 trimestres la empresa promedió un crecimiento desestacionaliza de las ventas de 148 televisores por trimestre. Si se supone que los datos de ventas de los últimos 16 trimestres es un indicador bastante bueno del futuro, esta ecuación se puede utilizar para proyectar el patrón de tendencia de los próximos trimestres. Por ejemplo, si en esta ecuación se sustituye $t = 17$, obtenemos la proyección de la tendencia desestacionalizada para el siguiente trimestre, T_{17}.

$$T_{17} = 5.10 + 0.148\,(17) = 7.616$$

Por tanto, al utilizar los datos desestacionalizados, el pronóstico de tendencia lineal es 7 616 televisores para el próximo trimestre (periodo 17). Del mismo modo, los pronósticos de tendencia desestacionalizada para los próximos tres trimestres (periodos 18, 19 y 20) son 7 764, 7 912 y 8 060 televisores, respectivamente.

Ajustes estacionales

El último paso para obtener un pronóstico cuando existe tanto un patrón de tendencia como un patrón estacional, es usar el índice estacional a efecto de ajustar la proyección de tendencia desestacionalizada. Volviendo al ejemplo de las ventas de televisores, tenemos una proyección de la tendencia desestacionalizada para los próximos cuatro trimestres. Ahora es necesario ajustar el pronóstico para el efecto estacional. El índice estacional para el primer trimestre del año 5 ($t = 17$) es 0.93, por lo que se obtiene el pronóstico trimestral al multiplicar el pronóstico desestacionalizado basado en la tendencia ($T_{17} = 7\,616$) por el índice estacional (0.93). Por tanto, el pronóstico para el siguiente trimestre es $7\,616(0.93) = 7\,083$. En la tabla 18.25 se presentan los pronósticos para los trimestres 17 a 20. El cuarto trimestre, de alto volumen de ventas, tiene un pronóstico de 9 188 unidades, y el segundo trimestre, de volumen bajo de ventas, tiene como pronóstico 6 522 unidades.

TABLA 18.25 Pronósticos trimestrales para la serie de tiempo de ventas de televisores

Año	Trimestre	Pronóstico para la tendencia desestacionalizada	Índice estacional	Pronóstico trimestral
5	1	7 616	0.93	(7 616)(0.93) = 7 083
	2	7 764	0.84	(7 764)(0.84) = 6 522
	3	7 912	1.09	(7 912)(1.09) = 8 624
	4	8 060	1.14	(8 060)(1.14) = 9 188

Modelos basados en datos mensuales

En el ejemplo anterior de ventas de televisores se utilizaron datos trimestrales para ilustrar el cálculo de los índices estacionales. Sin embargo, muchas empresas prefieren los pronósticos mensuales. En tales casos, el procedimiento presentado en esta sección se aplica con modificaciones menores. Primero, en lugar de un promedio móvil de cuatro trimestres se usa uno de 12 meses; segundo, se calculan los índices estacionales de 12 meses en lugar de índices estacionales de cuatro trimestres. Aparte de estos cambios, los procedimientos de cálculo y pronóstico son idénticos.

Patrón cíclico

En términos matemáticos, el modelo multiplicativo de la ecuación (18.14) se puede ampliar para incluir el componente cíclico.

$$Y_t = \text{Tendencia}_t \times \text{Cíclico}_t \times \text{Estacional}_t \times \text{Irregular}_t \qquad \textbf{(18.15)}$$

El componente cíclico, al igual que el estacional, se expresa como un porcentaje de la tendencia. Como se mencionó en la sección 18.1, este componente se debe a los ciclos multianuales en la serie de tiempo. Es semejante al componente estacional, pero a lo largo de periodos más prolongados. Sin embargo, debido a la extensión de tiempo involucrado, con frecuencia es difícil obtener suficientes datos relevantes para estimar el componente cíclico. Otra dificultad radica en que estos ciclos suelen tener longitudes variables. Como es tan difícil identificar y/o separar los efectos cíclicos de los efectos de tendencia a largo plazo, en la práctica estos efectos a menudo se combinan y se les llama *componente combinado* de tendencia-ciclo. Se deja la discusión adicional del tema para libros más especializados sobre métodos de elaboración de pronósticos.

NOTAS Y COMENTARIOS

1. Existen varios métodos para calcular los índices estacionales. En esta sección se calculó cada índice estacional promediando los valores estacional-irregular correspondientes. Otro método, y el único utilizado por Minitab, es la mediana de los valores estacional-irregulares, como el índice estacional.
2. A menudo se realizan ajustes en el calendario antes de desestacionalizar una serie de tiempo. Por ejemplo, si una serie se compone de valores de las ventas mensuales, el valor de las ventas de febrero podrá ser menor que el de cualquier otro mes, simplemente porque hay menos días en febrero. Para tener en cuenta este factor, primero se divide el valor de las ventas de cada mes entre el número de días del mes para obtener un promedio diario. Dado que el número promedio de días en un mes es de aproximadamente 365/12 = 30.4167, entonces se multiplican los promedios diarios por 30.4167 para obtener valores ajustados mensualmente. Para los ejemplos y ejercicios de este capítulo se puede suponer que ya se ha realizado cualquier ajuste necesario al calendario.

Ejercicios

Métodos

AUTOevaluación

35. Considere los datos de la siguiente serie de tiempo.

Trimestre	Año 1	Año 2	Año 3
1	4	6	7
2	2	3	6
3	3	5	6
4	5	7	8

a) Construya una gráfica de series de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Proporcione los valores del promedio móvil de cuatro trimestres y los valores del promedio móvil centrado para esta serie de tiempo.
c) Calcule los índices estacionales y los índices estacionales ajustados para los cuatro trimestres.

36. Remítase al ejercicio 35.
a) Desestacionalice la serie de tiempo utilizando los índices estacionales ajustados calculados en el inciso *c*) del ejercicio 35.
b) Calcule la ecuación de regresión de tendencia lineal para los datos desestacionalizados utilizando Minitab o Excel.
c) Calcule el pronóstico de tendencia desestacionalizada para los trimestres del año 4.
d) Utilice los índices estacionales para ajustar los pronósticos de tendencia desestacionalizada calculados en el inciso *c*).

Aplicaciones

37. A continuación se presentan los datos de las ventas por trimestre (Quarter) del número de ejemplares vendidos para un libro de texto universitario en los últimos tres años (Year 1, 2 y 3).

WEB archivo
TextSales

Quarter	Year 1	Year 2	Year 3
1	1 690	1 800	1 850
2	940	900	1 100
3	2 625	2 900	2 930
4	2 500	2 360	2 615

a) Construya una gráfica de serie de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Para esta serie de tiempo, proporcione los valores del promedio móvil de cuatro trimestres y del promedio móvil centrado.
c) Calcule los índices estacionales y los índices estacionales ajustados de los cuatro trimestres.
d) ¿Cuándo obtiene la editorial el mayor índice estacional? ¿Parece razonable este resultado? Explique.
e) Desestacionalice la serie de tiempo.
f) Calcule la ecuación de tendencia lineal para los datos desestacionalizados y pronostique las ventas utilizando la ecuación de tendencia lineal.
g) Modifique los pronósticos de tendencia lineal utilizando los índices estacionales ajustados calculados en el inciso *c*).

38. A continuación se presentan los gastos ($) por mes (Month) del mantenimiento de césped a lo largo de tres años (Year 1, 2 y 3) para un edificio de seis departamentos en el sur de Florida.

WEB archivo
AptExp

Month	Year 1	Year 2	Year 3
January	170	180	195
February	180	205	210
March	205	215	230
April	230	245	280
May	240	265	290
June	315	330	390
July	360	400	420
August	290	335	330
September	240	260	290
October	240	270	295
November	230	255	280
December	195	220	250

a) Construya una gráfica de serie de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Identifique los índices estacionales mensuales para los tres años de gastos de mantenimiento del césped del edificio de apartamentos al sur de Florida. Utilice el cálculo del promedio móvil de 12 meses.
c) Desestacionalice la serie de tiempo.
d) Calcule la ecuación de tendencia lineal para los datos desestacionalizados.
e) Calcule los pronósticos de tendencia desestacionalizada y después ajuste los pronósticos de tendencia usando los índices estacionales para obtener un pronóstico de los gastos mensuales en el año 4.

39. En el sur de California, los especialistas en el control de la contaminación atmosférica monitorean cada hora la cantidad de ozono, dióxido de carbono y dióxido de nitrógeno en el aire. Los datos para esta serie de tiempo por hora presentan estacionalidad, por lo que los niveles de contaminación muestran ciertos patrones según la hora del día. Los siguientes niveles de dióxido de nitrógeno se observaron en el centro de la ciudad para 12 horas, de las 6:00 de la mañana a las 6:00 de la tarde, los días 15, 16 y 17 de julio.

WEB archivo
Pollution

Julio 15	25	28	35	50	60	60	40	35	30	25	25	20
Julio 16	28	30	35	48	60	65	50	40	35	25	20	20
Julio 17	35	42	45	70	72	75	60	45	40	25	25	25

a) Construya una gráfica de serie de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Identifique los índices estacionales por hora para las 12 lecturas de cada día.
c) Desestacionalice la serie de tiempo.
d) Utilice Minitab o Excel para calcular la ecuación de tendencia lineal de los datos desestacionalizados.
e) Calcule los pronósticos de tendencia desestacionalizada de las 12 horas del 18 de julio y después ajuste el pronóstico de tendencia con los índices estacionales obtenidos en *b*).

40. El consumo de energía eléctrica se mide en kilowatts-hora (kWh). La empresa local de servicios públicos ofrece un programa de ahorro en el que los clientes comerciales participantes pagan tarifas muy favorables con la condición de que reduzcan su consumo de energía cuando la entidad pública se los solicite. La empresa Timko Products redujo su consumo a partir del mediodía del jueves. Para evaluar el ahorro de energía, la empresa proveedora de energía tiene que estimar el consumo normal de energía de Timko. El periodo de reducción abarcó desde el mediodía hasta las 8:00 de la noche. Los datos sobre el consumo de energía eléctrica de esta empresa en las 72 horas anteriores son los siguientes, e incluyen los periodos (Time Period) del lunes (Monday), martes (Tuesday), miércoles (Wednesday) y jueves (Thursday).

WEB archivo
Power

Time Period	Monday	Tuesday	Wednesday	Thursday
12–4 a.m.	—	19 281	31 209	27 330
4–8 a.m.	—	33 195	37 014	32 715
8–12 noon	—	99 516	119 968	152 465
12–4 p.m.	124 299	123 666	156 033	
4–8 p.m.	113 545	111 717	128 889	
8–12 midnight	41 300	48 112	73 923	

a) ¿Se observa un efecto estacional dentro del periodo de 24 horas?
b) Calcule los índices estacionales para los seis periodos de 4 horas.
c) Utilice la tendencia ajustada por los índices estacionales para estimar el consumo normal de Timko en el periodo en que realizó el ahorro.

Resumen

En este capítulo se presentó una introducción a los métodos básicos del análisis de series de tiempo y pronóstico. Primero se indicó que el patrón subyacente en la serie de tiempo a menudo puede ser identificado construyendo una gráfica de serie de tiempo. Se distinguen varios tipos de patrón de datos, entre ellos un patrón horizontal, un patrón de tendencia y un patrón

estacional. Los métodos de elaboración de pronósticos estudiados se basan en que estos patrones están presentes en la serie de tiempo.

Se mostró cómo se utilizan los promedios móviles y el suavizamiento exponencial para desarrollar un pronóstico de una serie de tiempo con un patrón horizontal. El método de promedios móviles consiste en calcular el promedio de los valores de datos pasados, y después usar ese promedio como pronóstico para el siguiente periodo. En el método de suavizamiento exponencial se usa un promedio ponderado de los valores pasados de la serie de tiempo para calcular un pronóstico. Estos métodos también se adaptan bien cuando un patrón horizontal cambia a un nivel diferente y se reanuda un patrón horizontal.

Un factor importante para determinar qué método de elaboración de pronósticos utilizar involucra la exactitud del método. Se estudiaron tres medidas de exactitud de los pronósticos: el error absoluto medio (EAM), el cuadrado medio debido al error (CME), y el error porcentual absoluto medio (EPAM). Cada una de estas medidas está diseñada para determinar qué tan bien un método de elaboración de pronósticos particular es capaz de reproducir los datos disponibles de la serie de tiempo. Al seleccionar un método que tiene la mejor exactitud de los datos ya conocidos, se espera que aumente la probabilidad de obtener mejores pronósticos para periodos futuros.

Para una serie de tiempo que sólo tiene tendencia lineal a largo plazo, se demostró que puede utilizarse la regresión simple de la serie de tiempo para hacer proyecciones de su tendencia. También se estudió cómo una extensión del suavizamiento exponencial simple, conocido como *suavizamiento exponencial lineal de Holt*, se utiliza para pronosticar una serie de tiempo con tendencia lineal a largo plazo. Para una serie de tiempo con una tendencia curvilínea o no lineal, se demostró cómo la regresión múltiple permite ajustar los datos a una ecuación de tendencia cuadrática o a una ecuación de tendencia exponencial.

Para una serie de tiempo con un componente estacional, se demostró cómo utilizar las variables ficticias en un modelo de regresión múltiple a efecto de desarrollar una ecuación de regresión estimada con efectos estacionales. Luego se amplió el método de regresión para incluir situaciones en las que la serie de tiempo contiene tanto el efecto estacional como el efecto de tendencia lineal, y se mostró cómo combinar el método de la variable ficticia para el manejo de la estacionalidad con el método de regresión de la serie de tiempo para el manejo de la tendencia lineal.

En la última sección del capítulo se vio cómo la descomposición de la serie de tiempo se utiliza para separar o descomponer ésta en sus componentes estacional y de tendencia, para después desestacionalizarla. Se mostró cómo calcular los índices estacionales para un modelo multiplicativo, cómo utilizar los índices estacionales para desestacionalizar una serie de tiempo y cómo utilizar el análisis de regresión con los datos desestacionalizados para estimar el componente de tendencia. El último paso en el desarrollo de un pronóstico cuando existe tanto el componente de tendencia como el estacional es utilizar los índices estacionales para ajustar las proyecciones de tendencia.

Glosario

Constante de suavizamiento Parámetro del modelo de suavizamiento exponencial que proporciona el peso atribuido al valor más reciente de la serie de tiempo en el cálculo del valor pronosticado.

Cuadrado medio debido al error (CME) o error cuadrático medio Promedio de la suma de los errores de pronóstico cuadrados.

Descomposición de una serie de tiempo Método de series de tiempo que se utiliza para separar o descomponer una serie de tiempo en componentes estacional y de tendencia.

Error absoluto medio (EAM) Promedio de los valores absolutos de los errores de pronóstico.

Error de pronóstico Diferencia entre el valor real de la serie de tiempo y su pronóstico.

Error porcentual absoluto medio (EPAM) Promedio de los valores absolutos de los errores de pronóstico porcentuales.

Gráfica de serie de tiempo Presentación gráfica de las relaciones entre el tiempo y la variable de la serie de tiempo. El tiempo se muestra en el eje horizontal y los valores de una serie de tiempo en el eje vertical.

Modelo aditivo En este modelo, el valor real de una serie de tiempo en el periodo t se obtiene al sumar los valores de los componentes de tendencia, estacional e irregular.
Modelo multiplicativo En este modelo, el valor real de una serie de tiempo en el periodo t se obtiene al multiplicar los valores de los componentes de tendencia, estacional y el componente irregular.
Patrón cíclico Este patrón se presenta si la gráfica de una serie de tiempo muestra alternativamente una secuencia de puntos por debajo y por arriba de una línea de tendencia que tiene una duración de más de un año.
Patrón de tendencia Existe si la gráfica de la serie de tiempo presenta cambios o movimientos graduales hacia valores relativamente más altos o más bajos durante un largo periodo.
Patrón estacional Es aquel patrón que existe si la gráfica de la serie de tiempo presenta un patrón de repetición en periodos sucesivos. Éstos se presentan a menudo en intervalos de un año, que es de donde proviene el nombre de *patrón estacional*.
Patrón horizontal Se obtiene cuando los datos fluctúan alrededor de una media constante.
Promedios móviles Método de elaboración de pronósticos que utiliza el promedio de los valores de los k datos más recientes para pronosticar una serie de tiempo del periodo siguiente.
Promedios móviles ponderados Método de elaboración de pronósticos que consiste en seleccionar un peso diferente para los valores de los k datos más recientes de la serie de tiempo y luego calcular el promedio ponderado de los valores. La suma de los pesos debe ser 1.
Serie de tiempo Secuencia de observaciones sobre una variable medida en puntos sucesivos en el tiempo o en periodos sucesivos.
Serie de tiempo desestacionalizada Serie de tiempo de la cual ha sido eliminado el efecto estacional al dividir cada observación de la serie de tiempo original entre el índice estacional correspondiente.
Serie de tiempo estacionaria Serie de tiempo cuyas propiedades estadísticas son independientes del tiempo. Para una serie de tiempo estacionaria, el proceso de generación de datos tiene una media constante y la variabilidad de la serie de tiempo es constante en el tiempo.
Suavizamiento exponencial Método de elaboración de pronósticos que utiliza un promedio ponderado de los valores pasados de la serie de tiempo como un pronóstico; es un caso especial del método de promedios móviles ponderados en el que se selecciona un solo peso, el de la observación más reciente.
Suavizamiento exponencial lineal Extensión del suavizamiento exponencial simple que utiliza dos constantes de suavizamiento para que los pronósticos puedan obtener una serie de tiempo con una tendencia lineal.

Fórmulas clave

Pronóstico de promedio móvil de orden *k*

$$F_{t+1} = \frac{\sum(\text{los } k \text{ valores más recientes de los datos})}{k} = \frac{Y_t + Y_{t-1} + \ldots + Y_{t-k+1}}{k} \qquad (18.1)$$

Pronóstico de suavizamiento exponencial

$$F_{t+1} = \alpha Y_t + (1 - \alpha)F_t \qquad (18.2)$$

Ecuación de tendencia lineal

$$T_t = b_0 + b_1 t \qquad (18.4)$$

donde

$$b_1 = \frac{\sum_{t=1}^{n}(t - \bar{t})(Y_t - \bar{Y})}{\sum_{t=1}^{n}(t - \bar{t})^2} \qquad (18.5)$$

$$b_0 = \bar{Y} - b_1\bar{t} \qquad (18.6)$$

Suavizamiento exponencial lineal de Holt

$$L_t = \alpha Y_t + (1 - \alpha)(L_{t-1} + b_{t-1}) \quad (18.7)$$

$$b_t = \beta(L_t - L_{t-1}) + (1 - \beta)\, b_{t-1} \quad (18.8)$$

$$F_{t+k} = L_t + b_t\, k \quad (18.9)$$

Ecuación de tendencia cuadrática

$$T_t = b_0 + b_1 t + b_2 t^2 \quad (18.10)$$

Ecuación de tendencia exponencial

$$T_t = b_0 (b_1)^t \quad (18.11)$$

Modelo de descomposición aditiva

$$Y_t = \text{Tendencia}_t + \text{Estacional}_t + \text{Irregular}_t \quad (18.13)$$

Modelo de descomposición multiplicativa

$$Y_t = \text{Tendencia}_t \times \text{Estacional}_t \times \text{Irregular}_t \quad (18.14)$$

Ejercicios complementarios

41. La demanda semanal (en algunos casos) de una determinada marca de detergente para lavavajillas automática en una cadena de tiendas de abarrotes ubicada en Columbus, Ohio, es la siguiente.

Semana	Demanda	Semana	Demanda
1	22	6	24
2	18	7	20
3	23	8	19
4	21	9	18
5	17	10	21

a) Construya una gráfica de series de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Utilice un promedio móvil de tres semanas y obtenga un pronóstico para la semana 11.
c) Utilice el suavizamiento exponencial con una constante de suavizamiento de $\alpha = 0.2$ para desarrollar un pronóstico sobre la semana 11.
d) ¿Cuál de los dos métodos prefiere usted? ¿Por qué?

42. En la tabla siguiente se presentan los porcentajes invertidos en acciones de un portafolio a lo largo de nueve trimestres de 2007 a 2009.

Trimestre	Acciones %
1o.–2007	29.8
2o.–2007	31.0
3o.–2007	29.9
4o.–2007	30.1
1o.–2008	32.2
2o.–2008	31.5
3o.–2008	32.0
4o.–2008	31.9
1o.–2009	30.0

a) Construya una gráfica de serie de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Utilice el suavizamiento exponencial para pronosticar esta serie de tiempo. Considere las constantes de suavizamiento de $\alpha = 0.2$, 0.3 y 0.4. ¿Con cuál valor de la constante de suavizamiento se obtienen los pronósticos más exactos?
c) ¿Cuál es el pronóstico del porcentaje de acciones en un portafolio típico para el segundo trimestre de 2009?

43. United Dairies, Inc. es el proveedor de leche de varias compañías de abarrotes en el condado de Dade, Florida. Los directivos de la empresa desean contar con un pronóstico que proporcione la cantidad de litros de leche que se venden por semana. Los datos de ventas de las 12 semanas anteriores son los siguientes.

Semana	Ventas	Semana	Ventas
1	2 750	7	3 300
2	3 100	8	3 100
3	3 250	9	2 950
4	2 800	10	3 000
5	2 900	11	3 200
6	3 050	12	3 150

a) Construya una gráfica de serie de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Utilice el método de suavizamiento exponencial con $\alpha = 0.4$ para obtener un pronóstico de la demanda en la semana 13.

44. Para evitar un cargo mensual por servicio en una cuenta corriente que devenga intereses, el cliente debe mantener un saldo promedio diario mínimo. Se llevó a cabo un estudio en 2008 de 249 bancos y casas de ahorro de las 25 principales áreas metropolitanas de Estados Unidos, en el que se mostró que es necesario mantener un saldo medio de $3 462 para evitar un cargo mensual por servicio. Con un cargo promedio mensual de $11.97 y una tasa de interés promedio de sólo 0.24%, los clientes con cuenta de cheques que devengan intereses no están recibiendo mucho valor por ofrecer al banco una línea de crédito igual al saldo promedio mensual necesario para evitar el cargo mensual por servicio (página web de Bankrate, 27 de octubre de 2008). La siguiente tabla muestra el saldo promedio mínimo de 2001 a 2008 requerido para evitar un cargo mensual por servicio.

Año	Saldo ($)
2001	2 435
2002	2 593
2003	2 258
2004	2 087
2005	2 294
2006	2 660
2007	3 317
2008	3 462

a) Construya una gráfica de serie de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Utilice Minitab o Excel para obtener una ecuación de tendencia lineal de esta serie de tiempo. Calcule una estimación del saldo promedio requerido para evitar cargos mensuales por servicio para 2009.
c) Utilizando Minitab o Excel, obtenga una ecuación de tendencia cuadrática de esta serie de tiempo. Calcule un estimado del saldo promedio requerido para evitar cargos mensuales por servicio para 2009.
d) ¿Qué método ofrece pronósticos más precisos para los datos históricos con base en el CME?
e) ¿Recomendaría que con estos datos el pronóstico para 2009 se obtuviera a partir de una ecuación de tendencia lineal o de una ecuación de tendencia cuadrática? Explique.

45. El Garden Avenue Seven vende los discos compactos (CD) de sus interpretaciones musicales. La tabla siguiente presenta las ventas (Sales) en unidades por mes (Month) de los últimos 18 meses. El gerente del grupo desea un método preciso para pronosticar las ventas futuras.

WEB archivo
CDSales

Month	Sales	Month	Sales	Month	Sales
1	293	7	381	13	549
2	283	8	431	14	544
3	322	9	424	15	601
4	355	10	433	16	587
5	346	11	470	17	644
6	379	12	481	18	660

a) Construya una gráfica de serie de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Utilice el suavizamiento exponencial con $\alpha = 0.3$, 0.4 y 0.5. ¿Qué valor de α proporciona pronósticos más exactos?
c) Utilice la proyección de tendencia para ofrecer un pronóstico. ¿Cuál es el valor del CME?
d) ¿Qué método de elaboración de pronósticos le recomendaría al gerente? ¿Por qué?

46. Mayfair Department Store se encuentra en Davenport, Iowa, y desea determinar la pérdida de ventas que registró durante los meses de julio y agosto, en los que tuvo que cerrar debido a los daños causados por la inundación del río Mississippi. Los datos de ventas de enero a junio son los siguientes.

Mes	Ventas ($ miles)	Mes	Ventas ($ miles)
Enero	185.72	Abril	210.36
Febrero	167.84	Mayo	255.57
Marzo	205.11	Junio	261.19

a) Utilice el suavizamiento exponencial, con $\alpha = 0.4$, a efecto de obtener un pronóstico para julio y agosto (*Sugerencia*. Use el pronóstico de julio como ventas reales de julio al pronosticar agosto.) Exponga un comentario sobre el uso del suavizamiento exponencial para pronosticar más de un periodo en el futuro.
b) Utilice la proyección de tendencia para pronosticar las ventas de julio y agosto.
c) La compañía de seguros de Mayfair propuso una liquidación de $240 000 por la pérdida en julio y agosto. ¿Es esto suficiente? Si no lo es, ¿qué cantidad recomendaría como contraoferta?

47. Canton Supplies, Inc. es una empresa de servicios que emplea a unas 100 personas. Los gerentes de Canton Supplies se preocupan por cumplir con las obligaciones mensuales en efectivo y desean obtener un pronóstico de los requerimientos mensuales de efectivo. Debido a un cambio reciente en la política de operación, únicamente se consideraron relevantes los últimos siete meses.

Mes	1	2	3	4	5	6	7
Efectivo requerido ($ miles)	205	212	218	224	230	240	246

a) Construya una gráfica de serie de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Utilice el método de suavizamiento exponencial lineal de Holt con $\alpha = 0.6$ y $\beta = 0.4$ para pronosticar el efectivo requerido en cada uno de los próximos dos meses.
c) Utilice Minitab o Excel para obtener una ecuación de tendencia lineal que pronostique el efectivo requerido para cada uno de los próximos dos meses.
d) ¿Recomendaría el método de suavizamiento exponencial lineal de Holt con $\alpha = 0.6$ y $\beta = 0.4$, o la ecuación de tendencia lineal a efecto de pronosticar el efectivo requerido para cada uno de los próximos dos meses? Explique.

48. Costello Music Company ha estado en el negocio por cinco años. Durante ese tiempo las ventas aumentaron de 12 pianos en el primer año a 76 en el último año. Fred Costello, propietario de la empresa, desea obtener un pronóstico de ventas de pianos para el próximo año. Los siguientes son los datos históricos.

Año	1	2	3	4	5
Ventas	12	28	34	50	76

a) Construya una gráfica de series de tiempo. ¿Qué tipo de patrón existe en los datos?
b) Desarrolle una ecuación de tendencia lineal para la serie de tiempo. ¿Cuál es el crecimiento promedio anual en ventas que la empresa ha registrado por año?
c) Pronostique las ventas para los años 6 y 7.

49. Considere el problema de Costello Music Company del ejercicio 48. Los siguientes son los datos de las ventas por trimestre (Quarter 1, 2, 3 y 4) para 5 años (Year), incluyendo el total de ventas anuales (Total Yearly Sales).

WEB archivo
PianoSales

Year	Quarter 1	Quarter 2	Quarter 3	Quarter 4	Total Yearly Sales
1	4	2	1	5	12
2	6	4	4	14	28
3	10	3	5	16	34
4	12	9	7	22	50
5	18	10	13	35	76

a) Utilice las siguientes variables ficticias para obtener una ecuación de regresión estimada que considere los efectos estacionales y de tendencia lineal en los datos: Qtr1 = 1 si el trimestre es 1, 0 en caso contrario; Qtr2 = 1 si el trimestre es 2, 0 en caso contrario, y Qtr3 = 1 si el trimestre es 3, 0 en caso contrario.
b) Calcule los pronósticos trimestrales para el próximo año.

50. Consulte el problema de Costello Music Company del ejercicio 49.
a) Utilizando la descomposición de series de tiempo, calcule los índices estacionales para los cuatro trimestres.
b) ¿Cuándo experimenta Costello Music el mayor efecto estacional? ¿Parece razonable este resultado? Explique.

51. Remítase a la serie de tiempo de la empresa Costello Music del ejercicio 49.
a) Desestacionalice los datos y utilice la serie de tiempo desestacionalizada para identificar la tendencia.
b) Utilice los resultados del inciso *a*) a efecto de obtener un pronóstico trimestral para el próximo año con base en la tendencia.
c) Utilice los índices estacionales obtenidos en el ejercicio 50 para ajustar los pronósticos del inciso *b*) con objeto de tomar en cuenta el efecto estacional.

52. Durante los últimos siete años, Hudson Marine ha sido un distribuidor autorizado de radios náuticos C&D. La tabla siguiente presenta el número de radios que se venden por año.

Año	1	2	3	4	5	6	7
Cantidad vendida	35	50	75	90	105	110	130

a) Trace una gráfica de serie de tiempo. ¿Existe una tendencia lineal?
b) Utilice Minitab o Excel para obtener la ecuación de tendencia lineal de esta serie de tiempo.
c) Use la ecuación del inciso *b*) para obtener un pronóstico de ventas anuales en el año 8.

53. Consulte el problema de Hudson Marine del ejercicio 52. Suponga que los valores de ventas trimestrales para los siete años de datos históricos son los siguientes.

WEB archivo
HudsonMarine

Year	Quarter 1	Quarter 2	Quarter 3	Quarter 4	Total Yearly Sales
1	6	15	10	4	35
2	10	18	15	7	50
3	14	26	23	12	75
4	19	28	25	18	90
5	22	34	28	21	105
6	24	36	30	20	110
7	28	40	35	27	130

a) Utilice las siguientes variables ficticias para obtener una ecuación de regresión estimada en la que se tome en cuenta cualquier estación del año y los efectos de tendencia lineal en los datos: Qtr1 = 1 si es el trimestre 1, 0 en caso contrario; Qtr2 = 1 si es el trimestre 2, 0 en caso contrario; Qtr3 = 1 si es el trimestre 3, 0 en caso contrario.
b) Calcule los pronósticos trimestrales para el próximo año.

54. Consulte el problema de Hudson Marine del ejercicio 53.
a) Calcule los valores del promedio móvil centrado de esta serie de tiempo.
b) Trace una gráfica de la serie de tiempo que presente tanto el promedio móvil centrado como la serie de tiempo original. Exponga su opinión acerca de las diferencias entre la gráfica de la serie de tiempo original y la serie de tiempo con promedio móvil centrado.
c) Calcule los índices estacionales para los cuatro trimestres.
d) ¿Cuándo experimenta Hudson Marine el mayor efecto estacional? ¿Parece razonable este resultado? Explique.

55. Continúe con los datos de Hudson Marine del ejercicio 53.
a) Desestacionalice los datos y utilice la serie de tiempo desestacionalizada para identificar la tendencia.
b) Utilice los resultados del inciso *a*) y obtenga un pronóstico trimestral para el año siguiente con base en la tendencia.
c) Utilice los índices estacionales obtenidos en el ejercicio 54 para ajustar los pronósticos obtenidos en el inciso *b*) tomando en cuenta el efecto estacional.

Caso a resolver 1 Pronóstico de ventas de alimentos y bebidas

El Vintage Restaurant, en la isla Captiva, cerca de Fort Myers, Florida, es operado por su propietaria Karen Payne, y acaba de cumplir tres años de funcionamiento. Desde que inauguró su negocio, Karen ha tratado de establecer una reputación para Vintage como un restaurante de alta calidad que se especializa en mariscos frescos. Gracias a sus esfuerzos y los de su personal, el restaurante se ha convertido en uno de los mejores y con mayor crecimiento en la isla.

Para mejorar la planificación del crecimiento del restaurante en el futuro, Karen tiene que desarrollar un sistema que le permita pronosticar las ventas mensuales de alimentos y bebidas con hasta un año de anticipación. La tabla 18.26 muestra el valor de las ventas de alimentos y bebidas (en miles de $) para cada mes (Month) de los tres primeros años de funcionamiento: primer año (First Year), segundo año (Second Year) y tercer año (Third Year).

Informe gerencial

Elabore un análisis de los datos de las ventas de Vintage Restaurant. Prepare un informe para Karen que resuma sus hallazgos, pronósticos y recomendaciones. Incluya lo siguiente.

1. Una gráfica de serie de tiempo. Comente acerca del patrón principal en la serie de tiempo.
2. Un análisis de la estacionalidad de los datos. Indique el índice estacional para cada mes y comente sobre las ventas mensuales en las estaciones bajas y altas. ¿Los índices estacionales tienen sentido intuitivo? Comente.
3. Desestacionalice la serie de tiempo. ¿Existe alguna tendencia en la serie de tiempo desestacionalizada?
4. Utilizando el método de descomposición de una serie de tiempo, pronostique las ventas de enero a diciembre del cuarto año.
5. Utilizando el método de regresión con las variables ficticias, pronostique las ventas de enero a diciembre del cuarto año.
6. En el apéndice de su informe proporcione tablas con el resumen de sus cálculos y sus gráficas.

Suponga que en enero del cuarto año las ventas resultan ser de $295 000. ¿Cuál fue su error de pronóstico? Si el error es grande, a Karen puede confundirle esta diferencia entre el pronóstico y el valor de las ventas reales. ¿Qué puede hacer usted para resolver la incertidumbre del procedimiento de elaboración de pronósticos?

TABLA 18.26 Ventas de alimentos y bebidas del restaurante Vintage ($ miles)

Month	First Year	Second Year	Third Year
January	242	263	282
February	235	238	255
March	232	247	265
April	178	193	205
May	184	193	210
June	140	149	160
July	145	157	166
August	152	161	174
September	110	122	126
October	130	130	148
November	152	167	173
December	206	230	235

WEB archivo
Vintage

Caso a resolver 2 Elaboración del pronóstico de pérdidas de ventas

Carlson Department Store sufrió graves daños cuando un huracán azotó el 21 de agosto la zona donde se encuentra establecida. La tienda fue cerrada durante cuatro meses (de septiembre a diciembre), y Carlson está involucrada en una disputa con su compañía de seguros sobre el monto de las ventas perdidas durante el tiempo en que la tienda permaneció cerrada. Los dos temas clave que deben ser resueltos son: 1) el importe de las ventas que Carlson habría hecho si no hubiese ocurrido el huracán, y 2) si Carlson tiene derecho a alguna compensación por el exceso de ventas debido al aumento de actividad comercial generado en la zona después del huracán. El condado recibió más de $8 000 millones de ayuda federal por desastres y en pagos por seguros. El resultado fue el aumento de las ventas en las grandes tiendas departamentales y en muchos otros negocios.

En la tabla 18.27 se presentan los datos de las ventas de Carlson en los 48 meses anteriores al huracán y en la tabla 18.28 las ventas totales en el mismo periodo de todas las tiendas departamentales del condado, así como las ventas totales en el condado durante los cuatro meses en que Carlson Department Store permaneció cerrada. Los directivos de la tienda le pedirán que analice estos datos y que obtenga una estimación de la pérdida de ventas que sufrió de septiembre a diciembre. También se le pedirá que determine si se puede solicitar un pago por las ventas

TABLA 18.27 Ventas de Carlson Department Store ($ millones)

Month	Year 1	Year 2	Year 3	Year 4	Year 5
January		1.45	2.31	2.31	2.56
February		1.80	1.89	1.99	2.28
March		2.03	2.02	2.42	2.69
April		1.99	2.23	2.45	2.48
May		2.32	2.39	2.57	2.73
June		2.20	2.14	2.42	2.37
July		2.13	2.27	2.40	2.31
August		2.43	2.21	2.50	2.23
September	1.71	1.90	1.89	2.09	
October	1.90	2.13	2.29	2.54	
November	2.74	2.56	2.83	2.97	
December	4.20	4.16	4.04	4.35	

WEB archivo
CarlsonSales

TABLA 18.28 Ventas de las tiendas departamentales en el condado ($ millones)

WEB archivo
CountySales

Month	Year 1	Year 2	Year 3	Year 4	Year 5
January		46.80	46.80	43.80	48.00
February		48.00	48.60	45.60	51.60
March		60.00	59.40	57.60	57.60
April		57.60	58.20	53.40	58.20
May		61.80	60.60	56.40	60.00
June		58.20	55.20	52.80	57.00
July		56.40	51.00	54.00	57.60
August		63.00	58.80	60.60	61.80
September	55.80	57.60	49.80	47.40	69.00
October	56.40	53.40	54.60	54.60	75.00
November	71.40	71.40	65.40	67.80	85.20
December	117.60	114.00	102.00	100.20	121.80

extra relacionadas con el huracán. Si este caso se puede resolver, Carlson tiene derecho a una indemnización por el exceso de ventas que hubiera ganado por encima de las ventas normales.

Informe gerencial

Redacte un informe para los directivos de Carlson Department Store que resuma sus hallazgos, pronósticos y recomendaciones. Incluya lo siguiente:

1. Una estimación de las ventas que la tienda habría registrado de no haberse producido el huracán.
2. Una estimación de las ventas que habría tenido el condado de no haberse producido el huracán.
3. Una estimación de la pérdida de ventas de Carlson Department Store de septiembre a diciembre.

Además, utilice las ventas reales en las tiendas departamentales del condado desde septiembre hasta diciembre y la estimación del inciso 2) para argumentar a favor o en contra del exceso de ventas relacionadas con el huracán.

Apéndice 18.1 Elaboración de pronósticos con Minitab

En este apéndice se muestra cómo utilizar Minitab para obtener pronósticos mediante los siguientes métodos: promedios móviles, suavizamiento exponencial, proyección de tendencias, suavizamiento exponencial lineal de Holt y descomposición de una serie de tiempo.

Promedios móviles

WEB archivo
Gasoline

Para mostrar cómo utilizar Minitab en la elaboración de pronósticos con el método de promedios móviles, se calculará un pronóstico para la serie de tiempo de la venta de gasolina de la tabla 18.1 y de la figura 18.1. Los datos de las ventas en las 12 semanas se ingresan en la columna 2 de la hoja de cálculo. Los siguientes pasos se utilizan para obtener un pronóstico de promedio móvil de tres semanas para la semana 13.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Time Series**.
Paso 3. Elija **Moving Average**.
Paso 4. Cuando el cuadro de diálogo de Moving Average aparezca:
Introduzca C2 en el cuadro **Variable**.
Introduzca 3 en el cuadro **MA length**.

Seleccione **Generate forecasts**.
Introduzca 1 en el cuadro **Number of forecasts**.
Introduzca 12 en el cuadro **Starting from origin**.
Haga clic en **OK**.

Las medidas de exactitud de pronóstico para la semana 13 se muestran en la ventana de la sección. En los resultados de Minitab, el error absoluto medio se etiqueta MAD y el cuadrado medio debido al error o error cuadrático medio, MSD.

Suavizamiento exponencial

WEB archivo
Gasoline

Para mostrar cómo utilizar Minitab con objeto de obtener un pronóstico de suavizamiento exponencial, se recurrirá nuevamente a los datos presentados en la tabla 18.1 y en la figura 18.1 a efecto de obtener un pronóstico de las ventas para la semana 13 de la serie de tiempo de las ventas de gasolina. Los datos de las ventas para las 12 semanas se introducen en la columna 2 de la hoja de cálculo. Los siguientes pasos se utilizan para obtener un pronóstico sobre la semana 13 mediante una constante de suavizamiento de $\alpha = 0.2$.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Time Series**.
Paso 3. Elija **Single Exp Smoothing**.
Paso 4. Cuando el cuadro de diálogo de Single Exponential Smoothing aparezca:
Introduzca C2 en el cuadro **Variable**.
Seleccione la opción **Use** para Weight to Use in Smoothing.
Introduzca 0.2 en el cuadro Use.
Seleccione **Generate forecasts**.
Introduzca 1 en el cuadro **Number of forecasts**.
Introduzca 12 en el cuadro **Starting from origin**.
Seleccione **Options**.
Paso 5. Cuando el cuadro de diálogo de Single Exponential Smoothing-Options aparezca:
Introduzca 1 en el cuadro **Use average of first K observations**.
Haga clic en **Ok**.
Paso 6. Cuando el cuadro de diálogo de Single Exponential Smoothing aparezca:
Haga clic en **OK**.

Las medidas de la exactitud del pronóstico y el pronóstico de suavizamiento exponencial de la semana 13 se muestran en la ventana de la sección. En los resultados de Minitab,* el error absoluto medio se etiqueta MAD y el cuadrado medio debido al error o error cuadrático medio, MSD.

Proyección de tendencia

WEB archivo
Bicycle

Para mostrar cómo Minitab permite obtener pronósticos mediante la proyección de tendencias, se emplea un pronóstico para la serie de tiempo de las ventas de bicicletas de la tabla 18.3 y de la figura 18.3. El número de años se introduce en la columna 1 y los datos de las ventas en la columna 2 de la hoja de cálculo. Con los pasos siguientes se obtiene un pronóstico para el año 11 con la proyección de tendencia.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Time series**.
Paso 3. Elija **Trend Analysis**.

* El valor de MSD que proporciona Minitab no es el mismo que el valor del CME que aparece en la sección 18.3. Minitab utiliza 17 como pronóstico para la semana 1, así que para calcular el MSD utiliza los datos de las 12 semanas. En la sección 18.3 se calcula el CME utilizando sólo los datos para la semana 2 a 12 porque no se contaba con los valores del pasado que permitiera obtener un pronóstico para la semana 1.

Paso 4. Cuando el cuadro de diálogo Trend Analysis aparezca:
Introduzca C2 en el cuadro **Variable**.
Elija **Linear** como tipo de modelo.
Seleccione **Generate forecasts**.
Introduzca 1 en el cuadro **Number of forecasts**.
Introduzca 10 en el cuadro **Starting from origin**.
Haga clic en **Ok**.

La ecuación para la tendencia lineal, medidas de exactitud del pronóstico y los pronósticos para el año siguiente se muestran en la ventana de la sesión. En los resultados de Minitab, el error absoluto medio se etiqueta MAD y el error cuadrático medio MSD. Para obtener pronósticos de una tendencia cuadrática o una tendencia exponencial, seleccione **Quadratic** de **Exponential growth** en vez de **Linear** en el paso 4.

Suavizamiento exponencial lineal de Holt

WEB archivo
Bicycle

Para mostrar cómo utilizar Minitab para elaborar pronósticos con el método de suavizamiento exponencial lineal de Holt, se desarrolla nuevamente un pronóstico para la serie de tiempo de las ventas de bicicletas de la tabla 18.3 y la figura 18.3. En Minitab, el método de suavizamiento exponencial lineal de Holt se refiere como suavizamiento exponencial doble (Double Exponential Smoothing). El número de años se introduce en la columna 1 y los datos de ventas en la columna 2 de la hoja de cálculo. Los pasos siguientes pueden utilizarse para pronosticar las ventas del año 11 mediante el suavizamiento exponencial lineal de Holt con $\alpha = 0.1$ y $\beta = 0.2$.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Time Series**.
Paso 3. Elija **Double Exp Smoothing**.
Paso 4. Cuando el cuadro de diálogo Double Exponential Smoothing aparezca:
Introduzca C2 en el cuadro **Variable**.
Seleccione la opción **Use** para Weights to Use in Smoothing.
Introduzca 0.1 en el cuadro **level**.
Introduzca 0.2 en el cuadro **trend**.
Seleccione **Generate forecasts**.
Introduzca 1 en el cuadro **Number of forecasts**.
Introduzca 10 en el cuadro **Starting from origin**.
Haga clic en **Ok**.

Las medidas de exactitud de los pronósticos y los pronósticos de suavizamiento exponencial lineal de Holt para el año 11 se muestran en la ventana de la sesión. En los resultados de Minitab, el error absoluto medio se etiqueta MAD y el cuadrado medio debido al error o error cuadrático medio, MSD.

Descomposición de una serie de tiempo

WEB archivo
TVSales

El uso de Minitab para pronosticar una serie de tiempo con tendencia y estacionalidad mediante la descomposición de series de tiempo se muestra desarrollando un pronóstico para la serie de tiempo de las ventas de televisores de la tabla 18.6 y de la figura 18.6. En Minitab, el usuario tiene la opción de elegir un modelo de descomposición multiplicativa o aditiva. En la sección 18.6 se ilustra cómo utilizar el método multiplicativo. El número de años se introduce en la columna 1, los valores trimestrales en la columna 2 y los datos de las ventas en la columna 3 de la hoja de cálculo. Con los siguientes pasos se obtiene un pronóstico para el siguiente trimestre.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Time Series**.
Paso 3. Elija **Decomposition**.
Paso 4. Cuando el cuadro de diálogo Decomposition aparezca:
Introduzca C3 en el cuadro **Variable**.
Introduzca 4 en el cuadro **Season Length**.

Seleccione **Multiplicative** para Method Type.
Elija **Trend plus Seasonal** para Model Components.
Seleccione **Generate forecasts**.
Introduzca 1 en el cuadro **Number of forecasts**.
Introduzca 16 en el cuadro **Starting from origin**.
Haga clic en **OK**.

Los índices estacionales,[†] las medidas de exactitud de los pronósticos y los pronósticos para el próximo trimestre se muestran en la ventana de la sesión. En los resultados de Minitab, el error absoluto medio se etiqueta MAD y el cuadrado medio debido al error o error cuadrático medio, MSD.

Apéndice 18.2 Elaboración de pronósticos con Excel

En este apéndice se muestra cómo puede utilizarse Excel para elaborar pronósticos con los tres métodos de elaboración de pronósticos: promedios móviles, suavizamiento exponencial y proyección de tendencia.

Promedios móviles

WEB archivo
Gasoline

En la aplicación de Excel para obtener pronósticos mediante el método de promedios móviles se utiliza un pronóstico para la serie de tiempo de las ventas de gasolina de la tabla 18.1 y de la figura 18.1. Los datos de las ventas para las 12 semanas se ingresan en las filas de la 2 a la 13 de la columna B de la hoja de cálculo. Los siguientes pasos pueden utilizarse para elaborar un promedio móvil de tres semanas.

Paso 1. Haga clic en **Data** de la barra de herramientas.
Paso 2. En el grupo **Analysis**, hag clic en **Data Analysis**.
Paso 3. Elija **Moving Average** de la lista de Analysis Tools.
Haga clic en **Ok**.
Paso 4. Cuando el cuadro de diálogo Moving Average aparezca:
Introduzca B2:B13 en el cuadro **Input Range**.
Ingrese 3 en el cuadro **Interval**.
Introduzca C2 en el cuadro **Output Range**.
Haga clic en **OK**.

Los promedios móviles de tres semanas aparecerán en la columna C de la hoja de cálculo. El pronóstico para la semana 4 aparece al lado del valor de las ventas para la semana 3, y así sucesivamente. Los pronósticos para el periodo de otra longitud se calculan fácilmente introduciendo un valor diferente en el cuadro **Interval**.

Suavizamiento exponencial

WEB archivo
Gasoline

Para el uso de Excel en el suavizamiento exponencial, nuevamente se desarrolla un pronóstico para la serie de tiempo de las ventas de gasolina de la tabla 18.1 y de la figura 18.1. Los datos de las ventas de las 12 semanas se introducen en las filas 2 a la 13 de la hoja de cálculo de la columna B. Los siguientes pasos se utilizan para elaborar pronósticos con una constante de suavizamiento de $\alpha = 0.2$.

Paso 1. Haga clic en **Data** de la barra de herramientas.
Paso 2. En el grupo **Analysis**, haga clic en **Data Analysis**.
Paso 3. Elija **Exponential Smoothing** de la lista de Analysis Tools.
Haga clic en **OK**.
Paso 4. Cuando el cuadro de diálogo Exponential Smoothing aparezca:
Introduzca B2:B13 en el cuadro **Input Range**.
Introduzca 0.8 en el cuadro **Damping factor**.

[†] Los resultados difieren ligeramente de los que se muestran en la tabla 18.12 debido a que Minitab calcula los índices estacionales con la mediana de los valores estacional-irregulares.

Introduzca C2 en el cuadro de **Output Range**.
Haga clic en **OK**.

Los pronósticos de suavizamiento exponencial aparecerán en la columna C de la hoja de cálculo. Observe que el valor ingresado en el cuadro de factor de suavización es $1 - \alpha$; los pronósticos en otras constantes de suavizamiento se pueden calcular fácilmente introduciendo un valor distinto a $1 - \alpha$ en el cuadro de factor de suavización.

Proyección de tendencia

WEB archivo
Bicycle

Para mostrar cómo se utiliza Excel para la proyección de tendencias, se utilizarán los pronósticos de la serie de tiempo de las ventas de bicicletas de la tabla 18.3 y de la figura 18.3. Los datos, con sus correspondientes etiquetas en la fila 1, se ingresan en las filas 1 a la 11 de la columna A y B de la hoja de cálculo. Los siguientes pasos pueden utilizarse para obtener un pronóstico para el año 11 con la proyección de tendencias.

Paso 1. Seleccione una celda vacía en la hoja de cálculo.
Paso 2. Seleccione la barra de herramientas **Formulas**.
Paso 3. En el grupo **Function Library**, haga clic en **Insert Function**.
Paso 4. Cuando el cuadro de diálogo Insert Function aparezca:
Elija **Statistical** en el cuadro Or select a category box.
Elija **Forecast** en el cuadro Select a function.
Haga clic en **OK**.
Paso 5. Cuando el cuadro de diálogo de Forecast Arguments aparezca:
Introduzca 11 en el cuadro **x**.
Introduzca B2:B11 en el cuadro **Known y's**.
Introduzca A2:A11 en el cuadro **Known x's**.
Haga clic en **OK**.

En este caso el pronóstico para el año 11 es 32.5 y aparecerá en la celda elegida en el paso 1.

Apéndice 18.3 Elaboración de pronósticos con StatTools

En este apéndice se muestra cómo StatTools se puede utilizar para obtener un pronóstico empleando tres métodos de elaboración de pronósticos: los promedios móviles, el suavizamiento exponencial y el suavizamiento exponencial lineal de Holt.

Promedios móviles

WEB archivo
Gasoline

Para mostrar cómo StatTools se puede utilizar para obtener pronósticos mediante el método de promedios móviles se empleará un pronóstico para la serie de tiempo de las ventas de gasolina de la tabla 18.1 y de la figura 18.1. Inicie usando el Data Set Manager para crear una base de datos de StatTools para estos datos utilizando el procedimiento descrito en el apéndice del capítulo 1. Con los pasos siguientes se obtendrá el pronóstico del promedio móvil de tres semanas para la semana 13.

Paso 1. Haga clic en la barra de herramientas **StatTools**.
Paso 2. En **Analysis Group**, haga clic en **Time Series and Forecasting**.
Paso 3. Elija la opción **Forecast**.
Paso 4. Cuando el cuadro de diálogo StatTools-Forecast aparezca:
En la sección de **Variables** seleccione **Sales**.
Elija **Forecast Settings**.
En la sección **Method**, seleccione **Moving Average**.
En la sección **Parameters**, introduzca 3 en el cuadro **Span**.
Seleccione **Time Scale**.

Seleccione **None** en la sección **Seasonal Period**.
Seleccione **Integer** en la sección **Label Style**.
Haga clic en **OK**.

El siguiente resultado se mostrará en una nueva hoja de cálculo: tres medidas de exactitud de los pronósticos, la gráfica de la serie de tiempo que muestra los datos originales, los pronósticos y el error de pronóstico, así como una tabla con los pronósticos y los errores de pronóstico. Observe que StatTools utiliza el término "Means Abs Error" para identificar el valor del EAM, "Root Mean Sq Err" para identificar la raíz cuadrada del valor del CME y "Mean Abs Per% Err" para el valor del EPAM.

Suavizamiento exponencial

WEB archivo
Gasoline

Para mostrar cómo se utiliza StatTools en la elaboración de un pronóstico de suavizamiento exponencial, se empleará nuevamente un pronóstico para las ventas de la semana 13 de la serie de tiempo de las ventas de gasolina mostrado en la tabla 18.1 y en la figura 18.1. Use el Data Set Manager para crear una base de datos de StatTools mediante el procedimiento descrito en el apéndice del capítulo 1. Para obtener un pronóstico con una constante de suavizamiento de $\alpha = 0.2$ se efectúan los siguientes pasos.

Paso 1. Haga clic en la barra de herramientas **StatTools**.
Paso 2. En **Analysis Group**, haga clic en **Time Series and Forecasting**.
Paso 3. Elija la opción **Forecast**.
Paso 4. Cuando el cuadro de diálogo de StatTools-Forecast aparezca:
Seleccione **Sales** en la sección **Variables**.
Elija **Forecast Settings**.
Seleccione **Exponential Smoothing (Simple)** en la sección **Method**.
Elimine la marca de verificación del cuadro **Optimize Parameters**.
Introduzca 0.2 en el cuadro **Level (a)** en la sección **Parameters**.
Seleccione la barra de Time Scale.
Elija **None** en la sección **Seasonal Period**.
Seleccione **Integer** en la sección **Label Style**.
Haga clic en **OK**.

El siguiente resultado aparecerá en una nueva hoja de cálculo: las tres medidas de exactitud de los pronósticos, las gráficas de series de tiempo de las ventas de bicicletas mostrando los datos originales, los pronósticos y el error de pronóstico, así como una tabla que muestre los pronósticos y los errores de pronóstico. Observe que StatTools utiliza el término "Mean Abs Err" para identificar el valor del EAM, "Root Mean Sq Err" para identificar la raíz cuadrada del valor de CME y "Mean Abs Per% Err" para el valor del EPAM.

Suavizamiento exponencial lineal de Holt

WEB archivo
Bicycle

Para mostrar cómo utilizar StatTools en una tendencia de proyección, se obtendrá un pronóstico de la serie de tiempo de las ventas de bicicletas de la tabla 18.3 y de la figura 18.3 mediante el uso del suavizamiento exponencial lineal de Holt. Comience por utilizar el Data Set Manager para crear una base de datos de StatTools mediante el procedimiento descrito en el apéndice del capítulo 1. Con los siguientes pasos se obtendrá un pronóstico que emplee las constantes de suavizamiento $\alpha = 0.1$ y $\beta = 0.2$.

Paso 1. Haga clic en la barra de herramientas de **StatTools**.
Paso 2. En el **Analysis Group**, haga clic en **Time Series and Forecasting**.
Paso 3. Elija la opción **Forecast**.
Paso 4. Cuando el cuadro de diálogo de StatTools-Forecast aparezca:
Seleccione **Sales** en la sección **Variables**.
Elija **Forecast Settings**.
En la sección **Method**, seleccione **Exponential Smoothing (Holt's)**.
Elimine la marca de verificación del cuadro **Optimize Parameters**.

En la sección **Parameters**, introduzca 0.1 en el cuadro **Level (a)**.
En la sección **Parameters**, introduzca 0.2 en el cuadro **Trend (b)**.
Seleccione la etiqueta **Time Scale**.
En la sección **Seasonal Period**, elija **None**.
En la sección **Label Style**, seleccione **Integer**.
Haga clic en **OK**.

El siguiente resultado se mostrará en una nueva hoja de cálculo: las tres medidas de exactitud de los pronósticos; las gráficas de series de tiempo que muestran los datos originales, los pronósticos y los errores de pronóstico, y una tabla con los pronósticos y los errores de pronóstico. Observe que StatTools utiliza el término "Mean Abs Err" para denotar el valor del EAM, "Root Mean Sq Err" para identificar la raíz cuadrada del CME y "Mean Abs Per% Err" para identificar el valor del EPAM. El resultado de StatTools difiere ligeramente de los resultados mostrados en la sección 18.4 debido a que este programa utiliza un método diferente para calcular la estimación de la pendiente en el periodo 1. Con bases de datos más grandes, la elección de los valores iniciales no es crítica.

CAPÍTULO 19

Métodos no paramétricos

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: WEST SHELL REALTORS

19.1 LA PRUEBA DE SIGNOS
- Prueba de hipótesis acerca de una mediana poblacional
- Prueba de hipótesis con muestras pareadas

19.2 PRUEBA DE RANGOS CON SIGNO DE WILCOXON

19.3 PRUEBA DE MANN-WHITNEY-WILCOXON

19.4 PRUEBA DE KRUSKAL-WALLIS

19.5 CORRELACIÓN DE RANGOS

ESTADÍSTICA *en* LA PRÁCTICA

WEST SHELL REALTORS*

CINCINNATI, OHIO

West Shell Realtors fue fundada en 1958 con una oficina y un equipo de ventas de tres personas. En 1964 la empresa comenzó un programa de expansión a largo plazo durante el cual casi cada año abrió nuevas oficinas. Con los años, West Shell creció hasta convertirse en uno de los mayores corredores de bienes raíces de Greater Cincinnati, y ahora cuenta con oficinas en el sudeste de Ohio, el sudeste de Indiana y el norte de Kentucky.

A compañías de bienes raíces como West Shell el análisis estadístico les sirve para monitorear el curso de las ventas. Cada mes se presenta un informe de cada una de las oficinas de West Shell, así como del total de la empresa. Para mantener informados a los gerentes de las distintas oficinas y a los altos directivos acerca del progreso y los problemas de la organización, son esenciales el resumen estadístico sobre la cantidad total de ventas, el número de unidades vendidas y el precio medio de venta por unidad.

Además de los resúmenes mensuales sobre el curso de las operaciones, la empresa utiliza consideraciones estadísticas para guiar sus planes y estrategias corporativas. West Shell ha puesto en marcha una estrategia de expansión planificada. Cada vez que, debido a este plan de expansión, se quiere abrir una nueva oficina de ventas, la empresa debe abordar la cuestión de su ubicación. Los precios de venta, las tasas de facturación total en dólares y los volúmenes de ventas pronosticados son los tipos de datos utilizados para evaluar y comparar lugares alternos.

En una ocasión West Shell identificó dos suburbios como primeros candidatos para una nueva oficina: Clifton y Roselawn. Al comparar las dos zonas se consideró una variedad de factores, incluidos los precios de las casas.



Para ser competitivo, WestShell utiliza el análisis estadístico en las ventas de casas. © Cortesía de Coldwell Banker West Shell.

Se tomaron muestras de 25 ventas en el área de Clifton y 18 ventas en el área de Roselawn, y fue elegida la suma de rangos como una prueba estadística adecuada para las diferencias en el patrón de precios de venta. En el nivel de significancia de 0.05, la prueba de Mann-Whitney-Wilcoxon no permitió rechazar la hipótesis nula de que las dos poblaciones de precios de venta eran idénticas. Por tanto, West Shell se enfocó en otros criterios diferentes de los precios de venta de las casas en el proceso de selección del lugar.

En este capítulo se mostrará cómo aplicar las pruebas estadísticas no paramétricas como la de Mann-Whitney-Wilcoxon. También se discutirá la interpretación adecuada de dichas pruebas.

* Los autores agradecen a Rodney Fightmaster, de West Shell Realtors, por proporcionar este artículo para *Estadística en la práctica.*

A los métodos estadísticos de inferencia presentados previamente en este libro se les conoce como **métodos paramétricos**. Éstos comienzan con un supuesto sobre la distribución de probabilidad de la población, que a menudo establece que la población tiene una distribución normal. Con base en este supuesto, los expertos en estadística son capaces de obtener la distribución de muestreo que permite hacer inferencias sobre uno o más parámetros de la población, como la media poblacional μ o la desviación estándar poblacional σ. Por ejemplo, en el capítulo 9 se presentó un método para hacer inferencias sobre la media poblacional basado en el supuesto de que la población tenía un distribución de probabilidad normal con parámetros desconocidos μ y σ. Al utilizar la desviación estándar muestral s para estimar la desviación estándar pobla-

cional σ, el estadístico de prueba para hacer una inferencia sobre la media poblacional ha demostrado tener una distribución t. Como resultado, la distribución t se utilizó para calcular los intervalos de confianza y realizar pruebas de hipótesis sobre una media poblacional normalmente distribuida.

En este capítulo se presentan **métodos no paramétricos** para realizar inferencias sobre la población sin necesidad de un supuesto sobre la forma específica de la distribución de probabilidad poblacional. Por esta razón, estos métodos no paramétricos se llaman **métodos de distribución libre**.

La mayoría de los métodos estadísticos conocidos como métodos paramétricos requieren datos cuantitativos, mientras que los no paramétricos permiten inferencias basadas en cualquier tipo de datos cualitativos o cuantitativos. Sin embargo, los cálculos en los métodos no paramétricos se realizan generalmente con datos categóricos. Así, siempre que los datos sean cuantitativos, se transformarán en categóricos con el fin de realizar la prueba no paramétrica. En la primera sección del capítulo se muestra cómo la distribución binomial utiliza dos tipos de datos para hacer una inferencia sobre la media poblacional. En las siguientes tres secciones se describe cómo los datos ordenados por rango se utilizan en pruebas no paramétricas acerca de dos o más poblaciones. En la sección final se manejan datos ordenados por rango para calcular la correlación de rangos de dos variables.

19.1 La prueba de signos

La **prueba de signos** es un método no paramétrico versátil para pruebas de hipótesis que utiliza la distribución binomial con $p = 0.50$ como distribución de muestreo. No requiere un supuesto acerca de la distribución de la población. En esta sección se presentan dos aplicaciones de la prueba de signos: la que involucra la prueba de hipótesis acerca de una mediana poblacional y la que involucra la prueba de muestras pareadas acerca de la diferencia entre dos poblaciones.

Prueba de hipótesis acerca de una mediana poblacional

En el capítulo 9 se describe cómo llevar a cabo pruebas de hipótesis sobre la mediana poblacional. En esta sección se estudia cómo la prueba de signos se utiliza para realizar una prueba de hipótesis acerca de la mediana poblacional. Si se considera una población donde no hay un valor en los datos que sea exactamente igual a la mediana, ésta es la medida de tendencia central que divide a la población de manera que 50% de los valores son mayores a la mediana y 50% de los valores son menores a la mediana. Cada vez que una distribución de la población está sesgada, a menudo se prefiere la mediana como la mejor medida de tendencia central para la población. La prueba de signos establece un procedimiento no paramétrico para probar la hipótesis sobre el valor de la mediana poblacional.

Con el fin de demostrar la prueba de signos, se consideran las ventas semanales de papas fritas Cape May de la cadena de tiendas Lawler Grocery Store. Los directivos de Lawler tomaron la decisión de manejar un producto nuevo de papas fritas basado en la estimación del fabricante de que las ventas promedio deberían ser de $450 semanales por tienda. Después de manejar el producto por tres meses, los directivos de la cadena solicitaron la siguiente prueba de hipótesis sobre la mediana poblacional de las ventas semanales.

$$H_0: \text{mediana} = 450$$
$$H_a: \text{mediana} \neq 450$$

Los datos de las ventas de una semana en 10 tiendas de Lawer seleccionadas al azar se presentan en la tabla 19.1.

TABLA 19.1 Ventas de una semana de papas fritas Cape May de 10 tiendas de Lawler Grocery

Número de tienda	Ventas semanales ($)	Número de tienda	Ventas semanales ($)
56	485	63	474
19	562	39	662
36	415	84	380
128	860	102	515
12	426	44	721

Las observaciones igual al valor hipotético se descartan y el análisis procede con las observaciones que tengan un signo + o un signo −.

Al realizar la prueba de signos se compara cada observación muestral con el valor hipotético de la mediana poblacional. Si la observación es mayor que el valor hipotético, se anota un signo "+". Si la observación es menor que el valor hipotético, se anota un signo "−". Si una observación es exactamente igual al valor de la hipótesis, se elimina de la muestra y se procede a analizar la muestra de menor tamaño utilizando sólo las observaciones en las que se ha anotado el signo positivo o el signo negativo. Esta conversión de los datos muestrales en un signo positivo o un signo negativo es el que le da al método no paramétrico su nombre: la *prueba de signos*.

Considere los datos muestrales de la tabla 19.1. La primera observación, 485, es mayor que la mediana hipotética de 450; se registra un signo positivo. La segunda observación, 562, es mayor que la mediana hipotética de 450; se anota un signo más. Al continuar con las 10 observaciones, la muestra de la tabla 19.2 presenta los signos correspondientes. Observe que hay 7 signos positivos y tres signos negativos.

La asignación de los signos ha hecho de la situación una aplicación de distribución binomial. El tamaño de la muestra $n = 10$ es el número de eventos. Existen dos posibles resultados por evento, un signo positivo o un signo negativo, y los eventos son independientes. Sea p la probabilidad de un signo positivo. Si la mediana de la población es 450, p sería igual a 0.50, ya que debería haber 50% de signos positivos y 50% de signos negativos en la población. Por tanto, en términos de probabilidad binomial p, las hipótesis de la prueba de signos sobre la mediana de la población

$$H_0: \text{mediana} = 450$$
$$H_a: \text{mediana} \neq 450$$

se convierten en las siguientes hipótesis sobre la probabilidad binomial p.

$$H_0: p = 0.50$$
$$H_a: p \neq 0.50$$

TABLA 19.2 Datos muestrales para la prueba de signos de las ventas de Lawler sobre la mediana de la población

Número de tienda	Ventas semanales ($)	Signo	Número de tienda	Ventas semanales ($)	Signo
56	485	+	63	474	+
19	562	+	39	662	+
36	415	−	84	380	−
128	860	+	102	515	+
12	426	−	44	721	+

TABLA 19.3
Probabilidades binomiales con $n = 10$ y $p = 0.50$

Número de signos positivos	Probabilidad
0	0.0010
1	0.0098
2	0.0439
3	0.1172
4	0.2051
5	0.2461
6	0.2051
7	0.1172
8	0.0439
9	0.0098
10	0.0010

En la tabla 5 del apéndice B se proporcionan probabilidades binomiales cuando el tamaño de la muestra es menor o igual que 20. Se puede usar Excel o Minitab para encontrar las probabilidades binomiales para cualquier tamaño de muestra.

Si H_0 no puede ser rechazada, no se concluirá que p es diferente de 0.50 y, por tanto, no es posible que la mediana de la población sea diferente de 450. Sin embargo, si H_0 es rechazada, se concluye que p no es igual a 0.50 y, por ende, la mediana de la población no es igual a 450.

Con $n = 10$ tiendas o eventos y $p = 0.50$, se utiliza la tabla 5 del apéndice B para obtener las probabilidades binomiales del número de signos positivos bajo el supuesto de que H_0 es verdadera. Estas probabilidades se presentan en la tabla 19.3 y en la figura 19.1, las cuales muestran una representación gráfica de esta distribución binomial.

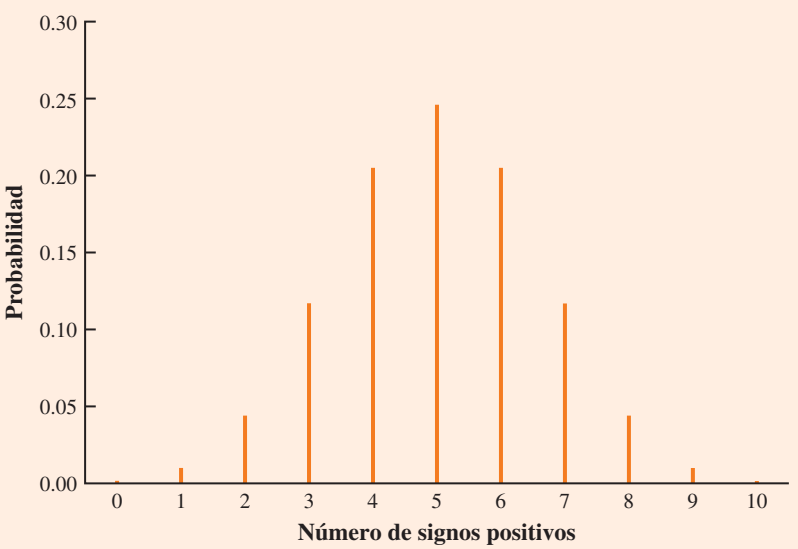
Procederemos a demostrar que la distribución binomial se utiliza para probar la hipótesis acerca de la mediana poblacional. Se utilizará 0.10 como nivel de significancia para la prueba. Puesto que el número observado de signos positivos para los datos muestrales, 7, está en la parte superior de la distribución binomial, comenzamos por calcular la probabilidad de obtener 7 o más signos positivos. Ésta es la probabilidad de obtener 7, 8, 9 o 10 signos positivos. La suma de estas probabilidades, que se muestra en la tabla 19.3, es $0.1172 + 0.0439 + 0.0098 + 0.0010 = 0.1719$. Puesto que se está utilizando una prueba de hipótesis de dos colas, esta probabilidad en la cola superior se duplica para obtener el valor-$p = 2(0.1719) = 0.3438$. Con un valor-$p > \alpha$, H_0 no puede ser rechazada. En términos de la probabilidad binomial p, H_0: $p = 0.50$ no es rechazada y, por tanto, tampoco la hipótesis de que la mediana de la población es \$450.

En este ejemplo, la prueba de hipótesis acerca de la mediana de la población se formuló como una prueba de dos colas. Sin embargo, las pruebas de signo de una cola acerca de la mediana de la población también son posibles. Por ejemplo, podríamos haber formulado la hipótesis como una prueba de cola superior, de modo que las hipótesis nula y alternativa se escribirían de la siguiente manera:

$$H_0: \text{mediana} \leq 450$$
$$H_a: \text{mediana} > 450$$

El valor-p correspondiente es igual a la probabilidad binomial de que el número de signos positivos encontrados en la muestra sea mayor o igual que 7. Este valor-p de una cola hubiera sido $0.1172 + 0.0439 + 0.0098 + 0.0010 = 0.1719$. Si la muestra se convierte en una prueba de cola inferior, el valor-p habría sido la probabilidad de obtener 7 o menos signos positivos.

FIGURA 19.1 Distribución de muestreo binomial para el número de signos positivos cuando $n = 10$ y $p = 0.50$



La aplicación que se acaba de describir utiliza la distribución binomial con $p = 0.50$. Con las probabilidades binomiales indicadas en la tabla 5 del apéndice B se calcula el valor-p cuando el tamaño de la muestra es de 20 o menos. Si el tamaño de la muestra es más grande, se utiliza la distribución normal como aproximación de la distribución binomial para calcular el valor-p, lo que hace que los cálculos sean más rápidos y fáciles. Una aplicación para una muestra grande de la prueba de signos se ilustra en el siguiente ejemplo.

Hace un año, el precio promedio de una casa nueva era de $236 000. Sin embargo, una caída reciente en la economía obligó a las compañías de bienes raíces a utilizar datos muestrales de las ventas recientes inmobiliarias para determinar si la mediana poblacional del precio de una casa nueva es menor hoy que hace un año. La prueba de hipótesis sobre la mediana poblacional del precio de una casa nueva es la siguiente.

$$H_0: \text{mediana} \geq 236\,000$$
$$H_a: \text{mediana} < 236\,000$$

Se utilizará 0.05 como nivel de significancia para llevar a cabo esta prueba.

WEB archivo
HomeSales

La muestra aleatoria de las ventas recientes de 61 casas reveló que 22 fueron vendidas en más de $236 000, 38 en menos de $236 000, y sólo una casa en $236 000. Después de eliminar la casa que se vendió en el precio mediano hipotético de $236 000, la prueba de signos prosigue con 22 signos positivos, 38 signos negativos y una muestra de 60 casas.

La hipótesis nula de que la mediana poblacional es mayor o igual que $236 000 se expresa mediante la hipótesis de distribución binomial $H_0: p \geq 0.50$. Si H_0 fuera verdadera como igualdad, se esperaría $0.50(60) = 30$ casas con signo positivos. El resultado muestral que presenta 22 signos positivos está en la cola inferior de la distribución binomial. Por tanto, el valor-p es la probabilidad de tener 22 o menos signos positivos cuando $p = 0.50$. Si bien es posible calcular las probabilidades binomiales exactas para 0, 1, 2, . . . al 22, así como la suma de estas probabilidades, para hacer que este cálculo sea más fácil se utilizará la aproximación a la distribución normal de la distribución binomial. Para esta aproximación, la media y la desviación estándar de la distribución normal son las siguientes.

APROXIMACIÓN NORMAL DE LA DISTRIBUCIÓN DE MUESTREO DEL NÚMERO DE SIGNOS POSITIVOS CUANDO
$H_0: p = 0.50$

$$\text{Media: } \mu = 0.50n \quad \textbf{(19.1)}$$
$$\text{Desviación estándar: } \sigma = \sqrt{0.25n} \quad \textbf{(19.2)}$$

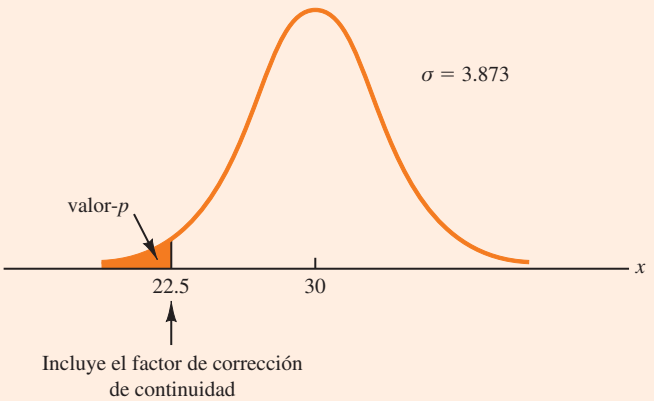
Forma de la distribución: aproximadamente normal para $n > 20$

Mediante las ecuaciones (19.1) y (19.2) con $n = 60$ casas y $p = 0.50$, la distribución de muestreo del número de signos positivos se puede aproximar a una distribución normal con

$$\mu = 0.50n = 0.50(60) = 30$$
$$\sigma = \sqrt{0.25n} = \sqrt{0.25(60)} = 3.873$$

Se utilizará ahora la distribución normal para aproximar la probabilidad binomial de 22 o menos signos positivos. Antes de continuar, recuerde que la distribución de probabilidad binomial es discreta y la de probabilidad normal es continua. Para tener esto en cuenta, la probabilidad binomial de 22 se calcula por el intervalo de probabilidad normal de 21.5 a 22.5. Al hecho de sumarle y restarle 0.5 a 22 se le llama *factor de corrección de continuidad*. Por tanto, para

FIGURA 19.2 Aproximación de la distribución normal del valor-p para la prueba de signos acerca de la mediana del precio de las casas nuevas



calcular el valor-p de 22 o menos signos positivos se utiliza la distribución normal con $\mu = 30$ y $\sigma = 3.873$ a efecto de determinar la probabilidad de que la variable aleatoria normal, x, tenga un valor menor o igual a 22.5. En la figura 19.2 se muestra una gráfica para este valor-p.

Utilizando esta distribución normal, calculamos el valor-p de la siguiente forma.

$$\text{Valor-}p = P(x \leq 22.5) = P\left(z \leq \frac{22.5 - 30}{3.873}\right) = P(z \leq -1.94)$$

Utilizando las áreas de la tabla para una distribución de probabilidad normal, se observa que con la probabilidad acumulada para $z = -1.94$ obtenemos el valor-$p = 0.0262$. Con $0.0262 < 0.05$, rechazamos la hipótesis nula y se concluye que la mediana del precio de una casa nueva es menor que la mediana del precio de $236 000 de hace un año.

NOTAS Y COMENTARIOS

1. Los ejemplos que ilustran una prueba de hipótesis acerca de una mediana poblacional involucran datos de ventas semanales y de precios de casas. Las distribuciones de probabilidad para este tipo de variables no suelen ser simétricas, y a menudo la mayoría está sesgada a la derecha. En dichos casos, la mediana de la población, más que la media poblacional, se convierte en una medida preferida de tendencia central. En general, cuando la población no es simétrica, la prueba de signos no paramétrica para la mediana poblacional es con frecuencia la prueba estadística más apropiada.
2. Para calcular una estimación del intervalo de confianza de la mediana poblacional se utiliza la distribución de muestreo binomial para la prueba de signos. Sin embargo, los cálculos son bastante complejos y rara vez se hacen a mano. Paquetes estadísticos como Minitab se pueden utilizar para obtener un intervalo de confianza de la mediana poblacional. Este procedimiento de Minitab se describe en el apéndice 19.1. Para el ejemplo de los precios de las casas manejado en esta sección, con Minitab se obtiene el intervalo de confianza de 95% para el precio mediano de una casa nueva, $183 000 a $231 000.

Prueba de hipótesis con muestras pareadas

En el capítulo 10 se describió un diseño experimental de muestras pareadas, donde cada una de las n unidades experimentales tenía un par de observaciones, una de la población 1 y otra de la población 2. Al utilizar los datos cuantitativos, y suponiendo que las diferencias entre el par de observaciones pareadas eran normalmente distribuidas, la distribución t se utilizó para hacer una inferencia acerca de las diferencias entre las medias de las dos poblaciones.

En el siguiente ejemplo se utilizará la prueba no paramétrica de signos para analizar datos muestrales pareados. A diferencia del procedimiento de distribución t, el cual requiere datos cuantitativos y el supuesto de que las diferencias se distribuyen normalmente, la prueba de signos permite analizar datos categóricos y cuantitativos, y no requiere ninguna hipótesis sobre la distribución de las diferencias. Este tipo de diseño de muestras pareadas se obtiene de la investigación de mercados, cuando a n clientes potenciales se les pide que comparen dos marcas de un producto como café, bebidas refrescantes o detergentes. Sin tener que obtener una medida cuantitativa de las preferencias de los individuos por cada marca, a cada persona se le pide que indique la marca de su preferencia. Considere el siguiente ejemplo.

La empresa Sun Coast Farms elabora un producto de jugo de naranja llamado *Citrus Valley*. La competencia principal de Citrus Valley proviene de la producción de un jugo de naranja conocido como *Tropical Orange*. En la comparación de la preferencia de un consumidor entre las dos marcas, 14 personas recibieron muestras sin identificación de los dos productos de jugo de naranja. La primera marca que cada individuo probó fue seleccionada al azar. Si eligió Citrus Valley como la preferida, se registraba un signo positivo. Si eligió Tropical Orange, se registraba un signo negativo. Si la persona no pudo expresar diferencia en la preferencia entre los dos productos, no se anotaba ningún signo. En la tabla 19.4 se presentan los datos de los 14 sujetos del estudio.

Después de eliminar a las dos personas que no expresaron una preferencia por la marca, los datos se convierten en una prueba de signos con 2 signos positivos y 10 signos negativos para $n = 12$ personas que expresaron su preferencia por una de las dos marcas. Siendo p la proporción de la población de los clientes que prefieren el jugo de naranja Citrus Valley, se desea probar la hipótesis de que no existe diferencia entre las preferencias para las dos marcas de la siguiente manera.

$$H_0: p = 0.50$$
$$H_a: p \neq 0.50$$

Si H_0 no puede ser rechazada, no podemos concluir que existe una diferencia en la preferencia por las dos marcas. Sin embargo, si H_0 es rechazada, concluimos que las preferencias del consumidor difieren para las dos marcas. Se utilizará un nivel de significancia de 0.05 para esta prueba de hipótesis.

Se efectuará la prueba de signos exactamente como se hizo antes en esta sección. La distribución de muestreo para el número de signos positivos es una distribución binomial con $p = 0.50$ y $n = 12$. En la tabla 5 del apéndice B obtenemos las probabilidades binomiales para el

TABLA 19.4 Datos de preferencias de la prueba de Sun Coast Farms

Individuo	Marca preferida	Signo	Individuo	Marca preferida	Signo
1	Tropical Orange	−	8	Tropical Orange	−
2	Tropical Orange	−	9	Tropical Orange	−
3	Citrus Valley	+	10	Sin preferencia	
4	Tropical Orange	−	11	Tropical Orange	−
5	Tropical Orange	−	12	Citrus Valley	+
6	Sin preferencia		13	Tropical Orange	−
7	Tropical Orange	−	14	Tropical Orange	−

número de signos positivos, como se muestra en la tabla 19.5. Bajo el supuesto de que H_0 es verdadera, se esperaría que $0.50n = 0.50(12) = 6$ signos positivos. Los resultados están en la cola inferior de la distribución binomial con sólo dos signos positivos en la muestra. Para calcular el valor-p de esta prueba de dos colas, primero se determina la probabilidad de 2 o menos signos positivos y luego se duplica este valor. Usando las probabilidades binomiales de 0, 1 y 2 mostradas en la tabla 19.5, el valor-p es $2(0.0002 + 0.0029 + 0.0161) + 0.0384$. Con $0.0384 < 0.05$, H_0 es rechazada. La prueba demuestra que las preferencias del consumidor difieren significativamente para las dos marcas de jugo de naranja. A Sun Coast Farms se le debe notificar este resultado y concluir que el producto competidor, Tropical Orange, es el más preferido. Sun Coast Farms puede entonces formular una estrategia para solucionar este problema.

TABLA 19.5
Probabilidades binomiales con $n = 12$ y $p = 0.50$

Número de signos positivos	Probabilidad
0	0.0002
1	0.0029
2	0.0161
3	0.0537
4	0.1208
5	0.1934
6	0.2256
7	0.1934
8	0.1208
9	0.0537
10	0.0161
11	0.0029
12	0.0002

Al igual que en otros usos de la prueba de signos, las pruebas de una cola se pueden utilizar dependiendo de la aplicación. Además, como el tamaño de la muestra llega a ser grande, la aproximación a la distribución normal de la distribución binomial facilitará los cálculos, como se mostró antes en esta sección. Mientras que la prueba de signos de Sun Coast Farms para muestras pareadas utiliza datos de preferencia categórica, también se puede aplicar la prueba de signos para muestras pareadas con datos cuantitativos. Esto sería particularmente útil si las diferencias idénticas no se distribuyen normalmente y son sesgadas. En este caso, a la diferencia positiva se le asigna un signo positivo y a la diferencia negativa un signo negativo, mientras que la diferencia de cero es descartada de la muestra. Los cálculos de la prueba de signos se realiza como ya se ha explicado.

Ejercicios

Métodos

AUTOevaluación

1. Se realizará la siguiente prueba de hipótesis.

$$H_0: \text{mediana} \geq 150$$
$$H_a: \text{mediana} < 150$$

Una muestra de 30 proporciona 22 observaciones mayores que 150, 3 observaciones iguales a 150 y 5 observaciones menores de 150. Utilice $\alpha = 0.01$. ¿Cuál es su conclusión?

AUTOevaluación

2. Diez personas participaron en una prueba de degustación en la que se consideraron dos marcas de un producto. Los resultados de la muestra revelaron que 7 prefirieron la marca A, 2 la marca B y 1 no pudo establecer su preferencia. Con $\alpha = 0.05$, pruebe si hay una diferencia significativa en las preferencias de las dos marcas. ¿Cuál es su conclusión?

Aplicaciones

3. La mediana del número de trabajadores de medio tiempo en restaurantes de comida rápida de una determinada ciudad fue 18 el año pasado. Las autoridades municipales piensan que la contratación de empleados de medio tiempo puede estar en aumento. En una muestra de nueve restaurantes de comida rápida se determinó que en siete se emplea a más de 18 trabajadores; en otro exactamente a 18, y en un restaurante a menos de 18 trabajadores de medio tiempo. ¿Se puede concluir que la mediana de número de trabajadores contratados de medio tiempo se ha incrementado? Realice la prueba con $\alpha = 0.05$.

4. Los activos netos de los 50 fondos de inversión en acciones más grandes muestran una mediana de $15 000 millones (*The Wall Street Journal*, 2 de marzo de 2009). La siguiente es una muestra de 10 de los 50 fondos de inversión de bonos más grandes.

Fondo de inversión en bonos	Activos netos	Fondo de inversión en bonos	Activos netos
Fidelity Intl Bond	6.1	T Rowe Price New Income	6.9
Franklin CA TF	11.7	Vanguard GNMA	15.0
American Funds	22.4	Oppenheimer Intl Bond	6.6
Vanguard Short Term	9.6	Dodge & Cox Income	14.5
PIMCO: Real Return	4.9	iShares: TIPS Bond	9.6

Utilizando la mediana, ¿se puede concluir que los fondos de inversión en bonos son menores y tienen menos activos netos que los fondos de inversión en acciones? Utilice $\alpha = 0.05$.

a) ¿Cuáles son las hipótesis para esta prueba?

b) ¿Cuál es el valor-*p*? ¿Qué concluye?

5. El ingreso mediano anual de suscriptores de la revista *Shutterbug* es de $75 000 (sitio web de Home Theater, 18 de agosto de 2008). Una muestra de 300 suscriptores de la revista *Popular Photography & Imaging* determinó que 165 de ellos tenían ingresos anuales de más de $75 000, mientras que 135 tenían ingresos menores de $75 000. ¿Se puede concluir que el ingreso mediano anual de los suscriptores de *Popular Photography & Imaging* difiere del ingreso mediano anual de los suscriptores de *Shutterbug*? Utilice $\alpha = 0.05$.

WEB archivo
ChicagoIncome

6. El ingreso mediano anual de las familias que viven en Estados Unidos es de $56 200 (*The New York Times Almanac,* 2008). Se presentan los ingresos anuales en miles de dólares para una muestra de 50 familias que viven en Chicago, Illinois. Utilice los datos de la muestra para ver si se puede concluir que las familias que viven en Chicago tienen un ingreso mediano anual de más de $56 200. Utilice $\alpha = 0.05$. ¿Cuál es su conclusión?

66.3	60.2	49.9	75.4	73.7
65.7	61.1	123.8	57.3	48.5
74.0	146.3	92.2	43.7	86.9
59.7	64.2	56.2	48.9	109.6
39.8	60.9	79.7	42.3	52.6
60.9	43.5	61.7	54.7	95.2
70.4	43.8	57.8	83.5	56.5
51.3	42.9	87.5	43.6	67.2
48.7	79.1	61.9	53.4	56.2
57.0	49.6	109.5	42.1	74.6

7. ¿Los splits o partición de acciones son benéficos para los accionistas? SNL Financial estudió durante un periodo de 18 meses el split de acciones en el sector bancario. En una muestra de 20 splits, 14 incrementaron el valor de la inversión, 4 lo disminuyeron y 2 no tuvieron cambios. Lleve a cabo una prueba de signos para determinar si se puede concluir que el split es benéfico para los titulares de acciones de los bancos.

a) ¿Cuáles son las hipótesis nula y alternativa?

b) ¿Cuál es su conclusión utilizando $\alpha = 0.05$?

8. Una encuesta realizada por el Pew Research Center preguntó a los adultos si el lugar ideal para vivir sería uno con un ritmo de vida más acelerado o uno con un ritmo de vida más lento (*USA Today*, 13 de febrero de 2009). Una muestra preliminar de 16 encuestados mostró que cuatro prefieren un ritmo de vida más rápido, 11 prefieren un lugar con un ritmo de vida más lento y uno dijo que no le importaba.

a) ¿Estos datos son suficientes para concluir que existe una diferencia entre las preferencias por un ritmo de vida más rápido o por un ritmo de vida más lento? Utilice $\alpha = 0.05$. ¿Cuál es su conclusión?

b) Tomando en cuenta la muestra completa de los 16 encuestados, ¿cuál es el porcentaje de las personas a las que les gustaría tener un ritmo de vida más rápido? ¿Cuál es el porcentaje de los que les gustaría un ritmo más lento? ¿Qué recomendación tiene para el estudio?

9. En una encuesta realizada durante la recesión de 2008 se formuló una serie de preguntas a 600 adultos sobre el estado de la economía y el futuro de sus hijos. Una pregunta fue: "¿Espera que sus hijos tengan una vida mejor de la que usted ha tenido, una vida peor, o una vida igual que la suya?". De los encuestados, 242 respondieron que una vida mejor, 310 una vida peor y 48 dijo que una vida igual a la suya. Utilice la prueba de signos y $\alpha = 0.05$ para determinar si existe una diferencia entre el número de adultos que siente que sus hijos tendrán una vida mejor en comparación con una vida peor. ¿Cuál es su conclusión?

10. Nielsen Media Research estableció que *American Idol* y *Dancing with the Stars* son los dos programas de televisión con el nivel de audiencia mayor en el horario estelar (*USA Today*, 14 de abril de 2008). En un estudio local sobre el programa de televisión preferido, se le pidió a 750 personas que indicaran cuál era su favorito en el horario de mayor audiencia: 330 eligieron *American Idol*, 270 *Dancing with the Stars* y 150 otro programa. Utilice 0.05 como nivel de significancia para probar la hipótesis de que no hay diferencia en la preferencia por *American Idol* y *Dancing with the Stars*. ¿Cuál es su conclusión?

11. La competencia en el mercado de las computadoras personales es intensa. Una muestra señaló que de 450 compras de computadoras, 202 fueron de la marca A, 175 de la marca B y 73 de otras marcas. Utilice 0.05 como nivel de significancia para probar si la hipótesis nula de la marca A y de la marca B tienen la misma participación en el mercado de las computadoras personales. ¿Cuál es su conclusión?

19.2 Prueba de rangos con signo de Wilcoxon

En el capítulo 10 se presentó un diseño experimental de muestras pareadas donde cada una de las n unidades experimentales proporcionó un par de observaciones: una de la población 1 y otra de la población 2. La prueba paramétrica para este experimento requiere datos cuantitativos y el supuesto de que las diferencias entre las observaciones pares se distribuyen normalmente. La distribución t se puede utilizar para hacer una inferencia entre las medias de las dos poblaciones.

Si la población de las diferencias es sesgada, se recomienda la prueba de signos de muestras pareadas presentada en la sección 19.1.

La **prueba de rangos con signo de Wilcoxon** es un procedimiento no paramétrico para el análisis de datos de un experimento de muestras pareadas. La prueba utiliza datos cuantitativos pero no necesita el supuesto de que las diferencias entre las observaciones pareadas se distribuyen normalmente. Sólo requiere el supuesto de que estas diferencias tienen una distribución simétrica. Esto ocurre cuando las formas de las dos poblaciones son iguales y se enfoca la atención en determinar si existe una diferencia entre sus medianas. La prueba de Wilcoxon se demostrará con el siguiente ejemplo.

Considere una empresa de manufactura que desea determinar si dos métodos de producción difieren en cuanto al tiempo para la realización de un trabajo. Mediante un diseño experimental de muestras pareadas, se eligió aleatoriamente a 11 trabajadores para terminar dos veces su tarea de producción una vez utilizando el método A y una vez con el método B. El método de producción que el trabajador utilizó por primera vez fue seleccionado aleatoriamente. El tiempo requerido para la realización con los dos métodos y las diferencias entre los tiempos para completar la tarea se muestran en la tabla 19.6 Una diferencia positiva indica que el método A requiere más tiempo; una diferencia negativa indica que el método B requiere más tiempo. ¿Los datos indican que los dos métodos de producción difieren significativamente en términos del

TABLA 19.6 Tiempo para la realización de una tarea de producción (minutos)

	Método		
Trabajador	A	B	Diferencia
1	10.2	9.5	0.7
2	9.6	9.8	−0.2
3	9.2	8.8	0.4
4	10.6	10.1	0.5
5	9.9	10.3	−0.4
6	10.2	9.3	0.9
7	10.6	10.5	0.1
8	10.0	10.0	0.0
9	11.2	10.6	0.6
10	10.7	10.2	0.5
11	10.6	9.8	0.8

tiempo que se requiere para realizar la tarea? Si se asume que las diferencias tienen una distribución simétrica, pero no necesariamente una distribución normal, se aplica la prueba de rangos con signo de Wilcoxon.

Los ejemplos en esta sección parten del punto de vista de que las dos poblaciones tienen la misma forma, y si difieren, es sólo en el lugar. Esto permite que las hipótesis para la prueba de rangos con signo de Wilcoxon se expresen en términos de las medianas de la población.

En particular, se utilizará la prueba de rangos con signo de Wilcoxon para la diferencia entre la mediana del tiempo de realización de la tarea para los dos métodos de producción. Las hipótesis son las siguientes.

$$H_0\text{: mediana para el método A} - \text{mediana para el método B} = 0$$
$$H_a\text{: mediana para el método A} - \text{mediana para el método B} \neq 0$$

Si H_0 no es rechazada, no podremos concluir que las medianas de los tiempos de realización de la tarea son diferentes. Sin embargo, si H_0 es rechazada, concluiremos que las medianas de los tiempos de realización de la tarea son diferentes. Se utilizará 0.05 como nivel de significancia para la prueba.

Las diferencias de 0 se descartan y el análisis continúa con el tamaño de muestra más pequeño que involucra las diferencias distintas de cero.

El primer paso en la prueba de rangos con signo de Wilcoxon es descartar la diferencia de cero para el trabajador 8 y luego calcular el valor absoluto de la diferencia para los 10 trabajadores restantes, como se muestra en la columna 3 de la tabla 19.7. A continuación se clasificarán esas diferencias absolutas de menor a mayor, como se observa en la columna 4. A la diferencia absoluta más pequeña, que es de 0.1 para el trabajador 7, se le asigna el rango 1. A la segunda menor diferencia absoluta de 0.2, obtenida para el trabajador 2, se le asigna el rango 2. Esta clasificación de las diferencias absolutas continúa con la diferencia absoluta mayor de 0.9 para el trabajador 6, al que se le asigna el rango 10. A las diferencias absolutas iguales de 0.4 para los trabajadores 3 y 5 se les asigna el rango promedio de 3.5. Del mismo modo, a las diferencias absolutas iguales de 0.5 para los trabajadores 4 y 10 se les asigna el rango promedio de 5.5.

La igualdad entre las diferencias absolutas se asigna al promedio de sus rangos.

Una vez que los rangos de las diferencias absolutas se han determinado, a cada rango se le especifica el *signo* de la diferencia original de los trabajadores. Los rangos de signo negativo se colocan en la columna 5 y los de signo positivo en la columna 6 (vea la tabla 19.7). Por ejemplo, la diferencia para el trabajador 1 fue positiva de 0.7 (vea la columna 2) y el rango de la diferencia absoluta fue 8 (vea la columna 4). Así, el rango del trabajador 1 se muestra con signo positivo en la columna 6. La diferencia para el trabajador 2 fue un negativo 0.2, que se muestra como un rango de signo negativo -2 en la columna 5. Al continuar con este proceso se generan los rangos de signos positivo y negativo como se muestran en la tabla 19.7.

TABLA 19.7 Rangos de las diferencias absolutas y rangos con signo para el tiempo necesario para realizar una tarea de producción

Trabajador	Diferencia	Diferencia absoluta	Rango	Rangos con signo	
				Negativo	Positivo
1	0.7	0.7	8		8
2	−0.2	0.2	2	−2	
3	0.4	0.4	3.5		3.5
4	0.5	0.5	5.5		5.5
5	−0.4	0.4	3.5	−3.5	
6	0.9	0.9	10		10
7	0.1	0.1	1		1
8	0.0				
9	0.6	0.6	7		7
10	0.5	0.5	5.5		5.5
11	0.8	0.8	9		9
				Suma de los rangos con signo	$T^+ = 49.5$

Sea T^+ la suma de los rangos con signo positivo, que es $T^+ = 49.5$. Para efectuar la prueba de rangos con signo de Wilcoxon se usará T^+ como el estadístico de prueba. Si las medianas de las dos poblaciones son iguales y el número de pares es 10 o más, la distribución de muestreo de T^+ se puede aproximar mediante una distribución normal de la siguiente manera.

DISTRIBUCIÓN DE MUESTREO DE T^+ PARA LA PRUEBA DE RANGOS CON SIGNO DE WILCOXON

$$\text{Media: } \mu_{T^+} = \frac{n(n+1)}{4} \tag{19.3}$$

$$\text{Desviación estándar: } \sigma_{T^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}} \tag{19.4}$$

Forma de distribución: aproximadamente normal para $n \geq 10$

Después de descartar la observación de una diferencia cero para el trabajador 8, el análisis continúa con los pares $n = 10$. Utilizando las ecuaciones (19.3) y (19.4), tenemos que

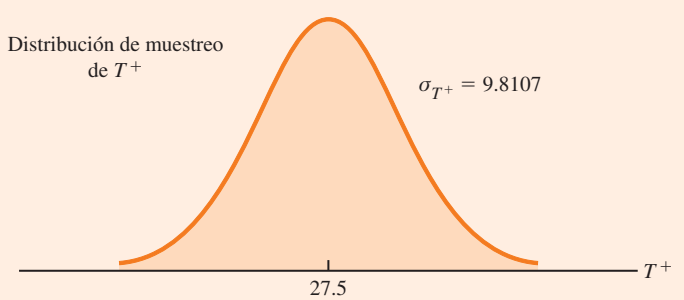
$$\mu_{T^+} = \frac{n(n+1)}{4} = \frac{10(10+1)}{4} = 27.5$$

$$\sigma_{T^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{10(10+1)(20+1)}{24}} = \sqrt{\frac{2310}{24}} = 9.8107$$

La figura 19.3 presenta la distribución de muestreo del estadístico de prueba T^+.

Se calcula el valor-p de dos colas para la hipótesis de que las medianas del tiempo para realizar la tarea con los dos métodos de producción son iguales. Puesto que el estadístico de prueba $T^+ = 49.5$ está en la cola superior de la distribución de muestreo, se empieza por calcular la probabilidad de la cola superior $P(T^+ \geq 49.5)$. Debido a que la suma de los rangos positivos T^+ es discreta y la distribución normal es continua, se obtiene la mejor aproximación incluyendo el factor de corrección de continuidad. Por tanto, la probabilidad discreta de $T^+ = 49.5$ se

FIGURA 19.3 Distribución de muestreo de T^+ para el ejemplo de tiempo de realización de una tarea de producción



aproxima por el intervalo de probabilidad normal, 49 a 50, y la probabilidad de que $T^+ \geq 49.5$ se aproxima por:

$$P(T^+ \geq 49.5) = P\left(z \geq \frac{49 - 27.5}{9.8107}\right) = P(z \geq 2.19)$$

Utilizando la tabla de distribución normal estándar y $z = 2.19$, se observa que el valor-p de dos colas $= 2(1 - 0.9857) = 0.0286$. Con el valor-$p \leq 0.05$, H_0 es rechazada, y se concluye que las medianas del tiempo de realización de la tarea para los dos métodos de producción no son iguales. Con T^+ en la cola superior de la distribución de muestreo, se observa que el método A toma mayor tiempo de realización. Es de esperar que la gerencia concluya que B es el método de producción más rápido o el más apropiado.

Las pruebas de rangos con signo de Wilcoxon de una cola son posibles. Por ejemplo, si al principio se hubiera buscado una evidencia estadística para concluir que el método A tomaba un tiempo mediano mayor para la realización de la tarea y que el método B tenía el tiempo mediano menor, se habría formulado la siguiente prueba de hipótesis de cola superior.

$$H_0\text{: mediana para el método A} - \text{mediana para el método B} \leq 0$$
$$H_a\text{: mediana para el método A} - \text{mediana para el método B} > 0$$

Al rechazar H_0 se llegaría a la conclusión de que el método A tiene el tiempo mediano mayor para completar la tarea y el método B tiene el tiempo mediano menor para completar la tarea. También es posible la prueba de hipótesis de cola inferior.

La prueba de rangos con signo de Wilcoxon puede utilizarse para probar la hipótesis acerca de la mediana de una población simétrica. Si la población es desigual, se prefiere la prueba de signos presentada en la sección 19.1.

Como nota final, en la sección 19.1 se mostró cómo la prueba de signos puede utilizarse tanto en una prueba de hipótesis para la mediana poblacional, como en una prueba de hipótesis con muestras pareadas. En esta sección se ha demostrado el uso de la prueba de rangos con signo de Wilcoxon para una prueba de hipótesis con muestras pareadas. Sin embargo, dicha prueba también puede utilizarse para una prueba no paramétrica de una mediana de la población, y no hace ningún supuesto acerca de que la distribución de la población tan solo es simétrica. Si este supuesto es apropiado, la prueba de rangos con signo de Wilcoxon es la prueba no paramétrica preferida para una mediana de la población. Sin embargo, si la población está sesgada, se prefiere la prueba de signos presentada en la sección 19.1. Con la prueba de rangos con signo de Wilcoxon se utilizan las diferencias entre las observaciones y el valor hipotético de la mediana poblacional en lugar de las diferencias entre las observaciones pareadas. Por lo demás, los cálculos son exactamente como se muestran en esta sección. En el ejercicio 17 se le pedirá que utilice esta prueba para llevar a cabo una prueba de hipótesis acerca de la mediana de una población simétrica.

NOTAS Y COMENTARIOS

1. La prueba de rangos con signo de Wilcoxon para una mediana poblacional se basa en el supuesto de que la población es simétrica. Bajo esta primera hipótesis, la mediana poblacional es igual a la media poblacional. Por tanto, la prueba de rangos con signo de Wilcoxon también puede utilizarse como una prueba acerca de la media de una población simétrica.
2. El procedimiento de rangos con signo de Wilcoxon también permite calcular un intervalo de confianza para la mediana de una población simétrica. Sin embargo, los cálculos son bastante complejos y rara vez se efectúan a mano. Se pueden utilizar paquetes estadísticos como el de Minitab para obtener un intervalo de confianza.

Ejercicios

Aplicaciones

En los siguientes ejercicios que involucran diferencias pareadas, se considera razonable asumir que las poblaciones que se comparan tienen aproximadamente la misma forma y que la distribución de las diferencias pareadas es aproximadamente simétrica.

AUTOevaluación

12. Se prueban dos aditivos de gasolina para determinar su efecto sobre el rendimiento de la gasolina en millas por galón en automóviles de pasajeros. Los siguientes son los resultados de la prueba para 12 automóviles (Car) en cada uno se han probado los dos aditivos (Additive). Utilice $\alpha = 0.05$ y la prueba de rangos con signo de Wilcoxon para determinar si hay una diferencia significativa entre las millas medianas por galón respecto de los aditivos.

WEB archivo
Additive

	Additive			Additive	
Car	1	2	Car	1	2
1	20.12	18.05	7	16.16	17.20
2	23.56	21.77	8	18.55	14.98
3	22.03	22.57	9	21.87	20.03
4	19.15	17.06	10	24.23	21.15
5	21.23	21.22	11	23.21	22.78
6	24.77	23.80	12	25.02	23.70

AUTOevaluación

13. En un estudio se utilizó una muestra de 10 hombres para probar los efectos de un relajante sobre el tiempo necesario para conciliar el sueño. Los datos siguientes corresponden al número de minutos que requiere cada individuo (Subject) para quedar dormido con o sin el relajante (Relaxant). Utilice 0.05 como nivel de significancia para determinar si el relajante reduce la mediana del tiempo necesario para quedar dormido. ¿Cuál es su conclusión?

WEB archivo
Relaxant

	Relaxant			Relaxant	
Subject	No	Yes	Subject	No	Yes
1	15	10	6	7	5
2	12	10	7	8	10
3	22	12	8	10	7
4	8	11	9	14	11
5	10	9	10	9	6

14. Los porcentajes de llegadas puntuales (Percent on Time) de vuelos en 2006 y 2007 fueron recabados aleatoriamente de 11 aeropuertos (Airport). Los datos se muestran en la parte superior de la página siguiente (página web de Research and Innovative Technology Administration, 29 de agosto de 2008). Utilice $\alpha = 0.05$ como nivel de significancia para probar la hipótesis de que no hay diferencia entre las medianas del porcentaje de llegadas a tiempo para los dos años. ¿Cuál es su conclusión?

WEB archivo
OnTime

Airport	Percent On Time 2006	2007
Boston Logan	71.78	69.69
Chicago O'Hare	68.23	65.88
Chicago Midway	77.98	78.40
Denver	78.71	75.78
Fort Lauderdale	77.59	73.45
Houston	77.67	78.68
Los Angeles	76.67	76.38
Miami	76.29	70.98
New York (JFK)	69.39	62.84
Orlando	79.91	76.49
Washington (Dulles)	75.55	72.42

15. Se realizó una prueba para dos servicios (Service) nocturnos de entrega de correo. Se crearon dos muestras idénticas para que a los dos servicios se les notificara la necesidad de reparto al mismo tiempo. Los tiempos necesarios para cada entrega (Delivery) se listan enseguida. ¿Los datos mostrados sugieren una diferencia en las medianas de los tiempos de entrega para los dos servicios? Utilice 0.05 como nivel de significancia para la prueba.

WEB archivo
Overnight

Delivery	Service 1	2
1	24.5	28.0
2	26.0	25.5
3	28.0	32.0
4	21.0	20.0
5	18.0	19.5
6	36.0	28.0
7	25.0	29.0
8	21.0	22.0
9	24.0	23.5
10	26.0	29.5
11	31.0	30.0

16. El PGA Players Championship se llevó a cabo en el Sedgefield Country Club en Greensboro, California del Norte, del 11 al 17 de agosto de 2008. Aquí se muestran los resultados de la primera (1st Round) y segunda (2nd Round) rondas de una muestra aleatoria de 11 golfistas (Golfer). Utilice $\alpha = 0.05$ para determinar si los promedios medianos de los golfistas para la primera y segunda rondas en el Players Championship difirieron de forma significativa. ¿Cuál es su conclusión?

WEB archivo
GolfScores

Golfer	1st Round	2nd Round
Marvin Laird	63	74
Jimmy Walker	70	73
Kevin Chappell	72	70

(*Continúa*)

Golfer	1st Round	2nd Round
Kevin Duke	65	71
Andrew Buckle	70	74
Paul Claxton	69	73
Larry Mize	72	71
Chris Riley	68	70
Bubba Watson	70	68
Carlos Franco	71	71
Richard Johnson	72	69

17. La prueba de aptitud académica SAT se compone de tres partes: lectura crítica, matemáticas y escritura. Cada parte de la prueba se califica en una escala de 200 hasta 800 puntos, con una mediana de aproximadamente 500 (*The World Almanac,* 2009). Se puede asumir que las puntuaciones para cada parte de la prueba son simétricas. Utilice los siguientes datos para probar la hipótesis de que la puntuación mediana de la población de estudiantes en la parte de escritura es 500. Utilice $\alpha = 0.05$. ¿Cuál es su conclusión?

WEB archivo
WritingScore

635	701	439	447	464
502	405	453	471	476
447	590	337	387	514

19.3 Prueba de Mann-Whitney-Wilcoxon

En el capítulo 10 se introdujo un procedimiento para conducir una prueba de hipótesis acerca de la diferencia entre las medias de dos poblaciones con dos muestras independientes, una de la población 1 y otra de la población 2. Esta prueba paramétrica requería datos cuantitativos y el supuesto de que ambas poblaciones tenían una distribución normal. En el caso de que las desviaciones estándar de la población σ_1 y σ_2 fueran desconocidas, con la desviación estándar muestral s_1 y s_2 se obtenían estimaciones para σ_1 y σ_2, y se utilizó la distribución t para hacer una inferencia acerca de la diferencia entre las medias de las dos poblaciones.

En esta sección se presenta una prueba no paramétrica de la diferencia entre dos poblaciones a partir de dos muestras independientes. Las ventajas de este procedimiento estriban en que se pueden utilizar ya sea los datos ordinales[1] o los datos cuantitativos, y no requiere el supuesto de que las poblaciones tengan una distribución normal. Versiones diferentes de esta prueba fueron desarrolladas conjuntamente por Mann y Whitney y también por Wilcoxon. Como resultado, a la prueba se le conoce como *prueba de Mann-Whitney* y *prueba de suma de rangos de Wilcoxon*. Las pruebas son equivalentes y ambas versiones ofrecen la misma conclusión. En esta sección nos referiremos a esta prueba no paramétrica como **prueba de Mann-Whitney-Wilcoxon (MWW)**.

[1] Los datos ordinales son datos categóricos que pueden tener un orden jerárquico. Esta escala de medición se estudia con más detalle en la sección 1.2 del capítulo 1.

La prueba de MWW inicia indicando la forma más general de las hipótesis nula y alternativa de la siguiente manera.

H_0: las dos poblaciones son idénticas
H_a: las dos poblaciones no son idénticas

La hipótesis alternativa de que las dos poblaciones no son idénticas requiere una aclaración. Si H_0 es rechazada, se está utilizando la prueba para concluir que las poblaciones no son idénticas y que la población 1 tiende a ofrecer un valor más pequeño o más grande que la población 2. La situación donde la población 1 tiende a ofrecer valores menores que la población 2 se muestra en la figura 19.4. Observe que no es necesario que todos los valores de la primera sean menores que todos los valores de la segunda. Sin embargo, la figura muestra correctamente la conclusión de que H_a es verdadera: las dos poblaciones no son idénticas y la 1 tiende a ofrecer valores más pequeños que la 2. En una prueba de dos colas se considera la hipótesis alternativa de que la población pueda ofrecer ya sea valores más pequeños o más grandes. Las versiones de la prueba de una cola pueden ser formuladas con la hipótesis alternativa de que la población 1 ofrece ya sea valores menores o mayores en comparación con la población 2.

En primer lugar se ilustra la prueba MWW utilizando muestras pequeñas con datos ordenados por rango. Esto le ayudará a comprender cómo se calcula el estadístico de suma de rangos y cómo se utiliza para determinar si la hipótesis nula de que las dos poblaciones son idénticas debe ser rechazada. Más adelante en la sección se introducirá una aproximación para muestras grandes basada en la distribución normal, que simplificará los cálculos requeridos para la prueba de MWW.

Consideremos las calificaciones del desempeño laboral de los empleados de 20 salas de cine y teatro de Showtime Cinemas. Durante una revisión, en un informe anual el gerente del teatro clasificó a 35 empleados en una escala de los mejores (calificación 1) a los peores (calificación 35). Sabiendo que los trabajadores de tiempo parcial eran principalmente estudiantes universitarios y de bachillerato, el gerente de distrito preguntó si había pruebas de una diferencia significativa en el desempeño de los universitarios comparado con los estudiantes de bachillerato. En términos de la población de ambos grupos de estudiantes que podría ser considerada para el empleo en el teatro, las hipótesis consignaron lo siguiente.

H_0: las poblaciones de estudiantes universitarios y de bachillerato son idénticas en términos de desempeño
H_a: las poblaciones de estudiantes universitarios y de bachillerato no son idénticas en términos de desempeño

Utilizaremos 0.05 como nivel de significancia para esta prueba.

Comenzaremos por seleccionar una muestra aleatoria de cuatro estudiantes universitarios y una muestra aleatoria de cinco estudiantes de bachillerato que trabajan en Showtime Cinemas. La calificación del gerente del teatro sobre el desempeño global de los 35 empleados se registró para cada uno de ellos, como se muestra en la tabla 19.8. El primer estudiante universitario seleccionado se clasificó en el lugar 15 en el informe anual de desempeño del gerente, el segundo estudiante universitario seleccionado se clasificó en tercer lugar, y así sucesivamente.

FIGURA 19.4 Las dos poblaciones no son idénticas. La población 1 tiende a ofrecer los valores más pequeños



TABLA 19.8 Calificación de desempeño de una muestra de estudiantes universitarios y una muestra de estudiantes de bachillerato que trabajan en Showtime Cinemas

Estudiante universitario	Calificación de desempeño del gerente	Estudiante de bachillerato	Calificación de desempeño del gerente
1	15	1	18
2	3	2	20
3	23	3	32
4	8	4	9
		5	25

Los datos de este ejemplo muestran cómo se puede utilizar la prueba de MWW con datos ordinales (ordenados por rango). El ejercicio 17 ofrece otra aplicación que utiliza este tipo de datos.

El siguiente paso del procedimiento MWW es clasificar la *combinación* de las muestras de menor a mayor. Puesto que hay un total de 9 sujetos, los datos de desempeño de la tabla 19.8 se clasifican de 1 a 9. El valor más bajo de 3 para el estudiante universitario 2 recibe un rango de 1, y el segundo valor más bajo, de 8, para el estudiante universitario 4, recibe un rango de 2. El valor más alto, 32, para el estudiante de bachillerato 3, recibe un rango de 9. La clasificación de rangos de las muestras combinadas para los 9 estudiantes se ilustra en la tabla 19.9.

A continuación se suman los rangos de cada muestra como se presentan en la tabla 19.9. El procedimiento de MWW puede utilizar la suma de los rangos de cualquiera de las muestras. Sin embargo, en esta aplicación de la prueba se continuará con la práctica común de utilizar la primera muestra, que es la de cuatro estudiantes universitarios. La suma de los rangos correspondiente será el estadístico de prueba W para la prueba de MWW. Esta suma, como se muestra en la tabla 19.9, $W = 4 + 1 + 7 + 2 = 14$.

Se explicará ahora por qué la suma de los rangos servirá para seleccionar entre las dos hipótesis H_0: las dos poblaciones son idénticas, y H_a: las dos poblaciones no son idénticas. Asumiendo que C denota a un estudiante universitario y H a un estudiante de bachillerato, suponga que los rangos de los nueve estudiantes tienen el siguiente orden, con los cuatro estudiantes universitarios obteniendo los cuatro rangos más bajos.

Rango	1	2	3	4	5	6	7	8	9
Estudiantes	C	C	C	C	H	H	H	H	H

Observe que esta permutación u orden separa las dos muestras con todos los estudiantes universitarios que tienen un rango inferior al de los estudiantes de bachillerato. Esta es una fuerte indicación de que las dos poblaciones no son idénticas. La suma de rangos para los estudiantes universitarios en este caso es $W = 1 + 2 + 3 + 4 = 10$.

TABLA 19.9 Rangos para las muestras combinadas de los nueve estudiantes de Showtime Cinemas

Estudiantes universitarios	Calificación de desempeño del gerente	Rango	Estudiantes de bachillerato	Calificación de desempeño del gerente	Rango
1	15	4	1	18	5
2	3	1	2	20	6
3	23	7	3	32	9
4	8	2	4	9	3
	Suma de rangos	14	5	25	8
				Suma de rangos	31

Ahora considere una clasificación donde los cuatro estudiantes universitarios tienen los cuatro rangos más altos.

Rango	1	2	3	4	5	6	7	8	9
Estudiantes	H	H	H	H	H	C	C	C	C

Observe que esta permutación u orden separa de nuevo las dos muestras, pero esta vez todos los estudiantes universitarios tienen un rango mayor que los de bachillerato. Esta es otra clara indicación de que las dos poblaciones no son idénticas. La suma de rangos para los estudiantes universitarios en este caso es $W = 6 + 7 + 8 + 9 = 30$. Así, se ve que la suma de los rangos de este grupo debe estar entre 10 y 30. Los valores de W cerca de 10 implican que los estudiantes universitarios tienen rangos más bajos que los estudiantes de bachillerato, mientras que los valores de W cerca de 30 implican que los primeros tienen rangos más altos que los segundos. Cualquiera de estos dos extremos sería una señal de que las dos poblaciones no son idénticas. Sin embargo, si lo fueran, se esperaría una combinación en el orden de C y H de modo que la suma de los rangos de W esté más cerca del promedio de los dos extremos, o sea, más cerca de $(10 + 30)/2 = 20$.

Asumiendo que las dos poblaciones son idénticas, se usa un programa de cómputo para calcular todos los órdenes posibles para los nueve estudiantes. Para cada orden se calcula la suma de los rangos de los estudiantes universitarios. Esto proporciona la distribución de probabilidad que muestra la distribución de muestreo exacta de W en la figura 19.5. Las probabilidades exactas relacionadas con el valor de W se resumen en la tabla 19.10. Si bien no se le pedirá que realice esta distribución de muestreo exacta, ésta se utilizará para probar la hipótesis de que las dos poblaciones de estudiantes son idénticas.

Usemos la distribución de muestreo de W de la figura 19.5 para calcular el valor-p para la prueba como se ha realizado en otras distribuciones de muestreo. La tabla 19.9 indica que la suma de rangos para los cuatro estudiantes universitarios es $W = 14$. Debido a que este valor de W está en la cola inferior de la distribución de muestreo, empezamos por calcular la probabilidad de la cola inferior $P(W \leq 14)$. Por tanto, tenemos

$$\begin{aligned} P(W \leq 14) &= P(10) + P(11) + P(12) + P(13) + P(14) \\ &= 0.0079 + 0.0079 + 0.0159 + 0.0238 + 0.0397 = 0.0952 \end{aligned}$$

FIGURA 19.5 Distribución de muestreo exacta de la suma de rangos para la muestra de estudiantes universitarios

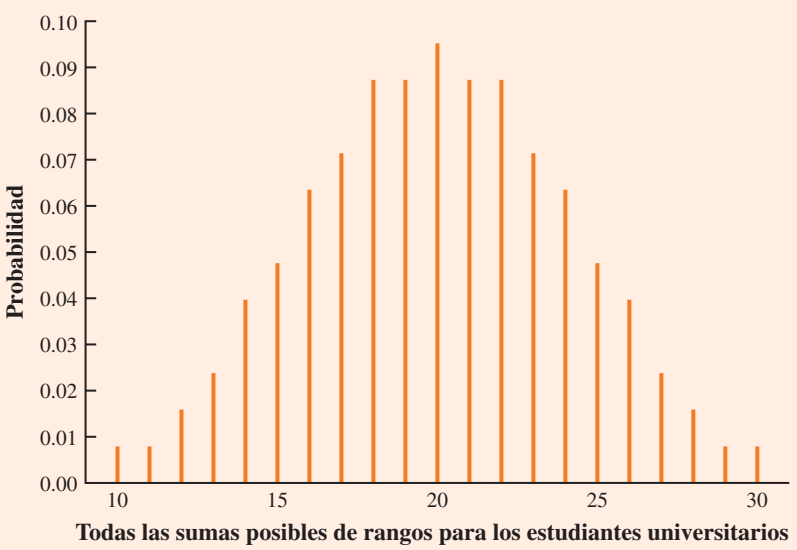


TABLA 19.10 Probabilidad de la distribución de muestreo exacta de la suma de rangos para la muestra de estudiantes universitarios

W	Probabilidad	*W*	Probabilidad
10	0.0079	20	0.0952
11	0.0079	21	0.0873
12	0.0159	22	0.0873
13	0.0238	23	0.0714
14	0.0397	24	0.0635
15	0.0476	25	0.0476
16	0.0635	26	0.0397
17	0.0714	27	0.0238
18	0.0873	28	0.0159
19	0.0873	29	0.0079
		30	0.0079

El valor-p de dos colas $= 2(0.0952) = 0.1904$. Con $\alpha = 0.05$ como nivel de significancia y con el valor-$p > 0.05$, la conclusión de la prueba de MWW es que no se rechaza la hipótesis nula de que las poblaciones de los estudiantes universitarios y los estudiantes de bachillerato son idénticas. Si bien las muestras de cuatro estudiantes universitarios y de los cinco de bachillerato no ofrecen evidencia estadística para concluir que hay una diferencia entre las dos poblaciones, este es un momento ideal para sugerir una opinión. Se debe considerar la conveniencia de estudiar muestras más grandes antes de formular una conclusión definitiva.

La mayoría de las aplicaciones de la prueba de MWW implica tamaños de muestras mayores al que se presentó en este primer ejemplo. Para tales aplicaciones se emplea una aproximación para muestras grandes de la distribución de muestreo de W basada en la distribución normal. De hecho, observe que el que la distribución de muestreo de W en la figura 19.5 indique una distribución normal, es una aproximación bastante buena para un tamaño de muestra tan pequeño como cuatro y cinco. Se utilizará el mismo procedimiento de rangos de las muestras combinadas que se manejó en el ejemplo anterior, pero con la aproximación de la distribución normal en lugar de la distribución de muestreo exacta de W para calcular el valor-p y llegar a una conclusión.

Se ilustrará el uso de la aproximación a la distribución normal para la prueba de MWW considerando la situación del Third National Bank. El gerente del banco supervisa los saldos en cuenta corriente de dos sucursales bancarias y se pregunta si las poblaciones de los saldos de cuenta en ambas sucursales son idénticas. Se toman dos muestras independientes de las cuentas corrientes con un tamaño muestral de $n_1 = 12$ en la sucursal 1 y $n_2 = 10$ en la sucursal 2. Los datos se muestran en la tabla 19.11.

La clasificación de las muestras combinadas toma algún tiempo. Las rutinas de computadora permiten hacer esta clasificación de manera rápida y eficiente.

Como antes, el primer paso en la prueba de MWW es clasificar los datos *combinados* del menor al mayor valor. Utilizando las 22 observaciones combinadas de la tabla 19.11, se determina que al valor más pequeño de \$750 (sucursal 2, cuenta 6) se le asigna un rango de 1. Al segundo valor más pequeño de \$800 (sucursal 2, cuenta 5), se le asigna un rango de 2. Al tercer valor más pequeño de \$805 (sucursal 1, cuenta 7), se le asigna un rango de 3, y así sucesivamente. En la clasificación de los datos combinados se puede determinar que dos o más valores son los mismos. En este caso, a los valores iguales se les asigna el rango *promedio* de sus posiciones en el conjunto de datos combinados. Por ejemplo, el saldo de \$950 se presenta tanto para la sucursal 1, cuenta 6, como para la sucursal 2, cuenta 4. En el conjunto de datos combinados, los dos valores de \$950 están en las posiciones 12 y 13 cuando los datos se ordenan de menor a mayor. Como resultado, a estas dos cuentas se les asigna el rango promedio $(12 + 13)/2 = 12.5$. La tabla 19.12 presenta los rangos asignados a las muestras combinadas.

TABLA 19.11 Saldos de cuenta para dos sucursales del Third National Bank

Sucursal 1		Sucursal 2	
Cuenta	**Saldo ($)**	**Cuenta**	**Saldo ($)**
1	1 095	1	885
2	955	2	850
3	1 200	3	915
4	1 195	4	950
5	925	5	800
6	950	6	750
7	805	7	865
8	945	8	1 000
9	875	9	1 050
10	1 055	10	935
11	1 025		
12	975		

TABLA 19.12 Rangos asignados para las muestras combinadas de saldos de cuentas

Sucursal	Cuenta	Saldo	Rango
2	6	750	1
2	5	800	2
1	7	805	3
2	2	850	4
2	7	865	5
1	9	875	6
2	1	885	7
2	3	915	8
1	5	925	9
2	10	935	10
1	8	945	11
1	6	950	12.5
2	4	950	12.5
1	2	955	14
1	12	975	15
2	8	1 000	16
1	11	1 025	17
2	9	1 050	18
1	10	1 055	19
1	1	1 095	20
1	4	1 195	21
1	3	1 200	22

Retroceda ahora a las dos muestras independientes y obtenga los rangos de la tabla 19.12 para cada saldo de cuenta. Estos resultados figuran en la tabla 19.13. El siguiente paso es sumar los rangos para cada muestra: 169.5 para la muestra 1 y 83.5 para la muestra 2. Como se

TABLA 19.13 Rango combinado de los datos en las dos muestras del Third National Bank

	Sucursal 1			Sucursal 2	
Cuenta	**Saldo ($)**	**Rango**	**Cuenta**	**Saldo ($)**	**Rango**
1	1 095	20	1	885	7
2	955	14	2	850	4
3	1 200	22	3	915	8
4	1 195	21	4	950	12.5
5	925	9	5	800	2
6	950	12.5	6	750	1
7	805	3	7	865	5
8	945	11	8	1 000	16
9	875	6	9	1 050	18
10	1 055	19	10	935	10
11	1 025	17		Suma de rangos	83.5
12	975	15			
	Suma de rangos	169.5			

mencionó antes, se continuará siempre con el procedimiento de utilizar la suma de los rangos de la muestra 1 como el estadístico de prueba de W. Por tanto, tenemos que $W = 169.5$. Cuando los dos tamaños de muestra son 7 o más, se puede utilizar una aproximación normal de la distribución de muestreo de W. Bajo el supuesto de que la hipótesis nula es verdadera y las poblaciones son idénticas, la distribución de muestreo del estadístico de prueba W es la siguiente.

DISTRIBUCIÓN DE MUESTREO DE W CON POBLACIONES IDÉNTICAS

$$\text{Media: } \mu_W = \tfrac{1}{2}n_1(n_1 + n_2 + 1) \qquad \textbf{(19.5)}$$

$$\text{Desviación estándar: } \sigma_W = \sqrt{\tfrac{1}{12}n_1n_2(n_1 + n_2 + 1)} \qquad \textbf{(19.6)}$$

Forma de distribución: aproximadamente normal siempre que $n_1 \geq 7$ y $n_2 \geq 7$

Dado el tamaño de las muestras $n_1 = 12$ y $n_2 = 10$, las ecuaciones (19.5) y (19.6) ofrecen las siguientes media y desviación estándar de la distribución de muestreo.

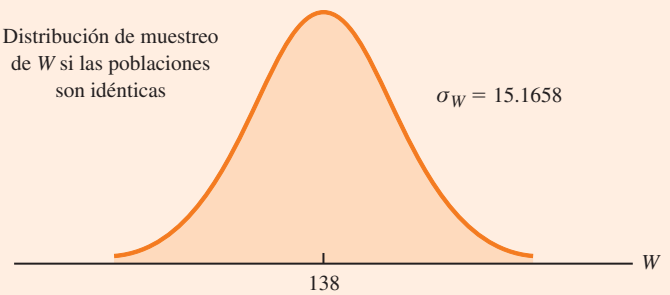
$$\mu_W = \tfrac{1}{2}n_1(n_1 + n_2 + 1) = \tfrac{1}{2}(12)(12 + 10 + 1) = 138$$

$$\sigma_W = \sqrt{\tfrac{1}{12}n_1n_2(n_1 + n_2 + 1)} = \sqrt{\tfrac{1}{12}(12)(10)(12 + 10 + 1)} = 15.1658$$

La figura 19.6 ilustra la distribución normal utilizada para la distribución de muestreo de W.

Se procederá con la prueba de MWW y se utilizará 0.05 como nivel de significancia para llegar a una conclusión. Dado que el estadístico de prueba W es discreto y la distribución normal es continua, se volverá a utilizar el factor de corrección de continuidad para la aproximación de

FIGURA 19.6 Distribución de muestreo de W para el ejemplo del Third National Bank



distribución normal. Con $W = 169.5$ en la cola superior de la distribución de muestreo, tenemos los siguientes cálculos del valor-p.

$$P(W \geq 169.5) = P\left(z \geq \frac{169 - 138}{15.1658}\right) = P(z \geq 2.04)$$

Utilizando la variable aleatoria normal estándar y $z = 2.04$, el valor-p de dos colas $= 2(1-0.9793) = 0.0414$. Con el valor-$p \leq 0.05$, H_0 es rechazada y se concluye que las dos poblaciones de saldos de cuentas no son idénticas. El valor de la cola superior del estadístico de prueba W indica que la población de saldos de cuentas en la sucursal 1 tiende a ser mayor.

Si se puede suponer que las dos poblaciones tienen la misma forma, la prueba MWW se convierte en una prueba sobre la diferencia entre las medianas de las dos poblaciones.

Como comentario final, conviene decir que algunas aplicaciones de la prueba de MWW hacen adecuado asumir que las dos poblaciones tienen formas idénticas, y si éstas son diferentes, es sólo por un cambio en la ubicación de las distribuciones. Si las dos poblaciones tienen la misma forma, la prueba de hipótesis puede expresarse en términos de la diferencia entre las dos medianas de las poblaciones. Cualquier diferencia entre las medianas se interpreta como un cambio de ubicación de una población comparada con la otra. En este caso, las tres formas de la prueba de MWW sobre la mediana de dos poblaciones son las siguientes.

Prueba de dos colas	Prueba de cola inferior	Prueba de cola superior
H_0: $\text{mediana}_1 - \text{mediana}_2 = 0$	H_0: $\text{mediana}_1 - \text{mediana}_2 \geq 0$	H_0: $\text{mediana}_1 - \text{mediana}_2 \leq 0$
H_a: $\text{mediana}_1 - \text{mediana}_2 \neq 0$	H_a: $\text{mediana}_1 - \text{mediana}_2 < 0$	H_a: $\text{mediana}_1 - \text{mediana}_2 > 0$

NOTAS Y COMENTARIOS

El procedimiento de Minitab para la prueba MWW se describe en el apéndice 19.1. Minitab parte del supuesto de que las dos poblaciones tienen la misma forma. En consecuencia, describe los resultados de la prueba en términos de una prueba acerca de las medianas de las dos poblaciones. Si no se siente cómodo al realizar el supuesto de la "misma forma", los resultados de Minitab siguen siendo aplicables. Sin embargo, es necesario que usted los interprete como una prueba de la hipótesis nula de que las dos poblaciones son idénticas.

Ejercicios

Aplicaciones

AUTOevaluación

18. Dos aditivos de combustible son evaluados para determinar su efecto en el millaje de la gasolina. Se aplicaron sendas pruebas a siete vehículos con el aditivo 1 y a nueve vehículos con el aditivo 2. Los datos siguientes muestran las millas por galón obtenidas con los dos aditivos. Utilice $\alpha = 0.05$ y la prueba de MWW para ver si existe una diferencia significativa entre el rendimiento de la gasolina con los dos aditivos.

Aditivo 1	Aditivo 2
17.3	18.7
18.4	17.8
19.1	21.3
16.7	21.0
18.2	22.1
18.6	18.7
17.5	19.8
	20.7
	20.2

AUTOevaluación

19. Las siguientes son muestras de los sueldos iniciales anuales de personas que ingresan a las carreras de contador público (Public Accountant) y de planificador financiero (Financial Planner). Los sueldos anuales se presentan en miles de dólares.

WEB archivo
AcctPlanners

Public Accountant	Financial Planner
50.2	49.0
58.8	49.2
56.3	53.1
58.2	55.9
54.2	51.9
55.0	53.6
50.9	49.7
59.5	53.9
57.0	51.8
51.9	48.9

a) Utilice 0.05 como nivel de significancia y la prueba de hipótesis de que no existe diferencia entre el sueldo inicial anual de los contadores públicos y los planificadores financieros. ¿Cuál es su conclusión?

b) ¿Cuáles son las medianas de los sueldos anuales de la muestra para las dos profesiones?

20. La brecha entre los ingresos de hombres y mujeres con el mismo nivel de educación se está reduciendo, pero no se ha cerrado. Los datos muestrales para siete hombres y siete mujeres con grado de licenciatura se listan a continuación. Los datos se presentan en miles de dólares.

Hombres	Mujeres
35.6	49.5
80.5	40.4
50.2	32.9
67.2	45.5
43.2	30.8
54.9	52.5
60.3	29.8

a) Cuál es la mediana del sueldo para los hombres y cuál para las mujeres?
b) Utilice $\alpha = 0.05$ y efectúe la prueba de hipótesis para distribuciones poblacionales idénticas. ¿Cuál es su conclusión?

21. Unisys mantiene una base de datos que proporciona información sobre los huracanes en el Atlántico/Caribe/Golfo de México (Atlantic/Caribbean/Gulf of Mexico) y el Océano Pacífico oriental (Eastern Pacific Ocean). A continuación se presentan las velocidades máximas del viento (Max Wind Speed) en nudos para una muestra de los huracanes (Storm) de las últimas cuatro temporadas (sitio web del clima de Unisys, abril de 2009).

WEB archivo
Hurricanes

Atlantic/Caribbean/Gulf of Mexico		Eastern Pacific Ocean	
Storm	**Max Wind Speed**	**Storm**	**Max Wind Speed**
Frances	125	Darby	105
Jeanne	110	Frank	75
Lisa	65	Isis	65
Emily	135	Hilary	90
Ophelia	80	Max	70
Rita	150	Bud	110
Wilma	150	Daniel	130
Ernesto	65	Sergio	95
Florence	80	Cosme	65
Helene	105	Flossie	120
Dean	145	Henriette	75
Karen	60	Ivo	70

Utilice $\alpha = 0.05$ y pruebe si la distribución de las velocidades es del viento de los huracanes es la misma para estas dos regiones. ¿Cuál es su conclusión?

22. *BusinessWeek* publica estadísticas anuales sobre las 1 000 empresas más grandes del mundo. La razón precio/utilidades (P/E) de una empresa (Company) es el precio actual de la acción dividido entre las utilidades por acción en los últimos 12 meses. En la tabla siguiente se muestra la relación P/E (P/E Ratio) para una muestra de 10 empresas japonesas (Japan) y 12 estadounidenses (United States). ¿La diferencia entre las razones P/E de estos dos países es significativa? Utilice la prueba de MWW y $\alpha = 0.01$ para apoyar su conclusión.

WEB archivo
JapanUS

Japan		United States	
Company	**P/E Ratio**	**Company**	**P/E Ratio**
Sumitomo Corp.	153	Gannet	19
Kinden	21	Motorola	24
Heiwa	18	Schlumberger	24
NCR Japan	125	Oracle Systems	43

(*Continúa*)

Japan		United States	
Company	**P/E Ratio**	**Company**	**P/E Ratio**
Suzuki Motor	31	Gap	22
Fuji Bank	213	Winn-Dixie	14
Sumintomo Chemical	64	Ingersoll-Rand	21
Seibu Railway	666	American Electric	14
Shiseido	33	Hercules	21
Toho Gas	68	Times Mirror	38
		WellPoint Health	15
		Northern States Power	14

23. Los registros de la policía muestran los números siguientes de delitos por día reportados para una muestra de días durante los meses de invierno y verano. Utilice 0.05 como nivel de significancia para determinar si existe una diferencia significativa entre los meses de invierno (Winter) y de verano (Summer) en términos del número de delitos reportados. ¿Cuál es su conclusión?

WEB archivo
PoliceRecords

Winter	Summer
18	28
20	18
15	24
16	32
21	18
20	29
12	23
16	38
19	28
20	18

24. Los hornos de microondas de determinada marca se venden en 10 tiendas de Dallas y 13 tiendas de San Antonio. Los datos se presentan a continuación. Utilice 0.05 como nivel de significancia y pruebe si los precios de los hornos son los mismos en las dos ciudades.

WEB archivo
Microwave

Dallas	San Antonio
445	460
489	451
405	435
485	479
439	475
449	445
436	429
420	434
430	410
405	422
	425
	459
	430

25. La National Football League (NFL) lleva a cabo su *draft* anual de los mejores jugadores de futbol colegial del país en abril de cada año. Antes del *draft*, varios noticieros deportivos pronostican cuáles jugadores serán seleccionados y el orden en que se les elegirá. Los mejores jugadores son seleccionados temprano en el *draft*. Para el *draft* 2009, se proyectaba que la mayoría de los jugadores seleccionados durante la primera ronda (sitio web SportProjection, 15 de marzo de 2009) serían de las universidades de la Conferencia del Sureste y de la Conferencia de la Costa del Atlántico. La universidad de los jugadores y el orden en que se proyectaba que serían elegidos para el draft se presenta a continuación para siete jugadores de cada conferencia.

Conferencia del Sureste		Conferencia de la Costa del Atlántico	
Universidad del jugador	**Orden de elección proyectado**	**Universidad del jugador**	**Orden de elección proyectado**
Georgia	1	Georgia Tech	3
Alabama	2	Wake Forest	6
Vanderbilt	14	Virginia	8
Florida	18	Wake Forest	23
Mississippi	20	Florida State	25
Mississippi	24	Maryland	26
Auburn	27	Virginia	29

Utilice el orden de proyección en el *draft* como un indicador de preferencia de los equipos de la NFL para las dos conferencias. Con la prueba de MWW determine si existe diferencia entre las preferencias por los jugadores de la NFL de estas dos conferencias. Utilice $\alpha = 0.05$ como nivel de significancia. ¿Cuál es el valor-p? ¿A qué conclusión llega?

19.4 Prueba de Kruskal-Wallis

En esta sección se amplían los procedimientos no paramétricos para probar la hipótesis en que participan tres o más poblaciones. En el capítulo 13 se consideró una prueba paramétrica para esta situación cuando se utilizaron los datos cuantitativos, y se asumió que en la población había distribuciones normales con las mismas desviaciones estándar. Con base en una muestra aleatoria independiente de cada población, se utilizó la distribución F para comprobar las diferencias entre las medias poblacionales.

La **prueba no paramétrica de Kruskal-Wallis** se basa en el análisis de muestras aleatorias independientes de cada una de las k poblaciones. Este procedimiento se puede utilizar ya sea con datos ordinales o con datos cuantitativos, y no requiere el supuesto de que las poblaciones tienen una distribución normal. La forma general de las hipótesis nula y alternativa es la siguiente.

H_0: todas las poblaciones son idénticas
H_a: no todas las poblaciones son idénticas

Si H_0 es rechazada, se concluirá que hay una diferencia entre las poblaciones con una o más poblaciones que tienden a dar valores más pequeños o más grandes en comparación con otras. Se demostrará la prueba de Kruskal-Wallis con el ejemplo que se menciona enseguida.

Williams Manufacturing Company contrata empleados para su equipo gerencial que provienen de tres universidades. Recientemente, el director del departamento de personal de la empresa comenzó a revisar el informe de desempeño anual en un intento por determinar si existen diferencias en el desempeño de los empleados de acuerdo con la universidad de procedencia. Se cuenta con los datos para muestras independientes de siete gerentes que se graduaron en la universidad A, seis que provienen de la universidad B y siete que egresaron de la universidad C. Estos datos se resumen en la tabla 19.14. La calificación del desempeño de cada gerente se registra en una escala de 0 a 100, donde 100 es la calificación más alta. Suponga que se desea probar si las tres poblaciones de gerentes son idénticas en términos de calificación del desempeño. Se utilizará un nivel de significancia de 0.05 para la prueba.

TABLA 19.14
Evaluación del desempeño de 20 empleados de Williams

Universidad		
A	**B**	**C**
25	60	50
70	20	70
60	30	60
85	15	80
95	40	90
90	35	70
80		75

El primer paso en el procedimiento de Kruskal-Wallis es clasificar las muestras combinadas de los valores más bajos a los más altos. Utilizando las 20 observaciones de la tabla 19.14, el nivel más bajo de 15 para el 4º gerente en la muestra de la universidad B recibe un rango 1. La más alta calificación de 95 para el 5º director la muestra de la universidad A recibe un rango de 20. Los datos de desempeño y sus rangos asignados se muestran en la tabla 19.15. Observe que se asignan rangos promedio para igualar las calificaciones de desempeño de 60, 70, 80 y 90. La tabla 19.15 también muestra la suma de los rangos para cada una de las tres muestras.

El estadístico de la prueba de Kruskal-Wallis utiliza la suma de los rangos de las tres muestras y se calcula de la siguiente manera.

ESTADÍSTICO DE PRUEBA DE KRUSKAL-WALLIS

$$H = \left[\frac{12}{n_T(n_T + 1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i}\right] - 3(n_T + 1) \qquad \textbf{(19.7)}$$

donde

k = número de poblaciones

n_i = número de observaciones en la muestra i

$n_T = \sum_{i=1}^{k} n_t$ = número total de observaciones en todas las muestras

R_i = suma de los rangos para la muestra i

TABLA 19.15 Rangos para las tres muestras combinadas

Universidad A	Rango	Universidad B	Rango	Universidad C	Rango
25	3	60	9	50	7
70	12	20	2	70	12
60	9	30	4	60	9
85	17	15	1	80	15.5
95	20	40	6	90	18.5
90	18.5	35	5	70	12
80	15.5	Suma de rangos	27	75	14
Suma de rangos	95			Suma de rangos	88

Kruskal y Wallis fueron capaces de demostrar que, bajo el supuesto de la hipótesis nula de que las poblaciones son idénticas, la distribución de muestreo de H se puede aproximar mediante una distribución ji-cuadrada con $(k - 1)$ grados de libertad. Esta aproximación es aceptable si el tamaño de las muestras de cada una de las k poblaciones es mayor o igual a cinco. La hipótesis nula de poblaciones idénticas será rechazada si el estadístico de prueba H es grande. Como resultado, la prueba de Kruskal-Wallis se expresa siempre como una prueba de cola superior. El cálculo del estadístico de prueba para los datos muestrales de la tabla 19.15 es el siguiente.

Los tamaños de las muestras son

$$n_1 = 7 \qquad n_2 = 6 \qquad n_3 = 7$$

y

$$n_T = \sum_{i=1}^{3} n_t = 7 + 6 + 7 = 20$$

Al utilizar la suma de rangos para cada muestra, el valor del estadístico de prueba de Kruskal-Wallis es el siguiente.

$$H = \left[\frac{12}{n_T(n_T + 1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i}\right] - 3(n_T + 1) = \frac{12}{20(21)}\left[\frac{(95)^2}{7} + \frac{(27)^2}{6} + \frac{(88)^2}{7}\right] - 3(20 + 1) = 8.92$$

Ahora se puede utilizar la tabla de distribución ji-cuadrada (tabla 3 del apéndice B) para determinar el valor-p de la prueba. Al utilizar $k - 1 = 3 - 1 = 2$ grados de libertad, se determina que $\chi^2 = 7.378$ tiene un área de 0.025 en la cola superior de la distribución ji-cuadrada y $\chi^2 = 9.21$ tiene un área de 0.01 en la cola superior de la distribución ji-cuadrada. Con $H = 8.92$ entre 7.378 y 9.21, se concluye que el área en dicha cola está entre 0.025 y 0.01. Debido a que esta es una prueba de cola superior, se llega a la conclusión de que el valor-p está entre 0.025 y 0.01. Al utilizar Minitab o Excel se encuentra que el valor-p exacto para $\chi^2 = 8.92$ es 0.0116. Ya que el valor-$p \leq \alpha = 0.05$, H_0 es rechazada y se concluye que las tres poblaciones no son iguales. El desempeño de las tres poblaciones no es idéntico y difiere de manera significativa dependiendo de la universidad. Debido a que la suma de rangos es relativamente baja para la muestra de los gerentes que provienen de la universidad B, sería razonable para la empresa ya sea reducir su reclutamiento de esta institución, o al menos evaluar a sus graduados más a fondo antes de decidir una contratación.

Si se puede hacer el supuesto de que las poblaciones tienen la misma forma, la prueba de Kruskal-Wallis se convierte en una prueba acerca de las medianas de las k poblaciones.

Como comentario final, observamos que en algunas aplicaciones de la prueba de Kruskal-Wallis puede ser apropiado hacer el supuesto de que las poblaciones tienen formas idénticas y, si difieren, será sólo por un cambio en la ubicación de una o más de las poblaciones. Si se supone que las poblaciones k tienen la misma forma, la prueba de hipótesis se puede afirmar en términos de las medianas poblacionales. En este caso, las hipótesis para la prueba de Kruskal-Wallis se escribirían de la siguiente manera.

H_0: $\text{mediana}_1 = \text{mediana}_2 = \cdots = \text{mediana}_k$

H_a: no todas las medianas son iguales

NOTAS Y COMENTARIOS

1. El ejemplo de esta sección utiliza datos cuantitativos sobre las calificaciones de desempeño de los empleados para efectuar la prueba de Kruskal-Wallis. Este examen también se podría haber utilizado si los datos fueran los valores ordinales de los 20 empleados en términos de desempeño. En este caso, la prueba utilizaría los datos ordinales directamente. No es necesario el paso de convertir los datos cuantitativos en datos ordenados por rango. En el ejercicio 30 se ilustra esta situación.

2. El procedimiento de Minitab para la prueba de Kruskal-Wallis se describe en el apéndice 19.1. Minitab asume que las poblaciones tienen la misma forma. Como resultado, describe la prueba de Kruskal-Wallis como una prueba de las diferencias entre las medianas de las poblaciones. Si no se siente cómodo al realizar el supuesto de la "misma forma", puede seguir utilizando Minitab. Sin embargo, tendrá que interpretar los resultados como una prueba de hipótesis nula en la que todas las poblaciones son idénticas.

Ejercicios

Aplicaciones

AUTOevaluación

26. Una muestra de 15 consumidores ofrece las siguientes calificaciones para tres productos diferentes. A cinco de los consumidores se les asignó aleatoriamente probar y calificar cada producto. Utilice la prueba de Kruskal-Wallis y $\alpha = 0.05$ para determinar si existe una diferencia significativa entre las respectivas calificaciones.

Producto		
A	**B**	**C**
50	80	60
62	95	45
75	98	30
48	87	58
65	90	57

WEB archivo
TestPrepare

27. Se evalúan tres programas de preparación para un examen de admisión. Las calificaciones obtenidas por una muestra de 20 personas que utilizaron los programas de preparación arrojaron los siguientes datos. Utilice la prueba de Kruskal-Wallis para determinar si existe una diferencia significativa entre los tres programas. Utilice $\alpha = 0.05$.

Programa		
A	**B**	**C**
540	450	600
400	540	630
490	400	580
530	410	490
490	480	590
610	370	620
	550	570

AUTOevaluación

28. El entrenamiento durante 40 minutos de una de las siguientes actividades durante tres días a la semana favorece la pérdida de peso. Los siguientes datos muestrales de tres diferentes actividades presentan el número de calorías quemadas durante 40 minutos de ejercicio. ¿Estos datos indican diferencias en la cantidad de calorías quemadas en las tres actividades? Utilice 0.05 como nivel de significancia. ¿Cuál es su conclusión?

Natación	Tenis	Ciclismo
408	415	385
380	485	250
425	450	295
400	420	402
427	530	268

WEB archivo
CruiseShips

29. La revista *Condé Nast Traveler* realiza un estudio anual entre sus lectores con el fin de calificar los 80 cruceros más importantes del mundo (*Condé Nast Traveler*, febrero de 2008). Con 100 como la calificación (Rating) más alta posible, se lista la siguiente puntuación global para una muestra de los barcos (Ship) de Holland America, Princess y Royal Caribbean. Utilice la prueba de Kruskal-Wallis con $\alpha = 0.05$ para determinar si en general las calificaciones entre las tres líneas de cruceros difieren significativamente. ¿Cuál es su conclusión?

Holland America		Princess		Royal Caribbean	
Ship	**Rating**	**Ship**	**Rating**	**Ship**	**Rating**
Amsterdam	84.5	Coral	85.1	Adventure	84.8
Maasdam	81.4	Dawn	79.0	Jewel	81.8
Ooterdam	84.0	Island	83.9	Mariner	84.0
Volendam	78.5	Princess	81.1	Navigator	85.9
Westerdam	80.9	Star	83.7	Serenade	87.4

30. Una empresa grande envía a muchos de sus gerentes de primer nivel a un curso de capacitación sobre habilidades de supervisión. Cuatro distintos centros de desarrollo gerencial ofrecen este curso. El director de recursos humanos desea saber si existe una diferencia entre la calidad de la formación impartida en los cuatro centros. Se eligió una muestra aleatoria independiente de cinco empleados de cada centro de capacitación. Los empleados se clasifican de 1 a 20 en términos de habilidades de supervisión. Un rango de 1 fue asignado al empleado con las mejores habilidades. Los rangos se listan a continuación. Utilice $\alpha = 0.05$ y compruebe si existe una diferencia significativa entre la calidad de la capacitación impartida en los cuatro programas.

Curso			
A	**B**	**C**	**D**
3	2	19	20
14	7	16	4
10	1	9	15
12	5	18	6
13	11	17	8

31. Los dulces más vendidos son a menudo los más ricos en calorías. Suponga que los datos que se presentan enseguida reflejan el contenido calórico de las muestras de M&M's, Kit Kat y Milky Way II. Pruebe si hay diferencias significativas entre el contenido calórico de estos tres dulces. Con un nivel de significancia de 0.05, ¿cuál es su conclusión?

M&M's	Kit Kat	Milky Way II
230	225	200
210	205	208
240	245	202
250	235	190
230	220	180

19.5 Correlación de rangos

El coeficiente de correlación producto-momento de Pearson presentado en el capítulo 3 es una medición de la asociación lineal entre dos variables a partir de datos cuantitativos. En esta sección se presenta una medida de correlación de la asociación entre dos variables para datos ordinales u ordenados por rango. El **coeficiente de correlación por rangos de Spearman** se ha desarrollado para este propósito.

COEFICIENTE DE CORRELACIÓN POR RANGOS DE SPEARMAN

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 + 1)} \qquad (19.8)$$

donde

n = número de observaciones en la muestra
x_i = rango de la observación i con respecto a la primera variable
y_i = rango de la observación i con respecto a la segunda variable
$d_i = x_i - y_i$

Ilustremos el uso del coeficiente de correlación por rangos de Spearman. Una empresa desea determinar si las personas que prometían el más alto potencial al momento de ser contratadas resultaron tener el registro de mayores ventas. Para investigarlo, el director de recursos humanos revisó los registros originales de las entrevistas de trabajo, expediente académico y cartas de recomendación de 10 miembros de la fuerza de ventas. Después de la revisión, el director clasificó a las 10 personas con base en su potencial de éxito en el momento de ser contratadas y asignó el rango 1 a la persona que tuvo el mayor potencial. A continuación se recabaron los datos de las ventas reales de cada vendedor durante sus primeros dos años de empleo. Sobre la base de los registros de ventas reales se obtuvo una segunda calificación de los 10 sujetos con base en el desempeño en las ventas. En la tabla 19.16 se presentan los rangos basados en el potencial (Ranking of Potential) y en el desempeño real de cada vendedor (Salesperson). En la tercera columna se listan las ventas de dos años (Two-Year Sales) en unidades (units) y en la cuarta columna el rango con base en las ventas de dos años (Ranking According to Two-Year Sales).

Calculemos el coeficiente de correlación por rangos de Spearman para los datos de la tabla 19.16, los cuales se resumen en la tabla 19.17. En primer lugar, se calcula la diferencia entre los dos rangos de cada vendedor, d_i, como se muestra en la columna 4. La suma de d_i^2 en la columna 5 es 44. Este valor y el tamaño de muestra $n = 10$ se utilizan para calcular el coeficiente de correlación por rangos $r_s = 0.733$ mostrado en la tabla 19.17.

El coeficiente de correlación por rangos de Spearman oscila entre -1.0 y $+1.0$, y su interpretación es igual al coeficiente de correlación producto-momento de Pearson para datos cuantitativos. El coeficiente de correlación por rangos cerca de $+1.0$ indica una relación positiva

TABLA 19.16 Ventas potenciales y datos de las ventas reales de dos años

WEB archivo
PotentialActual

Salesperson	Ranking of Potential	Two-Year Sales (units)	Ranking According to Two-Year Sales
A	2	400	1
B	4	360	3
C	7	300	5
D	1	295	6
E	6	280	7
F	3	350	4
G	10	200	10
H	9	260	8
I	8	220	9
J	5	385	2

TABLA 19.17 Cálculo del coeficiente de correlación por rangos de Spearman para el potencial y el desempeño de ventas

Vendedor	x_i = Rango del potencial	y_i = Rango del desempeño en ventas	$d_i = x_i - y_i$	d_i^2
A	2	1	1	1
B	4	3	1	1
C	7	5	2	4
D	1	6	−5	25
E	6	7	−1	1
F	3	4	−1	1
G	10	10	0	0
H	9	8	1	1
I	8	9	−1	1
J	5	2	3	9
				$\sum d_i^2 = 44$

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2 + 1)} = 1 - \frac{6(44)}{10(100 - 1)} = 0.733$$

entre los rangos de dos variables, mientras que el coeficiente cercano a −1.0 indica una fuerte relación negativa entre los rangos de las dos variables. Un coeficiente de 0 indica que no hay asociación entre los rangos de las dos variables. En el ejemplo, $r_s = 0.733$ indica una correlación positiva entre los rangos basada en el potencial y los rangos en función del desempeño en ventas. Las personas que mejor califican en el potencial al momento de ser contratadas tienden a tener un rango más alto en el desempeño de las ventas a dos años.

En este punto, es posible que se desee utilizar la correlación muestral por rangos r_s para hacer una inferencia acerca del coeficiente de correlación por rango de la población ρ_s. Para esto ponemos a prueba las siguientes hipótesis.

$$H_0: \rho_s = 0$$
$$H_a: \rho_s \neq 0$$

Bajo el supuesto de que la hipótesis nula es verdadera y el coeficiente de correlación por rangos de la población es 0, la siguiente distribución de muestreo de r_s puede utilizarse para realizar la prueba.

DISTRIBUCIÓN DE MUESTREO DE r_s

$$\text{Media: } \mu_{r_s} = 0 \qquad \textbf{(19.9)}$$

$$\text{Desviación estándar: } \sigma_{r_s} = \sqrt{\frac{1}{n-1}} \qquad \textbf{(19.10)}$$

Forma de distribución: aproximadamente normal siempre que $n \geq 10$

El coeficiente de correlación muestral por rangos entre el potencial y el desempeño de ventas es $r_s = 0.733$. Al utilizar la ecuación (19.9) tenemos $\mu_{r_s} = 0$, y con la ecuación (19.10) tenemos $\sigma_{r_s} = \sqrt{1/(10-1)} = 0.333$. Con la distribución de muestreo de r_s aproximada por una distribución normal, la variable aleatoria normal estándar z se convierte en el estadístico de prueba con

$$z = \frac{r_s - \mu_{r_s}}{\sigma_{r_s}} = \frac{0.733 - 0}{0.333} = 2.20$$

Utilizando la tabla de probabilidad normal estándar y $z = 2.20$ se determina que el valor-p de dos colas $= 2(1 - 0.9861) = 0.0278$. Con un nivel de significancia 0.05, el valor-$p \leq \alpha$. Por tanto, se rechaza la hipótesis nula de que el coeficiente de correlación por rangos poblacional sea cero. El resultado muestra que existe una correlación por rangos significativa entre el potencial del vendedor al momento de ser contratado y el desempeño de las ventas reales.

NOTAS Y COMENTARIOS

El coeficiente de correlación por rangos de Spearman proporciona el mismo valor que se obtiene mediante el procedimiento del coeficiente de correlación producto-momento de Pearson con los datos ordenados por rango. En los apéndices 19.1 y 19.2 se muestra cómo pueden utilizarse las herramientas de correlación de Minitab y Excel para el coeficiente de correlación producto-momento a efecto de calcular el coeficiente de correlación por rangos de Spearman.

Ejercicios

Métodos

AUTOevaluación

32. Considere el siguiente conjunto de clasificaciones o rangos para una muestra de 10 elementos.

Elemento	x	y	Elemento	x	y
1	10	8	6	2	7
2	6	4	7	8	6
3	7	10	8	5	3
4	3	2	9	1	1
5	4	5	10	9	9

a) Calcule el coeficiente de correlación por rangos de Spearman para los datos.
b) Utilice $\alpha = 0.05$ y pruebe la significancia de la correlación por rangos. ¿Cuál es su conclusión?

33. Considere los dos siguientes conjuntos de clasificaciones de seis artículos.

	Caso uno			Caso dos	
Artículo	Primera clasificación	Segunda clasificación	Artículo	Primera clasificación	Segunda clasificación
A	1	1	A	1	6
B	2	2	B	2	5
C	3	3	C	3	4
D	4	4	D	4	3
E	5	5	E	5	2
F	6	6	F	6	1

Observe que en el primer caso las clasificaciones o rangos son idénticos, y en el segundo son exactamente lo contrario. ¿Qué valor debe esperar para el coeficiente de correlación por rangos de Spearman en cada caso? Explique. Calcule el coeficiente de correlación por rangos para cada uno.

AUTO evaluación

Aplicaciones

34. Los datos siguientes muestran clasificaciones o rangos de 11 estados para los gastos por alumno (Expenditure per Student) (clasificando 1 como el más alto hasta 11 como el más bajo) y el cociente estudiante-profesor (Student-Teacher Ratio) (con 1 como el más bajo hasta 11 como el más alto).

WEB archivo
Student

State	Expenditure per Student	Student-Teacher Ratio
Arizona	9	10
Colorado	5	8
Florida	4	6
Idaho	2	11
Iowa	6	4
Louisiana	11	3
Massachusetts	1	1
Nebraska	7	2
North Dakota	8	7
South Dakota	10	5
Washington	3	9

a) ¿Cuál es la correlación por rangos entre el gasto por estudiante y la relación estudiante-profesor? Comente.

b) Al nivel $\alpha = 0.05$ ¿parece haber una relación entre el gasto por estudiante y la razón estudiante-profesor?

35. Un estudio nacional de Harris Interactive, Inc. evaluó las empresas de alta tecnología y su reputación. La tabla siguiente muestra cómo se clasifica la reputación de 10 empresas de alta tecnología y cómo se la clasifica en el porcentaje de encuestados que dijeron que comprarían acciones de la empresa. Se prevé una correlación por rangos positiva, ya que parece razonable esperar que una empresa (Company) con una reputación (Reputation) más grande tenga las acciones más cotizadas o compras de acciones (Stock Purchase) más altas.

Company	Reputation	Stock Purchase
Microsoft	1	3
Intel	2	4
Dell	3	1

(*Continúa*)

WEB archivo
Techs

Company	Reputation	Stock Purchase
Lucent	4	2
Texas Instruments	5	9
Cisco Systems	6	5
Hewlett-Packard	7	10
IBM	8	6
Motorola	9	7
Yahoo	10	8

a) Calcule la correlación por rangos entre la reputación y la compra de acciones
b) Pruebe si la correlación por rangos es significancia positiva. ¿Cuál es el valor-*p*?
c) Con $\alpha = 0.05$, ¿cuál es su conclusión?

36. A continuación se presenta la clasificación de una muestra de golfistas (Golfer) profesionales respecto del *driving distance* como del *putting*. ¿Cuál es la correlación por rangos entre el *driving distance* y el *putting* para estos jugadores? Utilice 0.10 como nivel de significancia y pruebe la significancia de la correlación por rangos.

WEB archivo
ProGolfers

Golfer	Driving Distance	Putting
Fred Couples	1	5
David Duval	5	6
Ernie Els	4	10
Nick Faldo	9	2
Tom Lehman	6	7
Justin Leonard	10	3
Davis Love III	2	8
Phil Mickelson	3	9
Greg Norman	7	4
Mark O'Meara	8	1

37. En una universidad, una organización estudiantil entrevistó tanto a los alumnos actuales como a los recién egresados para obtener información sobre la calidad de la enseñanza. Al analizar las respuestas, se llega a la siguiente clasificación sobre la habilidad de los profesores (Professor) para la enseñanza. ¿Coincide la clasificación de los estudiantes actuales (Current Students) con la de los recién egresados (Recent Graduates)? Utilice $\alpha = 0.10$ y pruebe la significancia de la correlación por rangos.

WEB archivo
Professors

Professor	Current Students	Recent Graduates
1	4	6
2	6	8
3	8	5
4	3	1
5	1	2
6	2	3
7	5	7
8	10	9
9	7	4
10	9	10

Resumen

En este capítulo se presentaron procedimientos estadísticos que se clasifican como métodos no paramétricos. Debido a que los métodos se pueden aplicar tanto a los datos categóricos como a los cuantitativos, y debido a que no requieren supuestos acerca de la distribución de la población, amplían el número de situaciones que pueden ser sometidas a análisis estadístico.

La prueba de signos es un procedimiento no paramétrico para probar hipótesis acerca de una mediana de la población o con muestras pareadas. Los datos deben resumirse en dos categorías: una denotada por un signo positivo y otra denotada por un signo negativo. La prueba de rangos con signo de Wilcoxon analiza las muestras pareadas de dos poblaciones cuando se dispone de datos cuantitativos. No se requiere ningún supuesto que no sea que la distribución de las diferencias pareadas sea simétrica. La prueba de rangos con signo de Wilcoxon se utiliza para determinar si la mediana poblacional de diferencias pareadas es cero. Esta prueba también se utiliza para hacer inferencias acerca de la mediana de una población simétrica.

La prueba de Mann-Whitney-Wilcoxon es un procedimiento no paramétrico para la diferencia entre dos poblaciones a partir de dos muestras independientes. Es una alternativa para la prueba paramétrica de t para la diferencia entre las medias de dos poblaciones. Se obtienen los rangos combinados para los datos de las dos muestras y el estadístico de prueba para la prueba de MWW es la suma de los rangos de la primera muestra. En la mayoría de las aplicaciones, los tamaños de muestra son lo suficientemente grandes como para utilizar una aproximación normal con el factor de corrección de continuidad en la realización de la prueba de hipótesis. Si no se realiza ningún supuesto acerca de la población, el procedimiento MWW prueba si las dos poblaciones son idénticas. Si puede hacer el supuesto de que las dos poblaciones tengan la misma forma, la prueba ofrece una inferencia acerca de las diferencias entre las medianas de las dos poblaciones.

La prueba de Kruskal-Wallis extiende la prueba de MWW a tres o más poblaciones. Es una alternativa al análisis paramétrico de la prueba de varianza para las diferencias entre las medias de tres o más poblaciones con distribución normal. La prueba de Kruskal-Wallis no requiere ningún supuesto acerca de la distribución de las poblaciones y utiliza la hipótesis nula de que las k poblaciones son idénticas. Si se puede hacer el supuesto de que las poblaciones tienen la misma forma, la prueba ofrece una inferencia acerca de las diferencias entre las medianas de las k poblaciones. En la última sección del capítulo se introdujo el coeficiente de correlación por rangos de Spearman como una medida de asociación entre dos variables basada en datos ordenados por rango.

Glosario

Coeficiente de correlación por rangos de Spearman Medida de correlación de la asociación entre dos variables basada en datos ordenados por rangos.

Métodos de distribución libre Métodos estadísticos que no realizan supuestos acerca de la probabilidad de distribución de la población.

Métodos no paramétricos Métodos estadísticos que no requieren un supuesto acerca de la forma de la distribución de probabilidad de una población; a menudo se conocen como *métodos de distribución libre*. Varios de ellos pueden aplicarse a los datos categóricos y a los cuantitativos.

Métodos paramétricos Métodos estadísticos que comienzan con un supuesto acerca de la distribución de probabilidad de de la población que a menudo establece que la población tiene una distribución normal. Entonces puede derivarse una distribución de muestreo para el estadístico de prueba y utilizarse para hacer una inferencia acerca de uno o más parámetros de la población, como la media poblacional μ o la desviación estándar poblacional Σ.

Prueba de Kruskal-Wallis Prueba no paramétrica de las diferencias entre tres o más poblaciones a partir del análisis sobre muestras independientes de cada población. La hipótesis nula es que las poblaciones son idénticas. Si se puede suponer que las poblaciones tienen la misma forma, esta prueba proporciona una inferencia acerca de las diferencias entre las medianas de las poblaciones.

Prueba de Mann-Whitney-Wilcoxon (MWW) Prueba no paramétrica de la diferencia entre dos poblaciones basada en una muestra independiente de cada población. La hipótesis nula establece que las dos poblaciones tienen la misma forma. Si puede plantearse tal supuesto, esta prueba proporciona una inferencia sobre la diferencia entre las medianas de las dos poblaciones.

Prueba de rangos con signo de Wilcoxon Prueba no paramétrica de la diferencia entre las medianas de dos poblaciones basada en muestras pareadas. El procedimiento utiliza datos cuantitativos y está basado en el supuesto de que la distribución de las diferencias es simétrica. Los datos de diferencias pareadas se utilizan para hacer una inferencia acerca de las medianas de las dos poblaciones. Esta prueba también se utiliza para hacer inferencias acerca de la mediana de una población simétrica.

Prueba de signos Prueba de hipótesis no paramétrica acerca de una mediana de la población o que identifica las diferencias entre dos poblaciones basadas en muestras pareadas. Los datos se resumen en dos categorías, la denotada por un signo positivo o por un signo negativo, y la distribución binomial con $p = 0.50$ proporciona la distribución de muestreo para el estadístico de prueba.

Fórmulas clave

Prueba de signos: aproximación normal

$$\text{Media: } \mu = 0.50n \tag{19.1}$$

$$\text{Desviación estándar: } \sigma = \sqrt{0.25n} \tag{19.2}$$

Prueba de rangos con signo de Wilcoxon: aproximación normal

$$\text{Media: } \mu_{T^+} = \frac{n(n+1)}{4} \tag{19.3}$$

$$\text{Desviación estándar: } \sigma_{T^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}} \tag{19.4}$$

Prueba de Mann-Whitney-Wilcoxon: aproximación normal

$$\text{Media: } \mu_W = \tfrac{1}{2}n_1(n_1 + n_2 + 1) \tag{19.5}$$

$$\text{Desviación estándar: } \sigma_W = \sqrt{\tfrac{1}{12}n_1n_2(n_1 + n_2 + 1)} \tag{19.6}$$

Estadístico de prueba de Kruskal-Wallis

$$H = \left[\frac{12}{n_T(n_T+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right] - 3(n_T + 1) \tag{19.7}$$

Coeficiente de correlación por rangos de Spearman

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2+1)} \tag{19.8}$$

Ejercicios complementarios

38. En una encuesta se planteó la siguiente pregunta: ¿Está a favor o en contra de proporcionar vales libres de impuestos o deducciones fiscales a los padres que envían a sus hijos a escuelas privadas? De 2 010 personas entrevistadas, 905 se manifestaron a favor de la propuesta, 1 045 se opusieron y 60 no dieron ninguna opinión. ¿Los datos indican una diferencia significativa en las preferencias por el apoyo financiero a los padres que envían a sus hijos a escuelas privadas? Utilice 0.05 como nivel de significancia.
39. Debido a una reciente disminución en el mercado de la vivienda, la mediana del precio nacional de venta de casas unifamiliares es $180 000 (Asociación Nacional de Agentes Inmobiliarios, enero de 2009). Suponga que los siguientes datos se obtuvieron de las muestras de ventas recientes para viviendas unifamiliares en San Luis y Denver.

Área metropolitana	Menos de $180 000	Igual a $180 000	Más de $180 000
San Luis	32	2	18
Denver	13	1	27

a) ¿La mediana del precio de venta en San Luis es significativamente inferior a la mediana nacional de $180 000? Utilice el estadístico de prueba con $\alpha = 0.05$ para apoyar su conclusión.

b) ¿La mediana del precio de venta en Denver es significativamente más alta que la mediana nacional de $180 000? Utilice el estadístico de prueba con $\alpha = 0.05$ para apoyar su conclusión.

40. A 12 amas de casa se les pidió que estimaran el precio de venta al menudeo de dos modelos de refrigeradores. Sus estimaciones se muestran en la tabla siguiente. Utilice estos datos y realice una prueba con nivel de significancia 0.05 para determinar si existen diferencias entre los dos modelos (Model 1 y 2) en términos de las percepciones de las amas de casa (Homemaker) sobre el precio de venta.

WEB archivo
Refrigerators

Homemaker	Model 1	Model 2	Homemaker	Model 1	Model 2
1	$850	$1 100	7	$900	$1 090
2	960	920	8	890	1 120
3	940	890	9	1 100	1 200
4	900	1 050	10	700	890
5	790	1 120	11	810	900
6	820	1 000	12	920	900

41. Un estudio fue diseñado para evaluar el aumento de peso con un nuevo alimento para aves. Una muestra de 12 pollos fue utilizada durante un periodo de seis semanas. El peso de cada pollo se registró antes y después del periodo de prueba. Las diferencias entre los pesos anteriores y posteriores de los 12 pollos son las siguientes: 1.5, 1.2, −0.2, 0.0, 0.5, 0.7, 0.8, 1.0, 0.0, 0.6, 0.2 y −0.01. Una diferencia positiva indica un aumento de peso y una diferencia negativa expresa una pérdida de peso. Utilice 0.05 como nivel de significancia para determinar si el nuevo alimento proporciona un aumento de peso significativo en las aves.

42. Los datos siguientes indican pesos de productos para los mismos artículos fabricados en dos líneas de producción distintas. Pruebe la diferencia entre los pesos del producto de las dos líneas (Line 1 y Line 2). Utilice $\alpha = 0.05$.

WEB archivo
ProductWeights

Line 1	Line 2
13.6	13.7
13.8	14.1
14.0	14.2
13.9	14.0
13.4	14.6
13.2	13.5
13.3	14.4
13.6	14.8
12.9	14.5
14.4	14.3
	15.0
	14.9

43. Un cliente desea determinar si existe una diferencia significativa en el tiempo requerido para completar una evaluación de programa con tres métodos diferentes de uso común. A continuación se presentan los tiempos (en horas) requeridos por cada uno de los 18 evaluadores para llevar a cabo la evaluación del programa.

Utilice $\alpha = 0.05$ y realice una prueba para ver si existe una diferencia significativa en el tiempo requerido para los tres métodos (Method 1, 2 y 3).

WEB archivo
Methods

Method 1	Method 2	Method 3
68	62	58
74	73	67
65	75	69
76	68	57
77	72	59
72	70	62

44. Una muestra de 20 ingenieros que han estado empleados en una empresa durante tres años ha sido ordenada por rangos con base en su potencial gerencial. Algunos ingenieros han asistido a cursos de desarrollo gerencial impartidos por la empresa (Company Program), otros asisten a cursos impartidos por una universidad local (Off-Site Program) y el resto no tomó ningún curso (No Program). Utilice las clasificaciones siguientes y $\alpha = 0.025$ para probar una diferencia significativa en el potencial gerencial de los tres grupos.

WEB archivo
Programs

No Program	Company Program	Off-Site Program
16	12	7
9	20	1
10	17	4
15	19	2
11	6	3
13	18	8
	14	5

45. A continuación se presentan las calificaciones de cuatro profesores (Instructor) universitarios. Utilice $\alpha = 0.05$ y pruebe si existe una diferencia significativa en las calificaciones. ¿Cuál es su conclusión?

WEB archivo
Evaluations

Instructor			
Black	Jennings	Swanson	Wilson
88	87	88	80
80	78	76	85
79	82	68	56
68	85	82	71
96	99	85	89
69	99	82	87
	85	84	
	94	83	
		81	

46. Los 15 alumnos de una muestra recibieron los siguientes rangos (Rank) por su desempeño en los exámenes de mitad de semestre (Midterm) y final de un curso de estadística. Calcule el coeficiente de correlación por rangos de Spearman y pruebe si hay una correlación significativa con $\alpha = 0.10$. ¿Cuál es el valor-p y cuál es su conclusión?

WEB archivo
Exams

Rank		Rank		Rank	
Midterm	**Final**	**Midterm**	**Final**	**Midterm**	**Final**
1	4	6	2	11	14
2	7	7	5	12	15
3	1	8	12	13	11
4	3	9	6	14	10
5	8	10	9	15	13

47. Nielsen Research ofrece las clasificaciones semanales de programas de televisión de difusión nacional. Las calificaciones de los 84 programas de horario estelar que se transmiten por las cuatro cadenas de televisión (Network) más importantes (ABC, CBS, FOX y NBC) para la semana del 14 al 20 abril de 2008 se proporcionan en el archivo NielsenResearch. Los rangos de las clasificaciones van de 1 a 103. A continuación se muestran las clasificaciones (Rating) para 12 programas (Program) del archivo (con los días y horarios en que salieron al aire los programas con múltiples episodios). ¿Estos datos sugieren que la puntuación de las cuatro televisoras difiere de manera significativa? Utilice la prueba de Kruskal-Wallis con un nivel de significancia de 0.10. ¿Cuál es el valor-p y cuál es su conclusión?

WEB archivo
NielsenResearch

Program	Network	Rating
20/20	ABC	60
30 Rock	NBC	44
48 Hours Mystery (Sat. 10:00 p.m.)	CBS	51
48 Hours Mystery (Sat. 9:00 p.m.)	CBS	78
48 Hours Mystery (Tues. 10:00 p.m.)	CBS	63
60 Minutes	CBS	13
According to Jim (Tues. 8:00 p.m.)	ABC	89
According to Jim (Tues. 8:30 p.m.)	ABC	91
American Dad (Sun. 7:30 p.m.)	FOX	100
American Dad (Sun. 9:30 p.m.)	FOX	65
American Idol (Tues. 8:00 p.m.)	FOX	1
American Idol (Wed. 9:00 p.m.)	FOX	2

Apéndice 19.1 Métodos no paramétricos con Minitab

Minitab se puede utilizar para todos los métodos no paramétricos presentados en este capítulo.

Prueba de signos para una prueba de hipótesis acerca de una mediana poblacional

WEB archivo
HomeSales

Se ilustra una prueba de hipótesis acerca de una mediana poblacional utilizando los datos de precios de venta de casas nuevas de la sección 19.1. Los precios aparecen en la columna C1 de la hoja de cálculo de Minitab llamada HomeSales. Los siguientes pasos se utilizan para probar la hipótesis H_0: mediana $\geq$ \$236 000 contra H_a: mediana $<$ \$236 000.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Nonparametrics**.
Paso 3. Elija **1-Sample Sign**.
Paso 4. Cuando el cuadro de diálogo 1-Sample Sign aparezca:
Introduzca C1 en el cuadro **Variables**.
Seleccione **Test Median**.
Introduzca el valor hipotético 23600 en el cuadro **Test Median**.
Seleccione **less than** desde el menú **Alternative**.
Haga clic en **OK**.

Minitab proporciona el valor-*p*, así como una estimación puntual de la mediana poblacional.

Este procedimiento también se utiliza para obtener una estimación por intervalo de la mediana de la población. En el paso 4, seleccione **Confidence interval** en lugar de **Test median**, introduzca el nivel de confianza en **Confidence level** y haga clic en **OK**. Para tamaños de muestra mayores de 50, Minitab utiliza una aproximación normal a la distribución muestral binomial con el factor de corrección de continuidad tanto para la prueba de hipótesis como para los cálculos del intervalo de confianza.

Prueba de los signos para una prueba de hipótesis con muestras pareadas

WEB archivo
SunCoast

En el procedimiento de Minitab de la prueba de los signos para una prueba de hipótesis con muestras pareadas se utiliza un código numérico para un signo positivo, un signo negativo y los datos de ninguna preferencia. Este procedimiento se ilustra con la prueba de hipótesis del ejemplo de Sun Coast Farms de la sección 19.1. El archivo de datos Suncoast muestra que la columna C1 contiene números de identificación de cada una de las 14 personas que participan en la prueba de sabor, y la columna C2 utiliza el número + 1 para cada signo positivo, a − 1 para cada signo negativo y 0 para los individuos sin preferencia. Si la hipótesis nula de "ninguna preferencia" es verdadera, la mediana de la población de +1, −1 y 0 será igual a cero. Por tanto, se siguen los pasos subsecuentes para probar una mediana poblacional con la mediana hipotética igual a cero. Para la prueba de hipótesis de Sun Coast Farms se utiliza una prueba de dos colas de la siguiente manera.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Nonparametrics**.
Paso 3. Elija **1-Sample Sign**.
Paso 4. Cuando el cuadro de diálogo 1-Sample Sign aparezca:
Introduzca C2 en el cuadro **Variables**.
Seleccione **Test Median**.
Introduzca el valor hipotético 0 en el cuadro **Test Median**.
Seleccione **not equal** desde el menú **Alternative**.
Haga clic en **OK**.

Prueba de rangos con signo de Wilcoxon con muestras pareadas

WEB archivo
MatchedSample

Los siguientes pasos permiten probar la hipótesis acerca de la diferencia entre dos medianas poblacionales a partir de datos de muestras pareadas. Para ilustrarlo, se utilizarán los datos del tiempo que toma completar una tarea de producción de la sección 19.2. El archivo de datos MatchedSample proporciona los tiempos de producción para el método A en la columna C1, los tiempos de producción para el método B en la columna C2 y las diferencias en la columna C3. Los pasos siguientes permiten probar la hipótesis H_0: mediana $= 0$ y H_a: mediana $\neq 0$ para las diferencias poblacionales.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Nonparametrics**.

Paso 3. Elija **1-Sample Wilcoxon**.
Paso 4. Cuando el cuadro de diálogo 1-Sample Wilcoxon aparezca:
Introduzca C3 en el cuadro **Variables**.
Seleccione **Test Median**.
Introduzca el valor hipotético 0 en el cuadro **Test Median**.
Seleccione **not equal** del menú **Alternative**.
Haga clic en **OK**.

Observe que el procedimiento de Minitab utiliza las diferencias de los datos pareados de la columna C3. Aunque el archivo muestra los tiempos para cada método de producción en las columnas C1 y C2, estos datos no se utilizan para obtener el resultado de Minitab.

El mismo procedimiento se utiliza para probar una hipótesis acerca de la mediana de una población *simétrica*. Ingrese los datos reales en cualquier columna de la hoja de cálculo y siga los pasos anteriores. Introduzca el valor hipótetico de la mediana poblacional en el cuadro **Test Median** y seleccione la hipótesis alternativa deseada en el cuadro **Alternative**. Los resultados se obtienen haciendo clic en **OK**. Para realizar esta prueba no es necesario ingresar la diferencia de los datos. La rutina de Minitab hará los cálculos automáticamente. Recuerde que esta prueba es válida sólo para la mediana de una población simétrica.

Prueba de Mann-Whitney-Wilcoxon

WEB archivo
ThirdNational

Los siguientes pasos permiten probar la hipótesis de que dos poblaciones son idénticas con dos muestras independientes, una de cada población. Se utilizará el ejemplo del Third National Bank de la sección 19.3 para ilustrar el procedimiento. El archivo de datos ThirdNational provee los 12 saldos de la cuenta de la sucursal 1 en la columna C1, y los 10 saldos de la cuenta de la sucursal 2 en la columna C2. Los pasos siguientes aplicarán el procedimiento de Minitab para probar H_0: las dos poblaciones son idénticas, contra H_a: las dos poblaciones no son idénticas.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Nonparametrics**.
Paso 3. Elija **Mann-Whitney**.
Paso 4. Cuando el cuadro de diálogo Mann-Whitney aparezca:
Introduzca C1 en el cuadro **First sample**.
Introduzca C2 en el cuadro **Second sample**.
Seleccione **not equal** del menú **Alternative**.
Haga clic en **OK**.

Minitab informará el valor del estadístico de prueba y el valor-*p* correspondiente. Puesto que Minitab supone automáticamente que las dos poblaciones tienen la misma forma, la salida describe el resultado en términos de la diferencia entre las medianas de las dos poblaciones. Observe que el resultado también proporciona una estimación puntual y una estimación por intervalo de confianza de la diferencia entre las medianas. La letra griega η (*eta*) se utiliza con frecuencia para denotar la mediana de una población, por lo que los resultados de Minitab utilizan ETA1 y ETA2 como abreviaturas para las dos medianas poblacionales.

Prueba de Kruskal-Wallis

WEB archivo
Williams

Los pasos siguientes permiten probar una hipótesis acerca de que tres o más poblaciones son idénticas a partir de muestras independientes, una de cada población. Se utilizarán como ejemplo los datos de Williams Manufacturing Company de la sección 19.4. El archivo de datos Williams proporciona la universidad de la que proviene el empleado (A, B o C) en la columna C1 y la calificación anual de desempeño en la columna C2. La terminología de Minitab se referirá a la universidad como el factor y a la calificación de desempeño como la respuesta. Los pasos siguientes se aplicarán al procedimiento de Minitab para probar H_0: todas las poblaciones son idénticas, contra H_a: no todas las poblaciones son idénticas. Si es posible suponer que las poblaciones tienen la misma forma, la hipótesis se puede establecer en términos de las medianas de la población.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Nonparametrics**.
Paso 3. Elija **Kruskal-Wallis**.
Paso 4. Cuando el cuadro de diálogo Kruskal-Wallis aparezca:
Introduzca C2 en el cuadro **Response**.
Introduzca C1 en el cuadro **Factor**.
Haga clic en **OK**.

Correlación por rangos de Spearman

WEB archivo
PotentialActual

El coeficiente de correlación por rangos de Spearman es el mismo que el coeficiente de correlación de Pearson calculado para datos ordinales u ordenados por rango. Así que se puede obtener el coeficiente de correlación de Spearman con los datos ordenados por rango mediante el procedimiento de Minitab que permite calcular el coeficiente de correlación de Pearson. Se utilizarán los datos de las ventas potenciales y las ventas reales del periodo de dos años presentados en la sección 19.5. El archivo de datos PotentialActual proporciona la clasificación del potencial de cada empleado en la columna C2 y la clasificación de las ventas reales de un periodo de dos años por empleado en la columna C3. Los siguientes pasos de Minitab sirven para calcular la correlación por rangos de Spearman para las dos variables.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Basic Statistics**.
Paso 3. Elija **Correlation**.
Paso 4. Cuando el cuadro de diálogo Correlation aparezca:
Introduzca C2 C3 en el cuadro **Variables**.
Elimine **Display p-values**.
Haga clic en **OK**.

En el resultado de Minitab se obtiene un valor de 0.733 para el coeficiente de correlación de Pearson. Dado que los datos ya estaban ordenados por rango, éste es también el coeficiente de correlación por rangos de Spearman. Sin embargo, el valor-p para el coeficiente de Pearson no es el apropiado para los datos ordenados por rango y no debe interpretarse como el valor-p para el coeficiente de Spearman.

Apéndice 19.2 Métodos no paramétricos con Excel

Excel carece de procedimientos no paramétricos en su paquete de Data Analysis, pero se mostrará cómo utilizar la función BINOMDIST para realizar una prueba de los signos y cómo utilizar el procedimiento Data Analysis para calcular un coeficiente de correlación por rango. El complemento de SatTools, para Excel se puede utilizar para la prueba de signos de Wilcoxon y en la prueba de Mann-Whitney-Wilcoxon (vea el apéndice 19.3).

Prueba de signos

La prueba de signos utiliza una distribución de muestreo binomial con $p = 0.50$ para conducir una prueba de hipótesis acerca de una mediana poblacional o una prueba de hipótesis con muestras pareadas. La función BINOMDIST de Excel permite calcular probabilidades binomiales exactas para estas pruebas. Puesto que las probabilidades de BINOMDIST son exactas, no hay necesidad de calcular la aproximación a la distribución normal cuando se utiliza Excel para la prueba de signos.

Sea x = número de signos positivos
n = tamaño de muestra de las observaciones con un signo positivo o un signo negativo

La función BINOMDIST se puede utilizar de la siguiente manera:

$$\text{Probabilidad de cola inferior} = \text{BINOMDIST}(x, n, 0.50, \text{True})$$
$$\text{Probabilidad de cola superior} = 1 - \text{BINOMDIST}(x - 1, n, 0.50, \text{True})$$

Se puede observar que en la expresión de la probabilidad en la cola inferior, la función BINOMDIST proporciona la probabilidad binomial acumulada de *menor o igual que x*. El 0.50 en la función es el valor de $p = 0.50$ y el término True se utiliza para obtener la probabilidad binomial acumulada. La probabilidad en la cola superior es 1 − (la probabilidad acumulada). Observe que, dado que la distribución binomial es discreta, $(x - 1)$ se utiliza en el cálculo de probabilidades de cola superior. Por ejemplo, la de $P(x \geq 7) = 1 - P(x \leq 6)$.

Al utilizar Excel para la prueba de hipótesis del ejemplo de Lawler Grocery Store acerca de una mediana de la población, tenemos 7 signos positivos y tres signos negativos para la muestra de las 10 tiendas. El número de signos positivos se presentó en la cola superior con $P(x \geq 7)$ dada por la función

$$= 1 - \text{BINOMDIST}(x - 1, n, .50, \text{True}) = 1 - \text{BINOMDIST}(6, 10, .50, \text{True}) = 0.1719$$

Como es una prueba de hipótesis de dos colas, tenemos que el valor-$p = 2(0.1719) = 0.3438$.

En la sección 19.1 se considera también la prueba de cola inferior acerca de la mediana poblacional del precio de una casa nueva:

$$H_0\text{: mediana} \geq 236\,000$$
$$H_a\text{: mediana} < 236\,000$$

Después de eliminar la casa que se vendió por exactamente $236 000, la muestra proporciona 22 signos positivo y 38 signos negativos para la muestra de 60 viviendas. Al tratarse de una prueba de cola inferior, el valor-p está dado por la probabilidad de cola inferior $P(x \leq 22)$, que es la siguiente:

$$=\text{BINOMDIST}(x, n, .50, \text{True}) = \text{BINOMDIST}(22, 60, .50, \text{True}) = 0.0259$$

Al utilizar la función BINOMDIST tenemos la capacidad para calcular el valor-p exacto para cualquier aplicación de la prueba de signos.

Correlación por rangos de Spearman

WEB archivo
PotentialActual

Excel no tiene un procedimiento específico para calcular el coeficiente de correlación por rangos de Spearman. Sin embargo, éste es el mismo que el coeficiente de correlación de Pearson siempre y cuando se utilicen los datos ordenados por rangos. Como resultado, se puede calcular el coeficiente de correlación por rangos de Spearman aplicando el procedimiento del coeficiente de correlación de Pearson de Excel al ordenar los datos por rango. Esto se ilustra utilizando los datos de las ventas potenciales y reales de dos años de la sección 19.5 el archivo de datos PotentialActual proporciona la clasificación obtenida para 10 personas en términos de su potencial en la columna B y su calificación en términos de las ventas reales durante dos años en la columna C. Con los pasos siguientes se obtiene el coeficiente de correlación por rangos de Spearman.

Paso 1. Haga clic en la barra de herramientas **Data**.
Paso 2. En el grupo **Analysis** haga clic en **Data Analysis**.
Paso 3. Elija **Correlation** de la lista de **Analysis Tools**.

Paso 4. Cuando el cuadro de diálogo Correlation Dialog aparezca:
Introduzca B1:C1 en el cuadro **Input Range**.
Seleccione **Grouped by Columns**.
Elija **Labels in First Row**.
Seleccione **Output Range**.
Introduzca D1 en el cuadro **Output Range**.
Haga clic en **OK**.

El coeficiente de correlación por rangos de Spearman aparecerá en la celda E3.

Apéndice 19.3 Métodos no paramétricos con StatTools

En este apéndice se muestra cómo usar StatTools para la prueba de signos de Wilcoxon y la prueba de Mann-Whitney-Wilcoxon.

Prueba de rangos con signo de Wilcoxon para muestras pareadas

WEB archivo
MatchedSample

Los siguientes pasos permiten probar la hipótesis acerca de la diferencia entre dos medianas poblacionales con base en muestras pareadas. Para ilustrar el procedimiento se utilizarán los datos del tiempo para completar la tarea de producción de la sección 19.2. El archivo de datos MatchedSample proporciona los tiempos de producción para el método A en la columna A, los tiempos de producción para el método B en la columna B y las diferencias entre los dos métodos en la columna C. Comenzamos por usar el Data Set Manager para crear un conjunto de datos de StatTools con el procedimiento descrito en el apéndice del capítulo 1. Con los siguientes pasos se prueba la hipótesis H_0: mediana $= 0$ y H_a: mediana $\neq 0$ para la población de las diferencias.

Paso 1. Haga clic en la barra de herramientas **StatTools**.
Paso 2. En **Analyses Group**, seleccione **Nonparametric Tests**.
Paso 3. Elija **Wilcoxon Signed-Rank Test**.
Paso 4. Cuando el cuadro de diálogo de Wilcoxon Sign-Rank Test aparezca:
Seleccione **One-Sample Analysis** en el cuadro **Analysis Type**.
Verifique la variable **Difference**.
Introduzca 0 en el cuadro **Null Hypothesis Value**.
Seleccione **Not Equal to Null Value** en el cuadro **Alternative Hypothesis**.
Haga clic en **OK**.

El mismo procedimiento también se puede usar para probar una hipótesis sobre la mediana de una población simétrica. Introduzca los datos en cualquier columna de la hoja de cálculo. A continuación siga los pasos anteriores. Ingrese el valor hipotético de la mediana poblacional en el cuadro de diálogo **Null Hypothesis Value** y seleccione las hipótesis alternativas deseadas en el cuadro **Alternative Hypothesis**. Haga clic en **OK** para obtener los resultados. Para esta prueba no tenemos que introducir los datos de las diferencias, porque la rutina de StatTools efectuará los cálculos automáticamente. Pero recuerde, esta prueba es sólo válida para la mediana de una población simétrica.

Prueba de Man-Whitney-Wilcoxon

WEB archivo
ThirdNational

Los siguientes pasos pueden utilizarse para probar la hipótesis de que dos poblaciones son idénticas con dos muestras independientes, una para cada población. Se utilizará el ejemplo de Third National Bank manejado en la sección 19.3. El archivo de datos ThirdNational proporciona los 12 saldos de cuentas de la sucursal 1 en la columna A y los 10 saldos de cuentas de la sucursal 2 en la columna B. Comience con el Data Set Manager para crear una base de datos de StatToools con el procedimiento descrito en el apéndice del capítulo 1. Los siguientes pasos se

utilizan para probar la hipótesis H_0: las dos poblaciones son idénticas y H_a: las dos poblaciones no son idénticas.

Paso 1. Hag clic en la barra de herramientas de **StatTools**.
Paso 2. En **Analyses Group**, seleccione **Nonparametric Tests**.
Paso 3. Elija **Mann-Whitney Test**.
Paso 4. Cuando el cuadro de diálogo Mann-Whitney Test aparezca:
Seleccione **General Version** en el cuadro **Analysis Type**.
Verifique la variable **Branch 1**.
Verifique la variable **Branch 2**.
Seleccione **Either distribution smaller (Two-Tailed Test)** en el cuadro **Alternative Hypothesis**.
Haga clic en **OK**.
Paso 5. Cuando el cuadro de diálogo StatTools aparezca:
Haga clic en **OK**.
Cuando el cuadro de diálogo Choose Variable Ordering aparezca:
Haga clic en **OK**.

Si desea realizar el supuesto de que las dos poblaciones tienen la misma forma, seleccione **Median Version** en el cuadro **Analysis Type**. Los resultados de las pruebas serán los mismos, pero indicando las hipótesis acerca de las diferencia entre las dos medianas de la población.

CAPÍTULO 20

Métodos estadísticos para el control de calidad

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: DOW CHEMICAL COMPANY

20.1 FILOSOFÍAS Y MARCOS DE REFERENCIA
Malcolm Baldrige National Quality Award
ISO 9000
Six Sigma

20.2 CONTROL ESTADÍSTICO DE PROCESOS
Gráficas de control
Gráfica $\bar{x}$: media y desviación estándar del proceso conocidas
Gráfica $\bar{x}$: media y desviación estándar del proceso desconocidas
Gráfica *R*
Gráfica *p*
Gráfica *np*
Interpretación de las gráficas de control

20.3 MUESTREO DE ACEPTACIÓN
KALI, Inc.: un ejemplo de muestreo de aceptación
Cálculo de la probabilidad de aceptación de un lote
Selección de un plan de muestreo de aceptación
Planes de muestreo múltiple

ESTADÍSTICA *en* LA PRÁCTICA

DOW CHEMICAL COMPANY*
FREEPORT, TEXAS

En 1940, Dow Chemical Company compró un terreno de 800 acres en Texas, en la costa del Golfo, para construir una planta de producción de magnesio. La planta original ha crecido hasta cubrir más de 5 000 acres y engloba uno de los complejos petroquímicos más grandes del mundo. Entre los productos de Dow Texas Operations se encuentran el magnesio, estireno, plásticos, adhesivos, solventes, glicol y cloro. Algunos productos se fabrican únicamente para ser utilizados en otros procesos, pero muchos terminan como ingredientes esenciales de productos farmacéuticos, pastas dentales, alimentos para perros, mangueras, refrigeradores, envases de cartón para leche, bolsas para basura, shampoos y muebles.

Dow's Texas Operations produce más de 30% del magnesio en el mundo, el cual es un metal extremadamente ligero que se emplea en diversos productos, como raquetas de tenis y rines de "magnesio". El Departamento de Magnesio fue el primer grupo de Texas Operations que capacitó a su personal técnico y a sus gerentes para utilizar el control estadístico de la calidad. Algunas de las primeras aplicaciones fueron en los procesos químicos.

En una aplicación que implica el funcionamiento de un secador, se tomaban muestras del producto a intervalos periódicos, se calculaba el valor promedio de cada muestra y se registraba en una gráfica llamada *gráfica* $\bar{x}$. Estas gráficas permitían a los analistas de Dow monitorear las tendencias en los productos que pudieran indicar que el proceso no se estaba desarrollando correctamente. En una ocasión los analistas empezaron a observar que las medias muestrales presentaban valores que no correspondían a un proceso



El control estadístico de la calidad ha permitido a Dow Chemical mejorar sus métodos de producción y sus productos. © PR Newswire Dow Chemical USA/AP Images,

desarrollado dentro de los límites previstos. En un examen más cuidadoso de las gráficas de control y de la operación misma, los analistas encontraron que las variaciones podían deberse a algún problema relacionado con el operador. Después de capacitarlo nuevamente, las gráficas $\bar{x}$ indicaron una mejora significativa en la calidad del proceso.

En cualquier parte en que Dow aplica el control estadístico de la calidad se logra una mejora. Se han conseguido ahorros documentados de cientos de miles de dólares por año, y continuamente se han descubierto nuevas aplicaciones.

En este capítulo se mostrará cómo elaborar una gráfica $\bar{x}$ como las empleadas por Dow. Estas gráficas son parte del control estadístico de la calidad conocido como *control estadístico de procesos*. También se verán métodos de control de calidad que se utilizan en situaciones en que la decisión de aceptar o rechazar un conjunto de artículos se basa únicamente en una muestra.

*Los autores agradecen a Clifford B. Wilson, director técnico del Departamento de Magnesio de Dow Chemical Company, por aportar este artículo para *Estadística en la práctica*.

Visión de ASQ: "Al hacer de la calidad una prioridad mundial, un método fundamental y una ética personal, la American Society for Quality (ASQ) se convierte en una comunidad de todos que busca conceptos de calidad, tecnología y herramientas para mejorarse a sí misma y su mundo." (Página web ASQ).

La American Society for Quality (ASQ) define la calidad como "la totalidad de rasgos y características de un producto o un servicio relacionada con la capacidad de satisfacer determinada necesidad". En otras palabras, la calidad mide si un producto o un servicio cumple con las necesidades del cliente. Las organizaciones reconocen que para ser competitivos en la economía global actual deben esforzarse por lograr un alto nivel de calidad. Como resultado, ponen mayor énfasis en los métodos de monitoreo y mantenimiento de la calidad.

Hoy el enfoque orientado al cliente, que es fundamental para las organizaciones de alto desempeño, ha transformado el ámbito de aplicación de las cuestiones de calidad, desde la simple eliminación de los defectos en una línea de producción, hasta el desarrollo de estrategias de calidad corporativa con base amplia. La ampliación del alcance de la calidad, naturalmente, conduce al concepto de **calidad total (TQ)**.

La calidad total (*total quality*, TQ) es un sistema de gestión enfocado en las personas que aspiran a incrementar la satisfacción del cliente a un costo real más bajo. TQ es un sistema de enfoque

> total (no de un área separada o un programa de trabajo) y parte integral de una estrategia de alto nivel; funciona horizontalmente por medio de funciones y departamentos; involucra a todos los empleados, desde el más alto hasta el más bajo, y se extiende hacia atrás y hacia adelante, incluyendo la cadena de proveedores y la cadena de clientes. TQ hace hincapié en el aprendizaje y la adaptación al cambio continuo como claves para el éxito de una organización.[1]

Independientemente de cómo se aplica en diferentes organizaciones, la calidad total se basa en tres principios fundamentales: el enfoque en los clientes y grupos de interés; la participación y el trabajo en equipo en todas las organizaciones, y un enfoque en la mejora continua y el aprendizaje. En la primera sección del capítulo se presenta una breve introducción a las tres estructuras en el manejo de la calidad: el Malcolm Baldrige Quality Award, ISO 9000 y la filosofía Six Sigma. En las dos últimas secciones se presentan dos herramientas estadísticas para monitorear la calidad: el control estadístico de procesos y el muestreo de aceptación.

20.1 Filosofías y marcos de referencia

Después de la Segunda Guerra Mundial, el Dr. W. Edwards Deming se convirtió en consultor de la industria japonesa. Se le atribuye haber convencido a los altos directivos de utilizar el control estadístico de la calidad.

A comienzos del siglo XX las prácticas de control de calidad se limitaban a la inspección de productos terminados y a la eliminación de artículos defectuosos. Pero todo cambió como resultado de los primeros esfuerzos de un joven ingeniero llamado Walter A. Shewhart. Después de completar su doctorado en física en la Universidad de California en 1917, el Dr. Shewhart se unió a la Western Electric Company, trabajando en el Departamento de Ingeniería de Inspección. En 1924 preparó un memorando que incluía un conjunto de principios que son las bases para lo que hoy se conoce como *control del proceso*. Su memorando también contenía un gráfico que sería reconocido como *diagrama de control estadístico*. Continuó con su trabajo sobre la calidad en Bell Telephone Laboratories, hasta que se retiró en 1956. El Dr. Shewhart reunió las disciplinas de la estadística, la ingeniería y la economía y, al hacerlo, cambió el curso de la historia industrial. Es conocido como el padre del *control estadístico de la calidad* y fue el primer miembro honorario de la ASQ.

Otras dos personas que han tenido gran influencia en el concepto de la calidad son el Dr. W. Edwards Deming, un estudiante del Dr. Shewhart, y Joseph Juran. Estos científicos ayudaron a la educación de los japoneses en la gestión de la calidad poco después de la Segunda Guerra Mundial. Aunque la calidad es tarea de todos, Deming enfatizó que el enfoque debe ser dirigido por los gerentes. Desarrolló una lista de 14 puntos que a su juicio representan sus responsabilidades clave. Por ejemplo, declaró que los directivos debían dejar la dependencia de la inspección en masa; poner fin a la práctica de hacer negocios únicamente en función del precio; buscar la mejora continua en todos los procesos de producción y servicios; fomentar un entorno orientado al trabajo en equipo, y eliminar los objetivos, consignas y metas de trabajo que recomiendan cuotas numéricas. Tal vez lo más importante estriba en que los gerentes deben crear un ambiente de trabajo en el que se mantiene un compromiso con la calidad y la productividad en todo momento.

Juran propuso una definición sencilla de calidad: la *aptitud para el uso*. Su método se enfoca en tres procesos: planeación de la calidad, control de la calidad y mejora de la calidad. En contraste con la filosofía de Deming, que requería un cambio cultural importante en la organización, los programas de Juran fueron diseñados para mejorar la calidad al trabajar dentro del sistema organizacional actual. Sin embargo, las dos filosofías se parecen en que ambas se centran en la necesidad de que los altos directivos participen y hagan hincapié en la importancia de la mejora continua, la capacitación y el uso de técnicas de control de calidad.

Muchas otras personas jugaron un papel importante en el movimiento de la calidad, como Philip B. Crosby, A. V. Feigenbaum, Karou Ishikawa y Genichi Taguchi. Libros más especializados dedicados exclusivamente a la calidad proporcionan detalles de las contribuciones de cada uno. Las aportaciones de todos los involucrados en el movimiento de la calidad ayudaron a definir un conjunto de mejores prácticas y dieron lugar a numerosos premios y programas de

[1] J. R. Evans y W. M. Lindsay, *The management and control of quality*, 6a. ed., Cincinnati, OH., South-Western, 2005, pp. 18–19.

certificación. Los dos programas más significativos son el Malcolm Baldrige National Award, de Estados Unidos, y el proceso internacional de certificación ISO 9000. En los últimos años también se ha incrementado el uso de Six Sigma, una metodología para mejorar el desempeño de la organización que se basa en la recolección de datos y en el análisis estadístico.

El Malcolm Baldrige National Quality Award

El Instituto de Normas y Tecnología (NIST) del Departamento de Comercio de Estados Unidos administra el Baldrige National Quality Program. Se puede obtener más información en el sitio web del NIST.

El Malcolm Baldrige National Quality Award es entregado por el presidente de Estados Unidos a las organizaciones que apliquen y se les considere destacadas en siete áreas: liderazgo; planeación estratégica; enfoque en el cliente y el mercado; gestión de medición, análisis y administración del conocimiento; enfoque en los recursos humanos; gestión de procesos, y resultados de negocios. El Congreso estadounidense estableció el programa de premios en 1987 para reconocera las organizaciones por sus logros en calidad y desempeño, y para crear una mayor conciencia sobre la importancia de este factor como ventaja competitiva. El premio lleva por nombre *Malcolm Baldrige* en honor de quien se desempeñó como secretario de Comercio desde 1981 hasta su muerte en 1987.

2004 fue el último año para Baldrige Stock Study, debido al aumento en el número de empresas sin fines de lucro o de propiedad privada.

Desde la entrega de los primeros premios en 1988, el Baldrige National Quality Program (BNQP) ha crecido en magnitud e impacto. Desde ese año se han distribuido aproximadamente 2 millones de ejemplares de los criterios, así como un número significativo de reproducciones a gran escala realizadas por las organizaciones, además del acceso electrónico. Por octavo año consecutivo, el índice accionario hipotético formado por empresas estadounidenses que cotizan en bolsa y que han recibido el Baldrige Award, supera al Standard & Poor's 500. En 2003, el "Índice Baldrige" superó al S&P 500 por 4.4 a 1. En la ceremonia del Baldrige Award 2003, Bob Barnett, vicepresidente ejecutivo de Motorola, Inc., dijo: "Ingresamos a la entrega de premios no con la idea de ganar, sino con el objetivo de recibir la evaluación de los examinadores Baldrige. Esa evaluación fue integral, profesional e intuitiva, por lo que es quizás la consulta más rentable y con mayor valor agregado disponible en cualquier lugar en el mundo actual."

ISO 9000

Las normas del ISO 9000 se revisan periódicamente para mejorar el estándar de calidad.

El ISO 9000 es una serie de cinco normas internacionales publicadas en 1987 por la Organización Internacional de Estandarización (ISO) con sede en Ginebra, Suiza. Las empresas pueden utilizar las normas a efecto de determinar qué se necesita para mantener un sistema de calidad con un desempeño eficiente. Por ejemplo, las normas describen la necesidad de un sistema de calidad eficaz que garantice que los equipos de medición y prueba sean calibrados con regularidad, y mantener un sistema de registro adecuado. El registro ISO 9000 establece si una empresa cumple con su propio sistema de calidad. En general, esta certificación abarca menos de 10% de los criterios del Baldrige Award.

Six Sigma

A finales de la década de 1980 Motorola reconoció la necesidad de mejorar la calidad de sus productos y servicios. Su objetivo era lograr un nivel de calidad tan bueno que por cada millón de operaciones no se presentaran más de 3.4 defectos. A este nivel de calidad se le conoce como *nivel de calidad six sigma*, y a la metodología de investigación creada para este objetivo se le denomina **Six Sigma**.

Una organización puede comprender dos tipos de proyectos Six Sigma:

- DMAIC (definir, medir, analizar, mejorar y controlar) para ayudar a rediseñar los procesos existentes.
- DFSS (diseño para Six Sigma) para diseñar nuevos productos, procesos o servicios.

Para contribuir a rediseñar los procesos existentes, Six Sigma hace un fuerte énfasis en el análisis estadístico y en la medición cuidadosa. Hoy es una herramienta importante para ayudar a las organizaciones a alcanzar niveles Baldrige de desempeño de negocios y de la calidad en

los procesos. Muchos examinadores Baldridge ven a Six Sigma como el método ideal para la aplicación de programas de mejoramiento Baldridge.

Six Sigma: límites y defectos por millón de oportunidades En la terminología de Six Sigma, un defecto es cualquier error que le llega al cliente. El proceso define el desempeño de calidad como defectos por millón de operaciones (dpmo). Como se ha indicado antes, Six Sigma representa un nivel de calidad de por lo menos 3.4 dpmo. Para ilustrar cómo se mide este nivel, se considerará la situación de KJW Packaging.

KJW opera una línea de producción en la que se llenan cajas de cereales. El proceso de llenado tiene una media de $\mu = 16.05$ onzas y una desviación estándar de $\alpha = 0.10$ onzas. Además, se supone que el peso de llenado tiene una distribución normal. La distribución de estos pesos se muestra en la figura 20.1. Suponga que los gerentes consideran que de 15.45 hasta 16.65 onzas son los límites aceptables de calidad para el proceso de llenado. Por tanto, cualquier caja de cereal que contenga menos de 15.45 o más de 16.65 onzas será considerada con defecto. Al utilizar Excel o Minitab se puede demostrar que 99.9999998% de las cajas llenas tendrá entre $16.05 - 6(0.10) = 15.45$ y$16.05 + 6(0.10) = 16.65$ onzas. En otras palabras, sólo 0.0000002% de las cajas llenas contendrá menos de 15.45 o más de 16.65 onzas. Así que la probabilidad de obtener una caja de cereales defectuosa durante el proceso de llenado parece ser muy pequeña, porque en promedio sólo dos cajas en 10 millones estarán en esta condición.

Desde los primeros trabajos con Six Sigma Motorola se convenció de que la media en un proceso de cambio puede significar en promedio hasta 1.5 desviaciones estándar. Por ejemplo, suponga que la media del proceso de KJW aumenta 1.5 desviaciones estándar o $1.5(0.10) = 0.15$ onzas. Con este cambio, la distribución normal del peso de llenado estaría centrada en $\mu = 16.05 + 0.15 = 16.20$ onzas. Con una media del proceso de $\mu = 16.05$ onzas, la probabilidad de obtener una caja de cereal con más de 16.65 onzas es extremadamente pequeña. Pero ¿cómo cambiaría la probabilidad si la media del proceso se desplazara hasta $\mu = 16.20$ onzas? La figura 20.2 muestra que para este caso el límite de calidad superior de 16.65 es de 4.5 desviaciones estándar a la derecha de la nueva media $\mu = 16.20$onzas. Al utilizar esta media y Excel o Minitab, encontramos que la probabilidad de obtener una caja con más de 16.65 onzas es de 0.0000034. Por tanto, si la media del proceso se desplaza hacia arriba 1.5 desviaciones estándar, aproximadamente $1\,000\,000(0.0000034) = 3.4$ cajas de cereales superará el límite superior de 16.65 onzas. En la terminología de Six Sigma, se dice que el nivel de calidad del proceso es

FIGURA 20.1 Distribución normal de los pesos de llenado de las cajas de cereal con una media del proceso $\mu = 16.05$

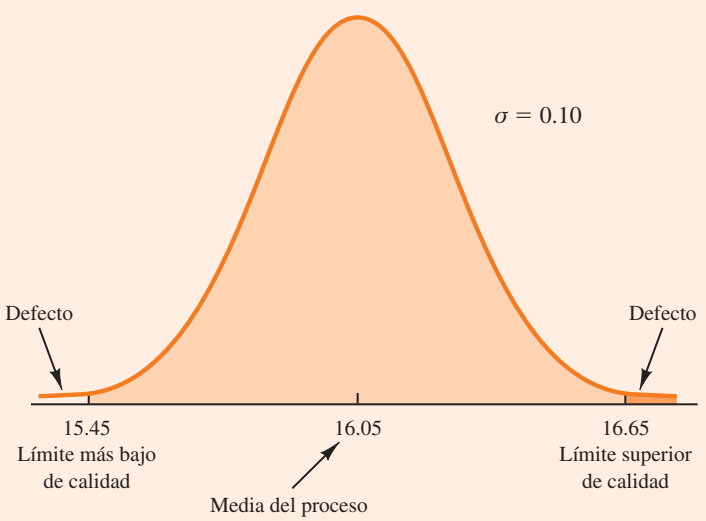
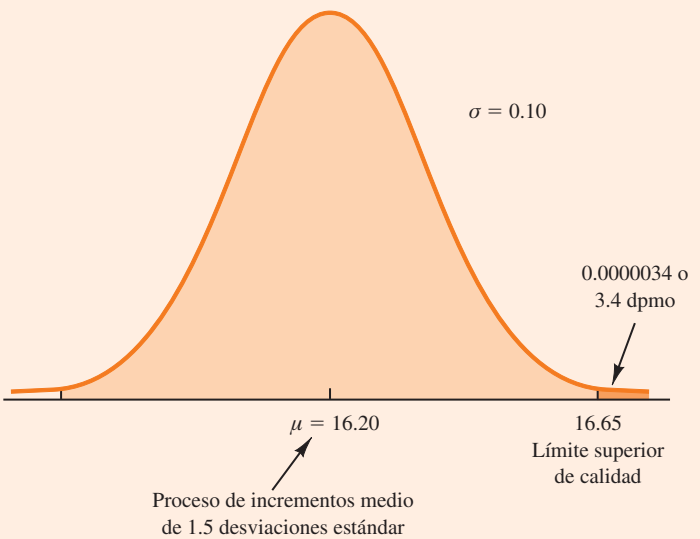


FIGURA 20.2 Distribución normal de los pesos de llenado de las cajas de cereales con una media del proceso $\mu = 16.20$



de 3.4 defectos por millón de oportunidades. Si los gerentes de KJE consideran que 15.4 hasta 16.65 onzas son límites aceptables de calidad para el proceso de llenado, el de KJW sería considerado un proceso Six Sigma. Así, si la media del proceso medio se mantiene dentro de 1.5 desviaciones estándar de su valor objetivo $\mu = 16.05$ onzas, se puede esperar un máximo de sólo 3.4 defectos por millón de cajas llenas.

Las organizaciones que desean alcanzar y mantener un nivel de de calidad Six Sigma deben hacer hincapié en los métodos de control y mantenimiento de la calidad. La *garantía de calidad* se refiere a todo el sistema de políticas, procedimientos y directrices establecido por una organización para lograr y mantener este indicador. La garantía de calidad consiste en dos funciones principales: ingeniería de calidad y control de calidad. El objeto de la *ingeniería de calidad* debe incluir la calidad en el diseño de productos y procesos, y para identificar problemas de calidad antes de la producción. El **control de calidad** consiste en una serie de inspecciones y acciones dirigidas a determinar si las normas de calidad se cumplen. Si no se cumplen, se aplican acciones correctivas o preventivas para lograr y mantener la conformidad. En las siguientes dos secciones se presentarán dos métodos estadísticos utilizados en el control de calidad. El primero, el *control estadístico de procesos*, utiliza representaciones gráficas conocidas como *gráficas de control* para monitorear un proceso; el objetivo es determinar si éste puede continuar o si se deben aplicar acciones correctivas para lograr el nivel de calidad deseado. El segundo método, el muestreo de aceptación, se utiliza en situaciones en que la decisión de aceptar o rechazar un grupo de elementos debe basarse en la calidad encontrada en una muestra.

20.2 Control estadístico de procesos

En esta sección se estudiarán los procedimientos de control de calidad dentro de un proceso de producción mediante el cual los artículos se fabrican de forma continua. Con base en un muestreo e inspección de la producción, se tomará la decisión ya sea de continuar el proceso de producción o de ajustarlo para que los artículos o bienes que se fabriquen logren estándares de calidad aceptables.

La mejora continua es uno de los conceptos más importantes del movimiento de gestión de la calidad total. El uso más importante de una gráfica de control se centra en la mejora del proceso.

A pesar de los altos estándares de calidad manejados en la operación de fabricación y producción, las herramientas del equipo invariablemente se desgastan, las vibraciones desajustan las máquinas, los materiales de compra están defectuosos y los operadores humanos cometen errores. Cualquiera o todos estos factores pueden dar lugar a una producción de mala calidad. Por fortuna, los procedimientos están disponibles para dar seguimiento a la producción, a efecto de que una pobre calidad sea detectada a tiempo y el proceso de producción se pueda ajustar o corregir.

Si la variación en la calidad de producción se debe a **causas imputables**, como el desgaste de las herramientas, una configuración incorrecta de máquinas, materias primas de baja calidad o un error del operador, el proceso debe ajustarse o corregirse tan pronto como sea posible. Sin embargo, si la variación se debe a lo que se llama **causas comunes**, es decir, variaciones que se presentan de forma aleatoria en los materiales, la temperatura, la humedad, etc., causas que no pueda controlar el fabricante, no es necesario ajustar el proceso. El objetivo principal del control estadístico de procesos es determinar si las variaciones de la producción se deben a causas imputables o a causas comunes.

Cuando las causas imputables son detectadas, se concluye que el *proceso está fuera de control.* En este caso, se aplicarán acciones correctivas para hacer que el proceso regrese a un nivel de calidad aceptable. Sin embargo, si la variación en el resultado de un proceso de producción se debe únicamente a causas comunes, se concluye que el proceso está bajo *control estadístico*, o simplemente *bajo control*, en cuyo caso no es necesario efectuar cambios o ajustes.

Los procedimientos de control de procesos están estrechamente relacionados con los procedimientos de prueba de hipótesis estudiados en este libro. Las gráficas de control permiten realizar sobre la marcha pruebas de hipótesis de que el proceso está bajo control.

Los procedimientos estadísticos para el control de procesos se basan en la metodología de pruebas de hipótesis presentadas en el capítulo 9. La hipótesis nula H_0 se formula en términos de un proceso de producción que está bajo control. La hipótesis alternativa H_a se plantea en términos de un proceso de producción que está fuera de control. En la tabla 20.1 se muestra cómo se toman las decisiones correctas para continuar un proceso bajo control, y el ajuste para un proceso fuera de control. Sin embargo, al igual que con otros procedimientos de prueba de hipótesis, también es posible cometer errores tipo 1 (ajuste de un proceso bajo control) y errores de tipo II (que permiten continuar con un proceso fuera de control).

Gráficas de control

Una **gráfica de control** proporciona una base para decidir si las variaciones en el producto se deben a causas comunes (bajo control) o a causas imputables (fuera de control). Siempre que se detecten situaciones de este último tipo, se realizarán los ajustes u otra acción correctiva que hagan recobrar el control del proceso.

Las gráficas de control basadas en datos que se pueden medir en una escala continua se llaman gráficas de control de variables. La gráfica $\bar{x}$ es de este tipo.

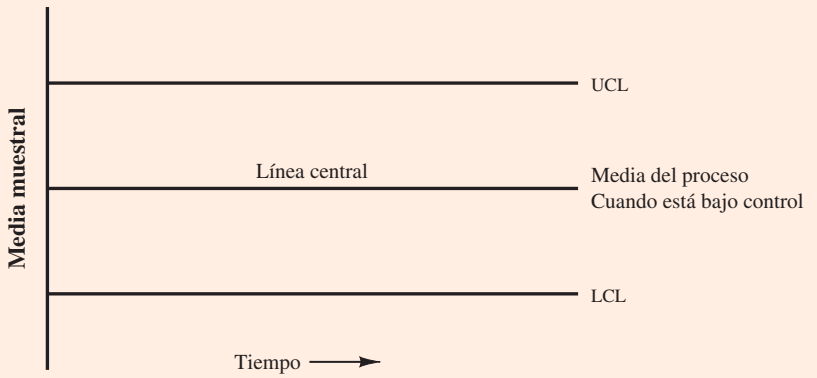
Las gráficas de control se clasifican por el tipo de datos que contienen. Una **gráfica $\bar{x}$** se utiliza cuando la calidad del producto de un proceso se mide en términos de una variable, como longitud, peso, temperatura, etc. En ese caso la decisión de continuar o de ajustar el proceso de producción se basará en el valor de la media encontrado en una muestra del producto. Para introducir algunos de los conceptos más comunes en todas las gráficas de control, se considerarán algunas características de una gráfica $\bar{x}$.

En la figura 20.3 se presenta la estructura general de una gráfica $\bar{x}$. La línea central de la tabla corresponde a la media del proceso cuando está *bajo control.* La línea vertical identifica

TABLA 20.1 Resultados del control estadístico de procesos

		Estado del proceso de producción	
		H_0 verdadera Proceso bajo control	**H_0 Falsa Proceso fuera de control**
Decisión	**Continuación del proceso**	Decisión correcta	Error tipo II (permite que continúe un proceso fuera de control)
	Ajuste del proceso	Error tipo I (ajuste de un proceso bajo control)	Decisión correcta

FIGURA 20.3 Estructura de una gráfica $\bar{x}$



la escala de medición para la variable de interés. Cada vez que se toma una muestra del proceso de producción, se calcula el valor de la media muestral $\bar{x}$ y se grafica el punto que indica el valor de $\bar{x}$ en la gráfica de control.

Las dos líneas etiquetadas como UCL y LCL son importantes para determinar si el proceso está bajo control o fuera de control. Las líneas se denominan *límite de control superior* y *límite de control inferior,* respectivamente. Estos límites se eligen de modo que cuando el proceso está bajo control, exista una probabilidad alta de que el valor de $\bar{x}$ esté dentro de estos dos límites. Los valores que están fuera proporcionan una fuerte evidencia estadística de que el proceso está fuera de control y se deberán aplicar acciones correctivas.

Con el tiempo, más y más puntos de datos se agregarán a la gráfica de control. El orden de los puntos será de izquierda a derecha como se realiza en la muestra del proceso. En esencia, cada vez que un punto se posiciona en una gráfica de control, se está llevando a cabo una prueba de hipótesis para determinar si el proceso está bajo control.

Además de la gráfica $\bar{x}$, se pueden utilizar otras gráficas de control para monitorear el rango de las mediciones en la muestra (**gráfica *R***), la proporción de defectos en la muestra (**gráfica *p***), y el número de artículos defectuosos en la muestra (**gráfica *np***). En cada caso, la gráfica tiene un LCL una línea central y un ULC similar a la gráfica $\bar{x}$ de la figura 20.3. La principal diferencia entre las gráficas radica en lo que el eje vertical mide; por ejemplo, en una gráfica p la escala de medición indica la proporción de artículos defectuosos en la muestra en lugar de la media muestral. En el estudio siguiente se ilustrará la construcción y uso de las gráficas $\bar{x}$, R, p y np.

Gráfica $\bar{x}$: media y desviación estándar del proceso conocidas

Para ilustrar la construcción de una gráfica $\bar{x}$, se reconsiderará la situación de la KJW Packaging. Recuerde que esta empresa opera una línea de producción en la que llena cajas de cereales. Cuando el proceso está funcionando correctamente, y por tanto el sistema está bajo control, el peso medio de llenado es $\mu = 16.05$ onzas, y la desviación estándar es $\sigma = 0.10$ onzas. Además, se supone que los pesos de llenado siguen una distribución normal. Ésta se muestra en la figura 20.4.

La distribución de muestreo de $\bar{x}$, como se presentó en el capítulo 7, se utiliza para determinar la variación que se puede esperar en $\bar{x}$ valores para un proceso que está bajo control. Realicemos primero una breve revisión de las propiedades de la distribución de muestreo de $\bar{x}$. En primer lugar, recuerde que el valor esperado o la media de $\bar{x}$ es igual a μ, la media del peso de llenado cuando la línea de producción está bajo control. Para muestras de tamaño n, la ecuación para la desviación estándar de $\bar{x}$, llamada *error estándar de la media*, es

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \textbf{(20.1)}$$

FIGURA 20.4 Distribución normal del peso de llenado de las cajas de cereales

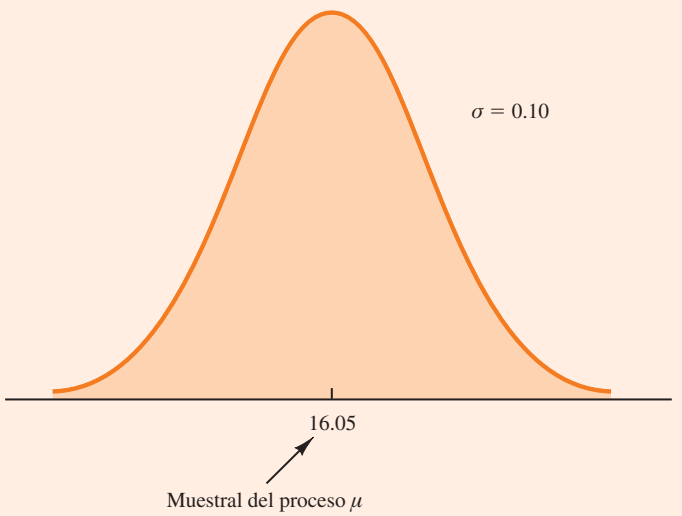
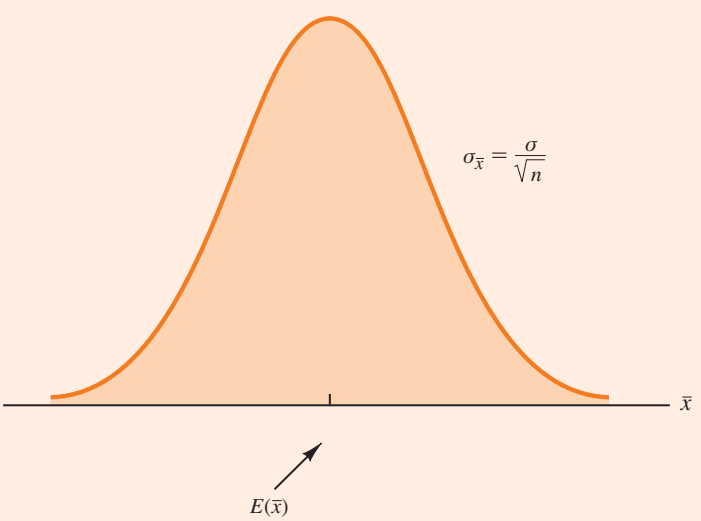


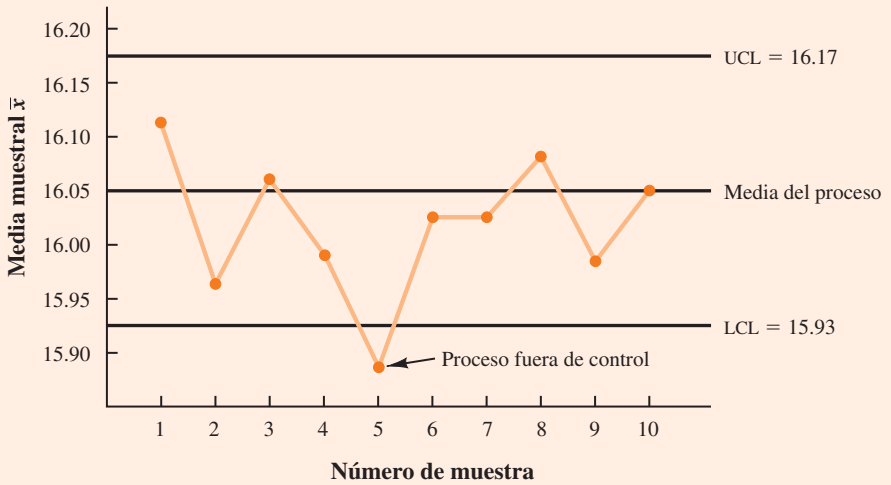
FIGURA 20.5 Distribución de muestreo de $\bar{x}$ para una muestra de n pesos de llenado



Además, debido a que los pesos de llenado tienen una distribución normal, la distribución de muestreo de $\bar{x}$ es normal para cualquier tamaño de muestra. Por tanto, es una distribución normal con media μ y desviación estándar $\sigma_{\bar{x}}$. Esta distribución se muestra en la figura 20.5.

La distribución de muestreo de $\bar{x}$ se utiliza para determinar qué valores de $\bar{x}$ son razonables cuando el proceso está bajo control. La práctica general en el control de calidad es definir qué tan razonable es cualquier valor de $\bar{x}$ que esté dentro de las 3 desviaciones estándar o errores estándar, arriba o abajo del valor de la media. Recuerde que al estudiar la distribución de probabilidad normal se vio que aproximadamente 99.7% de los valores de una variable aleatoria con distribución normal está dentro de ± 3 desviaciones estándar de su valor medio. Por tanto, si un valor de $\bar{x}$ se encuentra dentro del intervalo de $\mu - 3\sigma_{\bar{x}}$ a $\mu + 3\sigma_{\bar{x}}$, se admitirá que el proceso está bajo control. Los límites de control para una gráfica $\bar{x}$ se muestran a continuación.

FIGURA 20.6 Gráfica $\bar{x}$ para el proceso de llenado de las cajas de cereales



LÍMITES DE CONTROL PARA UNA GRÁFICA $\bar{x}$: MEDIA DEL PROCESO Y DESVIACIÓN ESTÁNDAR CONOCIDA

$$\text{UCL} = \mu + 3\sigma_{\bar{x}} \quad \textbf{(20.2)}$$

$$\text{LCL} = \mu - 3\sigma_{\bar{x}} \quad \textbf{(20.3)}$$

Retomando el ejemplo de KJW Packaging, el proceso de distribución de los pesos de llenado se muestra en la figura 20.4 y la distribución de muestreo de $\bar{x}$ en la figura 20.5. Suponga que un inspector de control de calidad toma periódicamente muestras de seis cajas de cartón y utiliza la media muestral del peso de llenado para determinar si el proceso está bajo control o fuera de control. Mediante la ecuación (20.1) se determina que el error estándar de la media es $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 0.10/\sqrt{6} = 0.04$. Por tanto, como la media del proceso es 16.05, los límites de control son UCL $= 16.05 + 3(0.04) = 16.17$ y LCL $= 16.05 - 3(0.04) = 15.93$. En la figura 20.6 se presenta la gráfica de control con los resultados de 10 muestras tomadas en un periodo de 10 horas. Para facilitar la lectura, las muestras número 1 hasta la 10 se listan bajo la gráfica.

Observe que la media de la quinta muestra de la figura 20.6 muestra que existe una fuerte evidencia de que el proceso está fuera de control. La quinta media muestral está debajo de LCL, lo que evidencia que existen causas imputables que ocasionan variación de la calidad del producto y que se está realizando un llenado insuficiente. Como resultado, la acción correctiva se aplica en este punto para que el proceso esté nuevamente bajo control. El hecho de que los demás puntos de la gráfica $\bar{x}$ estén dentro de los límites de control superior e inferior indica que la acción correctiva se ha realizado adecuadamente.

Gráfica $\bar{x}$: media y desviación estándar del proceso desconocidas

Con el ejemplo de KJW Packaging se ilustró cómo se desarrolla una gráfica $\bar{x}$ cuando la media y la desviación estándar del proceso son conocidas. En la mayoría de los casos el proceso de ambos indicadores se calcula utilizando muestras que se seleccionan cuando el proceso está bajo control. Por ejemplo, KJW podría seleccionar una muestra aleatoria de cinco cajas por la mañana y cinco cajas por la noche durante 10 días de operación bajo control. La media y la desviación

estándar de la muestra se calculan para cada subgrupo o muestra. El promedio total de ambas se utiliza para la elaboración de gráficas de control, tanto para la media como para la desviación estándar del proceso.

Es importante mantener el control tanto de la media como de la variabilidad de un proceso.

En la práctica, es más común monitorear la variabilidad del proceso mediante el uso del rango en lugar de la desviación estándar, ya que el rango es más fácil de calcular. Éste puede servir para obtener una buena estimación de la desviación estándar del proceso, por lo que se utiliza para construir los límites de control superior e inferior de la gráfica $\bar{x}$ con poco esfuerzo computacional. Para ilustrar esto, se tomará como ejemplo el caso de Jensen Computer Supplies, Inc.

Jensen Computer Supplies (JCS) fabrica discos para computadora de 3.5 pulgadas de diámetro; acaba de terminar el ajuste del proceso de su producción, de manera que éste opera bajo control. Suponga que fueron seleccionadas de cinco discos muestras aleatorias durante la primera hora de operación, cinco discos durante la segunda hora de operación, y así sucesivamente, hasta que se tienen 20 muestras. En la tabla 20.2 se presentan los diámetros de cada una de las muestra del disco, así como la media $\bar{x}_j$ y el rango R_j para cada una de las muestras (Sample Mean y Sample Range, respectivamente). También se listan el número de cada muestra (Sample Number) y las observaciones (Observations).

La estimación de la media del proceso μ está dada por la media muestral general.

MEDIA MUESTRAL GENERAL

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_k}{k} \qquad \textbf{(20.4)}$$

donde

$\bar{x}_j$ = media de la j-ésima muestra $j = 1, 2, \ldots, k$
k = número de muestras

TABLA 20.2 Datos del problema de Jensen Computer Supplies

WEB archivo
Jensen

Sample Number	Observations					Sample Mean $\bar{x}_j$	Sample Range R_j
1	3.5056	3.5086	3.5144	3.5009	3.5030	3.5065	0.0135
2	3.4882	3.5085	3.4884	3.5250	3.5031	3.5026	0.0368
3	3.4897	3.4898	3.4995	3.5130	3.4969	3.4978	0.0233
4	3.5153	3.5120	3.4989	3.4900	3.4837	3.5000	0.0316
5	3.5059	3.5113	3.5011	3.4773	3.4801	3.4951	0.0340
6	3.4977	3.4961	3.5050	3.5014	3.5060	3.5012	0.0099
7	3.4910	3.4913	3.4976	3.4831	3.5044	3.4935	0.0213
8	3.4991	3.4853	3.4830	3.5083	3.5094	3.4970	0.0264
9	3.5099	3.5162	3.5228	3.4958	3.5004	3.5090	0.0270
10	3.4880	3.5015	3.5094	3.5102	3.5146	3.5047	0.0266
11	3.4881	3.4887	3.5141	3.5175	3.4863	3.4989	0.0312
12	3.5043	3.4867	3.4946	3.5018	3.4784	3.4932	0.0259
13	3.5043	3.4769	3.4944	3.5014	3.4904	3.4935	0.0274
14	3.5004	3.5030	3.5082	3.5045	3.5234	3.5079	0.0230
15	3.4846	3.4938	3.5065	3.5089	3.5011	3.4990	0.0243
16	3.5145	3.4832	3.5188	3.4935	3.4989	3.5018	0.0356
17	3.5004	3.5042	3.4954	3.5020	3.4889	3.4982	0.0153
18	3.4959	3.4823	3.4964	3.5082	3.4871	3.4940	0.0259
19	3.4878	3.4864	3.4960	3.5070	3.4984	3.4951	0.0206
20	3.4969	3.5144	3.5053	3.4985	3.4885	3.5007	0.0259

La media muestral general de los datos de JCS presentados en la tabla 20.2 es $\bar{\bar{x}} = 3.4995$. Este valor será el eje central de la gráfica $\bar{x}$. El rango de cada muestra, que se denota R_j, es simplemente la diferencia entre el valor mayor y el valor menor de cada muestra. El rango promedio de las k muestras se calcula de la siguiente manera.

RANGO PROMEDIO

$$\bar{R} = \frac{R_1 + R_2 + \cdots + R_k}{k} \tag{20.5}$$

donde

R_j = rango de la j-ésima muestra $j = 1, 2, \ldots, k$
k = número de muestras

El rango promedio de los datos de JCS que se presenta en la tabla 20.2 es $\bar{R} = 0.0253$.

En la sección anterior se explicó que los límites de control superior e inferior de la gráfica $\bar{x}$ son

$$\bar{x} \pm 3\frac{\sigma}{\sqrt{n}} \tag{20.6}$$

Se utiliza la media muestral general $\bar{\bar{x}}$ para estimar μ, y los rangos muestrales para desarrollar una estimación de σ.

Por tanto, para obtener los límites de control de la gráfica $\bar{x}$, es necesario estimar μ y σ la media y la desviación estándar del proceso. La estimación de μ está dada por $\bar{\bar{x}}$. Una estimación de σ se obtiene utilizando los datos de los rangos.

Se puede demostrar que un estimador del proceso de desviación estándar σ es el promedio de los rangos divididos entre d_2, una constante que depende del tamaño de la muestra n. Es decir,

$$\text{Estimador de } \sigma = \frac{\bar{R}}{d_2} \tag{20.7}$$

En el *Manual on presentation of data and control chart analysis* de la *American Society for Testing and Materials* (ASTM) se obtienen los valores para d_2 mostrados en la tabla 20.3. Por ejemplo, cuando $n = 5$, $d_2 = 2.326$, y la estimación de σ es el rango promedio dividido entre 2.326. Si se sustituye $\bar{R}/d_2$ por σ en la fórmula (20.6), se pueden expresar los límites de control para la gráfica $\bar{x}$ como

$$\bar{\bar{x}} \pm 3\frac{\bar{R}/d_2}{\sqrt{n}} = \bar{\bar{x}} \pm \frac{3}{d_2\sqrt{n}}\bar{R} = \bar{\bar{x}} \pm A_2\bar{R} \tag{20.8}$$

Observe que $A_2 = 3/(d_2\sqrt{n})$ es una constante que sólo depende del tamaño de la muestra. Los valores de A_2 se obtienen de la tabla 20.3. Para $n = 5$, $A_2 = 0.577$, por lo que los límites de la gráfica de control $\bar{x}$ son

$$3.4995 \pm (0.577)(0.0253) = 3.4995 \pm 0.0146$$

Por tanto, UCL = 3.514 y LCL = 3.485.

TABLA 20.3 Factores para $\bar{x}$ y para las gráficas de control R

Observaciones en la muestra, n	d_2	A_2	d_3	D_3	D_4
2	1.128	1.880	0.853	0	3.267
3	1.693	1.023	0.888	0	2.574
4	2.059	0.729	0.880	0	2.282
5	2.326	0.577	0.864	0	2.114
6	2.534	0.483	0.848	0	2.004
7	2.704	0.419	0.833	0.076	1.924
8	2.847	0.373	0.820	0.136	1.864
9	2.970	0.337	0.808	0.184	1.816
10	3.078	0.308	0.797	0.223	1.777
11	3.173	0.285	0.787	0.256	1.744
12	3.258	0.266	0.778	0.283	1.717
13	3.336	0.249	0.770	0.307	1.693
14	3.407	0.235	0.763	0.328	1.672
15	3.472	0.223	0.756	0.347	1.653
16	3.532	0.212	0.750	0.363	1.637
17	3.588	0.203	0.744	0.378	1.622
18	3.640	0.194	0.739	0.391	1.608
19	3.689	0.187	0.734	0.403	1.597
20	3.735	0.180	0.729	0.415	1.585
21	3.778	0.173	0.724	0.425	1.575
22	3.819	0.167	0.720	0.434	1.566
23	3.858	0.162	0.716	0.443	1.557
24	3.895	0.157	0.712	0.451	1.548
25	3.931	0.153	0.708	0.459	1.541

Fuente. Adaptación de la tabla 27 de ASTM STP 15D, *ASTM Manual on presentation of data and control chart analysis*. Copyright 1976, American Society for Testing and Materials, Philadelphia, PA. Impreso con autorización.

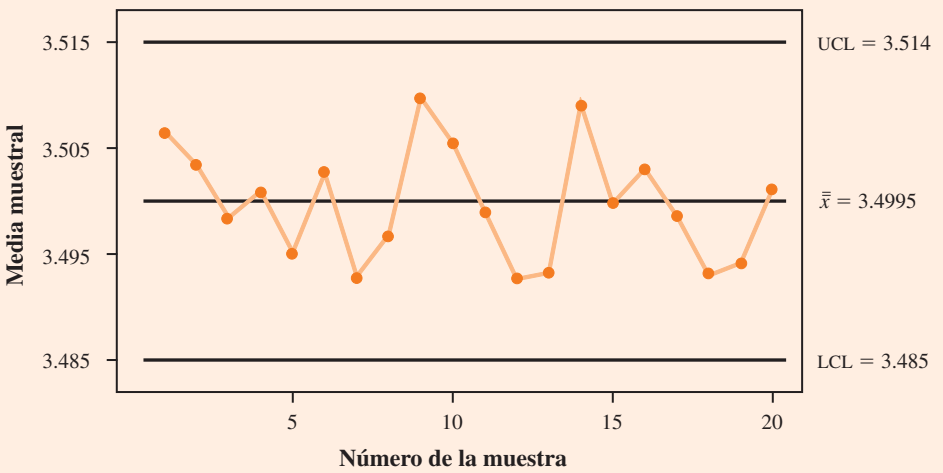
En la figura 20.7 se presenta la gráfica $\bar{x}$ para el problema de Jensen Computer Supplies. Se utilizaron los datos de la tabla 20.2 y la rutina de Minitab para obtenerla. La línea central se observa en la muestra media general $\bar{\bar{x}} = 3.4995$. El límite de control superior (UCL) es 3.514 y el de control inferior (LCL) es 3.485. La gráfica $\bar{x}$ muestra las 20 medias muestrales graficadas en el tiempo. Debido a que estás se encuentran dentro de los límites de control, confirman que la media del proceso ha estado bajo control durante el periodo de muestreo.

Gráfica *R*

Ahora se estudiarán las gráficas de rango (gráfica R) que se utilizan para controlar la variabilidad de un proceso. Para desarrollar una gráfica R se necesita considerar el rango de una muestra como una variable aleatoria con su media y desviación estándar propias. El rango promedio $\bar{R}$ proporciona una estimación de la media de esta variable aleatoria. Por otra parte, se puede demostrar que una estimación de la desviación estándar del rango es

$$\hat{\sigma}_R = d_3 \frac{\bar{R}}{d_2} \tag{20.9}$$

FIGURA 20.7 Gráfica $\bar{x}$ para el problema de Jensen Computer Supplies



donde d_2 y d_3 son constantes que dependen del tamaño de la muestra. Los valores de d_2 y d_3 se obtienen de la tabla 20.3. Por tanto, el UCL para la gráfica R está dado por

$$\bar{R} + 3\hat{\sigma}_R = \bar{R}\left(1 + 3\frac{d_3}{d_2}\right) \tag{20.10}$$

y el LCL es

$$\bar{R} - 3\hat{\sigma}_R = \bar{R}\left(1 - 3\frac{d_3}{d_2}\right) \tag{20.11}$$

Si se deja

$$D_4 = 1 + 3\frac{d_3}{d_2} \tag{20.12}$$

$$D_4 = 1 - 3\frac{d_3}{d_2} \tag{20.13}$$

Se pueden expresar los límites de control para la gráfica R como

$$\text{UCL} = \bar{R}\text{D}_4 \tag{20.14}$$

$$\text{LCL} = \bar{R}\text{D}_3 \tag{20.15}$$

Los valores de D_3 y D_4 también se obtienen de la tabla 20.3. Observe que $n = 5$, $D_3 = 0$ y $D_4 = 2.114$. Por tanto, con $\bar{R} = 0.0253$, los límites de control son

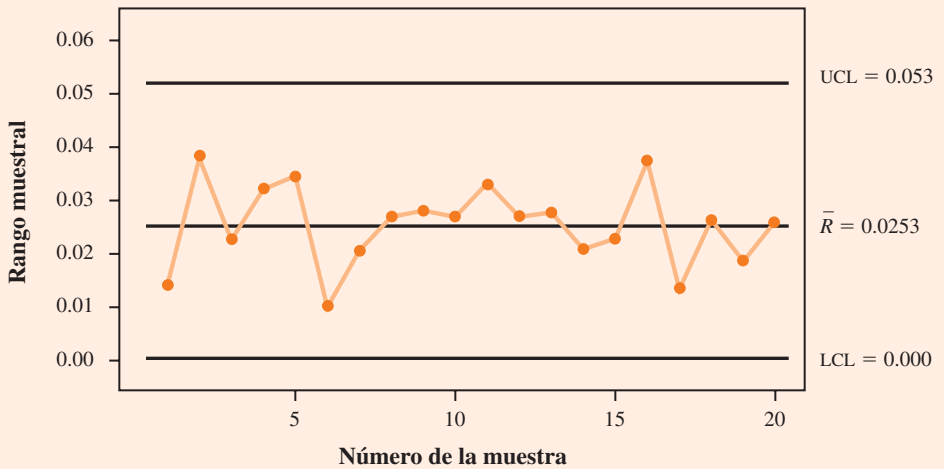
$$\text{UCL} = 0.0253(2.114) = 0.053$$

$$\text{LCL} = 0.0253(0) = 0$$

Si la gráfica R indica que el proceso está fuera de control, la gráfica $\bar{x}$ no debe interpretarse hasta que la gráfica R indique que la variabilidad del proceso está controlada.

La figura 20.8 muestra la gráfica R para el problema de Jensen Computer Supplies. Se utilizaron los datos de la tabla 20.2 y la gráfica de control de rutina de Minitab para obtenerla. La línea

FIGURA 20.8 Carta *R* para el problema de Jensen Computer Supplies



central se observa en la media general de los 20 rangos de la muestra, $\overline{R} = 0.0253$. El UCL es 0.053 y el LCL es 0.000. La gráfica *R* presenta los 20 rangos de la muestra trazada en el tiempo. Debido a que los 20 rangos están dentro de los límites de control, se confirma que la variabilidad del proceso estuvo bajo control durante el periodo de muestreo.

Gráfica *p*

A las gráficas de control basadas en datos que indican la presencia de un defecto o un número de defectos se les denomina gráficas de control de atributos. *Una gráfica p entra en esta clasificación.*

Se considera el caso en que la calidad del producto se mide tanto por los artículos no defectuosos como por los defectuosos. La decisión de continuar o de ajustar el proceso de producción se basará en $\overline{p}$, la proporción de artículos defectuosos que se encuentra en una muestra. A la gráfica de control utilizada para conocer la proporción de los datos defectuosos se le llama gráfica *p*.

Para ilustrar su elaboración, considere el uso de máquinas automáticas para la clasificación de las cartas que se utilizan en una oficina de correos. Estas máquinas escanean los códigos postales de las cartas y asignan cada una a la ruta de entrega adecuada. Aun cuando una máquina esté funcionando correctamente, algunas misivas no son asignadas de forma adecuada. Suponga que cuando una máquina está funcionando en forma apropiada, o bajo control, 3% de las cartas no es asignado correctamente. Así que *p*, la proporción de cartas no asignadas correctamente cuando el proceso está bajo control, es 0.03.

La distribución de muestreo de $\overline{p}$, como se presenta en el capítulo 7, se utiliza para determinar la variación que puede esperarse en los valores $\overline{p}$ para un proceso que está bajo control. Recuerde que el valor esperado de la media de $\overline{p}$ es *p*, la proporción de unidades defectuosas cuando el proceso está bajo control. Con muestras de tamaño *n*, la fórmula para la desviación estándar de $\overline{p}$, llamado *error estándar de la proporción*, es

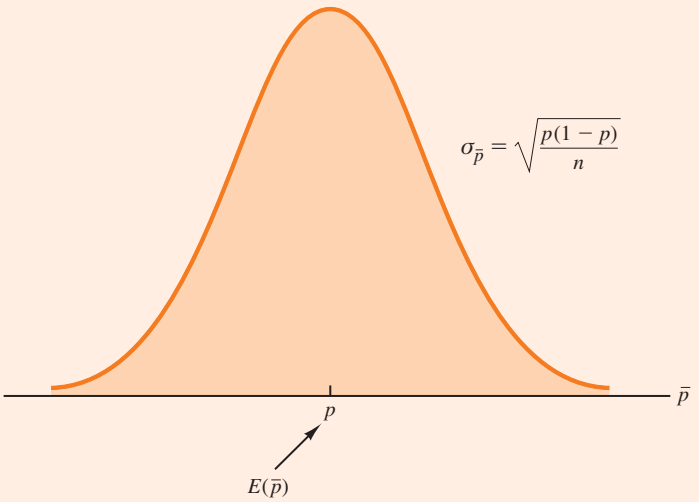
$$\sigma_{\overline{p}} = \sqrt{\frac{p(1-p)}{n}} \qquad \textbf{(20.16)}$$

Se estudió también en el capítulo 7 que la distribución de muestreo de $\overline{p}$ se puede aproximar por una distribución normal siempre que el tamaño de la muestra sea grande. Con $\overline{p}$ el tamaño de la muestra puede considerarse grande toda vez que las siguientes dos condiciones sean satisfechas.

$$np \geq 5$$

$$n(1-p) \geq 5$$

FIGURA 20.9 Distribución de muestreo de $\bar{p}$



En resumen, siempre que el tamaño de la muestra sea grande, la distribución de muestreo de $\bar{p}$ se puede aproximar mediante una distribución normal con la media p y la desviación estándar $\sigma_{\bar{p}}$. Esta distribución se presenta en la figura 20.9.

Para establecer los límites de control de una gráfica p se sigue el mismo procedimiento que se utilizó en el caso de una gráfica $\bar{x}$. Es decir, los límites para la gráfica de control se establecen a 3 desviaciones estándar, o errores estándar, por arriba y por abajo de la proporción de unidades defectuosas cuando el proceso está bajo control. Por tanto, se tienen los siguientes límites de control.

LÍMITES DE CONTROL PARA UNA GRÁFICA p

$$\text{UCL} = p + 3\sigma_{\bar{p}} \qquad \textbf{(20.17)}$$

$$\text{LCL} = p - 3\sigma_{\bar{p}} \qquad \textbf{(20.18)}$$

Con $p = 0.03$ y la muestra de tamaño $n = 200$, la ecuación (20.16) indica que el error estándar es

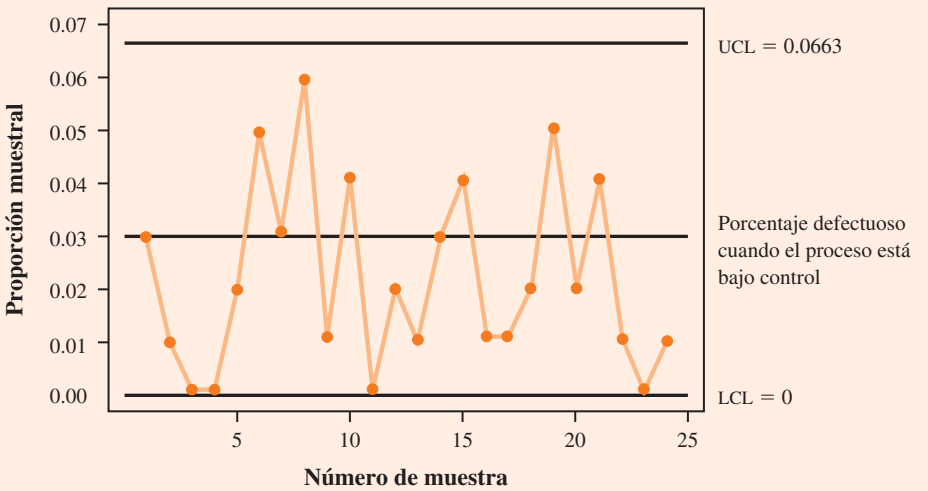
$$\sigma_{\bar{p}} = \sqrt{\frac{0.03(1 - 0.03)}{200}}$$

Por tanto, los límites de control son UCL $= 0.03 + 3(0.0121) = 0.0663$, y LCL $= 0.03 - 3(0.0121) = -0.0063$. Siempre que la ecuación (20.18) proporcione un valor negativo para LCL, éste es igual a cero en la gráfica de control.

La figura 20.10 ilustra la gráfica de control para el proceso de clasificación del correo. Los puntos graficados indican la proporción muestral defectuosa encontrada en la muestra de cartas tomadas del proceso. Todos los puntos están dentro de los límites de control, lo que indica que no existe evidencia alguna para concluir que el proceso de selección está fuera de control.

Si no se conoce la proporción de artículos defectuosos de un proceso bajo control, ese valor se puede estimar inicialmente utilizando datos de la muestra. Suponga, por ejemplo, que se toman k muestras diferentes, cada una de tamaño n, seleccionadas de un proceso que está bajo control. Se determina la fracción o proporción de artículos defectuosos en cada muestra.

FIGURA 20.10 Gráfica *p* para la proporción de defectos en el proceso de clasificación de las cartas



El tratamiento como una sola muestra grande de todos los datos obtenidos permite calcular la proporción de artículos defectuosos en todos los datos; ese valor puede utilizarse para estimar p, la proporción de artículos defectuosos observada cuando el proceso está bajo control. Note que esta estimación de p también permite calcular el error estándar de la proporción y determinar los límites de control superior e inferior.

Gráfica *np*

Una gráfica *np* no es una gráfica de control construida por el número de artículos defectuosos en una muestra. En este caso, n es el tamaño de la muestra y p es la probabilidad de observar un artículo defectuoso, cuando el proceso está bajo control. Siempre que el tamaño de la muestra sea grande, es decir, cuando $np \geq 5$ y $n(1 - p) \geq 5$, la distribución del número de artículos defectuosos observados en una muestra de tamaño n se puede aproximar por una distribución normal con media np y desviación estándar $\sqrt{np(1 - p)}$. Por tanto, en el ejemplo de clasificación de correspondencia, con $n = 200$ y $p = 0.03$, el número de artículos defectuosos observados en una muestra de 200 cartas se puede aproximar mediante una distribución normal con una media de $200(0.03) = 6$ y una desviación estándar de $\sqrt{200(0.03)(0.97)} = 2.4125$

Los límites de control para una gráfica *np* se fijan en tres desviaciones estándar arriba y abajo del número de artículos defectuosos esperado, cuando el proceso está bajo control. Por consiguiente, se tiene que los límites de control son los siguientes:

LÍMITES DE CONTROL EN UNA GRÁFICA *np*

$$\text{UCL} = np + 3\sqrt{np(1 - p)} \qquad \textbf{(20.19)}$$

$$\text{LCL} = np - 3\sqrt{np(1 - p)} \qquad \textbf{(20.20)}$$

En el ejemplo del proceso de clasificación de un correo, con $p = 0.03$ y $n = 200$, los límites de control son $\text{UCL} = 6 + 3(2.4125) = 13.2375$ y $\text{LCL} = 6 - 3(2.4125) = -1.2375$. Cuando LCL es negativo, LCL es igual a cero en la carta de control. Por tanto, si el número de cartas que no se asigna a las rutas correctas es mayor de 13, se concluye que el proceso está fuera de control.

La información que proporciona una gráfica *np* es equivalente a la proporcionada por la gráfica *p*, la única diferencia radica en que la gráfica *np* es la gráfica del *número* de artículos

defectuosos observados, mientras que la gráfica p es sobre la *proporción* de artículos defectuosos observados. Así, si se llega a la conclusión de que el proceso en particular está fuera de control sobre la base de la gráfica p, se concluye lo mismo sobre la base de una gráfica np.

Interpretación de las gráficas de control

La ubicación y el patrón de puntos en una gráfica de control permiten determinar, con una pequeña probabilidad de error, si un proceso está estadísticamente bajo control. La principal indicación de que un proceso está fuera de control es que un punto de los datos se encuentre fuera de los límites de control, como el punto 5 en la figura 20.6. Determinar ese punto es evidencia estadística de que el proceso está fuera de control, en cuyo caso la acción correctiva debe ser aplicada tan pronto como sea posible.

Además de los puntos que están fuera de los límites de control, existen ciertos patrones de los puntos dentro de tales límites que pueden representar señales de advertencia de problemas de control de calidad. Por ejemplo, suponga que todos los puntos de los datos están dentro de los límites de control, pero que un gran número está al lado de la línea central. El patrón puede indicar que existe un problema de equipo, que hay un cambio en los materiales, o que se ha presentado alguna otra causa imputable que ha ocasionado un cambio en la calidad. Se debe llevar a cabo una investigación cuidadosa en el proceso de producción para determinar si la calidad se ha modificado.

Aun cuando todos los puntos estén dentro de los límites de control superior e inferior, el proceso puede no estar bajo control. Las tendencias en los puntos de datos muestrales o secuencias inusualmente largas que pasen por encima o por debajo de la línea central también pueden indicar una situación fuera de control.

Otro patrón a observar en una gráfica de control es un cambio gradual, o tendencia, en el tiempo. Por ejemplo, por el desgaste de las herramientas, las dimensiones en la fabricación de una pieza poco a poco se apartan de sus niveles de diseño. Los cambios graduales en la temperatura o la humedad, el deterioro del equipamiento en general, la acumulación de suciedad o la fatiga del operador también pueden resultar en un patrón de tendencia en las gráficas de control. Seis o siete puntos en una fila que indiquen ya sea una tendencia de aumento o de disminución debe ser motivo de preocupación aunque los puntos de datos estén dentro de los límites de control. Cuando este patrón se presenta, el proceso deberá revisarse para evitar posibles modificaciones o cambios en la calidad. Será necesario aplicar las accioness correctivas para que el proceso esté de nuevo bajo control.

NOTAS Y COMENTARIOS

1. Dado que los límites de control para una gráfica $\bar{x}$ dependen del valor del rango promedio, estos límites no tienen mucho significado, a menos que la variabilidad del proceso esté bajo control. En la práctica, la gráfica R se elabora por lo general antes que la carta $\bar{x}$; si la primera indica que la variabilidad del proceso está bajo control, entonces se elabora la gráfica $\bar{x}$. Con la opción de Minitab Xbar-R se obtienen la gráfica $\bar{x}$ y la gráfica R al mismo tiempo. En el apéndice 20.1 se describen los pasos para este procedimiento.
2. La gráfica np se utiliza para monitorear un proceso en términos del número de defectos. El nivel de calidad Six Sigma de Motorola establece como objetivo producir no más de 3.4 defectos por millón de operaciones. Este objetivo implica $p = 0.0000034$.

Ejercicios

Métodos

1. En un proceso que está bajo control se tiene una media de $\mu = 12.5$ y una desviación estándar de $\sigma = 0.8$.
   a) Elabore una gráfica de control $\bar{x}$ para este proceso si se utilizaran muestras de tamaño 4.
   b) Repita el inciso *a*) para muestras de tamaño 8 y 16.
   c) ¿Qué sucede con los límites de la gráfica de control a medida que el tamaño de la muestra aumenta? Explique por qué es razonable.
2. Veinticinco muestras, cada una de tamaño 5, fueron seleccionadas de un proceso que está bajo control. La suma de todos los datos obtenidos fue de 677.5 libras.
   a) ¿Cuál es la estimación de la media del proceso (en términos de libras por unidad) cuando el proceso está bajo control?
   b) Elabore una gráfica de control $\bar{x}$ para este proceso si las muestras a utilizar son de tamaño 5. Suponga que la desviación estándar es 0.5 cuando el proceso está bajo control, y que la media del proceso es la estimación que se obtuvo en el inciso *a*).

3. De 25 muestras de 100 artículos, cada uno inspeccionado cuando el proceso se considera que funciona satisfactoriamente. En las 25 muestras se determinó que un total de 135 artículos estaba defectuoso.
   a) ¿Cuál es la estimación de la proporción de defectos cuando el proceso está bajo control?
   b) ¿Cuál sería el error estándar de la proporción si muestras de tamaño 100 se utilizaran para el control estadístico de procesos?
   c) Calcule los límites de control superior e inferior de la gráfica de control.

AUTOevaluación

4. En un proceso del que se toman 20 muestras, cada una de tamaño 8, resultan $\bar{\bar{x}} = 28.5$ y $\bar{R} = 1.6$. Calcule los límites de control superior e inferior para las gráficas $\bar{x}$ y R de este proceso.

Aplicaciones

5. La temperatura se utiliza para medir el resultado de un proceso de producción. Cuando éste está bajo control, la media del proceso es $\mu = 128.5$ y la desviación estándar es $\sigma = 0.4$.
   a) Elabore una gráfica $\bar{x}$ para este proceso si las muestras que se utilizarán son de tamaño 6.
   b) ¿El proceso está bajo control para una muestra de la que se obtienen los siguientes datos?

   128.8 128.2 129.1 128.7 128.4 129.2

   c) ¿El proceso está bajo control para una muestra de la que se obtienen los siguientes datos?

   129.3 128.7 128.6 129.2 129.5 129.0

6. En un proceso de control de calidad se supervisa el peso por paquete de un detergente para lavandería. Los límites de control se establecen en UCL = 20.12 onzas y LCL = 19.90 onzas. Se utilizan muestras de tamaño 5 en el proceso de muestreo e inspección. ¿Cuáles son la media del proceso y el proceso de desviación estándar para la operación de manufactura?

7. Goodman Tire and Rubber Company realiza pruebas periódicas de desgaste (Tread Wear) de los neumáticos en la banda de rodamiento bajo condiciones de una carretera simulada. Para estudiar y controlar el proceso de manufactura, 20 muestras (Sample), cada una con tres neumáticos radiales, fueron seleccionadas en diferentes turnos durante varios días de operación, con lo que se obtuvieron los siguientes resultados. Suponiendo que estos datos se recabaron cuando el proceso de manufactura operaba bajo control, desarrolle las gráficas de R y $\bar{x}$.

WEB archivo

Tires

Sample	Tread Wear*		
1	31	42	28
2	26	18	35
3	25	30	34
4	17	25	21
5	38	29	35
6	41	42	36
7	21	17	29
8	32	26	28
9	41	34	33
10	29	17	30
11	26	31	40
12	23	19	25
13	17	24	32
14	43	35	17
15	18	25	29
16	30	42	31
17	28	36	32
18	40	29	31
19	18	29	28
20	22	34	26

*Centésimas de pulgada

8. A lo largo de varias semanas de operación normal, o bajo control, 20 muestras, cada una de 150 paquetes de cuerdas sintéticas para raquetas de tenis, fueron analizadas para probar su resistencia a la rotura. Como resultado, un total de 141 paquetes de 3 000 que se probaron no se ajustan a las especificaciones del fabricante.
   a) ¿Cuál es la estimación de la proporción de defectos durante el proceso cuando el sistema está bajo control?
   b) Calcule los límites de control superior e inferior para una gráfica *p*.
   c) Con los resultados del inciso *b*), ¿a qué conclusión se debe llegar acerca del proceso si en las pruebas con una nueva muestra de 150 paquetes se encuentra que 12 están defectuosos? ¿Parece haber causas imputables en esta situación?
   d) Calcule los límites de control superior e inferior para una gráfica *np*.
   e) Responda el inciso *c*) utilizando los resultados del inciso *d*).
   f) ¿Qué gráfica de control sería preferible en esta situación? Explique.
9. Un proveedor de la industria automotriz produce pistones para varios modelos de automóviles. Se analizaron 20 muestras, cada una compuesta de 200 pistones, con conocimiento de que el proceso estaba funcionando correctamente. A continuación se presentan los números de pistones defectuosos que se hallaron en la muestra.

8	10	6	4	5	7	8	12	8	15
14	10	10	7	5	8	6	10	4	8

   a) ¿Cuál es la estimación de la proporción de unidades defectuosas cuando el proceso de fabricación del pistón está bajo control?
   b) Elabore una gráfica *p* para el proceso de fabricación asumiendo que cada muestra tiene 200 pistones.
   c) Con los resultados del inciso *b*), ¿a qué conclusión se debe llegar si una muestra de 200 pistones tiene 20 defectuosos?
   d) Calcule los límites de control superior e inferior para una gráfica *np*.
   e) Responda el inciso *c*) con los resultados del inciso *b*).

20.3 Muestreo de aceptación

En el muestreo de aceptación los elementos de interés pueden ser los embarques entrantes de materias primas, o partes compradas, así como productos terminados de montaje final. Suponga que se desea decidir si se acepta o rechaza un conjunto de artículos sobre la base de las características de calidad especificadas. En la terminología de control de calidad, el conjunto de artículos es un **lote**, y el **muestreo de aceptación** es un método estadístico que permite basar la decisión de aceptación o de rechazo en la inspección de una muestra de los elementos del lote.

En la figura 20.11 se ilustran los pasos generales que se siguen en el muestreo de aceptación. Después de recibir un lote se selecciona una muestra de artículos para su inspección. Los resultados de la inspección se comparan con las características de calidad especificadas. Si éstas se cumplen, el lote se acepta y se envía a producción o a los clientes. Si es rechazado, los gerentes tendrán que decidir su destino. En algunos casos la decisión puede ser quedarse con el lote y eliminar los artículos no aceptados o con los que se están inconforme. En otros casos, el lote podrá ser devuelto al proveedor a cuenta de éste; el trabajo extra y los costos cargados al proveedor pueden motivarlo buscar una alta calidad en su producción. Por último, si el lote rechazado se compone de artículos terminados, los productos deben ser desechados o revisados a fin de cumplir con los estándares de calidad aceptables.

El muestreo de aceptación tiene las siguientes ventajas sobre el 100% de inspección:

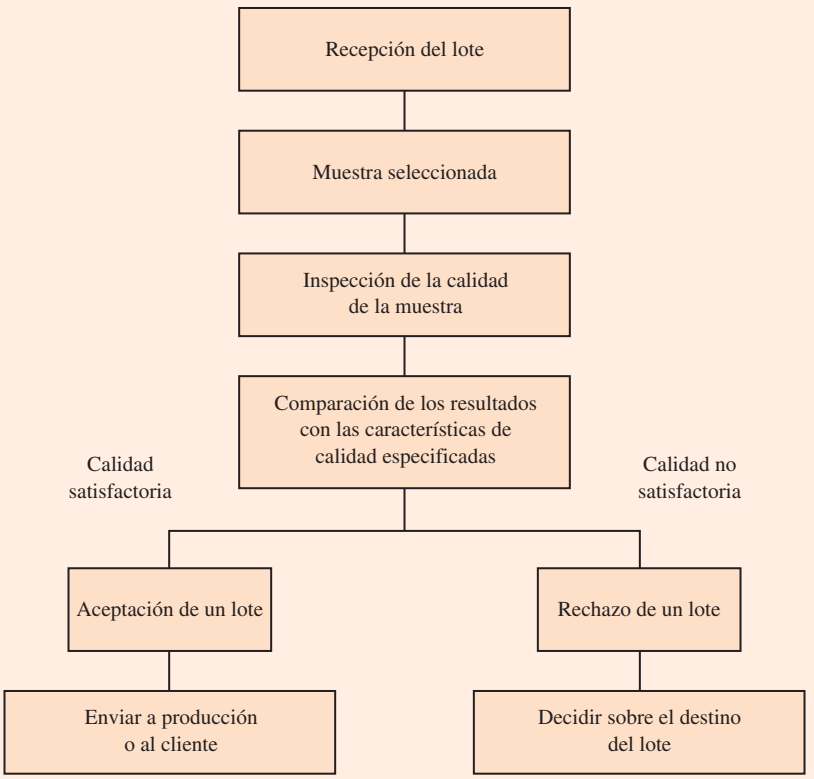
1. *Es por lo general menos costoso.*
2. *Reporta menos daños al producto debido a menor manipulación y menos pruebas.*
3. *Requieren menos inspectores.*
4. *El único enfoque posible es si debe utilizarse la prueba destructiva.*

El procedimiento estadístico de muestreo de aceptación se basa en la metodología de prueba de hipótesis presentada en el capítulo 9. Las hipótesis nula y alternativa se detallan a continuación.

H_0: lote de alta calidad
H_a: lote de baja calidad

En la tabla 20.4 se muestran los resultados del procedimiento de prueba de hipótesis. Observe que las decisiones correctas corresponden a la aceptación de un lote de alta calidad y al rechazo

FIGURA 20.11 Procedimiento de muestreo de aceptación



de un lote de baja calidad. Sin embargo, como con los demás procedimientos de prueba de hipótesis, hay que tener cuidado de no cometer un error tipo I (rechazar un lote de buena calidad) o un error tipo II (aceptar un lote de baja calidad).

La probabilidad de cometer un error tipo I crea un riesgo para el productor del lote y se le conoce como **riesgo del productor**. Por ejemplo, un riesgo del productor de 0.05 indica 5% de probabilidad de que un lote de alta calidad sea erróneamente rechazado. La probabilidad de cometer un error tipo II, por el contrario, crea un riesgo para el consumidor del lote y se conoce como **riesgo del consumidor**. Por ejemplo, un riesgo del consumidor de 0.10 significa que hay 10% de probabilidad de que un lote de baja calidad sea admitida por error y por tanto utilizado en la producción o entregado al cliente. Los valores específicos para los riesgos del productor

TABLA 20.4 Los resultados del muestreo de aceptación

		Estado del lote	
		H_0 verdadera Lote de alta calidad	**H_0 Falsa Lote de baja calidad**
Decisión	**Aceptación de un lote**	Decisión correcta	Error tipo II (aceptación de un lote de baja calidad)
	Rechazo de un lote	Error tipo I (rechazo de un lote de alta calidad)	Decisión correcta

y del consumidor pueden ser controlados por la persona que diseña el procedimiento de muestreo de aceptación. Para ilustrar la forma de asignar valores de riesgo se considerará el problema de KALI, Inc.

KALI, Inc.: un ejemplo de muestreo de aceptación

KALI, Inc. fabrica electrodomésticos que se venden bajo diferentes nombres comerciales. Sin embargo, la empresa no fabrica todos los componentes que utiliza en sus productos, sino que varios los compra directamente a los proveedores. Por ejemplo, uno de los componentes que compra es el que usa en aparatos de aire acondicionado para casas que es un protector de sobrecarga, un dispositivo que apaga el compresor si éste se sobrecalienta. El compresor puede verse seriamente dañado si el protector de sobrecarga no funciona correctamente y, por tanto, KALI está preocupada por la calidad de estos protectores. Una forma de asegurar la calidad consiste en poner a prueba todos los componentes recibidos por medio de un método que se conoce como *inspección al 100%*. Sin embargo, para determinar el buen funcionamiento de un protector de sobrecarga, el dispositivo debe ser sometido a pruebas que requieren mucho tiempo y que son costosas, y KALI no puede justificar la prueba de cada protector de sobrecarga que recibe.

A cambio, la empresa utiliza un plan de muestreo de aceptación para monitorear la calidad de los protectores de sobrecarga. Dicho plan requiere que los inspectores de control de calidad de KALI seleccionen y prueben una muestra de los protectores de cada embarque. Si en la muestra se encuentran pocas unidades defectuosas, el lote es probablemente de buena calidad y deberá ser aceptado. Sin embargo, si se detecta un gran número de unidadas defectuosas, el lote es probablemente de baja calidad y deberá ser rechazado.

Un plan de muestreo de aceptación consiste de una muestra de tamaño n y un criterio de aceptación c. El **criterio de aceptación** es el número máximo de artículos defectuosos que se puede encontrar en una muestra para que el lote se considere aceptable. Por ejemplo, para el problema de KALI, suponga que una muestra de 15 artículos será seleccionada de cada lote o pedido. Por otra parte, suponga que el gerente de control de calidad decide que el lote puede ser aceptado únicamente si ningún defecto se encuentra en las unidades de la muestra. En este caso, el plan de muestreo establecido por el gerente es $n = 15$ y $c = 0$.

Este plan de muestreo de aceptación es fácil de poner en práctica para el inspector de control de calidad. Él simplemente selecciona una muestra de 15 artículos, realiza las pruebas y llega a una conclusión basada en la siguiente regla de decisión.

- *Aceptar el lote* si encuentran cero artículos defectuosos.
- *Rechazar el lote* si se detectan uno o más defectos en los artículos.

Antes de implementar este plan de muestreo de aceptación, el gerente de control de calidad quiere evaluar los riesgos o posibles errores en el plan, el cual será llevado a cabo únicamente si tanto el riesgo del productor (error tipo I) como el riesgo del consumidor (error tipo II) están controlados en niveles razonables.

Cálculo de la probabilidad de aceptación de un lote

La clave para analizar tanto el riesgo del productor como el del consumidor es un tipo de análisis de "qué sucedería si". Es decir, suponga que un lote tiene algún porcentaje conocido de artículos defectuosos y se calcula la probabilidad de aceptarlo para un determinado plan de muestreo. Al variar el porcentaje previsto de los artículos defectuosos, se puede examinar el efecto del plan en ambos tipos de riesgos.

Comience por suponer que se ha recibido un gran cargamento de protectores de sobrecarga y que en el envío 5% está defectuoso. Para un envío o lote con 5% de artículos defectuosos, ¿cuál es la probabilidad de que $n = 15$, $c = 0$ en el plan de muestreo lleve a aceptar el lote? Debido a que cada protector de sobrecarga que se prueba puede estar o no defectuoso y el tamaño del lote es grande, el número de artículos defectuosos en una muestra de 15 tiene una

distribución binomial. A continuación se presenta la función de probabilidad binomial expuesta en el capítulo 5.

FUNCIÓN DE PROBABILIDAD BINOMIAL PARA UN MUESTREO DE ACEPTACIÓN

$$f(x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{(n-x)} \qquad \textbf{(20.21)}$$

donde

n = tamaño de la muestra
p = proporción de artículos defectuosos en el lote
x = número de artículos defectuosos en la muestra
$f(x)$ = probabilidad de x artículos defectuosos en la muestra

Para el plan de muestreo de aceptación de KALI, $n = 15$; por tanto, para un lote con 5% de defectos ($p = 0.05$) se tiene

$$f(x) = \frac{15!}{x!(15-x)!}(0.05)^x(1-0.05)^{(15-x)} \qquad \textbf{(20.22)}$$

Al utilizar la ecuación (20.22), $f(0)$ proporcionará la probabilidad de que cero de los protectores de sobrecarga estén defectuosos y que el lote sea aceptado. Respecto de la ecuación (20.22), recuerde que $0! = 1$. Por tanto, el cálculo de probabilidad para $f(0)$ es

$$f(0) = \frac{15!}{0!(15-0)!}(0.05)^0(1-0.05)^{(15-0)}$$

$$= \frac{15!}{0!(15)!}(0.05)^0(0.95)^{15} = (0.95)^{15} = 0.4633$$

Ahora ya se sabe que el plan de muestreo $n = 15$, $c = 0$ tiene una probabilidad de 0.4633 de aceptación de un lote con 5% de artículos defectuosos. Por tanto, debe existir una probabilidad correspondiente de $1 - 0.4633 = 0.5367$ para rechazar un lote con 5% de artículos defectuosos.

La probabilidad binomial también puede calcularse utilizando Excel o Minitab.

Las tablas de probabilidades binomiales (vea la tabla 5, anexo B) ayudan a reducir el esfuerzo de cálculo para determinar la probabilidad de aceptación de lotes. Las probabilidades binomiales seleccionadas para $n = 15$ y $n = 20$ figuran en la tabla 20.5. Usando esta tabla podemos determinar que si el lote contiene 10% de artículos defectuosos, existe una 0.2059 probabilidad de que el plan de muestreo $n = 15$, $c = 0$ indicará que es aceptable. En la tabla 20.6 se resume la probabilidad de que $n = 15$, $c = 0$ del plan de muestreo lleve a la aceptación de un lote con 1%, 2%, 3% . . . de artículos defectuosos.

Mediante el uso de las probabilidades de la tabla 20.6, se trazará una gráfica de la probabilidad de aceptar el lote con respecto al porcentaje de unidades defectuosas de un lote, como se muestra en la figura 20.12. Esta gráfica, o una curva, se llama la **curva característica de operación (OC)** para el plan de muestreo de aceptación $n = 15$, $c = 0$.

Tal vez deban considerarse otros planes de muestreo con diferentes tamaños n de muestra o distintos criterios de aceptación c. Considere primero el caso en que el tamaño de la muestra sigue siendo $n = 15$, pero el criterio de aceptación aumenta de $c = 0$ a $c = 1$. Es decir, ahora se aceptará el lote si se encuentran en la muestra 0 o 1 componente defectuoso. Para un lote con 5% de artículos defectuosos ($p = 0.05$), la tabla 20.5 indica que con $n = 15$ y $p = 0.05$, $f(0) = 0.4633$ y $f(1) = 0.3658$. Por tanto, existe una $0.4633 + 0.3658 = 0.8291$ probabilidad de que el plan de muestreo $n = 15$, $c = 1$ conduzca a la aceptación de un lote con 5% de artículos defectuosos.

TABLA 20.5 Selección de probabilidades binomiales para muestras de tamaño 15 y 20

		p								
n	***x***	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**
15	0	0.8601	0.7386	0.6333	0.5421	0.4633	0.2059	0.0874	0.0352	0.0134
	1	0.1303	0.2261	0.2938	0.3388	0.3658	0.3432	0.2312	0.1319	0.0668
	2	0.0092	0.0323	0.0636	0.0988	0.1348	0.2669	0.2856	0.2309	0.1559
	3	0.0004	0.0029	0.0085	0.0178	0.0307	0.1285	0.2184	0.2501	0.2252
	4	0.0000	0.0002	0.0008	0.0022	0.0049	0.0428	0.1156	0.1876	0.2252
	5	0.0000	0.0000	0.0001	0.0002	0.0006	0.0105	0.0449	0.1032	0.1651
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0019	0.0132	0.0430	0.0917
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0030	0.0138	0.0393
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0035	0.0131
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0007	0.0034
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0007
20	0	0.8179	0.6676	0.5438	0.4420	0.3585	0.1216	0.0388	0.0115	0.0032
	1	0.1652	0.2725	0.3364	0.3683	0.3774	0.2702	0.1368	0.0576	0.0211
	2	0.0159	0.0528	0.0988	0.1458	0.1887	0.2852	0.2293	0.1369	0.0669
	3	0.0010	0.0065	0.0183	0.0364	0.0596	0.1901	0.2428	0.2054	0.1339
	4	0.0000	0.0006	0.0024	0.0065	0.0133	0.0898	0.1821	0.2182	0.1897
	5	0.0000	0.0000	0.0002	0.0009	0.0022	0.0319	0.1028	0.1746	0.2023
	6	0.0000	0.0000	0.0000	0.0001	0.0003	0.0089	0.0454	0.1091	0.1686
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0020	0.0160	0.0545	0.1124
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0046	0.0222	0.0609
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0011	0.0074	0.0271
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0020	0.0099
	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0030
	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0008

Al continuar con estos cálculos se obtiene la figura 20.13, que presenta las curvas características de operación de cuatro planes alternos de muestreo de aceptación para el problema de KALI. Se consideraron muestras de tamaño 15 y 20. Observe que independientemente de la proporción de unidades defectuosas en el lote, el plan de muestreo $n = 15$, $c = 1$ proporciona la mayor probabilidad de aceptarlo. El plan de muestreo $n = 20$, $c = 0$ ofrece la menor probabilidad de aceptar el lote; sin embargo, el plan también prevé la más alta probabilidad de rechazarlo.

TABLA 20.6 Probabilidad de aceptación del lote para el problema de KALI $n = 15$ y $c = 0$

Porcentaje de defectos en el lote	**Probabilidad de aceptación del lote**
1	0.8601
2	0.7386
3	0.6333
4	0.5421
5	0.4633
10	0.2059
15	0.0874
20	0.0352
25	0.0134

FIGURA 20.12 Curva característica de operación para el plan de muestreo de aceptación $n = 15, c = 0$

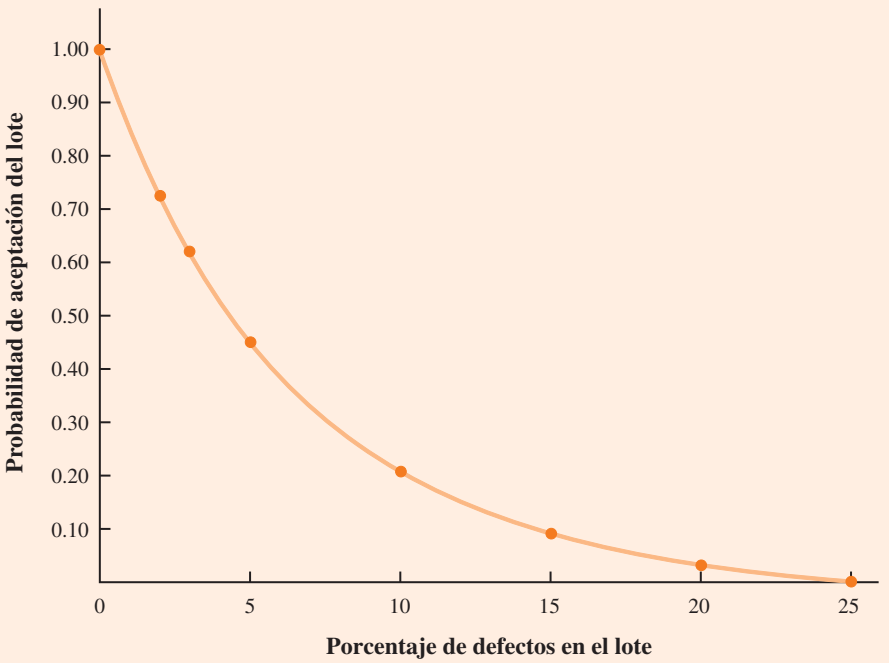
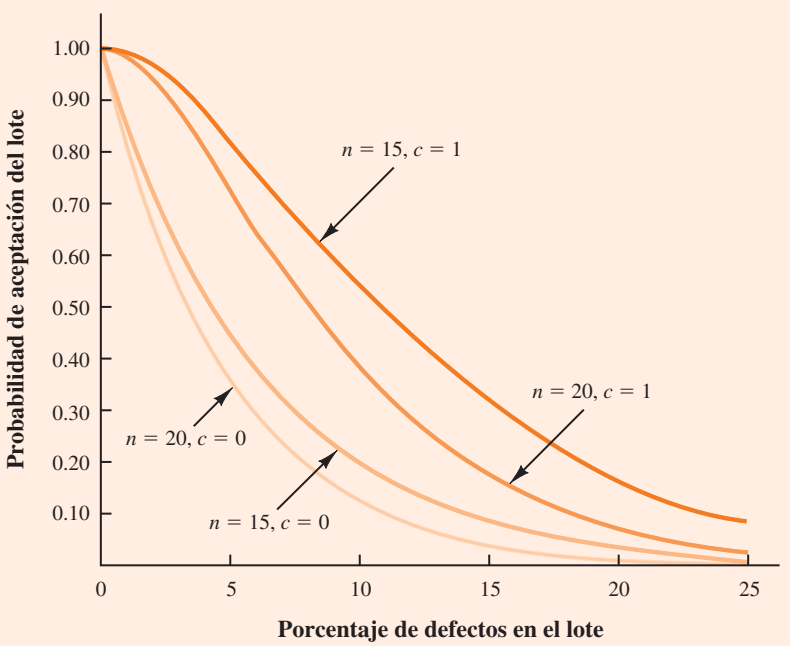


FIGURA 20.13 Curvas características de operación para cuatro planes de muestreo de aceptación



Selección de un plan de muestreo de aceptación

Ahora que se sabe cómo utilizar la distribución binomial para calcular la probabilidad de aceptación de un lote con determinada proporción de defectos, es posible elegir los valores de n y c que determinan el plan de muestreo de aceptación deseado, por ser la aplicación de estudio. Para desarrollar este plan, los gerentes deben especificar dos valores para la proporción de defectos de un lote: un valor, denotado p_0, para controlar el riesgo del productor, y otro valor, denotado p_1, para controlar el riesgo del consumidor.

Usaremos la siguiente notación.

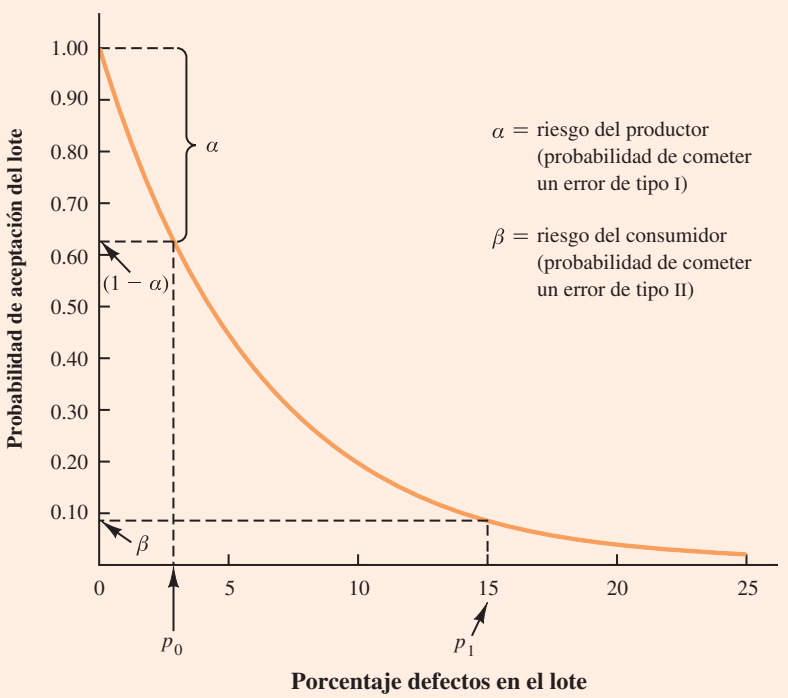
α = riesgo del productor; probabilidad de rechazar un lote con p_0 artículos defectuosos

β = riesgo del consumidor; la probabilidad de aceptar un lote con p_1 artículos defectuosos

Suponga que para el problema de KALI, los gerentes especifican que $p_0 = 0.03$ y $p_1 = 0.15$. De la curva OC para $n = 15$, $c = 0$ en la figura 20.14, vemos que $p_0 = 0.03$ provee un riesgo del productor de aproximadamente $1 - 0.63 = 0.37$, y $p_1 = 0.15$ proporciona un riesgo del consumidor de aproximadamente 0.09. Por tanto, si los gerentes están dispuestos a tolerar tanto una 0.37 probabilidad de rechazar un lote con 3% de artículos defectuosos (riesgo del productor) como una 0.09 probabilidad de aceptar un lote con 15% de artículos defectuosos (riesgo del consumidor), el plan de muestreo de aceptación $n = 15$, $c = 0$ será aceptable.

Suponga, sin embargo, que los gerentes desean que el riesgo del productor sea $\alpha = 0.10$ y el riesgo del consumidor sea $\beta = 0.20$. Observamos que ahora el plan de muestreo $n = 15$, $c = 0$ tiene un riesgo del consumidor mejor que el deseado, pero un riesgo del productor demasiado grande para aceptarlo. El hecho de que $\alpha = 0.37$ indica que 37% de los lotes será rechazado erróneamente cuando sólo 3% de los artículos que contienen son defectuosos. El riesgo del productor es demasiado alto y se tendrá que considerar un plan diferente de muestreo de aceptación.

FIGURA 20.14 Curva característica de operación para $n = 15$, $c = 0$ con $p_0 = 0.03$ y $p_1 = 0.15$

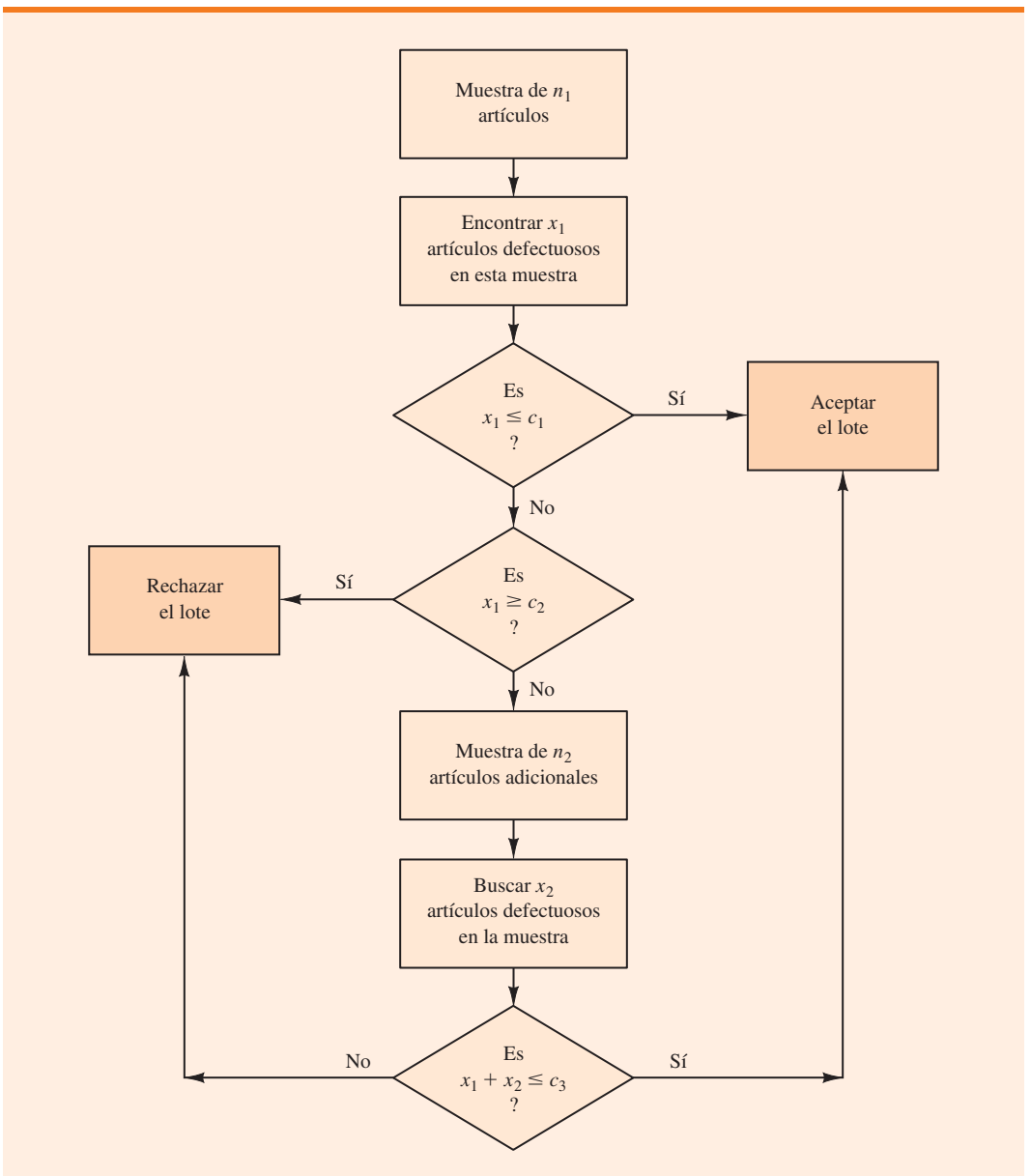


El ejercicio 13 al final de esta sección le pedirá calcular los riesgos del productor y el riesgo del consumidor para el plan de muestreo de $n = 20$ y $c = 1$.

En la figura 20.13 se observa que para $p_0 = 0.03$, $\alpha = 0.10$, $p_1 = 0.15$ y $\beta = 0.20$, el plan de muestreo de aceptación con $n = 20$ y $c = 1$ está más cerca de satisfacer los requerimientos de riesgo tanto del productor como del consumidor.

Como se muestra en esta sección, puede ser necesario considerar varios cálculos y varias curvas características de operación para determinar un plan de muestreo con el riesgo deseado para el productor y el consumidor. Por fortuna, ya exiten tablas de los planes de muestreo publicadas. Por ejemplo, la American Military Standard Table, MIL-STD-05D, ofrece información útil para el diseño de planes de muestreo de aceptación. Los libros más avanzados sobre control de calidad, como los que figuran en la bibliografía, describen el uso de dichas tablas. En estos libros también se estudia la función de los costos de muestreo para determinar un plan óptimo.

FIGURA 20.15 Plan de muestreo de aceptación de dos etapas



Planes de muestreo múltiple

El procedimiento del muestreo de aceptación que presentamos para el problema de KALI es un plan de *muestreo sencillo*. Se llama así porque se utiliza sólo una muestra o etapa de muestreo. Después de que se determina el número de componentes defectuosos en la muestra, debe tomarse una decisión de aceptar o rechazar el lote. Una alternativa al plan de muestreo sencillo es el **plan de muestreo múltiple**, en el que se utilizan dos o más etapas de muestreo. En cada etapa se toma una decisión entre tres posibilidades: dejar de muestrear y aceptar el lote, dejar de muestrear y rechazar el lote, o continuar con el muestreo. Aunque los muestreos son más complejos, múltiples planes suelen dar como resultado tamaños de la muestra más pequeños que los planes de muestreo sencillo con las mismas probabilidades α y β.

En la figura 20.15 se ilustra la lógica del plan de dos etapas, o de doble muestra. Inicialmente se selecciona una muestra de n_1 artículos. Si el número de elementos defectuosos x_1 es menor o igual que c_1, se acepta el lote. Si x_1 es mayor o igual que c_2, se rechaza el lote. Si x_1 está entre c_1 y c_2 ($c_1 < x_1 < c_2$), se toma una segunda muestra de n_2 elementos. Enseguida se determina la suma o el total de los componentes defectuosos a partir de la primera muestra (x_1) y la segunda muestra (x_2). Si $x_1 + x_2 \leq c_3$, se acepta el lote; de lo contrario se rechaza. El plan de doble muestra es más difícil porque el tamaño de las muestras n_1 y n_2 y los números de aceptación c_1, c_2 y c_3 deben cumplir tanto los riesgos deseados del productor como del consumidor.

NOTAS Y COMENTARIOS

1. El uso de la distribución binomial en el muestreo de aceptación se basa en el supuesto de grandes lotes. Si el tamaño del lote es pequeño, la distribución hipergeométrica es la adecuada. Expertos en el campo del control de calidad indican que la distribución de Poisson proporciona una buena aproximación para el muestreo de aceptación cuando el tamaño de la muestra es de por lo menos 16, el tamaño del lote es de por lo menos 10 veces el tamaño de la muestra, y p es menor que 0.1. Para los tamaños de muestras más grandes se puede utilizar la aproximación normal a la distribución binomial.
2. En las tablas de muestreo MIL-ST-105D, a p_0 se le llama *nivel de calidad aceptable* (AQL). En algunas tablas de muestreo, a p_1 se le denomina *porcentaje de defectos tolerable por lote* (LPTD) o *nivel de calidad rechazable* (RQL). Muchos planes de muestreo publicados también utilizan índices como el nivel de calidad indiferente (IQL) y el límite de la calidad promedio de salida (AOQL). Los libros más avanzados que se listan en la bibliografía proporcionan un estudio completo de estos índices.
3. En esta sección se ofrece una introducción a los *atributos de los planes de muestreo.* En estos planes, cada artículo de la muestra se clasifica como no defectuosa o defectuoso. En los *planes de muestreo de variables* se toma una muestra y se miden las características de calidad. Por ejemplo, en el caso de joyas de oro, una medida de la calidad podría ser la cantidad de oro que contienen. Una estadística sencilla como la cantidad promedio de oro en una muestra de joyas se calcula y compara con un valor permisible para determinar si el lote se debe aceptar o rechazar.

Ejercicios

Métodos

AUTOevaluación

10. Para un plan de muestreo de aceptación con $n = 25$ y $c = 0$, determine la probabilidad de aceptar un lote que tiene un porcentaje de defectos de 2. ¿Cuál es la probabilidad de aceptarlo si la tasa de defectos es de 6%?
11. Considere un plan de muestreo de aceptación con $n = 20$ y $c = 0$. Calcule el riesgo del productor para cada uno de los siguientes casos.
    a) El porcentaje de defectos que tiene el lote es de 2%.
    b) El porcentaje de defectos que tiene el lote es de 6%.
12. Repita el ejercicio 11 para el plan de muestreo de aceptación con $n = 20$ y $c = 1$. ¿Qué sucede con el riesgo del productor cuando aumenta el número de aceptación de c? Explique.

Aplicaciones

13. Consulte el problema de KALI presentado en esta sección. El gerente de control de calidad requirió un riesgo del productor de 0.10 cuando p_0 fue de 0.03 y un riesgo del consumidor de 0.20 cuando p_1 fue de 0.15. Considere un plan de muestreo de aceptación basado en una muestra de tamaño 20 y un número de aceptación de 1. Conteste las siguientes preguntas.
    a) ¿Cuál es el riesgo del productor si el plan de muestreo es $n = 20$, $c = 1$?
    b) ¿Cuál es el riesgo del consumidor con el mismo que en *a*) plan de muestreo?
    c) ¿El plan de muestreo $n = 20$, $c = 1$ satisface el riesgo requerido por el gerente de control de calidad? Comente.
14. Para inspeccionar un pedido de materia prima, el fabricante estudia muestras de tamaños 10, 15 y 20. Utilice las probabilidades binomiales de la tabla 5 del apéndice B para seleccionar un plan de muestreo que proporcione un riesgo del productor de $\alpha = 0.03$ cuando p_0 es 0.05, y un riesgo del consumidor de $\beta = 0.12$ cuando p_1 es 0.30.
15. Un fabricante nacional de relojes compra cristales de cuarzo a una empresa suiza. Los cristales son enviados en lotes de 1 000. El procedimiento de muestreo de aceptación utiliza 20 cristales elegidos al azar.
    a) Trace las curvas características de operación para los números de aceptación de 0.1, y 2.
    b) Si p_0 es 0.01 y $p_1 = 0.08$, ¿cuáles son los riesgos del productor y del consumidor con cada uno de los planes de muestreo del inciso *a*)?

Resumen

En este capítulo se estudió cómo utilizar el método estadístico en el control de calidad. En primer lugar se presentaron las gráficas de control de $\bar{x}$, R, p y np como ayudas gráficas para monitorear el proceso de calidad. En cada una de estas gráficas se establecen los límites de control; se seleccionan muestras periódicamente, y se grafican los puntos de los datos en la gráfica de control. Los puntos de datos fuera de los límites de control indican que el proceso está fuera de control y que debe aplicarse la acción correctiva. Los patrones de puntos de datos que se encuentran dentro de los límites de control también indican posibles problemas en el control de calidad y sugieren que se apliquen acciones correctivas.

También se consideró la técnica conocida como *muestreo de aceptación*. Con este procedimiento se selecciona una muestra y se inspecciona. El número de artículos defectuosos en la muestra proporciona la base para aceptar o rechazar el lote. El tamaño de la muestra y el criterio de aceptación se pueden ajustar para controlar tanto el riesgo del productor (error tipo I) como el riesgo del consumidor (error tipo II).

Glosario

Calidad total (TQ) Enfoque total del sistema para mejorar la satisfacción del cliente y reducir los costos reales por medio de una estrategia de mejora y aprendizaje continuos.
Causas comunes Variaciones normales o naturales en los resultados de un proceso que se deben únicamente a la casualidad. Ninguna acción correctiva es necesaria cuando las variaciones en los resultados de un proceso se deben a causas comunes.
Causas imputables Variaciones en los resultados de un proceso atribuibles a factores como desgaste de la maquinaria, una configuración incorrecta de la máquina, materia prima de baja calidad, errores del operador, y así sucesivamente. La acción correctiva se debe aplicar cuando se detectan las causas imputables en la variación del resultado.
Control de calidad Serie de inspecciones y mediciones que determinan si se cumplen los estándares de calidad.
Criterio de aceptación Número máximo de artículos defectuosos que se pueden encontrar en la muestra y que a pesar de ello indican que se acepte un lote.
Curva característica de operación (OC) Gráfica en la que se muestra la probabilidad de aceptar un lote en función del porcentaje de artículos defectuosos que contiene. Esta curva se utiliza para determinar si un plan de muestreo de aceptación particular satisface los requerimientos de riesgo tanto del productor como del consumidor.

Gráfica de control Herramienta gráfica que se utiliza para ayudar a determinar si un proceso está bajo control o fuera de control.
Gráfica *np* Gráfica de control que supervisa la calidad de los resultados de un proceso en términos del número de artículos defectuosos.
Gráfica *p* Gráfica de control que se utiliza para medir la calidad de los resultados de un proceso en términos de la proporción de unidades defectuosas.
Gráfica *R* Gráfica de control que mide la calidad del resultado de un proceso en términos del rango de una variable.
Gráfica $\bar{x}$ Gráfica de control que se utiliza cuando se mide la calidad de los resultados de un proceso en términos del valor medio de una variable, como longitud, peso, temperatura, etcétera.
Lote Conjunto de artículos, como embarques recibidos de materia prima o piezas compradas, así como bienes terminados para ensamblar un producto final.
Muestreo de aceptación Método estadístico en el que se utiliza el número de artículos defectuosos en una muestra para determinar si un lote debe ser aceptado o rechazado.
Plan de muestreo múltiple Forma de muestreo de aceptación en la que se utiliza más de una muestra o etapa. Con base en el número de artículos defectuosos que se encuentra en una muestra, se decide si se acepta o rechaza el lote, o si se continúa con el muestreo.
Riesgo del consumidor Riesgo de aceptar un lote de buena calidad; error tipo II.
Riesgo del productor Riesgo de rechazar un lote de buena calidad; error tipo I.
Six Sigma Metodología que utiliza la medición y el análisis estadístico para lograr un nivel de calidad tan bueno, que por cada millón de oportunidades no haya más de 3.4 defectos.

Fórmulas clave

Error estándar de la media

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{20.1}$$

Límites de control para una gráfica $\bar{x}$: media del proceso y desviación estándar conocida

$$\text{UCL} = \mu + 3\sigma_{\bar{x}} \tag{20.2}$$

$$\text{LCL} = \mu - 3\sigma_{\bar{x}} \tag{20.3}$$

Media muestral general

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_k}{k} \tag{20.4}$$

Rango promedio

$$\bar{R} = \frac{R_1 + R_2 + \cdots + R_k}{k} \tag{20.5}$$

Límites de control para una gráfica $\bar{x}$: media del proceso y desviación estándar desconocida

$$\bar{\bar{x}} \pm A_2\bar{R} \tag{20.8}$$

Límites de control para una gráfica *R*

$$\text{UCL} = \bar{R}\text{D}_4 \tag{20.14}$$

$$\text{LCL} = \bar{R}\text{D}_3 \tag{20.15}$$

Error estándar de la proporción

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{20.16}$$

Límites de control para una gráfica *p*

$$\text{UCL} = p + 3\sigma_{\bar{p}} \tag{20.17}$$

$$\text{LCL} = p - 3\sigma_{\bar{p}} \tag{20.18}$$

Límites de control para una gráfica *np*

$$\text{UCL} = np + 3\sqrt{np(1-p)} \tag{20.19}$$

$$\text{LCL} = np - 3\sqrt{np(1-p)} \tag{20.20}$$

Función de probabilidad binomial para un muestreo de aceptación

$$f(x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{(n-x)} \tag{20.21}$$

Ejercicios complementarios

16. Muestras de tamaño 5 proporcionaron las siguientes 20 medias muestrales para un proceso de producción que se considera bajo control.

95.72	95.24	95.18
95.44	95.46	95.32
95.40	95.44	95.08
95.50	95.80	95.22
95.56	95.22	95.04
95.72	94.82	95.46
95.60	95.78	

a) Con base en estos datos, ¿cuál es la estimación de la media cuando el proceso está bajo control?
b) Suponga que la desviación estándar del proceso es $\sigma = 0.50$. Desarrolle la gráfica de control $\bar{x}$ para este proceso de producción. Suponga que la media del proceso es la estimación desarrollada en el inciso *a*).
c) ¿Algunas de las 20 medias muestrales indican que el proceso está fuera de control?

17. En un proceso, los pesos de llenado se distribuyen normalmente con una media de 350 gramos y una desviación estándar de 15 gramos.
a) Obtenga los límites de control de la gráfica $\bar{x}$ para muestras de tamaño 10, 20 y 30.
b) ¿Qué le sucede a los límites de control a medida que aumenta el tamaño de la muestra?
c) ¿Qué ocurre cuando se comete un error tipo I?
d) ¿Qué sucede si se comete un error tipo II?
e) ¿Cuál es la probabilidad de cometer un error tipo I con muestras de tamaño 10, 20 y 30?
f) ¿Cuál es la ventaja de aumentar el tamaño de la muestra para los propósitos de una gráfica de control? ¿Qué probabilidad de error se reduce conforme este tamaño se incrementa?

18. Con 25 muestras de tamaño 5 se obtuvieron $\bar{\bar{x}} = 5.42$ y $\bar{R} = 2.0$. Calcule los límites de control de las gráficas $\bar{x}$ y R, y estime la desviación estándar del proceso.

19. Los datos de control de calidad que se presentan en la primera tabla de la página siguiente se obtuvieron de un proceso de manufactura en Kensport Chemical Company. Los datos muestran la temperatura en grados centígrados medida en cinco puntos durante un ciclo de manufactura. La empresa está interesada en emplear gráficas de control para monitorear la temperatura de su proceso de manufactura. Construya las gráficas $\bar{x}$ y R. ¿A qué conclusión se puede llegar acerca de la calidad del proceso?

Muestra	$\bar{x}$	R	Muestra	$\bar{x}$	R
1	95.72	1.0	11	95.80	0.6
2	95.24	.9	12	95.22	0.2
3	95.18	.8	13	95.56	1.3
4	95.44	.4	14	95.22	0.5
5	95.46	.5	15	95.04	0.8
6	95.32	1.1	16	95.72	1.1
7	95.40	.9	17	94.82	0.6
8	95.44	.3	18	95.46	0.5
9	95.08	.2	19	95.60	0.4
10	95.50	.6	20	95.74	0.6

20. Los siguientes datos se obtuvieron del proceso de producción de Master Blend Cofee y representan las cinco observaciones (Observations) de 10 muestras (Sample) de los pesos de llenado de latas de café de 3 libras. Utilice estos datos para construir las cartas $\bar{x}$ y R. ¿A qué conclusión se puede llegar sobre la calidad del proceso de producción?

WEB archivo
Coffee

	Observations				
Sample	1	2	3	4	5
1	3.05	3.08	3.07	3.11	3.11
2	3.13	3.07	3.05	3.10	3.10
3	3.06	3.04	3.12	3.11	3.10
4	3.09	3.08	3.09	3.09	3.07
5	3.10	3.06	3.06	3.07	3.08
6	3.08	3.10	3.13	3.03	3.06
7	3.06	3.06	3.08	3.10	3.08
8	3.11	3.08	3.07	3.07	3.07
9	3.09	3.09	3.08	3.07	3.09
10	3.06	3.11	3.07	3.09	3.07

21. Considere las siguientes situaciones. Opine acerca de si la situación podría ser motivo de preocupación debido a la calidad del proceso.
   a) En una gráfica p se tiene LCL $= 0$ y UCL $= 0.068$. Cuando el proceso está bajo control, la proporción de unidades defectuosas es 0.033. Represente gráficamente los siguientes siete resultados de la muestra: 0.035, 0.062, 0.055, 0.049, 0.058, 0.066 y 0.055. Comente.
   b) En una gráfica $\bar{x}$ se tiene LCL $= 22.2$ y UCL $= 24.5$. La media es $\mu = 23.35$ cuando el proceso está bajo control. Represente gráficamente los siguientes siete resultados de la muestra: 22,4, 22.6, 22.65, 23.2, 23.4, 23.85 y 24.1. Comente.

22. Dos veces al mes los gerentes de 1 200 comercios minoristas distintos ordenan pedidos de reposición de existencias a la casa matriz. La experiencia pasada demuestra que 4% de los pedidos resultó con uno o más errores, como el envío de artículos equivocados, cantidad del envío incorrecta y el artículo solicitado pero no enviado. Cada mes, muestras aleatorias de pedidos 200 son seleccionadas y se verifica su exactitud.
   a) Construya una gráfica de control para esta situación.
   b) Los datos de los últimos seis meses muestran los siguientes números de los pedidos con uno o varios errores: 10, 15, 6, 13, 8 y 17. Disponga estos datos en una gráfica de control. ¿Qué indica la gráfica acerca de este proceso de pedido?

23. Se está considerando un plan de muestreo de aceptación con $n = 10$, $c = 2$. Suponga que $p_0 = 0.05$ y $p_1 = 0.20$.
   a) Calcule los riesgos tanto del productor como del consumidor en este plan de muestreo de aceptación.
   b) ¿Estará inconforme el productor, el consumidor, o ambos, con el plan de muestreo propuesto?
   c) ¿Recomendaría algún cambio en su plan de muestreo?

24. Un plan de muestreo de aceptación fue diseñado con $n = 15$ y $c = 1$, con un riesgo del productor de 0.075.
    a) ¿Cuál es el valor de p_0 0.01, 0.02, 0.03, 0.04 o 0.05? ¿Qué significa este valor?
    b) ¿Cuál es el riesgo del consumidor asociado con este plan si p_1 es 0.25?
25. Un fabricante elabora productos alimenticios enlatados. Sea p la proporción de los lotes que no cumplen con las especificaciones de calidad del producto. El plan de muestreo de aceptación utilizará $n = 25$, $c = 0$.
    a) Calcule los puntos en la curva características de operación cuando $p = 0.01, 0.03, 0.10$ y 0.20.
    b) Trace la curva característica de operación.
    c) ¿Cuál es la probabilidad de que el plan de muestreo de aceptación rechace un lote que contenga 0.01 defectos?

Apéndice 20.1 Gráficas de control con Minitab

WEB archivo
Jensen

En este apéndice se describen los pasos necesarios para trazar las gráficas de control con Minitab utilizando los datos de Jensen Computer Supplies que se muestran en la tabla 20.2. El número de la muestra aparece en la columna C1, la primera observación en la columna C2, la segunda observación en la columna C3, y así sucesivamente. Los pasos siguientes describen cómo utilizar Minitab para producir al mismo tiempo tanto la gráfica $\bar{x}$ como la gráfica R.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Control Charts**.
Paso 3. Seleccione **Variables Charts for Subgroups**.
Paso 4. Elija **Xbar-R**.
Paso 5. Cuando el cuadro de diálogo Xbar-R Chart aparezca:
Seleccione **Observations for a subgroup are in one row of columns**.
En el cuadro que se encuentra debajo, introduzca C2-C6.
Seleccione **Xbar-R Options**.
Paso 6. Cuando el cuadro de diálogo Xbar-R-Options aparezca:
Seleccione la pestaña **Tests**.
Seleccione **Perform selected tests for special causes**.
Elija **1 point > K standard deviations from center line***.
Introduzca 3 en la caja **K**.
Haga clic en **OK**.
Paso 7. Cuando el cuadro de diálogo Xbar-R Chart aparezca:
Haga clic en **OK**.

Las gráficas $\bar{x}$ y R aparecen juntas en el resultado de Minitab. Las alternativas disponibles en el paso 3 del procedimiento anterior facilitan el acceso a una variedad de opciones de gráficas de control. Por ejemplo, se puede seleccionar que las gráficas $\bar{x}$ y R aparezcan por separado. Las opciones adicionales incluyen la gráfica p, la gráfica np, y otras.

Apéndice 20.2 Gráficas de control utilizando StatTools

WEB archivo
Jensen

En este apéndice se muestra cómo utilizar StatTools para elaborar las gráficas $\bar{x}$ y R con los datos de Jensen Computer Supplies mostrados en la tabla 20.2. La figura 20.16 es una hoja de cálculo de Excel con los datos de Jensen. Comience usando Data Set Manager para crear un

[1] Minitab proporciona pruebas adicionales para detectar causas especiales de variación, así como las situaciones fuera de control. El usuario puede seleccionar varios de estos análisis de forma simultánea.

conjunto de datos con StatTools mediante el procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes describen cómo utilizar StatTools para elaborar las gráficas de control.

Paso 1. Haga clic en la barra de herramientas de **StatTools**.
Paso 2. En el grupo **Analyses**, haga clic en **Quality Control**.
Paso 3. Elija la opción de **X/R Charts**.
Paso 4. Cuando el cuadro de diálogo StatTools-Xbar and R Control Charts aparezca:
Seleccione **X-Bar/R Chart** en el cuadro **Chart Type**.
En la sección **Variables**, elija **Observation 1, Observation 2, Observation 3, Observation 4** y **Observation 5**.
Haga clic en **OK**.

Aparecerá una gráfica $\bar{x}$ similar a la de la figura 20.7. Luego seguirá una gráfica R similar a la de la figura 20.8.

FIGURA 20.16 Hoja de cálculo de Excel para Jensen Computer Supplies

	A	B	C	D	E	F	G
1	Sample	Observation 1	Observation 2	Observation 3	Observation 4	Observation 5	
2	1	3.5056	3.5086	3.5144	3.5009	3.5030	
3	2	3.4882	3.5085	3.4884	3.5250	3.5031	
4	3	3.4897	3.4898	3.4995	3.5130	3.4969	
5	4	3.5153	3.5120	3.4989	3.4900	3.4837	
6	5	3.5059	3.5113	3.5011	3.4773	3.4801	
7	6	3.4977	3.4961	3.5050	3.5014	3.5060	
8	7	3.4910	3.4913	3.4976	3.4831	3.5044	
9	8	3.4991	3.4853	3.4830	3.5083	3.5094	
10	9	3.5099	3.5162	3.5228	3.4958	3.5004	
11	10	3.4880	3.5015	3.5094	3.5102	3.5146	
12	11	3.4881	3.4887	3.5141	3.5175	3.4863	
13	12	3.5043	3.4867	3.4946	3.5018	3.4784	
14	13	3.5043	3.4769	3.4944	3.5014	3.4904	
15	14	3.5004	3.5030	3.5082	3.5045	3.5234	
16	15	3.4846	3.4938	3.5065	3.5089	3.5011	
17	16	3.5145	3.4832	3.5188	3.4935	3.4989	
18	17	3.5004	3.5042	3.4954	3.5020	3.4889	
19	18	3.4959	3.4823	3.4964	3.5082	3.4871	
20	19	3.4878	3.4864	3.4960	3.5070	3.4984	
21	20	3.4969	3.5144	3.5053	3.4985	3.4885	
22							

CAPÍTULO 21

Análisis de decisiones

CONTENIDO

ESTADÍSTICA EN LA PRÁCTICA: OHIO EDISON COMPANY

21.1 FORMULACIÓN DEL PROBLEMA
Tablas de pagos
Árboles de decisión

21.2 TOMA DE DECISIONES CON PROBABILIDADES
Método del valor esperado
Valor esperado de la información perfecta

21.3 ANÁLISIS DE DECISIONES CON INFORMACIÓN MUESTRAL
Árbol de decisión
Estrategia de decisión
Valor esperado de la información muestral

21.4 CÁLCULO DE PROBABILIDADES MEDIANTE EL TEOREMA DE BAYES

ESTADÍSTICA *en* LA PRÁCTICA

OHIO EDISON COMPANY*

AKRON, OHIO

Ohio Edison Company es una empresa de FirstEnergy Corporation. Ohio Edison y su subsidiaria, Pennsylvania Power Company, suministran energía eléctrica a más de un millón de clientes en el centro y noreste de Ohio y en el oeste de Pensilvania. La mayor parte de la electricidad es generada mediante plantas de combustión de carbón. Debido a la evolución de los requerimientos sobre el control de la contaminación, Ohio Edison participa en un programa para remplazar el equipo de control de contaminación en la mayoría de sus plantas generadoras.

Para cumplir con los nuevos límites de emisión de dióxido de azufre en una de sus plantas de energía más grandes, la empresa decidió quemar carbón bajo en azufre en cuatro de las unidades más pequeñas de la planta así como instalar filtros de tela para controlar las emisiones de partículas. Dichos filtros usan miles de bolsas de tela para filtrar partículas y funcionan de la misma forma que una aspiradora casera.

Se consideró probable, aunque no seguro, que en las tres unidades más grandes de la planta se quemaría carbón con medio a alto contenido de azufre. Estudios preliminares redujeron la elección del equipo para partículas de estas unidades más grandes a filtros de tela y precipitadores electrostáticos (que eliminan las partículas suspendidas en los gases de combustión pasándolas a través de un campo eléctrico fuerte). Entre las incertidumbres que afectarían la elección final estaban la forma en que podían interpretarse algunas leyes y normas de calidad del aire, los futuros cambios potenciales en las leyes y normas de la calidad del aire, y las fluctuaciones en los costos de construcción.

Dada la complejidad del problema, el alto grado de incertidumbre relacionado con factores que influyen en la decisión y el impacto del costo para Ohio Edison, en el proceso de elección se utilizó el análisis de decisiones. Se desarrolló una descripción gráfica del problema, llamado *árbol de decisión*. La medida empleada para evaluar los resultados representados en el árbol de decisión consistió en los requerimientos de los ingresos anuales para las tres grandes unidades durante su vida restante. Las necesidades de ingresos representaron las cantidades de dinero que tendrían que ser recaudadas de los clientes de servicios públicos para recuperar los costos generados por la instalación



Las plantas de Ohio Edison suministran energía eléctrica a más de un millón de usuarios © Don Farrall Getty Images/ PhotoDisc.

de los nuevos equipos de control de la contaminación. Un análisis del árbol de decisión condujo a las conclusiones siguientes.

- El valor esperado de los ingresos anuales requeridos para los precipitadores electrostáticos representaba aproximadamente $1 millón menos que por los filtros de tela.
- Los filtros de tela tenían una mayor probabilidad de requerir altos ingresos que los precipitadores electrostáticos.
- Los precipitadores electrostáticos tenían casi 0.8 de probabilidad de tener menores necesidades de ingresos anuales.

Estas conclusiones llevaron a Ohio Edison a elegir los precipitadores electrostáticos para las unidades generadoras en cuestión. De no haberse llevado a cabo el análisis de decisiones, la toma de decisiones respecto de las partículas se habría basado principalmente en el costo de capital, una medida de decisión que favorecía la opción del equipo de filtro de tela. Se consideró que el análisis de decisiones identificó la opción tanto para los menores ingresos esperados como para el menor riesgo.

En este capítulo se presenta la metodología del análisis de decisiones que utilizó Ohio Edison. La atención se centra en mostrar cómo este análisis permite identificar la mejor alternativa de decisión dado un patrón de incertidumbre o lleno de riesgo en los eventos futuros.

* Agradecemos a los autores Thomas J. Madden y M. S. Hyrnick, de Ohio Edison Company, por proporcionar este material para *Estadística en la práctica*.

El análisis de decisiones puede utilizarse para desarrollar una estrategia óptima de decisión cuando quien decide se enfrenta a varias alternativas y a un patrón lleno de incertidumbre o de riesgo en torno de los eventos futuros. El estudio de este análisis inicia considerando los problemas que implican razonablemente pocas alternativas de decisión y razonablemente pocos eventos futuros. Se presentarán las tablas de pagos que proporcionan una estructura a los problemas de decisión. Después se presentarán árboles de decisión para mostrar el carácter secuencial de los problemas. Estos árboles permiten analizar los problemas más complejos e identificar una secuencia óptima de decisiones, a la que se le conoce como *estrategia óptima de decisión*. En la última sección se muestra cómo utilizar el teorema de Bayes, presentado en el capítulo 4, para calcular las probabilidades de los árboles de decisión. El apéndice del capítulo ofrece una introducción a PrecisionTree, un complemento de Excel que permite desarrollar y analizar los árboles de decisión.

En el sitio web de este libro se proporcionan instrucciones para descargar e instalar PrecisionTree .

21.1 Formulación del problema

El primer paso en el proceso de análisis de decisiones es formular el problema. Se inicia con una declaración verbal del problema. Después se identifican las alternativas de decisión, los eventos futuros inciertos, referidos como **eventos aleatorios**, y las **consecuencias** relacionadas con cada alternativa de decisión y cada resultado del evento. Comience por considerar el proyecto de construcción para Pittsburgh Development Corporation.

Pittsburgh Development Corporation (PDC) compró un terreno que será la sede de un complejo de condominios de lujo nuevos. El lugar ofrece una vista espectacular del centro de Pittsburgh y el Golden Triangle, formado por los ríos Allegheny y Monongahela, que se unen con el río Ohio. PDC planea vender cada condominio en un precio de entre $300 000 y $1 400 000.

La empresa ha encargado tres planos arquitectónicos preliminares para tres condominios de distintos tamaños, uno con 30 condominios, otro con 60 y otro con 90. El éxito financiero del proyecto depende tanto del tamaño del complejo como del evento aleatorio de la demanda inmobiliaria. El problema de decisión de PDC es elegir el tamaño del proyecto de los nuevos condominios de lujo que dará lugar a una mayor utilidad, dada la incertidumbre sobre la demanda.

Establecido el planteamiento del problema, es claro que la decisión es elegir el mejor tamaño para el complejo inmobiliario. PDC tiene las siguientes tres alternativas de decisión.

d_1 = un complejo pequeño de 30 condominios
d_2 = un complejo mediano de 60 condominios
d_3 = un complejo grande de 90 condominios

Un factor en la elección de la mejor alternativa de decisión es la incertidumbre relacionada con el evento aleatorio de la demanda. Cuando se preguntó acerca de la posible demanda de los condominios, el presidente de PDC reconoció una amplia gama de posibilidades, pero decidió que sería adecuado considerar dos posibles resultados de eventos aleatorios: una demanda alta y una demanda baja.

En el análisis de decisiones, a los posibles resultados de un evento aleatorio se les conoce como **estados de la naturaleza**. Éstos se definen como uno y sólo uno de los posibles estados que puedan presentarse. Para el problema de PDC, el evento aleatorio sobre la demanda de los condominios tiene dos estados de la naturaleza:

s_1 = una demanda alta de los condominios
s_2 = una demanda baja de los condominios

Los gerentes deberán elegir, primero, una alternativa de decisión (tamaño del complejo), a la cual seguirá un estado de la naturaleza (la demanda de los condominios), y finalmente se tendrá una consecuencia. En este caso, la consecuencia son las utilidades que obtendrá PDC.

Tablas de pagos

Dadas las tres alternativas de decisión y los dos estados de la naturaleza, ¿qué tamaño del complejo debe elegir PDC? Para responder esta pregunta, la empresa necesita conocer la consecuencia relacionada con cada alternativa de decisión y cada estado. En el análisis de decisiones, nos referimos a la consecuencia resultante de la combinación específica de una alternativa de decisión y un estado como un **pago**. La tabla que muestra los pagos para todas las combinaciones de las alternativas de decisión y los estados de la naturaleza es una **tabla de pagos**.

Los pagos pueden expresarse en términos de utilidad, costo, tiempo, distancia o cualquier otra medida apropiada para que el problema de decisión sea analizado.

Debido a que PDC desea elegir el tamaño del complejo que le proporcione una mayor utilidad, ésta se usa como consecuencia. La tabla de pagos junto con las utilidades expresadas en millones de dólares se muestra en la tabla 21.1. Observe, por ejemplo, que si se construye un complejo mediano y la demanda resulta ser alta, se tendrá una utilidad de $14 millones. Se utilizará la notación V_{ij} para denotar la utilidad relacionada con la alternativa de decisión i y el estado de la naturaleza j. Al utilizar la tabla 21.1, $V_{31} = 20$ significa que se tendrá un pago de $20 millones si la decisión es construir un complejo grande (d_3) y la demanda que presenta el estado de la naturaleza es alta (s_1). De manera similar, $V_{32} = -9$ significa una pérdida de 9 millones si la decisión es construir un complejo grande (d_3) y se presenta un estado de la naturaleza de baja demanda (s_2) .

Árboles de decisión

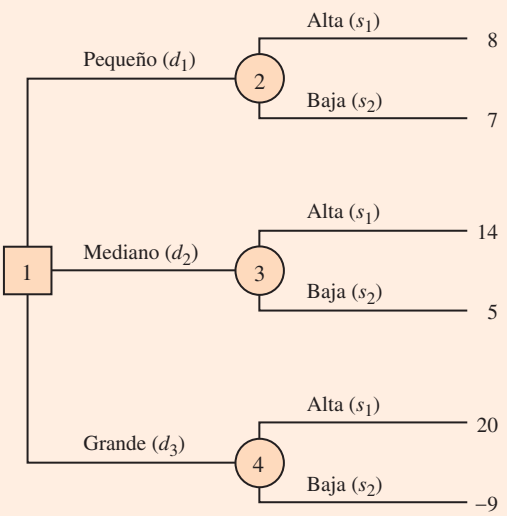
Un **árbol de decisión** muestra gráficamente el carácter secuencial del proceso de toma de decisiones. En la figura 21.1 se presenta un árbol de decisión para el problema de PDC, lo que demuestra el carácter o la progresión lógica que se presentará con el tiempo. En primer lugar, PDC deberá tomar una decisión con respecto al tamaño del condominio (d_1, d_2, d_3). Después de tomada ésta, se presentará ya sea el estado de la naturaleza s_1 o s_2. El número que aparece en cada punto extremo del árbol indica el pago relacionado con una secuencia en particular. Por ejemplo, el pago 8 que se encuentra más arriba indica que se espera una utilidad de $8 millones si PDC construye un complejo de condominios pequeño (d_1) y la demanda resulta ser alta (s_1). El pago siguiente, 7, significa que se espera una utilidad de $7 millones si PDC construye un complejo pequeño (d_1) y la demanda resulta ser baja (s_2). Por tanto, el árbol de decisión muestra gráficamente las secuencias de alternativas de decisión y los estados de la naturaleza que proporcionan los seis posibles pagos.

El árbol de la figura 21.1 tiene cuatro **nodos**, numerados del 1 al 4, que representan las decisiones y eventos aleatorios. Los cuadrados describen los **nodos de decisión** y los círculos representan los **nodos aleatorios**. Así, el nodo 1 es un nodo de decisión, y los nodos 2, 3 y 4 son aleatorios. Las **ramas** que salen del nodo de decisión corresponden a las alternativas de decisión. Las subramas que salen de cada nodo aleatorio corresponden a los estados de la naturaleza. Los pagos se muestran al final de las ramas de los estados. Ahora se vuelve a la pregunta: La

TABLA 21.1 Tabla de pagos para el proyecto del condominio de PDC (pagos en millones de dólares)

	Estado de la naturaleza	
Alternativa de decisión	**Demanda alta s_1**	**Demanda baja s_2**
Complejo pequeño, d_1	8	7
Complejo mediano, d_2	14	5
Complejo grande, d_3	20	−9

FIGURA 21.1 Árbol de decisión para el proyecto de condominio de PDC (pagos en millones de dólares)



persona que toma la decisión cómo puede utilizar la información de la tabla de pagos o del árbol de decisión para elegir la mejor alternativa de decisión?

NOTAS Y COMENTARIOS

1. Los expertos están de acuerdo en que el primer paso en la solución de un problema complejo es descomponerlo en una serie de subproblemas menores. Los árboles de decisión proporcionan una manera útil de demostrar cómo descomponer un problema y el carácter secuencial del proceso de decisión.
2. Las personas suelen ver el mismo problema desde distintas perspectivas. Por tanto, la discusión en torno al desarrollo de un árbol de decisión puede proporcionar una idea adicional sobre el problema en cuestión.

21.2 Toma de decisiones con probabilidades

Una vez definidas las alternativas de decisión y los estados de la naturaleza de los eventos aleatorios, se determinan las probabilidades de dichos estados. Para identificarlas, se puede utilizar el método clásico, el método de frecuencia relativa o el método subjetivo de asignar las probabilidades estudiados en el capítulo 4. Después de determinar las probabilidades adecuadas, se muestra cómo utilizar el **método del valor esperado** para identificar la mejor alternativa de decisión, o una decisión recomendada para el problema.

Método del valor esperado

Comience por definir el valor esperado de una alternativa de decisión. Sea

N = número de estados de la naturaleza
$P(s_j)$ = probabilidad del estado de la naturaleza s_j

Debido a que uno y sólo uno de los N estados de la naturaleza puede presentarse, las probabilidades deben satisfacer las dos condiciones siguientes.

Las probabilidades para los estados de la naturaleza deben satisfacer los requerimientos básicos para la asignación de probabilidades presentada en el capítulo 4.

$$P(s_j) \geq 0 \quad \text{por todos los estados de la naturaleza} \tag{21.1}$$

$$\sum_{j=1}^{N} P(s_j) = P(s_1) + P(s_2) + \cdots + P(s_N) = 1 \tag{21.2}$$

El **valor esperado (VE)** de una alternativa de decisión d_i es el siguiente.

VALOR ESPERADO

$$\text{VE}(d_i) = \sum_{j=1}^{N} P(s_j)V_{ij} \tag{21.3}$$

donde

V_{ij} = valor del pago para la alternativa de decisión d_i y el estado de la naturaleza s_j.

Es decir, el valor esperado de una alternativa de decisión es la suma de los pagos ponderados para la alternativa de decisión. El peso de ponderación para un pago es la probabilidad del estado de la naturaleza relacionado y por tanto la probabilidad de que el pago se presente. Vuelva al problema de PDC para ver cómo se puede aplicar el método del valor esperado.

La constructora es optimista sobre el potencial para el complejo de condominios de lujo de gran altura. Este optimismo la lleva a una evaluación subjetiva de la probabilidad 0.8 inicial de que la demanda será alta (s_1) y a la correspondiente probabilidad 0.2 de que la demanda será baja (s_2). Por tanto, $P(s_1) = 0.8$ y $P(s_2) = 0.2$. Utilizando los valores de pago de la tabla 21.1 y la ecuación (21.3), se calcula el valor esperado para cada una de las tres alternativas de decisión de la siguiente manera.

$$\begin{aligned}
\text{VE}(d_1) &= 0.8(8) + 0.2(7) = 7.8 \\
\text{VE}(d_2) &= 0.8(14) + 0.2(5) = 12.2 \\
\text{VE}(d_3) &= 0.8(20) + 0.2(-9) = 14.2
\end{aligned}$$

De esta manera, con el método del valor esperado se encuentra que el complejo de condominios grande, con un valor esperado de $14.2 millones, es la decisión recomendada.

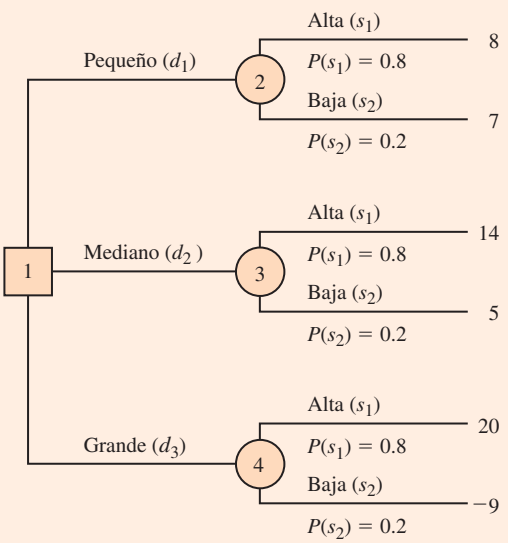
Los cálculos necesarios para identificar la alternativa de decisión con el mejor valor esperado pueden ser convenientemente efectuados con un árbol de decisión. En la figura 21.2 se muestra el árbol de decisión para el problema de PDC con las probabilidades en las ramas de los estados de la naturaleza. Al trabajar de atrás para adelante en el árbol, primero se calcula el valor esperado en cada nodo; es decir, en cada nodo se pondera cada posible pago multiplicado por la probabilidad de que ocurra. De este modo se obtiene el valor esperado para los nodos 2, 3 y 4 que se muestra en la figura 21.3.

Debido a que quien toma decisiones controla la rama que sale del nodo 1 de decisión y se está tratando de maximizar la utilidad esperada, la mejor alternativa de decisión en el nodo 1 es d_3. Por tanto, el análisis del árbol de decisión lleva a la recomendación de d_3 con un valor esperado de $14.2 millones. Observe que esta recomendación también se obtiene con el método del valor esperado junto con la tabla de pagos.

Existe software para apoyar la construcción de árboles de decisión más complejos.

Otros problemas de decisión pueden ser bastante más complejos que el de PDC, pero si un número razonable de alternativas de decisión y de estados de la naturaleza está presente, se podrá utilizar el método del árbol de decisión que se describe aquí. En primer lugar, dibuje un árbol que incluya nodos de decisión, nodos aleatorios y ramas que describan el carácter secuencial del problema. Si se utiliza el método del valor esperado, el siguiente paso es determinar las

FIGURA 21.2 Árbol de decisión del problema de PDC con las probabilidades de estado de la naturaleza

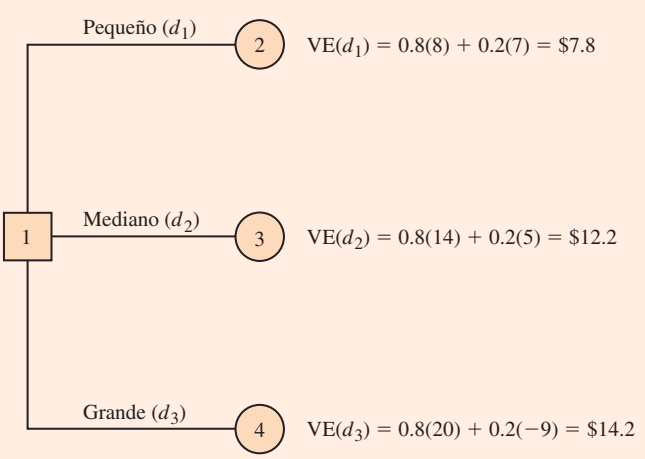


probabilidades de cada uno de los estados y calcular el valor esperado en cada nodo aleatorio. Después se elige la rama de decisión que lleva al nodo aleatorio con el mayor valor esperado. La alternativa de decisión que corresponde a esta rama es la decisión recomendada.

Valor esperado de la información perfecta

Suponga que PDC tiene la oportunidad de realizar un estudio de investigación de mercados que le ayudará a evaluar el interés del comprador por el proyecto del condominio y que proporcionará a los gerentes la información que podrían utilizar a efecto mejorar las evaluaciones de las probabilidades para los estados de la naturaleza. Para determinar el valor potencial de esta información, se empieza por suponer que el estudio podría proporcionar *información perfecta*

FIGURA 21.3 Aplicación del método del valor esperado utilizando árboles de decisión



sobre los estados de la naturaleza; es decir, se asume que por el momento PDC podría determinar con certeza, antes de tomar una decisión, qué estado ocurrirá. Para hacer uso de la información perfecta, se desarrollará una estrategia de decisión que PDC debe seguir una vez que sabe qué estado de la naturaleza ocurrirá. Una estrategia de decisión es simplemente una regla de decisión que especifica la alternativa a elegir después de disponer de la nueva información.

Para ayudar a determinar la estrategia de decisión que deberá seguir PDC, se reproduce la tabla de pagos 21.2 de PDC. Observe que si la inmobiliaria supiera con seguridad que se presentaría el estado s_1, la mejor alternativa de decisión sería d_3, cuyo pago sería de \$20 millones. Del mismo modo, si supiera con seguridad que se presentaría el estado s_2, la mejor alternativa de decisión sería d_1, con un pago de \$7 millones. Por consiguiente, se podría determinar la estrategia de decisión óptima si PDC contara con la información perfecta de la siguiente manera.

Si s_1, selecciona d_3 y obtiene un pago de \$20 millones
Si s_2, selecciona d_1 y obtiene un pago de \$7 millones

¿Cuál es el valor esperado de esta estrategia de decisión? Para calcularlo con la información perfecta, vuelva a las probabilidades originales de los estados de la naturaleza: $P(s_1) = 0.8$ y $P(s_2) = 0.2$. Por tanto, existe una probabilidad de 0.8 de que la información perfecta indicará el estado de la naturaleza s_1, y la alternativa de decisión resultante, d_3, dará una utilidad de \$20 millones. Del mismo modo, con una probabilidad de 0.2 para el estado de la naturaleza s_2, la alternativa para una óptima decisión d_1 proporcionará una utilidad de \$7 millones. Así que, al utilizar la ecuación (21.3) el valor esperado de la estrategia de decisión basada en la información perfecta es

$$0.8(20) + 0.2(7) = 17.4$$

A este valor esperado de \$17.4 millones se le conoce como *valor esperado con la información perfecta* (VECIP).

Antes en esta sección se demostró que la decisión recomendada al utilizar el método del valor esperado era la alternativa de decisión d_3, con un valor esperado de \$14.2 millones. Debido a que esta recomendación y el cálculo del valor esperado se efectuaron sin la ventaja de la información perfecta, a los \$14.2 millones se les conoce como *valor esperado sin la información perfecta* (VESIP).

Representará un valor de \$3.2 millones para PDC conocer el nivel de aceptación en el mercado antes de elegir una alternativa de decisión.

El valor esperado con la información perfecta es \$17.4 millones y el valor esperado sin la información perfecta es \$14.2 millones; por tanto, el valor esperado de la información perfecta (VEIP) de \$17.4 − \$14.2 = \$3.2 millones. En otras palabras, \$3.2 millones representan el valor esperado adicional que se puede obtener si se contara con la información de los estados de la naturaleza. En términos generales, una investigación de mercados no proporcionará información "perfecta"; sin embargo, si el estudio es bueno, la información recabada valdría una parte considerable de los \$3.2 millones. Dado que el VEIP es \$3.2 millones, PDC puede considerar seriamente un estudio de investigación de mercados con objeto de obtener más información acerca de los estados de la naturaleza.

TABLA 21.2 Tabla de pagos para el proyecto de condominio de PDC: (\$ millones)

	Estado	
Alternativa de decisión	**Demanda alta s_1**	**Demanda baja s_2**
Complejo pequeño, d_1	8	7
Complejo mediano, d_2	14	5
Complejo grande, d_3	20	−9

En general, el **valor esperado de la información perfecta** se calcula como sigue.

VALOR ESPERADO DE LA INFORMACIÓN PERFECTA

$$\text{VEIP} = |\text{VEcIP} - \text{VEsIP}| \qquad (21.4)$$

donde

VEIP = valor esperado de la información perfecta
VEcIP = valor esperado *con* la información perfecta acerca de los estados de la naturaleza
VEsIP = valor esperado *sin* la información perfecta acerca de los estados de la naturaleza

Observe el papel del valor absoluto en la ecuación (21.4). Para la minimización de los problemas, la información ayuda a reducir o bajar los costos, por lo que el valor esperado con la información perfecta es menor o igual que el valor esperado sin la información perfecta. En este caso, VEIP es la magnitud de la diferencia entre VEcIP y VEsIP, o el valor absoluto de la diferencia, como se muestra en la ecuación (21.4).

Ejercicios

Métodos

AUTOevaluación

1. La siguiente tabla de pagos muestra las utilidades para el análisis de decisiones de un problema con dos alternativas de decisión y tres estados de la naturaleza.

Alternativa de decisión	Estados de la naturaleza		
	s_1	s_2	s_3
d_1	250	100	25
d_2	100	100	75

   a) Construya un árbol de decisión para este problema.
   b) Suponga que quien debe tomar la decisión obtiene las probabilidades $P(s_1) = 0.65$; $P(s_2) = 0.15$, y $P(s_3) = 0.20$. Utilice el método del valor esperado para determinar la decisión óptima.

2. Una persona que debe tomar una decisión afronta cuatro alternativas de decisión y cuatro estados de la naturaleza, por lo que desarrolla la siguiente tabla de pagos.

Alternativa de decisión	Estados de la naturaleza			
	s_1	s_2	s_3	s_4
d_1	14	9	10	5
d_2	11	10	8	7
d_3	9	10	10	11
d_4	8	10	11	13

   La persona que toma la decisión obtiene la información que le permite realizar las evaluaciones de las probabilidades siguientes: $P(s_1) = 0.5$; $P(s_2) = 0.2$; $P(s_3) = 0.2$, y $P(s_1) = 0.1$.
   a) Utilice el método del valor esperado para determinar la solución óptima.
   b) Ahora suponga que las entradas de la tabla de pagos son los costos. Utilice el método del valor esperado para determinar la decisión óptima.

Aplicaciones

AUTO evaluación

3. Hudson Corporation está considerando tres opciones para el manejo de sus operaciones de procesamiento de datos: continuar con su personal, contratar a un proveedor externo para realizar la gestión (lo que se conoce como un *reclutador externo*) o utilizar una combinación de su personal y un proveedor externo. El costo de la operación depende de la demanda futura. El costo anual de cada opción (en miles de dólares) depende de la demanda de la siguiente forma.

	Demanda		
Opciones de personal	**Alta**	**Media**	**Baja**
Personal propio	650	650	600
Empresa externa	900	600	300
Combinación	800	650	500

a) Si las probabilidades para la demanda son 0.2, 0.5 y 0.3, respectivamente, ¿qué alternativa de decisión minimizará el costo esperado de operación del procesamiento de datos? ¿Cuál es el costo anual esperado de su recomendación?

b) ¿Cuál es el valor esperado de una información perfecta?

4. Myrtle Air Express decidió ofrecer un servicio directo desde Cleveland a Myrtle Beach. La gerencia debe decidir entre un servicio de primera a precio completo utilizando las nuevas flotas de jet de la empresa, y un servicio de descuento utilizando aviones regionales de menor capacidad. Es claro que la mejor elección depende de la reacción del mercado al servicio que ofrezca Myrtle Air. Los gerentes han desarrollado estimaciones de la aportación a las utilidades de cada tipo de servicio basado en los dos posibles niveles de demanda del servicio de Myrtle Beach: alta o baja. La siguiente tabla muestra las utilidades trimestrales estimadas (en miles de dólares).

	Demanda del servicio	
Servicio	**Alta**	**Baja**
Tarifa completa	$960	−$490
Descuento	$670	$320

a) ¿Qué decisión debe tomarse? ¿Cuál es el evento aleatorio, y cuál es la consecuencia para este problema? ¿Cuántas alternativas de decisión existen? ¿Cuántos resultados existen en el evento aleatorio?

b) Suponga que el gerente de Myrtle Air Express cree que la probabilidad de que la demanda sea alta es de 0.7 y de que sea baja es de 0.3. Utilice el método del valor esperado para determinar cuál es la decisión óptima.

c) Suponga que la probabilidad de que la demanda sea alta es 0.8 y de que sea baja es 0.2. ¿Cuál es la decisión óptima al utilizar el método del valor esperado?

5. La distancia de Potsdam a los grandes mercados y el servicio aéreo limitado han impedido atraer nueva industria a la ciudad. Air Express, un importante servicio de entrega nocturna, está considerando establecer un centro de distribución regional en Potsdam, pero no establecerá el centro a menos que aumente la longitud de la pista del aeropuerto local. Otro candidato para el nuevo desarrollo es Diagnostic Research, Inc. (DRI), uno de los principales fabricantes de equipo de pruebas médicas. DRI está considerando construir una nueva planta de fabricación. Ampliar la pista aérea no es un requerimiento para DRI, pero la comisión de planeación siente que ayudaría a convencer a DRI de establecer su nueva planta en Potsdam. Suponiendo que la

ciudad amplíe la longitud de la pista, la comisión de planeación de Potsdam considera que las probabilidades de la siguiente tabla son aplicables.

	Planta de DRI	Ninguna planta de DRI
Central de Air Express	0.30	0.10
Ninguna central de Air Express	0.40	0.20

Por ejemplo, la probabilidad de que Air Express estableciera un centro de distribución y DRI construyera una planta es de 0.30.

El ingreso anual estimado para la ciudad después de haber deducido el costo de la ampliación de la pista es el siguiente.

	Planta de DRI	Ninguna planta de DRI
Central de Air Express	\$600 000	\$150 000
Ninguna central de Air Express	\$250 000	−\$200 000

Si el proyecto de ampliación de la pista no se lleva a cabo, la comisión de planeación evalúa en 0.6 la probabilidad de que DRI estableciera su nueva planta en Potsdam; en este caso, el ingreso anual estimado para la ciudad sería de \$450 000. Si el proyecto de expansión de la pista no se lleva a cabo y DRI no establece su planta en Potsdam, el ingreso anual sería \$0, ya que no se incurrirá en ningún gasto y no se esperarán ingresos.

a) ¿Cuál es la decisión a tomar, el evento aleatorio y la consecuencia?
b) Calcule el ingreso anual esperado asociado con la alternativa de decisión para aumentar la longitud de la pista.
c) Calcule el ingreso anual esperado relacionado con la alternativa de decisión para no aumentar la longitud de la pista.
d) ¿El pueblo debería elegir aumentar la longitud de la pista? Explique su respuesta.
e) Suponga que las probabilidades relacionadas con la ampliación de la terminal aérea fueran las siguientes.

	Planta de DRI	Ninguna planta de DRI
Central de Air Express	0.40	0.10
Ninguna central de Air Express	0.30	0.20

¿Qué efecto, si hay alguno, tendría este cambio en las probabilidades sobre la decisión recomendada?

6. Seneca Hill Winery compró recientemente un terreno con el propósito de establecer un nuevo viñedo. El director está considerando dos variedades de uvas blancas para su nueva plantación: Chardonnay y Riesling. Las uvas Chardonnay se usarán para producir un vino seco Chardonnay, y las uvas Riesling para producir un vino Riesling semiseco. Esto toma aproximadamente cuatro años desde el momento en que se plantan las nuevas uvas hasta que puedan ser cosechadas. Esta extensión del tiempo crea un gran problema de incertidumbre con respecto a la futura demanda y hace que la decisión sobre el tipo de uvas a plantar sea difícil. Se están considerando tres posibilidades: sólo las uvas Chardonnay, sólo las uvas Riesling, o ambas. La gerencia de Seneca ha decidido que para los propósitos de planeación sería adecuado considerar sólo dos posibilidades de demanda para cada tipo de vino: la alta y la baja. Con las dos

posibilidades para cada tipo de vino fue necesario evaluar cuatro posibilidades. Apoyada en algunos pronósticos de publicaciones industriales, la gerencia desarrolló las siguientes evaluaciones de probabilidad.

	Demanda de Riesling	
Demanda de Chardonnay	**Baja**	**Alta**
Baja	0.05	0.50
Alta	0.25	0.20

Las proyecciones de ingresos muestran una contribución anual de $20 000 si Seneca Hill sólo planta uva Chardonnay y la demanda para este vino es baja, y de $70 000 si sólo planta dicha variedad y la demanda es alta. Si planta sólo uvas Riesling, la proyección anual de utilidades es de $25 000 si la demanda es baja para estas uvas, y de $45 000 si la demanda es alta. Si Seneca planta ambos tipos de uvas, las proyecciones anuales de utilidades se muestran en la siguiente tabla.

	Demanda de Riesling	
Demanda de Chardonnay	**Baja**	**Alta**
Baja	$22 000	$40 000
Alta	$26 000	$60 000

a) ¿Cuál es la decisión a tomar, el evento aleatorio y la consecuencia? Identifique las alternativas para las decisiones y los posibles resultados de los eventos aleatorios.
b) Elabore un árbol de decisión.
c) Utilice el método del valor esperado para recomendar qué alternativa debe seguir Seneca Hill Winery para maximizar la utilidad anual esperada.
d) Suponga que la gerencia está interesada en las probabilidades estimadas si la demanda para el vino Chardonnay es alta. Algunos suponen que es probable que la demanda para Riesling también sea alta en este caso. Suponga que la probabilidad de que la demanda sea alta para Chardonnay y baja para Riesling es de 0.05, y que la probabilidad de una demanda alta tanto para Chardonnay como para Riesling sea de 0.40. ¿Cómo cambiaría esto la decisión recomendada? Suponga que las probabilidades de que la demanda de Chardonnay sea baja se mantienen entre 0.05 y 0.50.
e) Otros miembros del equipo gerencial esperan que el mercado de Chardonnay se sature en algún punto en el futuro, causando la caída de los precios. Suponga que las proyecciones anuales de utilidades caigan a $50 000 si la demanda para Chardonnay es alta y sólo se siembra este tipo de uvas. Utilice las estimaciones de probabilidades iniciales y determine cuánto afectaría este cambio a la decisión óptima.

7. El Lake Placid Town Council ha decidido construir un nuevo centro comunitario para convenciones, conciertos y otros eventos públicos, pero existe una controversia importante respecto del tamaño apropiado. Muchos ciudadanos influyentes desean un centro grande que sea un escaparate para la zona, pero el alcalde cree que si la demanda no apoya dicho centro, la comunidad perdería una gran cantidad de dinero. A efecto de proporcionar una estructura para el proceso de decisión, el consejo ha reducido las alternativas de construcción a tres tamaños: pequeño, mediano y grande. Todos estuvieron de acuerdo en que el factor crítico al elegir el mejor tamaño es el número de personas que desean usar la nueva instalación. Un consultor de planeación regional proporcionó estimaciones de la demanda bajo tres escenarios: el peor de los casos, el caso base y el mejor de los casos. El peor de los casos corresponde a la situación en la que la derrama turística disminuya de forma significativa; el caso base indica una situación en la que Lake Placid continúa atrayendo la misma cantidad de visitantes, y el mejor

escenario indica un incremento turístico significativo. El consultor ha proporcionado las evaluaciones de probabilidades 0.10, 0.60 y 0.30 para el escenario del peor de los casos, el caso base y el mejor de los casos, respectivamente.

El consejo del pueblo ha sugerido utilizar el flujo de efectivo neto para un horizonte de planeación de cinco años como criterio para decidir el tamaño adecuado. Un consultor desarrolló las siguientes proyecciones del flujo de efectivo neto (en miles de dólares) sobre un horizonte de planeación de cinco años. Todos los costos, incluidos los honorarios del consultor, han sido considerados.

	Escenario de la demanda		
Tamaño del centro	El peor de los casos	Caso base	El mejor de los casos
Pequeño	400	500	660
Mediano	−250	650	800
Grande	−400	580	990

a) ¿Qué decisión deberá tomar Lake Placid con el método del valor esperado?

b) Calcule el valor esperado de la información perfecta. ¿Considera que será útil tratar de obtener información adicional con respecto a qué escenario es probable que se presente?

c) Suponga que la probabilidad del escenario del peor de los casos aumentara a 0.2, la probabilidad del escenario base disminuyera a 0.5, y la probabilidad del mejor escenario permaneciera en 0.3 ¿Qué efecto, si hubiera alguno, tendrían estos cambios sobre la decisión recomendada?

d) El consultor sugiere que un gasto de $150 000 en una campaña promocional sobre el horizonte de planeación efectivamente reduciría a cero la probabilidad del peor de los casos. Si se espera que la campaña aumente la probabilidad del escenario del mejor de los casos a 0.4, ¿es ésta una buena inversión?

21.3 Análisis de decisiones con información muestral

Al aplicar el método del valor esperado se mostró cómo la información de probabilidad acerca de los estados de la naturaleza afecta los cálculos del valor esperado y, por tanto, la recomendación de decisión. Suele suceder, con frecuencia, que quien toma la decisión cuente con las evaluaciones preliminares o **probabilidad previa** para los estados de la naturaleza, que son los mejores valores de probabilidad disponibles en ese momento. Sin embargo, para tomar la mejor decisión posible, puede desear obtener información adicional de dichos estados. Esta nueva información podría utilizarse para revisar o actualizar las probabilidades previas, de modo que la decisión final se base en probabilidades más certeras para los estados de la naturaleza. Con frecuencia, la información adicional se obtiene mediante experimentos diseñados para proporcionar **información muestral** acerca de los estados. El muestreo de materia prima, la prueba de productos y los estudios de investigación de mercados son ejemplos de experimentos (o estudios) que permiten a la gerencia revisar o actualizar las probabilidades del estado de la naturaleza. Estas revisiones se llaman **probabilidades posteriores**.

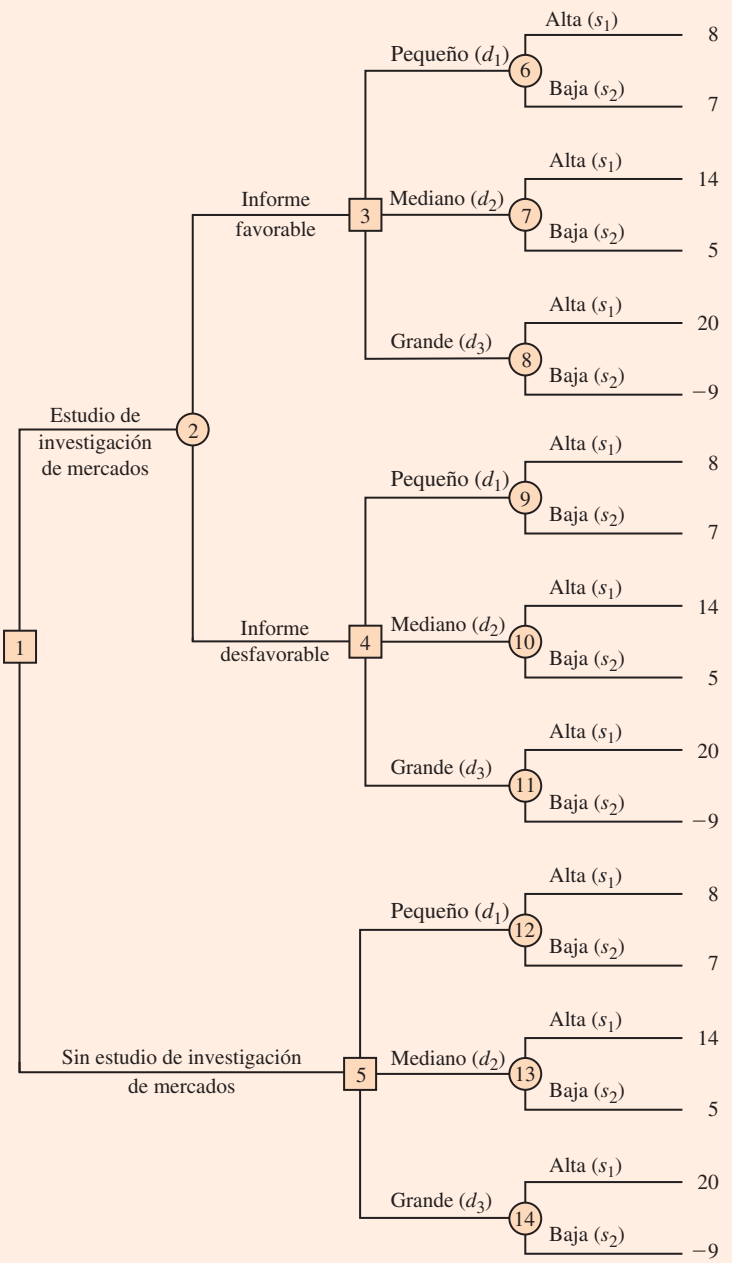
De regreso al problema de PDC, suponga que la gerencia considera un estudio de investigación de mercados de seis meses de duración para conocer más acerca de la aceptación potencial del proyecto de los condominios en el mercado inmobiliario. La gerencia prevé que el estudio proporcionará uno de los dos siguientes resultados.

1. Un informe favorable: un número significativo de las personas entrevistadas expresa interés por comprar un condominio de PDC.
2. Informe desfavorable: muy pocas de las personas entrevistadas expresan interés por comprar un condominio de PDC.

Árbol de decisión

En la figura 21.4 del árbol de decisión para el problema de PDC con información muestral se observa la secuencia lógica para la toma de decisiones y los eventos aleatorios. Primero, la gerencia de PDC debe decidir si el estudio de investigación de mercados debe llevarse a cabo. Si éste se realiza, debe estar preparada para tomar la decisión acerca del tamaño del proyecto de condominio si el informe de investigación de mercados es favorable y, probablemente, una decisión distinta acerca de dicho tamaño si el informe es desfavorable.

FIGURA 21.4 Árbol de decisión de PDC que incluye el estudio de investigación de mercados



En la figura 21.4, los cuadrados son los nodos de decisión y los círculos son los nodos aleatorios. En cada nodo de decisión, la rama del árbol que se tome depende de la decisión asumida. En cada nodo aleatorio, la rama que se tome depende de la probabilidad. Por ejemplo, el nodo de decisión 1 muestra que PDC debe primero determinar si lleva a cabo el estudio de investigación de mercados. Si éste se emprende, el nodo de probabilidad 2 indica que las ramas del informe tanto favorable como desfavorable no están bajo control de PDC y estarán determinadas por la casualidad. El nodo 3 es un nodo de decisión que indica que si el informe de investigación de mercados es favorable, PDC debe tomar la decisión de edificar un complejo pequeño, mediano o grande. El 4 es un nodo de decisión que muestra que si el informe de investigación de mercados es desfavorable, PDC debe decidir si construye un complejo pequeño, mediano o grande. El 5 es un nodo de decisión que denota que PDC debe determinar si construye el complejo pequeño, mediano o grande si el informe de investigación de mercados no se lleva a cabo. Los nodos 6 al 14 son aleatorios, y marcan que las ramas de los estados para la alta o la baja demanda serán determinadas por la casualidad.

En la sección 21.4 se explica cómo pueden desarrollarse estas probabilidades.

El análisis del árbol de decisión y la oportunidad de una estrategia óptima requieren que se conozcan las probabilidades correspondientes a todos los nodos aleatorios. PDC desarrolló las siguientes probabilidades.

Si el estudio de investigación de mercados se realiza,

$$P(\text{informe favorable}) = P(F) = 0.77$$
$$P(\text{informe desfavorable}) = P(U) = 0.23$$

Si el informe de investigación de mercados es favorable,

$$P(\text{demanda alta dado un informe favorable}) = P(s_1|F) = 0.94$$
$$P(\text{demanda baja dado un informe favorable}) = P(s_2|F) = 0.06$$

Si el informe de investigación de mercados es desfavorable,

$$P(\text{demanda alta dado un informe desfavorable}) = P(s_1|U) = 0.35$$
$$P(\text{demanda baja dado un informe desfavorable}) = P(s_2|U) = 0.65$$

Si el informe de investigación de mercados no se realiza, las probabilidades previas son aplicables.

$$P(\text{demanda alta}) = P(s_1) = 0.80$$
$$P(\text{demanda baja}) = P(s_2) = 0.20$$

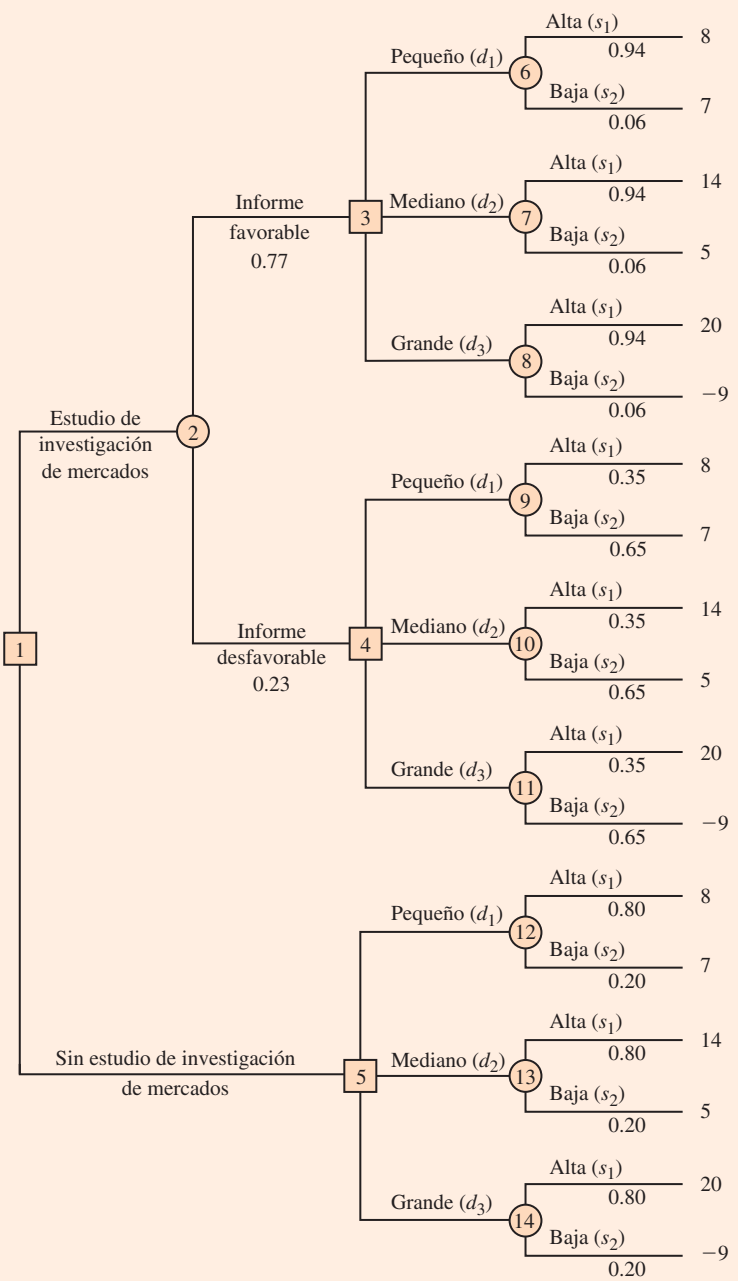
Las probabilidades anteriores se muestran en el árbol de decisión de la figura 21.5.

Estrategia de decisión

Una **estrategia de decisión** es una secuencia de decisiones y resultados aleatorios en la que las decisiones que se toman dependen de los resultados de los eventos aleatorios aún por determinar. El método para determinar la estrategia de una decisión óptima está basado en recorrer el árbol de decisión de atrás hacia adelante utilizando los siguientes pasos.

1. En los nodos aleatorios, calcule el valor esperado multiplicando el pago al final de cada rama por la correspondiente probabilidad.
2. En los nodos de decisión, seleccione la rama de decisión que lleve al mejor valor esperado. Éste se convertirá en el valor esperado en el nodo de decisión.

FIGURA 21.5 Árbol de decisión de PDC con probabilidades



Inicie el recorrido de atrás hacia adelante con el cálculo de los valores esperados de los nudos de probabilidad 6 a 14 que proporciona los siguientes resultados.

$$\begin{aligned}
\text{VE(nodo 6)} &= 0.94(8) + 0.06(7) = 7.94\\
\text{VE(nodo 7)} &= 0.94(14) + 0.06(5) = 13.46\\
\text{VE(nodo 8)} &= 0.94(20) + 0.06(-9) = 18.26\\
\text{VE(nodo 9)} &= 0.35(8) + 0.65(7) = 7.35\\
\text{VE(nodo 10)} &= 0.35(14) + 0.65(5) = 8.15\\
\text{VE(nodo 11)} &= 0.35(20) + 0.65(-9) = 1.15\\
\text{VE(nodo 12)} &= 0.80(8) + 0.20(7) = 7.80\\
\text{VE(nodo 13)} &= 0.80(14) + 0.20(5) = 12.20\\
\text{VE(nodo 14)} &= 0.80(20) + 0.20(-9) = 14.20
\end{aligned}$$

La figura 21.6 muestra el árbol de decisión reducido luego de calcular los valores esperados de estos nodos aleatorios.

Después se continúa con los nodos de decisión 3, 4 y 5. En cada uno, seleccione la rama de la alternativa de decisión que conduzca al mejor valor esperado. Por ejemplo, en el nodo 3 se puede elegir entre las ramas del complejo pequeño para la que el VE (nodo 6) = 7.94; la rama del complejo mediano para la que el VE (nodo 7) = 13.46, y la rama del complejo grande para la que el VE (nodo 8) = 18.26. Por tanto, se selecciona la rama de alternativa de decisión del complejo mayor y el valor esperado en el nodo 3 se convierten en VE (nodo 3) = 18.26.

Para el nodo 4, seleccione el mejor valor esperado de los nodos 9, 10 y 11. La mejor alternativa de decisión es la rama del complejo mediano, que proporciona el VE (nodo 4) = 8.15. En el nodo 5, seleccione el mejor valor esperado de los nodos 12, 13 y 14. La mejor alternativa de decisión es la rama del complejo grande que proporciona el VE (nodo 5) = 14.20. La figura 21.7 muestra el árbol de decisión reducido después de elegir las mejores opciones en los nodos 3, 4 y 5.

Ahora el valor esperado en el nodo aleatorio 2 puede calcularse de la siguiente forma.

$$\begin{aligned}
\text{VE(nodo 2)} &= 0.77\text{VE(nodo 3)} + 0.23\text{VE(nodo 4)}\\
&= 0.77(18.26) + 0.23(8.15) = 15.93
\end{aligned}$$

Este cálculo reduce el árbol de decisión a sólo dos ramas de decisión que salen del nodo 1 (vea la figura 21.8).

Finalmente, la decisión puede tomarse en el nodo de decisión 1, al elegir el mejor valor esperado desde los nodos 2 y 5. Esta acción lleva a la alternativa de decisión que conduce al estudio de investigación de mercados, el cual proporciona un valor esperado general de 15.93.

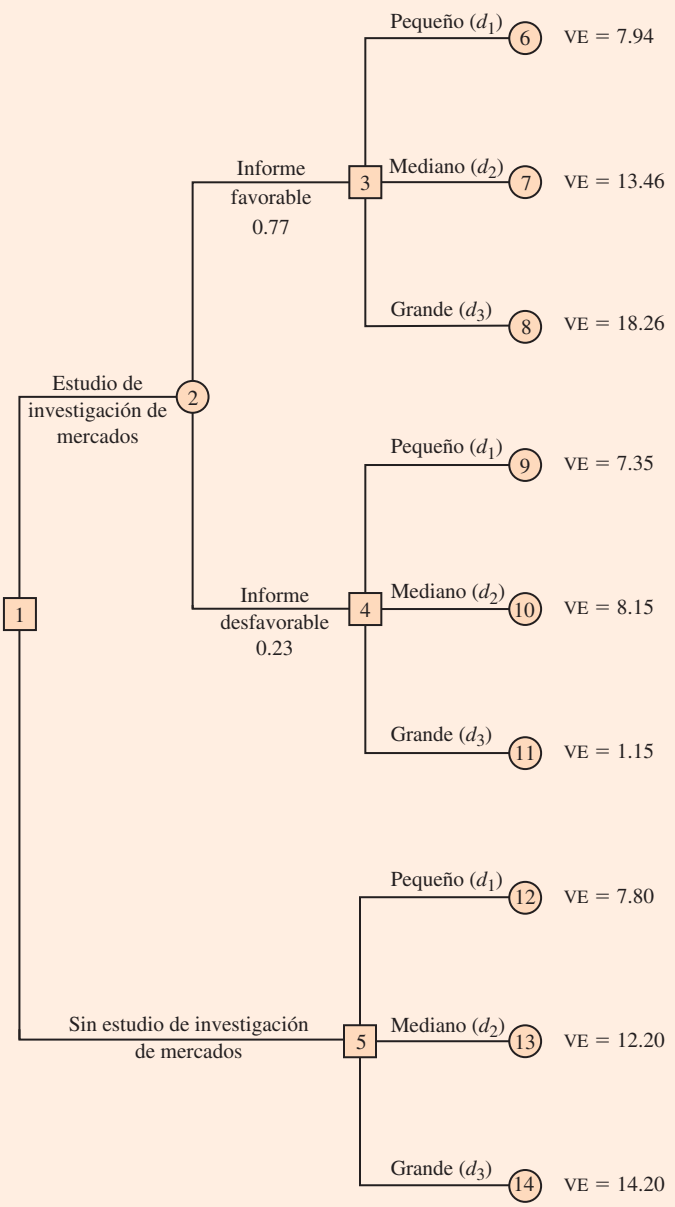
La decisión óptima para PDC es llevar a cabo el estudio de investigación de mercados y después manejar la siguiente estrategia de decisión.

Si la investigación de mercados es favorable, construir el complejo de condominios grande

Si la investigación de mercados es desfavorable, construir el complejo de condominios mediano

El análisis del árbol de decisión de PDC ilustra los métodos que pueden utilizarse para analizar los problemas más complejos de decisión secuencial. Primero, dibuje un árbol de decisión que consista de nodos de decisión, nodos aleatorios y las ramas que describan el carácter secuencial del problema. Determine las probabilidades para todos los resultados aleatorios. Después, al trabajar de atrás hacia adelante a través del árbol, calcule el valor esperado en todos los nodos aleatorios y seleccione la rama de mejor decisión en todos los nodos de decisión. La secuencia de las ramas de decisión óptima determina la estrategia de decisión óptima para el problema.

FIGURA 21.6 Árbol de decisión después de calcular los valores esperados en los nodos aleatorios 6 al 14

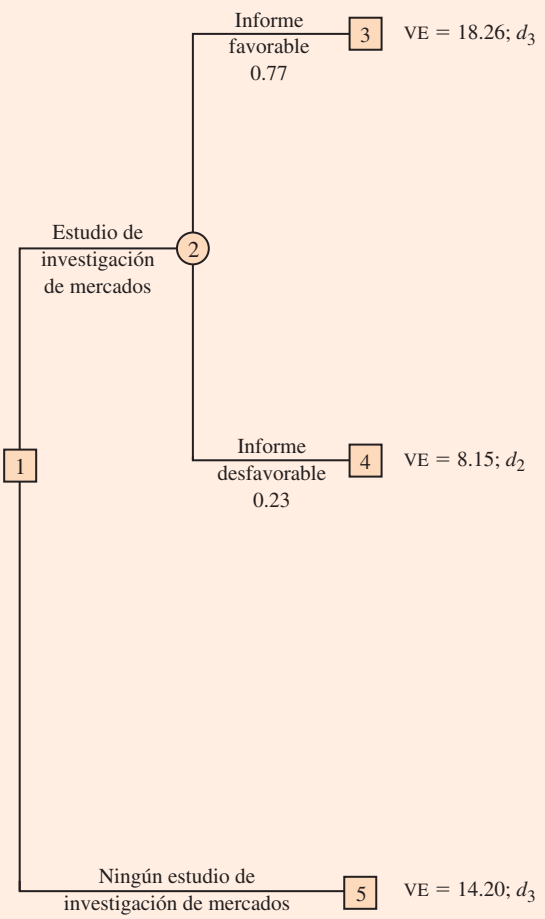


El VEIM = $1,73 millones sugiere que PDC deberá estar dispuesto a pagar hasta $1.73 millones para que se lleve a cabo el estudio de investigación de mercados.

Valor esperado de la información muestral

En el problema de PDC, el estudio de investigación de mercados es la información muestral utilizada para determinar la estrategia de decisión óptima. El valor esperado relacionado con el estudio es de $15.93. En la sección 21.3 se mostró que el mejor valor esperado si el estudio de investigación de mercados *no* se realiza es $14.20. Por tanto, se concluye que la diferencia, $15.93 − $14.20 = $1.73, es el **valor esperado de la información muestral (VEIM)**. En otras

FIGURA 21.7 Árbol de decisión de PDC después de elegir la mejor decisión en los nodos 3, 4 y 5.



palabras, se concluye que la realización del estudio añade $1.73 millones al valor esperado de PDC. En general, el valor esperado de la información muestral es la siguiente.

VALOR ESPERADO DE LA INFORMACIÓN MUESTRAL

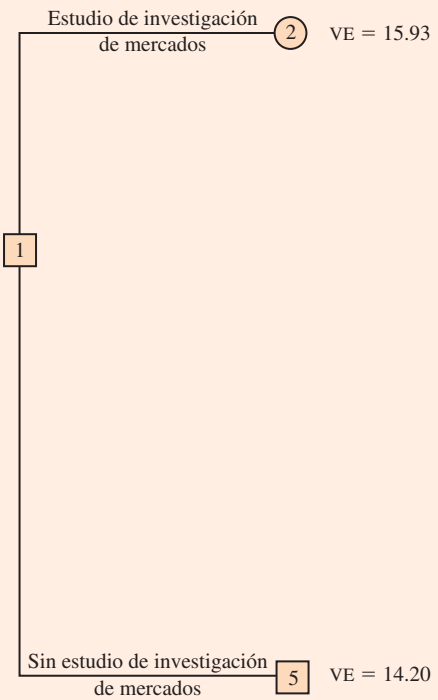
$$\text{VEIM} = |\text{VEcIM} - \text{VEsIM}| \qquad \textbf{(21.5)}$$

donde

VEIM = valor esperado de la información muestral
VEcIM = valor esperado *con* información muestral acerca de los estados de la naturaleza
VEsIM = valor esperado *sin* información muestral acerca de los estados de la naturaleza

Tome en cuenta el papel del valor absoluto en la ecuación (21.5). Para los problemas de minimización, el valor esperado con información muestral siempre es menor o igual que el

FIGURA 21.8 Árbol de decisión de PDC reducido a dos ramas de decisión



valor esperado sin información muestral. En este caso, VEIM es la magnitud de la diferencia entre VEcIM y VEsIM; por tanto, al tomar el valor absoluto de la diferencia como se muestra en la ecuación (21.5), se pueden manejar tanto los casos de maximización como de minimización con una sola ecuación.

Ejercicios

Métodos

AUTOevaluación

8. Considere una variación del árbol de decisión de PDC de la figura 21.5. La empresa debe primero decidir si lleva a cabo el estudio de investigación de mercados. Si dicho estudio se realiza, los resultados pueden ser favorables (F) o desfavorables (U). Suponga que existen sólo dos alternativas de decisión d_1 y d_2, y dos estados de la naturaleza s_1 y s_2. La tabla de pagos siguiente muestra las utilidades.

	Estado de la naturaleza	
Alternativa de decisión	s_1	s_2
d_1	100	300
d_2	400	200

a) Muestre el árbol de decisión.
b) Utilice las siguientes probabilidades. ¿Cuál es la estrategia de decisión óptima?

$$P(F) = 0.56 \quad P(s_1 \mid F) = 0.57 \quad P(s_1 \mid U) = 0.18 \quad P(s_1) = 0.40$$
$$P(U) = 0.44 \quad P(s_2 \mid F) = 0.43 \quad P(s_2 \mid U) = 0.82 \quad P(s_2) = 0.60$$

Aplicaciones

9. Un inversionista de bienes raíces tiene la oportunidad de comprar un terreno que actualmente es zona residencial. Si la junta del consejo del condado aprueba la solicitud de modificar el uso de esta propiedad a zona comercial en el lapso de un año, el inversionista podrá rentar el terreno a una tienda de descuento que desea abrir una nueva sucursal. Sin embargo, si el cambio de uso de suelo no es aprobado, el inversionista tendrá que vender la propiedad y resentir una pérdida. Las utilidades (en miles de dólares) se muestran en la siguiente tabla de pagos.

	Estado de la naturaleza	
	Aprobación al cambio de uso de suelo	**No aprobación al cambio de uso de suelo**
Alternativa de decisión	s_1	s_2
Comprar, d_1	600	−200
No comprar, d_2	0	0

a) Si la probabilidad de que sea aprobado el cambio de uso de suelo es 0.5, ¿Cuál decisión se recomienda? ¿Cuál es la utilidad esperada?

b) El inversionista puede tener una opción al comprar el terreno. Bajo la opción, tiene el derecho de comprarlo en cualquier momento durante los próximos tres meses mientras investiga la posible resistencia a la propuesta de cambio de uso de suelo del área residencial.

$$\text{Sea } H = \text{alta resistencia al cambio de uso de suelo}$$
$$L = \text{baja resistencia al cambio de uso de suelo}$$

$$P(H) = 0.55 \qquad P(s_1 \mid H) = 0.18 \qquad P(s_2 \mid H) = 0.82$$
$$P(L) = 0.45 \qquad P(s_1 \mid L) = 0.89 \qquad P(s_2 \mid L) = 0.11$$

¿Cuál es la estrategia de decisión óptima si el inversionista aprovecha la opción del periodo para investigar más acerca de la resistencia del área residencial antes de tomar la decisión sobre la compra?

c) Si la opción le costará al inversionista $10 000 más, ¿debería tomarla? ¿Por qué? ¿Cuál será la cantidad máxima que estaría dispuesto a pagar por la opción?

10. Dante Development Corporation está considerando concursar por un contrato para la construcción de un nuevo complejo de oficinas. La figura 21.9 muestra el árbol de decisión preparado por uno de los analistas de Dante. En el nodo 1, la empresa tiene que decidir si concursa por el contrato. El costo de preparar el concurso es de $200 000. La rama superior del nodo 2 muestra que la empresa tiene 0.8 de probabilidad de ganar el contrato si concursa. Si gana el contrato, tendrá que pagar $2 000 000 para convertirse en socio del proyecto. El nodo 3 indica que la empresa considerará entonces realizar un estudio de investigación de mercados para pronosticar la demanda sobre las unidades de oficinas antes de iniciar la construcción. El costo del estudio es de $15 000. El nodo 4 es un nodo aleatorio que muestra los posibles resultados del estudio.

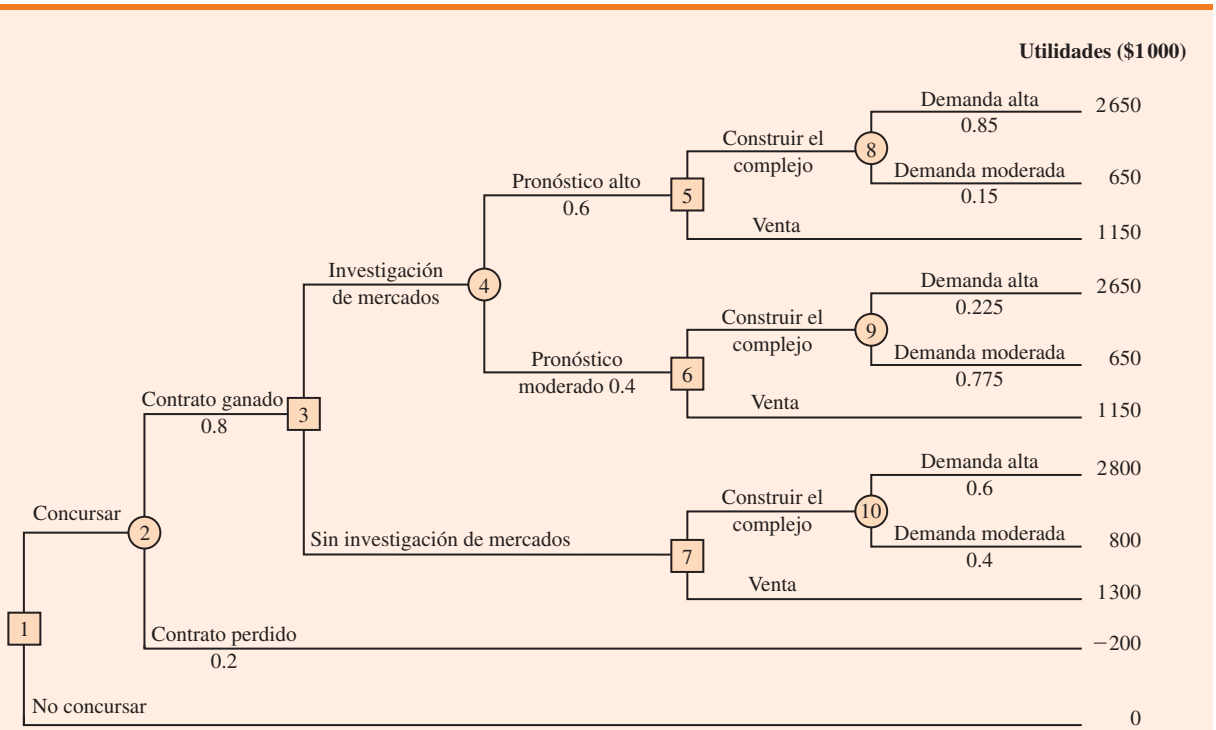
Los nodos 5, 6 y 7 son similares entre sí, dado que son los nodos de decisión para Dante, ya sea para construir el complejo de oficinas o vender los derechos del proyecto a otro desarrollador. La decisión de construir el complejo tendrá una utilidad de $5 000 000 si la demanda es alta y de 3 000 000 si es moderada. Si Dante elige vender sus derechos del proyecto a otro desarrollador, se estima que el ingreso por la venta será de $3 500 000. Las probabilidades mostradas en los nodos 4, 8 y 9 están basados en los resultados proyectados del estudio de investigación de mercados.

a) Verifique las proyecciones de utilidades de Dante mostradas al final de las ramas del árbol de decisión calculando los pagos de $2 650 000 y $650 000 para los dos primeros resultados.

b) ¿Cuál es la estrategia de decisión óptima para Dante, y cuál es la utilidad esperada para este proyecto?

c) ¿Qué costo tendría el estudio de investigación de mercados antes de que Dante modifique su decisión acerca de llevar a cabo el estudio?

FIGURA 21.9 Árbol de decisión para Dante Development Corporation



11. Hale's TV Productions está considerando producir un programa piloto para una serie de comedia con la esperanza de venderla a una cadena de televisión. La cadena puede decidir rechazar la serie, pero también comprar los derechos respectivos, ya sea por uno o dos años. En este punto, Hale tiene que producir el programa piloto y esperar la decisión de la cadena de televisión o transferir los derechos para el piloto y la serie a un competidor por $100 000. Las alternativas de decisión y las utilidades de Hale (en miles de dólares) son las siguientes.

	Estado de la naturaleza		
Alternativa de decisión	**Rechazo, s_1**	**1 año, s_2**	**2 años, s_3**
Producir un piloto, d_1	−100	50	150
Vender al competidor, d_2	100	100	100

Las probabilidades de los estados de la naturaleza son $P(s_1) = 0.2$, $P(s_2) = 0.3$ y $P(s_3) = 0.5$. Por una tarifa de consultoría de $5 000, una agencia revisará los planes para la serie de comedia e indicará las posibilidades globales de una reacción favorable de la cadena de televisión hacia la serie. Suponga que la revisión de la agencia resultará en una opción favorable (F) o desfavorable (U) y que las siguientes probabilidades son relevantes.

$$P(F) = 0.69 \quad P(s_1 \mid F) = 0.09 \quad P(s_1 \mid U) = 0.45$$
$$P(U) = 0.31 \quad P(s_2 \mid F) = 0.26 \quad P(s_2 \mid U) = 0.39$$
$$P(s_3 \mid F) = 0.65 \quad P(s_3 \mid U) = 0.16$$

a) Construya un árbol de decisión para este problema.
b) ¿Cuál es la decisión recomendada si la opinión de la agencia no es utilizada? ¿Cuál es el valor esperado?

c) ¿Cuál es el valor esperado de la información perfecta?
d) ¿Cuál es la estrategia de decisión óptima de Hale suponiendo que se utiliza la información de la agencia?
e) ¿Cuál es el valor esperado de la información de la agencia?
f) ¿Esta información vale los $5 000 pagados por honorarios? ¿Cuál es la cantidad máxima que Hale estaría dispuesto a pagar por la información?
g) ¿Cuál es la decisión recomendada?

12. Martin's Service Station está considerando participar en el negocio de quitanieve para la próxima temporada de invierno. La empresa puede comprar un accesorio para la cuchilla de su camión pick-up quitanieve o un nuevo camión para el trabajo pesado de retirar la nieve. Después de analizar la situación, Martin cree que cualquier alternativa será una inversión rentable si la nevada es muy fuerte. Las utilidades serán menores si ésta es moderada, y puede tener pérdidas si el resultado es una nevada ligera. En la siguiente tabla se aplican las utilidades y pérdidas.

	Estado de la naturaleza		
Alternativa de decisión	**Fuerte, s_1**	**Moderada, s_2**	**Ligera, s_3**
Aditamento, d_1	3 500	1 000	−1 500
Quitanieve nuevo, d_2	7 000	2 000	−9 000

Las probabilidades de los estados de la naturaleza son $P(s_1) = 0.4$, $P(s_2) = 0.3$ y $P(s_3) = 0.3$. Suponga que Martin decide esperar hasta septiembre antes de tomar una decisión final. Las valoraciones de las probabilidades relacionadas con un frío normal en septiembre (N) o un frío inesperado (U) son las siguientes.

$$P(N) = 0.8 \qquad P(s_1 \mid N) = 0.35 \qquad P(s_1 \mid U) = 0.62$$
$$P(U) = 0.2 \qquad P(s_2 \mid N) = 0.30 \qquad P(s_2 \mid U) = 0.31$$
$$P(s_3 \mid N) = 0.35 \qquad P(s_3 \mid U) = 0.07$$

a) Construya un árbol de decisión para este problema.
b) ¿Cuál es la decisión recomendada si Martin no espera hasta septiembre? ¿Cuál es el valor esperado?
c) ¿Cuál es el valor esperado de la información perfecta?
d) ¿Cuál es la estrategia de decisión óptima si Martin no toma la decisión sino hasta que se haya determinado el clima en septiembre? ¿Cuál es el valor esperado de esta estrategia de decisión?

13. Lawson's Department Store enfrenta la decisión de compra de un producto estacional para el cual la demanda puede ser alta, media o baja. El comprador para Lawson's puede ordenar 1, 2 o 3 lotes del producto antes de que la temporada inicie, pero no puede volver a ordenar más tarde. Las proyecciones de las utilidades (en miles de dólares) son las siguientes.

	Estado de la naturaleza		
Alternativa de decisión	**Demanda alta** s_1	**Demanda mediana** s_2	**Demanda baja** s_3
Orden de 1 lote, d_1	60	60	50
Orden de 2 lotes, d_2	80	80	30
Orden de 3 lotes, d_3	100	70	10

a) Si las probabilidades anteriores de los tres estados de la naturaleza son 0.3, 0.3 y 0.4, respectivamente, ¿cuál es la cantidad de lotes que se recomienda ordenar?
b) En cada reunión de ventas de pretemporada, el vicepresidente de ventas expresa una opinión personal respecto de la demanda potencial del producto. Debido al carácter entusiasta y optimista del vicepresidente, los pronósticos de las condiciones del mercado siempre han

sido "excelentes" (E) o "muy buenas" (M). Las probabilidades son las siguientes. ¿Cuál es la estrategia de decisión óptima?

$$P(E) = 0.7 \quad P(s_1 \mid E) = 0.34 \quad P(s_1 \mid M) = 0.20$$
$$P(M) = 0.3 \quad P(s_2 \mid E) = 0.32 \quad P(s_2 \mid M) = 0.26$$
$$P(s_3 \mid E) = 0.34 \quad P(s_3 \mid M) = 0.54$$

c) Calcule el VEIM y el VEIP. Discuta si la empresa debería considerar consultar a un experto que pudiera proporcionarle pronósticos independientes de las condiciones del mercado para el producto.

21.4 Cálculo de probabilidades mediante el teorema de Bayes

En la sección 21.3, las probabilidades para los nodos aleatorios del árbol de decisión de PDC fueron especificadas en la descripción del problema. Para determinar estas probabilidades no se requirieron cálculos. En esta sección se muestra cómo el **teorema de Bayes**, un tema abordado en el capítulo 4, puede utilizarse para calcular las probabilidades de los árboles de decisión.

El árbol de decisión de PDC se muestra nuevamente en la figura 21.10. Sea

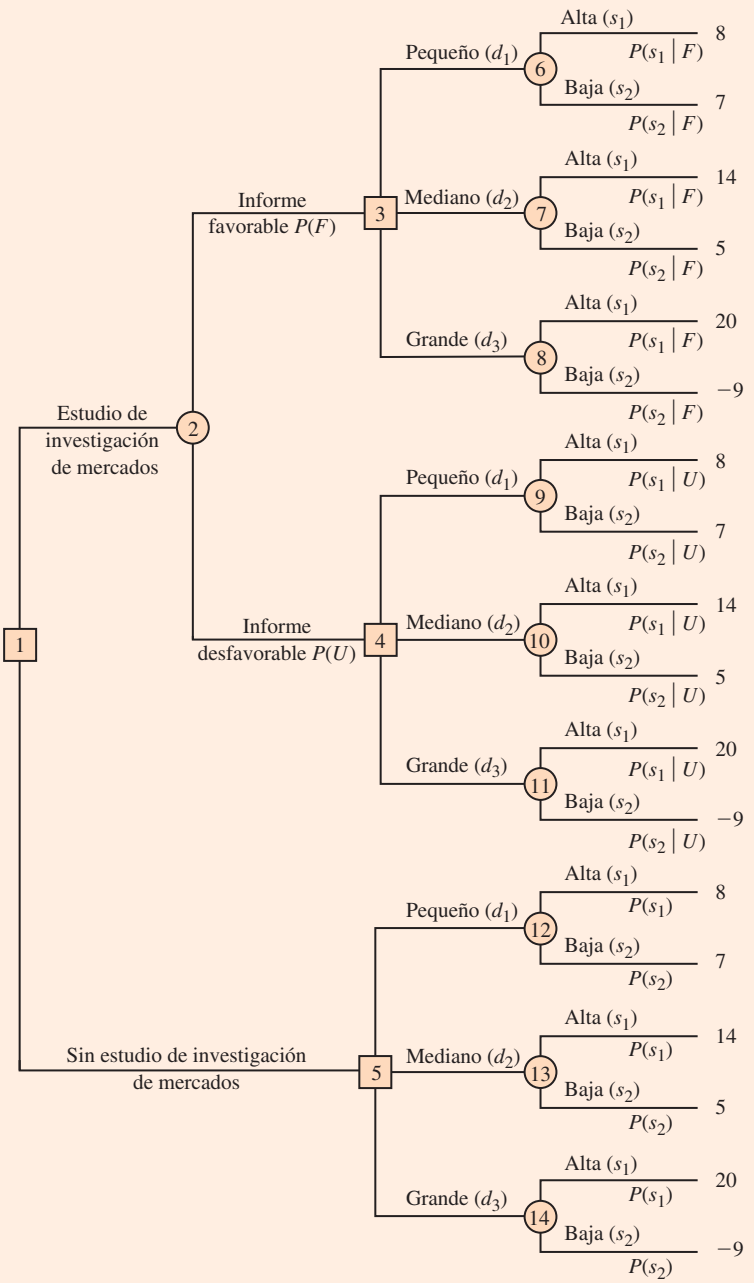
F = informe favorable de la investigación de mercados
U = informe desfavorable de la investigación de mercados
s_1 = demanda alta (estado de la naturaleza 1)
s_2 = demanda baja (estado de la naturaleza 2)

En el nodo aleatorio 2 se requiere conocer las probabilidades $P(F)$ y $P(U)$. En los nodos aleatorios 6, 7 y 8 se necesita conocer las probabilidades $P(s_1 \mid F)$, la probabilidad del estado de la naturaleza 1 dado un informe favorable de la investigación de mercados, y $P(s_2 \mid F)$, la probabilidad del estado de la naturaleza 2, dado un informe favorable de la investigación de mercados. $P(s_1 \mid F)$ y $P(s_2 \mid F)$ son referidos como *probabilidades posteriores* debido a que son probabilidades condicionales basadas en el resultado de la información muestral. En los nodos aleatorios 9, 10 y 11 se requiere conocer las probabilidades $P(s_1 \mid U)$ y $P(s_2 \mid U)$; observe que éstas también son probabilidades posteriores que denotan las probabilidades de los dos estados de la naturaleza *dado* que el informe de investigación de mercados es desfavorable. Finalmente, en los nodos aleatorios 12, 13 y 14 se requieren las probabilidades para los estados de la naturaleza, $P(s_1)$ y $P(s_2)$, si el estudio de investigación de mercados no se lleva a cabo.

Al realizar los cálculos de las probabilidades, es necesario conocer la evaluación de PDC para las probabilidades de los dos estados $P(s_1)$ y $P(s_2)$, la cual es previa, como se mencionó antes. Además, se debe conocer la **probabilidad condicional** de los resultados de la investigación de mercados (la información muestral) *dado* cada estado. Por ejemplo, se requiere conocer la probabilidad condicional de un informe favorable de la investigación de mercados considerando que existe una alta demanda para el proyecto de PDC; observe que esta probabilidad condicional de F dado el estado de la naturaleza s_1 se escribe $P(F \mid s_1)$. Para efectuar los cálculos respectivos se necesitarán las probabilidades condicionales para todos los resultados muestrales dados todos los estados de la naturaleza, es decir, $P(F \mid s_1)$, $P(F \mid s_2)$, $P(U \mid s_1)$ y $P(U \mid s_2)$. En el problema de PDC se supone que las estimaciones siguientes están disponibles para estas probabilidades condicionales.

	Estudio de investigación de mercados	
Estado de la naturaleza	***F*, favorable**	***U*, desfavorable**
Demanda alta, s_1	$P(F \mid s_1) = 0.90$	$P(U \mid s_1) = 0.10$
Demanda baja, s_2	$P(F \mid s_2) = 0.25$	$P(U \mid s_2) = 0.75$

FIGURA 21.10 Árbol de decisión de PDC



Observe que las estimaciones de probabilidad anteriores proporcionan un grado razonable de confianza en el estudio de investigación de mercados. Si el verdadero estado de la naturaleza es s_1, la probabilidad de que el informe del estudio sea favorable es 0.90, y la probabilidad de que sea desfavorable es 0.10. Si el verdadero estado es s_2, la probabilidad de que el informe de la investigación sea favorable es 0.25, y la probabilidad de que sea desfavorable es 0.75. La razón por la que la probabilidad de un potencialmente engañoso informe de investigación de mercados favorable sea 0.25 para el estado de la naturaleza s_2 radica en que cuando algunos compradores potenciales escuchen hablar acerca del nuevo proyecto de condominio, su entusiasmo los pueda

TABLA 21.3 Probabilidades del proyecto de condominio de PDC basadas en un informe favorable de la investigación de mercados

Estados de la naturaleza s_j	Probabilidades previas $P(s_j)$	Probabilidades condicionales $P(F \mid s_j)$	Probabilidades conjuntas $P(F \cap s_j)$	Probabilidades posteriores $P(s_j \mid F)$
s_1	0.8	0.90	0.72	0.94
s_2	0.2	0.25	0.05	0.06
	1.0		$P(F) = 0.77$	1.00

llevar a exagerar su verdadero interés por los condominios. La respuesta inicial favorable de un comprador potencial puede cambiar rápidamente a un "no gracias", cuando al final enfrente la realidad de firmar un contrato de compra y realizar un pago inicial.

En el siguiente análisis se presenta un planteamiento tabular como un método conveniente para calcular la probabilidad. Los cálculos para el problema de PDC basados en un informe favorable de la investigación de mercados (F) se resumen en la tabla 21.3. Los pasos utilizados para desarrollar esta tabla son los siguientes.

Paso 1. En la columna 1 introduzca los estados de la naturaleza. En la columna 2 anote las *probabilidades previas* de dichos estados. En la columna 3 introduzca las *probabilidades condicionales* que corresponden a un informe favorable de investigación de mercados (F) dado cada estado.

Paso 2. En la columna 4 calcule las **probabilidades conjuntas** multiplicando los valores de probabilidad previos de la columna 2, por los valores correspondientes de la probabilidad condicional de la columna 3.

Paso 3. Sume las probabilidades conjuntas en la columna 4 para obtener un informe favorable de investigación de mercados, $P(F)$.

Paso 4. Divida cada probabilidad conjunta en la columna 4 entre $P(F) = 0.77$ para obtener las *probabilidades posteriores* o revisadas $P(s_1 \mid F)$ y $P(s_2 \mid F)$.

En la tabla 21.3 se observa que la probabilidad de obtener un informe favorable de investigación de mercados es $P(F) = 0.77$. Además, $P(s_1 \mid F) = 0.94$ y $P(s_2 \mid F) = 0.06$. En concreto, observe que un informe de investigación favorable apuntará a una probabilidad posterior o revisada de 0.94 de que la demanda del mercado de condominios será alta, s_1.

El procedimiento de cálculo de probabilidad tabular se debe repetir en cada uno de los resultados posibles de la información muestral. Por tanto, la tabla 21.4 presenta los cálculos de las probabilidades del problema de PDC basados en un informe de investigación de mercados desfavorable. Observe que la probabilidad de obtener un informe desfavorable es $P(U) = 0.23$.

TABLA 21.4 Probabilidades para el proyecto de condominio de PDC basadas en un informe desfavorable de investigación de mercados

Estados de la naturaleza s_j	Probabilidades previas $P(s_j)$	Probabilidades condicionales $P(U \mid s_j)$	Probabilidades conjuntas $P(U \cap s_j)$	Probabilidades posteriores $P(s_j \mid U)$
s_1	0.8	0.10	0.08	0.35
s_2	0.2	0.75	0.15	0.65
	1.0		$P(U) = 0.23$	1.00

Si se obtiene un informe de esta índole, la probabilidad posterior de que la demanda sea alta, s_1, es 0.35, y de que se presente una demanda baja en el mercado, s_2, es 0.65. Las probabilidades de las tablas 21.3 y 21.4 se ilustraron en el árbol de decisión de PDC de la figura 21.5.

En el ejercicio 14 se le pide un cálculo de probabilidades posteriores.

El estudio en esta sección muestra una relación subyacente entre las probabilidades en varias ramas de un árbol de decisión. Sería inapropiado asumir las distintas probabilidades previas, $P(s_1)$ y $P(s_2)$, sin determinar cómo estos cambios alterarían $P(F)$ y $P(U)$, al igual que probabilidades posteriores $P(s_1 \mid F)$, $P(s_2 \mid F)$, $P(s_1 \mid U)$ y $P(s_2 \mid U)$.

Ejercicios

Métodos

AUTOevaluación

14. Suponga que se presenta una situación de una decisión con tres posibles estados de la naturaleza: s_1, s_2 y s_3. Las probabilidades previas son $P(s_1) = 0.2$, $P(s_2) = 0.5$ y $P(s_3) = 0.3$. Con la información muestral I, $P(I \mid s_1) = 0.1$, $P(I \mid s_2) = 0.05$ y $P(I \mid s_3) = 0.2$, calcule las probabilidades revisadas o posteriores: $P(s_1 \mid I)$, $P(s_2 \mid I)$ y $P(s_3 \mid I)$.

15. En la siguiente tabla de pagos para un problema de decisión con dos estados de la naturaleza y tres alternativas de decisión, las probabilidades previas para s_1 y s_2 son $P(s_1) = 0.8$ y $P(s_2) = 0.2$.

	Estado de la naturaleza	
Alternativa de decisión	s_1	s_2
d_1	15	10
d_2	10	12
d_3	8	20

a) ¿Cuál es la decisión óptima?
b) Determine el VEIP.
c) Suponga que se obtiene la información muestral I, con $P(I \mid s_1) = 0.20$ y $P(I \mid s_2) = 0.75$. Determine las probabilidades posteriores $P(s_1 \mid I)$ y $P(s_2 \mid I)$. Recomiende una alternativa de decisión basada en estas probabilidades.

Aplicaciones

16. Para ahorrar gastos, Rona y Jerry acordaron ir juntos hacia y desde el trabajo. Rona prefiere usar la avenida Queen City, un trayecto un poco más largo, pero más seguro. Aunque Jerry prefiere la autopista por ser más rápida, está de acuerdo con Rona en que deberían tomar la avenida Queen City en caso de que la autopista esté muy congestionada. La tabla de pagos siguiente proporciona el tiempo estimado en minutos para este recorrido.

	Estados de la naturaleza	
	Autopista sin tráfico	**Autopista congestionada**
Alternativas de decisión	s_1	s_2
Avenida Queen City, d_1	30	30
Autopista, d_2	25	45

Con base en su experiencia con los problemas de tráfico, Rona y Jerry acordaron que la probabilidad de que la autopista esté congestionada es de 0.15.

Además, están de acuerdo en que el clima parece afectar las condiciones del tráfico en la autopista. Sea

C = claro
O = nublado
R = lluvia

Las probabilidades condicionales aplicadas son las siguientes.

$$P(C \mid s_1) = 0.8 \quad P(O \mid s_1) = 0.2 \quad P(R \mid s_1) = 0.0$$
$$P(C \mid s_2) = 0.1 \quad P(O \mid s_2) = 0.3 \quad P(R \mid s_2) = 0.6$$

a) Utilice el teorema de Bayes para calcular la probabilidad revisada de cada condición climatológica y la probabilidad condicional de que la autopista esté despejada, s_1, o congestionada, s_2, dadas cada una de las condiciones climatológicas.
b) Muestre el árbol de decisión para este problema.
c) ¿Cuál es la estrategia de decisión óptima y cuál es el tiempo de viaje esperado?

17. Gorman Manufacturing Company tiene que decidir si fabrica un componente en su planta de Milán, Michigan, o si lo compra a un proveedor. La utilidad resultante depende de la demanda del producto. La siguiente tabla de pagos muestra la utilidad proyectada (en miles de dólares).

	Estado de la naturaleza		
	Demanda baja	**Demanda media**	**Demanda alta**
Alternativa de decisión	s_1	s_2	s_3
Fabricar, d_1	−20	40	100
Comprar, d_2	10	45	70

Las probabilidades de los estados de la naturaleza son: $P(s_1) = 0.35$, $P(s_2) = 0.35$ y $P(s_3) = 0.30$.

a) Utilice un árbol de decisión para recomendar una alternativa de decisión.
b) Utilice el VEIP para determinar si Gorman debe intentar obtener una mejor estimación de la demanda.
c) Un estudio de mercado de la demanda potencial del producto se espera que resulte ya sea en una condición favorable (F) o en una condición desfavorable (U). Las probabilidades condicionales son las siguientes.

$$P(F \mid s_1) = 0.10 \quad P(U \mid s_1) = 0.90$$
$$P(F \mid s_2) = 0.40 \quad P(U \mid s_2) = 0.60$$
$$P(F \mid s_3) = 0.60 \quad P(U \mid s_2) = 0.40$$

¿Cuál es la probabilidad de que el informe de investigación de mercados sea favorable?
d) ¿Cuál es la estrategia de decisión óptima para Gorman?
e) ¿Cuál es el valor esperado de la información que aportaría la investigación de mercados?

Resumen

El análisis de decisiones puede utilizarse para determinar la alternativa de decisión recomendable o una estrategia de decisión óptima cuando la persona que toma decisiones enfrenta un futuro incierto y un patrón lleno de riesgos en los eventos futuros. El objetivo del análisis de decisiones es identificar la mejor alternativa de decisión o la estrategia de decisión óptima dada

cierta información sobre los eventos inciertos y las posibles consecuencias o pagos. A los eventos futuros inciertos se les conoce como *eventos aleatorios* y a los resultados de los eventos aleatorios se les conoce como *estados de la naturaleza*.

En este capítulo se mostró cómo las tablas de pagos y los árboles de decisión permiten estructurar un problema de decisión y describir las relaciones entre las decisiones, los eventos aleatorios y las consecuencias. Con las estimaciones de probabilidad que proporcionan los estados de la naturaleza, se utilizó el método del valor esperado para identificar la alternativa de decisión o la estrategia de decisión recomendable.

En los casos en que la información muestral sobre los eventos aleatorios está disponible, existe una secuencia de decisiones que se puede tomar. En primer lugar, decidir si se obtiene o no la información muestral. Si la respuesta es afirmativa, deberá desarrollarse una estrategia de decisión óptima basada en información muestral específica. En este caso pueden utilizarse los árboles de decisión y el método del valor esperado para determinar la estrategia de decisión óptima.

El complemento de Excel PrecisionTree suele utilizarse para elaborar los árboles de decisión y resolver los problemas respectivos presentados en este capítulo. Las instrucciones para descargar e instalar el software PrecisionTree en su equipo se proporcionan en el sitio web del libro. En el apéndice al final del capítulo se incluye un ejemplo de cómo utilizar PrecisionTree para el problema del PDC de la sección 21.1.

Glosario

Árbol de decisión Representación gráfica de un problema de decisión que muestra el carácter secuencial del proceso de toma de decisiones.

Consecuencia Resultado obtenido cuando se elige una alternativa de decisión y se produce un evento aleatorio. A la medida de la consecuencia se le llama *pago*.

Estados de la naturaleza Posibles resultados de eventos aleatorios que afectan el pago asociado con una alternativa de decisión.

Estrategia de decisión Estrategia que involucra una secuencia de decisiones y los resultados aleatorios de proporcionar la solución óptima a un problema de decisión.

Eventos aleatorios Evento futuro incierto que afecta la consecuencia, o pagos, relacionada con una decisión.

Información muestral Nueva información obtenida mediante la investigación o la experimentación que permite una revisión o actualización de las probabilidades del estado de la naturaleza.

Método del valor esperado Método para elegir una alternativa de decisión que se basa en el valor esperado de cada alternativa. La alternativa de decisión recomendada es la que proporciona el mejor valor esperado.

Nodo Punto de intersección o unión en un diagrama de influencia o en un árbol de decisión.

Nodos aleatorios Nodos que indican los puntos en que se presentará un acontecimiento incierto.

Nodos de decisión Nodos que indican los puntos donde se toma la decisión.

Pagos Medida de la consecuencia de una decisión, como ingresos, costo o tiempo. Cada combinación de una alternativa de decisión y un estado de la naturaleza tiene un pago relacionado (consecuencia).

Probabilidades condicionales Probabilidad de un evento dado el resultado conocido de un caso (posiblemente) relacionado.

Probabilidades conjuntas Probabilidades de que ocurran simultáneamente la información muestral y un estado de la naturaleza determinado.

Probabilidades posteriores (revisadas) Probabilidades de los estados de la naturaleza después de la revisión de las probabilidades previas basada en la información muestral.

Probabilidades previas. Probabilidades de los estados de la naturaleza antes de obtener la información muestral.
Rama Líneas que muestran las alternativas que se derivan de los nodos de decisión y los resultados de los nodos aleatorios.
Tabla de pago Representación tabular de los pagos en un problema de decisión.
Teorema de Bayes Teorema que permite el uso de información muestral para revisar las probabilidades previas.
Valor esperado (VE) En un nodo aleatorio, promedio ponderado de los pagos. Los pesos son las probabilidades de los estados de la naturaleza.
Valor esperado de la información muestral (VEIM) Diferencia entre el valor esperado de una estrategia óptima basada en información muestral y el "mejor" valor esperado sin ningún tipo de información muestral.
Valor esperado de la información perfecta (VEIP) Valor esperado de la información que indicaría con exactitud a quien toma las decisiones el estado de la naturaleza que se presentará (es decir, información perfecta).

Fórmulas clave

Valor esperado

$$\text{VE}(d_i) = \sum_{j=1}^{N} P(s_j)V_{ij} \tag{21.3}$$

Valor esperado de la información perfecta

$$\text{VEIP} = |\text{VEcIP} - \text{VEsIP}| \tag{21.4}$$

Valor esperado de la información muestral

$$\text{VEIM} = |\text{VEcIM} - \text{VEsIM}| \tag{21.5}$$

Ejercicios complementarios

18. Un inversionista desea elegir uno de siete fondos de inversión para el próximo año. A continuación se listan los datos que muestran el porcentaje de rendimiento anual para cada fondo durante un periodo de cinco años. El supuesto es que uno de estos periodos de cinco años se presentará nuevamente durante el próximo año. Por tanto, los años A, B, C, D y E son los estados de la naturaleza para la decisión sobre los fondos de inversión.

	Estado de la naturaleza				
Fondo de inversión	**Año A**	**Año B**	**Año C**	**Año D**	**Año E**
Mayor capital social	35.3	20.0	28.3	10.4	−9.3
Capital social medio	32.3	23.2	−0.9	49.3	−22.8
Menor capital social	20.8	22.5	6.0	33.3	6.1
Sector de recursos energéticos	25.3	33.9	−20.5	20.9	−2.5
Sector salud	49.1	5.5	29.7	77.7	−24.9
Sector de tecnología	46.2	21.7	45.7	93.1	−20.1
Sector de bienes raíces	20.5	44.0	−21.1	2.6	5.1

a) Suponga que un analista financiero experimentado revisa los cinco estados de la naturaleza y ofrece las siguientes probabilidades: 0.1, 0.3, 0.1, 0.1 y 0.4. Utilizando el método

del valor esperado, ¿cuál es el fondo de inversión recomendado? ¿Cuál es el rendimiento anual esperado? Al utilizar este fondo de inversión, ¿cuáles son los rendimientos anuales mínimos y máximos?

b) Un inversionista conservador señala que el fondo de inversión de menor capital social es el único que no tiene posibilidad de pérdida. De hecho, si éste es elegido, el inversionista garantiza un rendimiento de por lo menos 6%. ¿Cuál es el rendimiento anual esperado para este fondo?

c) Teniendo en cuenta los fondos de inversión recomendados en los incisos *a*) y *b*), ¿cuál parece tener más riesgo? ¿Por qué? ¿El rendimiento anual esperado es mayor que para el fondo de inversión con más riesgo?

d) ¿Qué fondo de inversión le recomendaría a los inversionistas? Explique.

19. Warren Lloyd está interesado en alquilar un automóvil nuevo y ha contactado a tres concesionarias automotrices para la información de precios. Cada distribuidor ofrece a Warren un cierre de contrato a 36 meses de arrendamiento sin pago inicial, en el momento que lo firma. Cada contrato incluye un cargo mensual y un subsidio de millaje. Las millas adicionales tienen un cargo sobre una base por milla. El costo mensual por arrendamiento, el millaje subsidiado y el costo de millas adicionales son los siguientes.

Distribuidor	Costo mensual	Millaje subsidiado	Costo por milla adicional
Forno Automotive	$299	36 000	$0.15
Midtown Motors	$310	45 000	$0.20
Hopkins Automotive	$325	54 000	$0.15

Warren decidió elegir la opción de arrendamiento que minimice su costo total a 36 meses. La dificultad estriba en que no está seguro de cuántas millas recorrerá en los próximos tres años. Para efectos de tomar esta decisión, considera razonable suponer que conducirá 12 000 millas, 15 000 millas o 18 000 millas por año. Con este supuesto, Warren estima el costo de las tres opciones de arrendamiento. Por ejemplo, supone que el contrato de arrendamiento de Forno Automotive le costará $10 764 si maneja 12 000 millas, $ 12 114 si maneja 15 000 millas o $ 13 464 si recorre 18 000 millas por año.

a) ¿Cuál es la decisión y cuál es el evento aleatorio?

b) Construya una tabla de pagos.

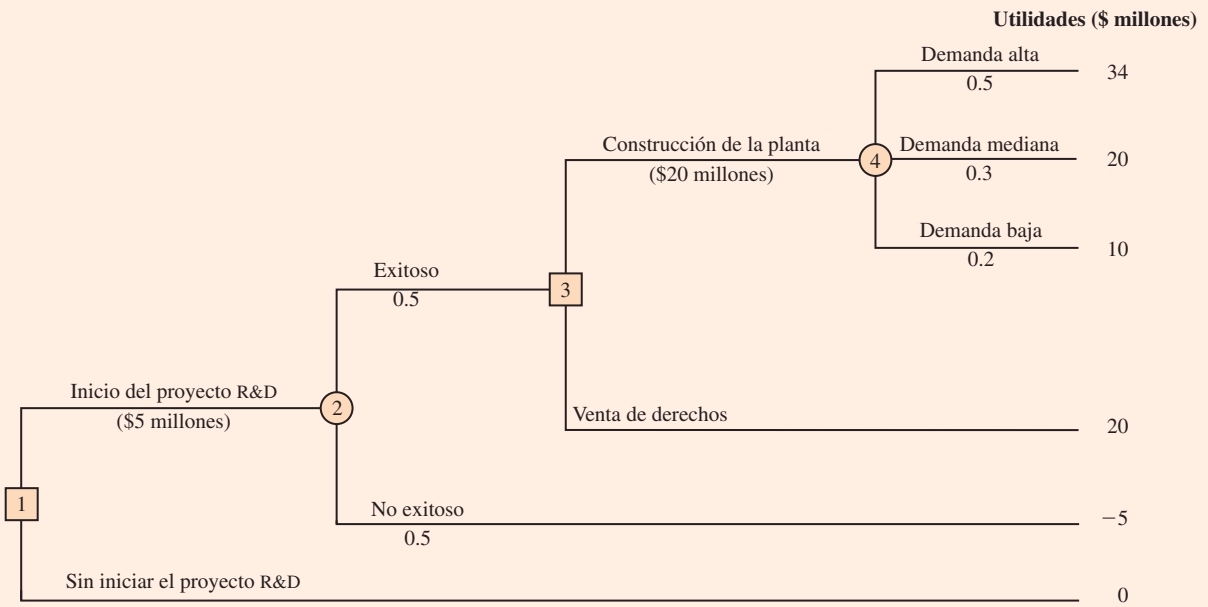
c) Suponga que las probabilidades de que Warren recorra 12 000, 15 000 y 18 000 millas por año son 0.5, 0.4 y 0.1 respectivamente. ¿Qué distribuidor debe elegir?

d) Suponga que después de mucho pensarlo, Warren llega a la conclusión de que las probabilidades de que maneje 12 000, 15 000 o 18 000 son 0.3, 0.4 y 03, respectivamente. ¿Qué distribuidor debe elegir?

20. Hemingway, Inc. está considerando invertir $5 millones para investigación y desarrollo (R&D) en un proyecto. Las utilidades proyectadas parecen ser prometedoras, pero el presidente de Hemingway está preocupado porque la probabilidad de que el proyecto de R&D sea exitoso es sólo de 0.50. En segundo lugar, el presidente sabe que incluso si el proyecto tiene éxito, se requerirá que la empresa construya una nueva planta de producción a un costo de $20 millones para la fabricación de los productos. Si la instalación es construida, la incertidumbre acerca de la demanda permanece y, por tanto, sobre la utilidad que se obtendrá. Otra opción indica que si el proyecto de R&D tiene éxito, la empresa podría vender los derechos del producto a un estimado de $25 millones. Con esta opción, la empresa no construirá la planta de producción de $20 millones.

En la figura 21.1 se muestra el árbol de decisión. Las utilidades proyectadas para cada resultado se presentan al final de las ramas. Por ejemplo, los ingresos proyectados para el resultado de la demanda alta son de $59 millones. Sin embargo, el costo del proyecto de R&D ($5 millones) y el costo de la planta de producción ($20 millones) muestran que la utilidad de este resultado será de $59 − $5 − $20 = $34 millones. También se presentan las probabilidades para los eventos aleatorios.

FIGURA 21.11 Árbol de decisión para Hemmingway, Inc.



a) Analice el árbol de decisión para determinar si la empresa debe comprometerse en el proyecto de R&D. Si lo hace, y si el proyecto tiene éxito, ¿qué debe hacer? ¿Cuál es el valor esperado de su estrategia?

b) ¿Cuál deberá ser el precio de venta para que la empresa considere vender los derechos de los productos?

21. Embassy Publishing Company recibió un manuscrito del sexto capítulo de un libro de texto para una nueva universidad. El editor de la división universitaria está familiarizado con el manuscrito y se estima que la probabilidad de que el libro tenga éxito es de 0.65. Si es exitoso, se tendrá una utilidad de $750 000. Si la empresa decide publicar el libro y no es exitoso, se producirá una pérdida de $250 000.

Antes de tomar la decisión de aceptar o rechazar el manuscrito, el editor considera la posibilidad de enviarlo a revisión. Este proceso proporciona una evaluación favorable (F) o desfavorable (U) del manuscrito. La experiencia en el proceso de revisión sugiere las probabilidades $P(F) = 0.7$ y $P(U) = 0.3$. Sea s_1 = el libro de texto tiene éxito, y s_2 = el libro de texto no tiene éxito. Las probabilidades iniciales del editor de s_1 y s_2 se revisarán en función de si la revisión es favorable o desfavorable. Las probabilidades de la revisión son las siguientes.

$$P(s_1 \mid F) = 0.75 \qquad P(s_1 \mid U) = 0.417$$
$$P(s_2 \mid F) = 0.25 \qquad P(s_2 \mid U) = 0.583$$

a) Construya un árbol de decisión suponiendo que la empresa deberá tomar primero la decisión de si debe enviar el manuscrito para revisión y después asumir la decisión de aceptarlo o rechazarlo.

b) Analice el árbol de decisión y determine la estrategia de decisión óptima para la editorial.

c) Si la revisión del manuscrito tiene un costo de $5 000, ¿cuál es su recomendación?
d) ¿Cuál es el valor esperado de la información perfecta? ¿Qué sugiere VEIP para la empresa?

Caso a resolver

Estrategia de defensa en una demanda

John Campbell, empleado de Manhattan Construction Company, afirma haberse lesionado la espalda como resultado de una caída mientras reparaba el techo de uno de los edificios de Eastview Apartments. En una demanda en la que reclama daños por $1 500 000 presentada contra Doug Reynolds, el propietario de Eastview Apartments, John afirma que el techo se había podrido en secciones y que su caída podría haberse evitado si el señor Reynolds le hubiera informado a Manhattan Construction sobre el problema. Reynolds notificó a Allied Insurance, su compañía de seguros, la situación de litigio. Allied debe defender al señor Reynolds y decidir qué acciones emprender con respecto a la demanda.

A raíz de algunos testimonios y una serie de discusiones entre las dos partes, John Campbell ofreció aceptar un acuerdo de $750 000. Por tanto, una opción es que Allied Insurance pague este monto para resolver el caso. Allied también está considerando una contraoferta de $400 000, con la esperanza de que John acepte una cantidad menor para evitar el tiempo y el costo de ir a juicio. Las primeras investigaciones de Allied revelan que John tiene un caso fuerte; le preocupa que pueda rechazar su contrapropuesta y solicitar un juicio con jurado. Los abogados de Allied pasaron algún tiempo explorando cuál puede ser la reacción de John si le hacen una contraoferta de $400 000.

Los abogados llegaron a la conclusión de que es adecuado considerar tres posibles resultados para representar la posible reacción de John a la contrapropuesta de $400 000: 1) John aceptará la contraoferta y el caso será cerrado; 2) John rechazará la contraoferta y optará porque un jurado decida el monto de liquidación, o 3) John hará una contraoferta a Allied de $600 000. Si John hace una contraoferta, Allied ha decidido que no hará más contraofertas; aceptará los $600 000 o irán a juicio.

Si el caso va a los tribunales, Allied considera que los tres posibles resultados son: 1) que el jurado rechace el reclamo de John y Allied no estará obligado a pagar los daños; 2) el jurado falla a favor de John y le otorga una indemnización de $750 000 por los daños, o 3) el jurado llega a la conclusión de que John tiene un caso fuerte y le otorga un monto total de $1 500 000.

Las consideraciones clave en cómo Allied desarrolla su estrategia para resolver el caso son las probabilidades relacionadas con la respuesta de John a la contraoferta de $400 000, y las probabilidades relacionadas con los tres posibles resultados en el juicio. Los abogados de Allied creen que la probabilidad de que John acepte una contraoferta de $400 000 es 0.10; la probabilidad de que la rechace es 0.40, y la probabilidad de que el mismo John plantee una contraoferta por 600 000 es de 0.50. Si el caso va a juicio, consideran que la probabilidad de que el jurado otorgue una indemnización por daños a John por $1 500 000 es de 0.30; la probabilidad de que conceda la indemnización por daños por $750 000 es de 0.50, y la probabilidad de que no otorgue ninguna indemnización es de 0.20.

Informe gerencial

Realice un análisis del problema que enfrenta Allied Insurance y prepare un informe en el que se resuman sus hallazgos y recomendaciones. Asegúrese de incluir los siguientes elementos:

1. Un árbol de decisión.
2. Una recomendación sobre si Allied debe aceptar la oferta inicial de John para resolver la demanda por $ 750 000.
3. La estrategia de decisión que Allied debe seguir si opta por hacer una contraoferta a John por $400 000.
4. Un perfil de riesgos para la estrategia recomendada.

Apéndice

Introducción a PrecisionTree

PrecisionTree es un complemento de Excel que se utiliza para desarrollar y analizar los árboles de decisión. En este apéndice se muestra cómo instalar y utilizar este programa para resolver el problema de PDC presentado en la sección 21.1.

Instalación y apertura de PrecisionTree

Las instrucciones para descargar e instalar PrecisionTree en el equipo se proporcionan en el sitio web de este libro. Después de instalar el software, realice los pasos siguientes para utilizarlo como complemento de Excel.

Paso 1. Haga clic en el botón **Start** de la barra de tareas y después señale **All Programs**.
Paso 2. Seleccione la carpeta titulada **Palisade Decision Tools**.
Paso 3. Haga clic en **PrecisionTree for Excel**.

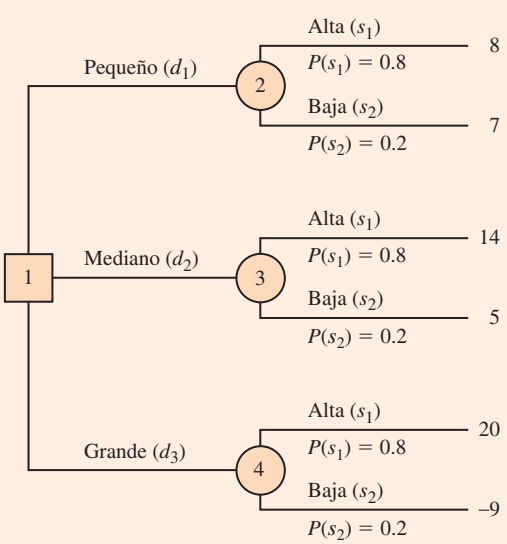
Estos pasos abren Excel y agregan la ficha PrecisionTree junto a la barra de complementos en la barra de opciones. Alternativamente, si ya está trabajando en Excel, con estos pasos PrecisionTree estará disponible.

Para empezar: un primer árbol de decisión

Suponga que PrecisionTree ha sido instalado y que se ha abierto un libro de Excel, y que una hoja de cálculo contendrá el árbol de decisión seleccionado. Para crear una versión de PrecisionTree del árbol de decisión de PDC (vea la figura 21.12), realice lo siguiente.

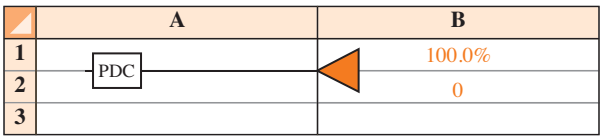
Paso 1. Haga clic en la ficha **PrecisionTree** en la barra de opciones.
Paso 2. En el grupo **Create New**, haga clic en **Decision Tree**.
Paso 3. Cuando el cuadro de diálogo PrecisionTree de Excel aparezca:
Haga clic en la celda **A1**.
Haga clic en **OK**.

FIGURE 21.12 Árbol de decisión de PDC



Paso 4. Cuando el cuadro de diálogo PrecisionTree-Model Settings aparezca:
Introduzca PDC en el cuadro **Name**.
Haga clic en **OK**.

Aparecerá un primer árbol con una nota final y sin ramas.



Agregar un nodo de decisión y ramas

Un primer árbol contiene un nombre y un nodo extremo en forma de triángulo. Recuerde que el árbol de decisión de PDC tiene un nodo de decisión con tres ramas, uno por cada alternativa de decisión (complejos pequeños, medianos y grandes). Los pasos siguientes muestran cómo cambiar el nodo extremo a un nodo de decisión y añadir las ramas alternativas del árbol.

Paso 1. Haga clic en la nota final en forma de triángulo.
Paso 2. Cuando el cuadro de diálogo PrecisionTree-Decision Tree Node Settings aparezca:
Haga clic en el botón **Decision** abajo de Node Type.
Haga clic en la ficha Branches.
Haga clic en **Add**.
Haga clic en **OK**.

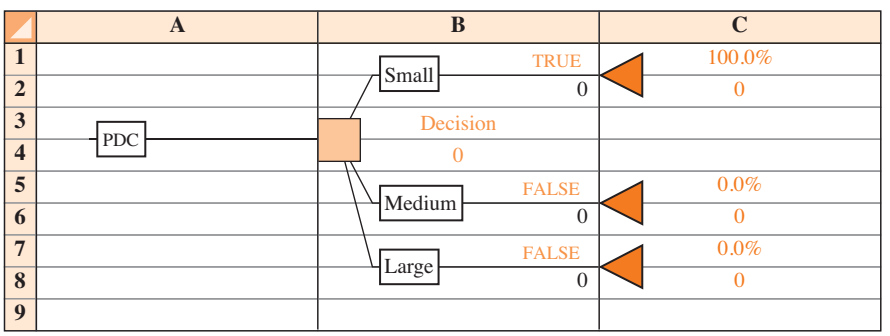
En el desarrollo aparecerá un árbol de decisión modificado con un nodo de decisión y tres ramas.

Nombrar las alternativas de decisión

Cada una de las tres ramas de decisión tiene el nombre genérico **branch**, seguido de un número para identificarla. Deseamos renombrar las ramas con **Small, Medium** y **Large**. Comience con Branch#1.

Paso 1. Haga clic en el nombre Branch#1.
Paso 2. Cuando el cuadro de diálogo PrecisionTree para Excel aparezca:
Remplace Branch#1 con **Small**.
Haga clic en **OK**.

Continúe aplicando los mismos dos pasos para nombrar las otras dos ramas de decisión. Después de etiquetar las ramas, el árbol de decisión PDC con las tres ramas aparecerá de la siguiente forma.

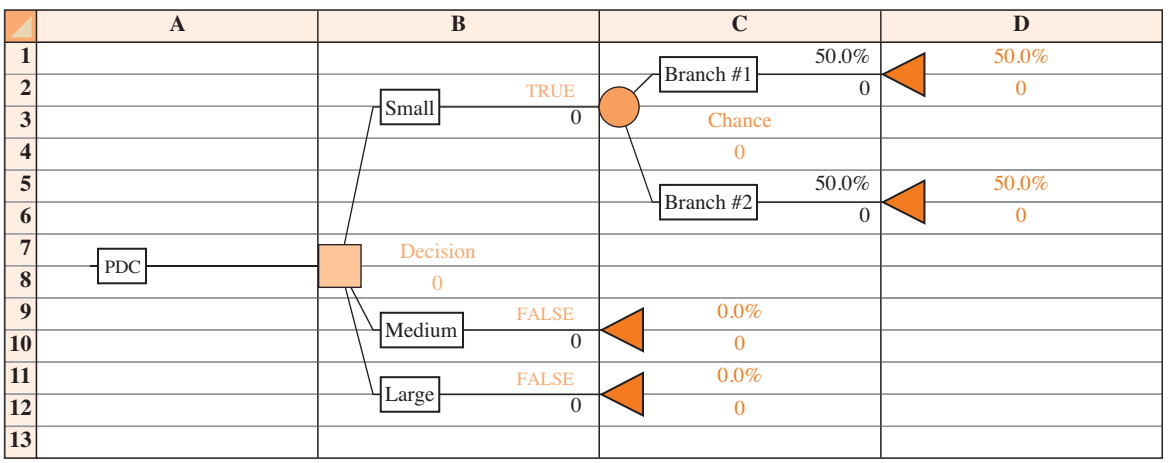


Agregar nodos aleatorios y ramas

En el problema de PDC, el evento aleatorio es la demanda de los condominios, que puede ser alta o baja. Por tanto, debe añadirse un nodo aleatorio con dos ramas al final de cada rama de una alternativa de decisión.

Paso 1. Haga clic en el nodo extremo de la rama Small de la alternativa de decisión.
Paso 2. Cuando el cuadro de diálogo PrecisionTree-Decision Tree Node Settings aparezca:
Haga clic en el botón **Chance** abajo de Node Type.
Haga clic en **OK**.

En el paso 2, el valor predeterminado para el número de ramas en el cuadro de diálogo Decision Tree Node Settings es 2. Como resultado, para el problema de PDC no es necesario especificar el número de ramas para el nodo aleatorio que se acaba de crear. El árbol de decisión aparece ahora como sigue.



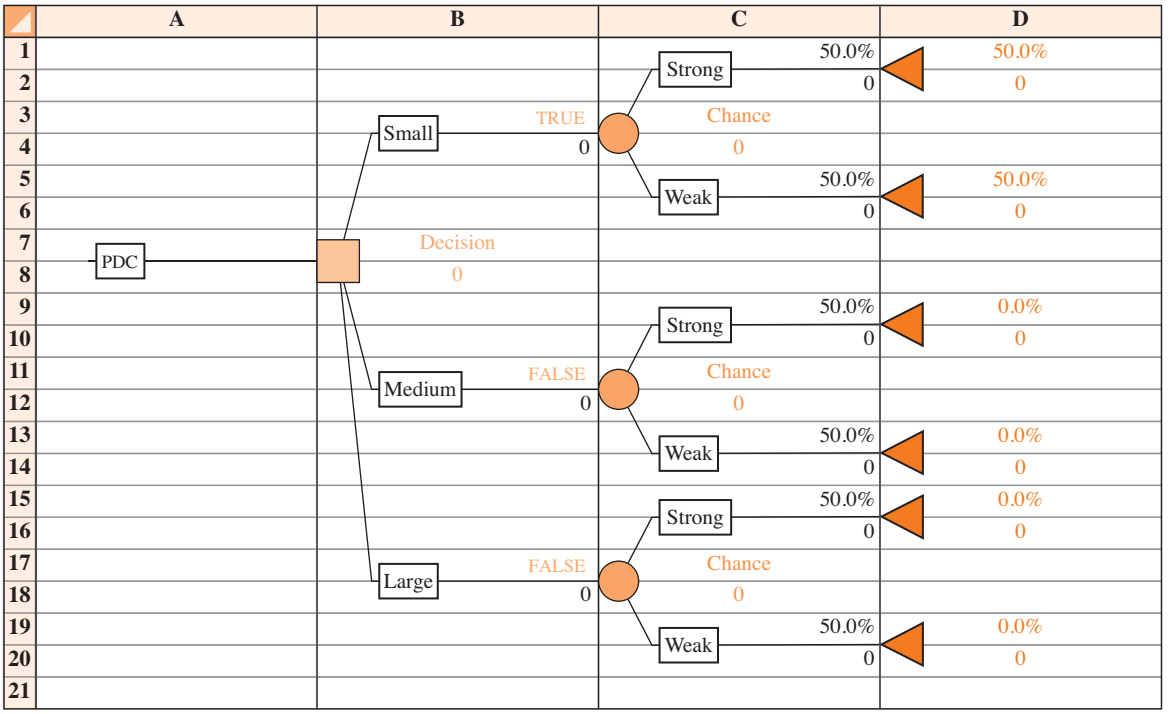
Ahora podemos cambiar el nombre de las ramas del nodo aleatorio como Strong (alta) y Weak (baja), utilizando el mismo procedimiento que se realizó para las ramas de decisión. Los nodos aleatorios ahora se pueden insertar en el extremo de las otras dos ramas de decisión de forma similar.* Realizar estos pasos lleva al árbol de decisión de PDC de la figura 21.13.

Inserción de probabilidades y pagos

PrecisionTree tiene la capacidad de insertar las probabilidades y los pagos en el árbol de decisión. En la figura 21.13 se observa que el programa asigna automáticamente la misma probabilidad de 0.5 (que se muestra como 50%) a cada rama de un nodo aleatorio. Para PDC, la probabilidad de una alta demanda es 0.8 y de una baja demanda es 0.2. Se pueden elegir las celdas C1, C5, C9, C13, C15 y C19 e insertar las probabilidades adecuadas. Se introducen los pagos para los resultados aleatorios en las celdas C2, C6, C10, C14, C16 y C20. Después de insertar las probabilidades y los pagos, el árbol de decisión de PDC aparece como se muestra en la figura 21.14.

* PrecisionTree también tiene la capacidad de copiar los nodos que pueden utilizarse para crear los otros dos nodos aleatorios. Sólo haga clic en el primer nodo aleatorio creado y en Copy SubTree. Después haga clic en el botón derecho en uno de los otros nodos extremos, y luego en Past SubTree. Realice lo mismo para el otro nodo extremo.

FIGURA 21.13 Desarrollo del árbol de decisión de PDC mediante PrecisionTree



Interpretación del resultado

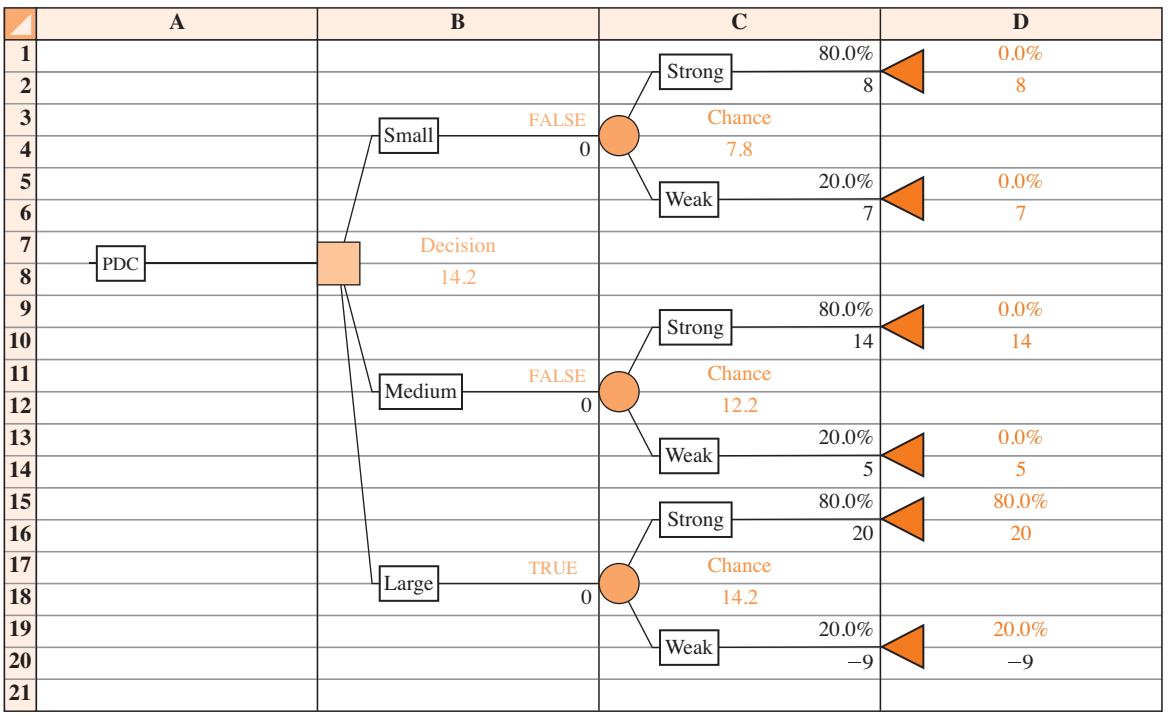
Cuando las probabilidades y los pagos son insertados, PrecisionTree realiza automáticamente en forma regresiva los cálculos necesarios para obtener los valores esperados y determinar la solución óptima. Las decisiones óptimas se identifican con la palabra **TRUE** (verdad) en la rama de decisiones. Las ramas sin decisión opcional se identifican con la palabra **FALSE** (falso). Observe que la palabra **TRUE** aparece en la rama de decisión grande. Por tanto, se recomienda el análisis de decisiones de que PDC debe construir el complejo de condominios grande. El valor esperado de esta decisión aparece justo a la derecha del nodo de decisión al inicio del árbol. Por tanto, advierta que el valor máximo esperado es $14.2 millones. Los valores esperados de las otras alternativas de decisión aparecen a la derecha de los nodos de oportunidad en el extremo de las ramas de las alternativas de decisión. Note que el valor esperado de la decisión de construir un complejo pequeño es de $7.8 millones y el de construir un complejo mediano es $12.2 millones.

Otras opciones

Hemos usado PrecisionTree con un objetivo de maximización. Este es el valor predeterminado. Si se tiene un árbol de decisión con un objetivo de minimización, siga los pasos que se indican a continuación:

Paso 1. Haga clic en el nombre del árbol de decisión (al principio del árbol).
Paso 2. Cuando el cuadro de diálogo PrecisionTree-Model Settings aparezca:
Haga clic en la ficha Calculation.
Seleccione **Minimum Payoff** en el cuadro **Optimum Path**.
Haga clic en **OK**.

FIGURA 21.14 Árbol de decisión de PDC con probabilidades y pagos



Apéndices

APÉNDICE A
Referencias y bibliografía

APÉNDICE B
Tablas

APÉNDICE C
Notación de suma

APÉNDICE D
Soluciones a las autoevaluaciones y respuestas a los ejercicios de números pares

APÉNDICE E
Uso de las funciones de Excel

APÉNDICE F
Cálculo de los valores-*p* utilizando Minitab y Excel

Apéndice A Referencias y bibliografía

General

Freedman, D., R. Pisani y R. Purves, *Statistics*, 4a. ed., W. W. Norton, 2007.

Hogg, R. V., J. W. McKean y A. T. Craig, *Introduction to Mathematical Statistics*, 6a. ed., Pearson, 2004.

Hogg, R.V. y E. A. Tanis, *Probability and Statistical Interference*, 7a. ed., Pearson, 2005.

Miller, I. y M. Miller, *John E. Freund's Mathematical Statistics*, 7a. ed., Pearson, 2003.

Moore, D. S., G. P. McCabe y B. Craig. *Introduction to the Practice of Statistics*, 6a. ed., Freeman, 2007.

Wackerly, D. D., W. Mendenhall y R. L. Scheaffer, *Mathematical Statistics with Applications*, 7a. ed., Cengage Learning, 2007.

Diseño de experimentos

Cochran, W. G. y G. M. Cox, *Experimental Design,* 2a. ed., Wiley, 1992.

Hicks, C. R. y K. V. Turner, *Fundamental Concepts in the Design of Experiments*, 5a. ed., Oxford University Press, 1999.

Montgomery, D. C., *Design and Analysis of Experiments*, 6a. ed., Wiley, 2004.

Winer, B. J., K. M. Michels y D. R. Brown, *Statistical Principles in Experimental Desing*, 3a. ed., McGraw-Hill, 1991.

Wu, C. F. Jeff y M. Hamada, *Experiments: Planning, Analysis and Parameter Optimization*, 2a. ed., Wiley, 2009.

Series de tiempo y elaboración de pronósticos

Bowerman, B. L. y R. T. O'Connell, *Forecasting and Time Series: An Applied Approach*, 3a. ed., Brooks/Cole, 2000.

Box, G. E. P., G. M. Jenkins y G. C. Reinsel, *Time Series Analysis: Forecasting and Control,* 4a. ed., Wiley, 2008.

Makridakis S, G., S. C. Wheelwright y R. J. Hyndman, *Forecasting Methods and Applications*, 3a. ed., Wiley, 1998.

Wilson, J. H., B. Keating y John Galt Solutions, Inc., *Business Forecasting with Accompanying Excel-Based Forecast X™*, 5a. ed., McGraw-Hill/Irwin, 2007.

Números índice

U.S. Department of Commerce, *Survey of Current Business.*

U.S. Department of Labor, Bureau of Labor Statistics, *CPI Detailed* Report.

U.S. Department of Labor, *Producer Price Indexes.*

Métodos no paramétricos

Conover, W. J., *Practical Nonparametric Statistics*, 3a. ed., Wiley, 1999.

Gibbons, J. D. y S. Chakraborti, *Nonparamettric Statistical Inference*, 4a. ed., CRC Press, 2003.

Higgins, J. J., *Introduction to Modern Nonparametric Statistics*, Thomson-Brooks/Cole, 2004.

Hollander, M. y D. A. Wolfe, *Non-Parametric Statistical Methods*, 2a. ed., Wiley, 1999.

Probabilidad

Hogg, R. V. y E. A. Tanis, *Probability and Statistical Interference*, 7a. ed., Pearson, 2005.

Ross, S. M, *Introduction to Probability Models*, 9a. ed., Elsevier, 2006.

Wackerly, D. D., W. Mendenhall y R. L. Scheaffer, *Mathematical Statistics with Applications,* 7a. ed., Cengage Learning, 2007.

Control de calidad

Evans, J. R. y W. M. Lindsay, *The Management and Control of Quality*, 6a. ed., South-Western, 2006.

Juran, J. M. y A. B. Godfrey, *Juran's Quality Handbook,* 5a. ed., McGraw-Hill, 1999.

Montgomery, D. C., *Introduction to Statistical Quality Control*, 6a. ed., Wiley, 2008.

Análisis de regresión

Chatterjee, S. y A. S. Hadi, *Regresion Analysis by Example*, 4a. ed., Wiley, 2006.

Draper, N. R. y H. Smith, *Applied Regression Analysis*, 3a. ed., Wiley, 1998.

Graybill, F. A. y H. K. Iyer, *Regression Analysis: Concepts and Applications*, Wadsworth, 1994.

Hosmer, D. W. y S. Lemeshow, *Applied Logistic Regression*, 2a. ed., Wiley, 2000.

Kleinbaum, D. G., L. L. Kupper y K. E. Muller, *Applied Regression Analysis and Multivariate Methods*, 4a. ed., Cengage Learning, 2007.

Neter, J., W. Wasserman, M. H. Kutner y C. Nashtsheim, *Applied Linear Statistical Models*, 4a. ed., McGraw-Hill, 1996.

Mendenhall, M., T. Sincich y T. R. Dye, *A Second Course in Statistics: Regression Analysis*, 6a. ed., Pearson, 1996.

Análisis de decisiones

Clemen, R. T. y T. Reilly, *Making Hard Decisions with Decision Tools,* Cengage Learning, 2004.

Goodwin, P., *Decision Analysis for Management Judgment*, 3a. ed., Wiley, 2004.

Pratt, J. W., H. Raiffa y R. Schlaifer, *Introduction to Statistical Decision Theory*, MIT Press, 1995.

Muestreo

Cochran, W. G., *Sampling Techniques,* 3a. ed., Wiley, 1977.

Hansen, M. H., W. N. Hurwitz, W. G. Madow y M. N. Hanson, *Sample Survey Methods and Theory*, Wiley, 1993.

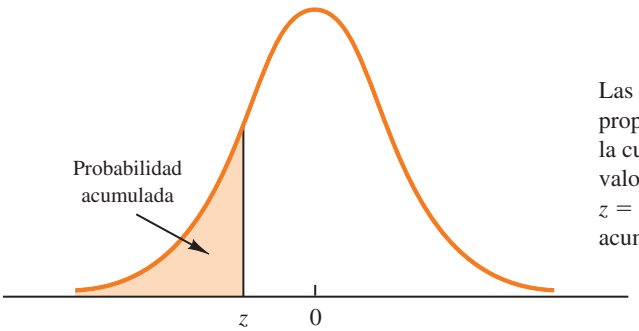
Kish, L., *Survey Sampling*, Wiley, 2008.

Levy, P. S. y S. Lemeshow, *Sampling of Populations: Methods and Applications*, 4a. ed., Wiley, 2008.

Scheaffer, R. L., W. Mendenhall y L. Ott, *Elementary Survey Sampling*, 6a. ed., Cengage Learning, 2005.

Apéndice B Tablas

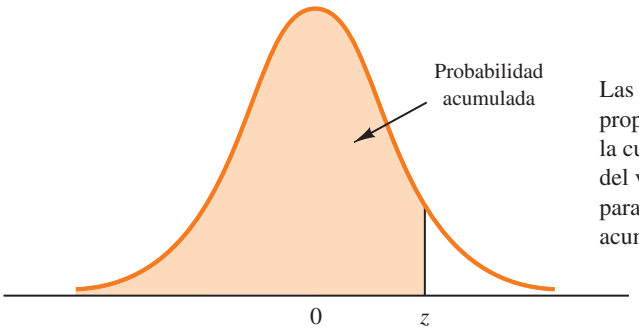
TABLA 1 Probabilidades acumuladas en la distribución normal estándar



Las entradas en la tabla proporcionan el área bajo la curva a la izquierda del valor de z. Por ejemplo, para $z = -0.85$, la probabilidad acumulada es 0.1977.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	.03228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

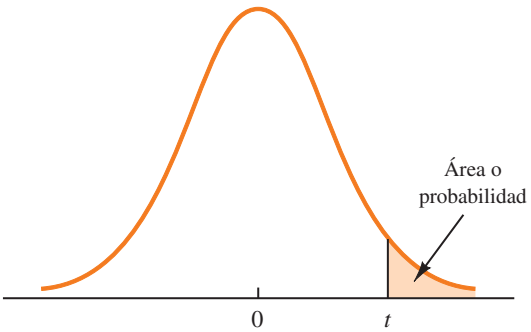
TABLA 1 Probabilidades acumuladas en la distribución normal estándar (*continuación*)



Las entradas en la tabla proporcionan el área bajo la curva a la izquierda del valor de z. Por ejemplo, para $z = 1.25$, la probabilidad acumulada es 0.8944.

z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

TABLA 2 Distribución *t*



Las entradas en la tabla proporcionan los valores de t para un área o probabilidad en la cola superior de la distribución t. Por ejemplo, con 10 grados de libertad y un área de 0.05 en la cola superior, $t_{0.05} = 1.812$.

Grados de libertad	Área en la cola superior					
	0.20	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
1	1.376	3.078	6.314	12.706	31.821	63.656
2	1.061	1.886	2.920	4.303	6.965	9.925
3	0.978	1.638	2.353	3.182	4.541	5.841
4	0.941	1.533	2.132	2.776	3.747	4.604
5	0.920	1.476	2.015	2.571	3.365	4.032
6	0.906	1.440	1.943	2.447	3.143	3.707
7	0.896	1.415	1.895	2.365	2.998	3.499
8	0.889	1.397	1.860	2.306	2.896	3.355
9	0.883	1.383	1.833	2.262	2.821	3.250
10	0.879	1.372	1.812	2.228	2.764	3.169
11	0.876	1.363	1.796	2.201	2.718	3.106
12	0.873	1.356	1.782	2.179	2.681	3.055
13	0.870	1.350	1.771	2.160	2.650	3.012
14	0.868	1.345	1.761	2.145	2.624	2.977
15	0.866	1.341	1.753	2.131	2.602	2.947
16	0.865	1.337	1.746	2.120	2.583	2.921
17	0.863	1.333	1.740	2.110	2.567	2.898
18	0.862	1.330	1.734	2.101	2.552	2.878
19	0.861	1.328	1.729	2.093	2.539	2.861
20	0.860	1.325	1.725	2.086	2.528	2.845
21	0.859	1.323	1.721	2.080	2.518	2.831
22	0.858	1.321	1.717	2.074	2.508	2.819
23	0.858	1.319	1.714	2.069	2.500	2.807
24	0.857	1.318	1.711	2.064	2.492	2.797
25	0.856	1.316	1.708	2.060	2.485	2.787
26	0.856	1.315	1.706	2.056	2.479	2.779
27	0.855	1.314	1.703	2.052	2.473	2.771
28	0.855	1.313	1.701	2.048	2.467	2.763
29	0.854	1.311	1.699	2.045	2.462	2.756
30	0.854	1.310	1.697	2.042	2.457	2.750
31	0.853	1.309	1.696	2.040	2.453	2.744
32	0.853	1.309	1.694	2.037	2.449	2.738
33	0.853	1.308	1.692	2.035	2.445	2.733
34	0.852	1.307	1.691	2.032	2.441	2.728

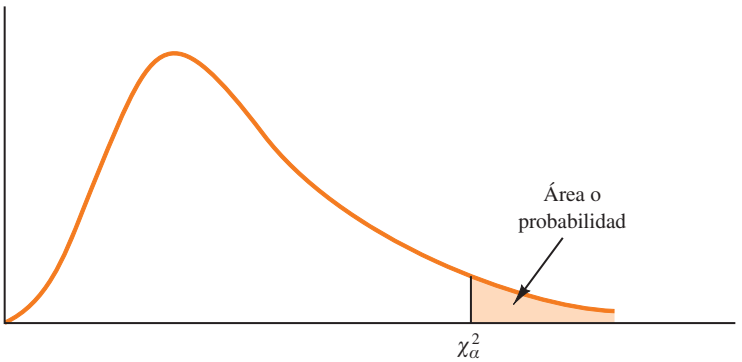
TABLA 2 Distribución *t* (*continuación*)

Grados de libertad	Área en la cola superior					
	0.20	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
35	0.852	1.306	1.690	2.030	2.438	2.724
36	0.852	1.306	1.688	2.028	2.434	2.719
37	0.851	1.305	1.687	2.026	2.431	2.715
38	0.851	1.304	1.686	2.024	2.429	2.712
39	0.851	1.304	1.685	2.023	2.426	2.708
40	0.851	1.303	1.684	2.021	2.423	2.704
41	0.850	1.303	1.683	2.020	2.421	2.701
42	0.850	1.302	1.682	2.018	2.418	2.698
43	0.850	1.302	1.681	2.017	2.416	2.695
44	0.850	1.301	1.680	2.015	2.414	2.692
45	0.850	1.301	1.679	2.014	2.412	2.690
46	0.850	1.300	1.679	2.013	2.410	2.687
47	0.849	1.300	1.678	2.012	2.408	2.685
48	0.849	1.299	1.677	2.011	2.407	2.682
49	0.849	1.299	1.677	2.010	2.405	2.680
50	0.849	1.299	1.676	2.009	2.403	2.678
51	0.849	1.298	1.675	2.008	2.402	2.676
52	0.849	1.298	1.675	2.007	2.400	2.674
53	0.848	1.298	1.674	2.006	2.399	2.672
54	0.848	1.297	1.674	2.005	2.397	2.670
55	0.848	1.297	1.673	2.004	2.396	2.668
56	0.848	1.297	1.673	2.003	2.395	2.667
57	0.848	1.297	1.672	2.002	2.394	2.665
58	0.848	1.296	1.672	2.002	2.392	2.663
59	0.848	1.296	1.671	2.001	2.391	2.662
60	0.848	1.296	1.671	2.000	2.390	2.660
61	0.848	1.296	1.670	2.000	2.389	2.659
62	0.847	1.295	1.670	1.999	2.388	2.657
63	0.847	1.295	1.669	1.998	2.387	2.656
64	0.847	1.295	1.669	1.998	2.386	2.655
65	0.847	1.295	1.669	1.997	2.385	2.654
66	0.847	1.295	1.668	1.997	2.384	2.652
67	0.847	1.294	1.668	1.996	2.383	2.651
68	0.847	1.294	1.668	1.995	2.382	2.650
69	0.847	1.294	1.667	1.995	2.382	2.649
70	0.847	1.294	1.667	1.994	2.381	2.648
71	0.847	1.294	1.667	1.994	2.380	2.647
72	0.847	1.293	1.666	1.993	2.379	2.646
73	0.847	1.293	1.666	1.993	2.379	2.645
74	0.847	1.293	1.666	1.993	2.378	2.644
75	0.846	1.293	1.665	1.992	2.377	2.643
76	0.846	1.293	1.665	1.992	2.376	2.642
77	0.846	1.293	1.665	1.991	2.376	2.641
78	0.846	1.292	1.665	1.991	2.375	2.640
79	0.846	1.292	1.664	1.990	2.374	2.639

TABLA 2 Distribución *t* (*continuación*)

Grados de libertad	Área en la cola superior					
	0.20	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
80	0.846	1.292	1.664	1.990	2.374	2.639
81	0.846	1.292	1.664	1.990	2.373	2.638
82	0.846	1.292	1.664	1.989	2.373	2.637
83	0.846	1.292	1.663	1.989	2.372	2.636
84	0.846	1.292	1.663	1.989	2.372	2.636
85	0.846	1.292	1.663	1.988	2.371	2.635
86	0.846	1.291	1.663	1.988	2.370	2.634
87	0.846	1.291	1.663	1.988	2.370	2.634
88	0.846	1.291	1.662	1.987	2.369	2.633
89	0.846	1.291	1.662	1.987	2.369	2.632
90	0.846	1.291	1.662	1.987	2.368	2.632
91	0.846	1.291	1.662	1.986	2.368	2.631
92	0.846	1.291	1.662	1.986	2.368	2.630
93	0.846	1.291	1.661	1.986	2.367	2.630
94	0.845	1.291	1.661	1.986	2.367	2.629
95	0.845	1.291	1.661	1.985	2.366	2.629
96	0.845	1.290	1.661	1.985	2.366	2.628
97	0.845	1.290	1.661	1.985	2.365	2.627
98	0.845	1.290	1.661	1.984	2.365	2.627
99	0.845	1.290	1.660	1.984	2.364	2.626
100	0.845	1.290	1.660	1.984	2.364	2.626
∞	0.842	1.282	1.645	1.960	2.326	2.576

TABLA 3 Distribución ji-cuadrada



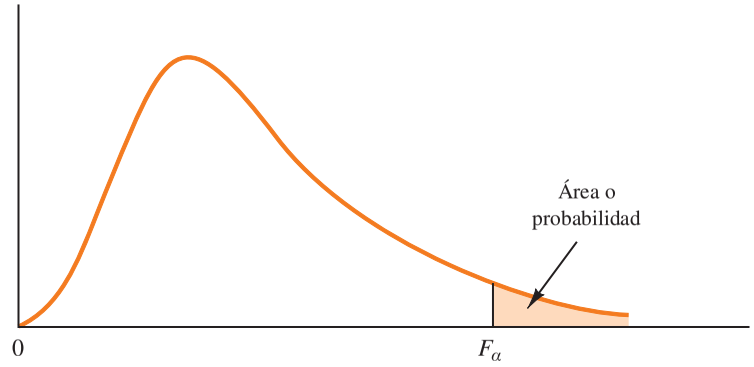
Las entradas en la tabla proporcionan los valores de χ^2_α, donde α es el área o probabilidad en la cola superior de la distribución ji-cuadrada. Por ejemplo, con 10 grados de libertad y un área de 0.01 en la cola superior, $\chi^2_{0.01} = 23.209$.

Grados de libertad	Área en la cola superior									
	0.995	**0.99**	**0.975**	**0.95**	**0.90**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.878	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.994
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.335

TABLA 3 Distribución ji-cuadrada (*continuación*)

Grados de libertad	Área en la cola superior									
	0.995	**0.99**	**0.975**	**0.95**	**0.90**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
35	17.192	18.509	20.569	22.465	24.797	46.059	49.802	53.203	57.342	60.275
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
45	24.311	25.901	28.366	30.612	33.350	57.505	61.656	65.410	69.957	73.166
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
55	31.735	33.571	36.398	38.958	42.060	68.796	73.311	77.380	82.292	85.749
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
65	39.383	41.444	44.603	47.450	50.883	79.973	84.821	89.177	94.422	98.105
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
75	47.206	49.475	52.942	56.054	59.795	91.061	96.217	100.839	106.393	110.285
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
85	55.170	57.634	61.389	64.749	68.777	102.079	107.522	112.393	118.236	122.324
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
95	63.250	65.898	69.925	73.520	77.818	113.038	118.752	123.858	129.973	134.247
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.170

TABLA 4 Distribución *F*



Las entradas en la tabla proporcionan los valores F_α, donde α es el área o probabilidad en la cola superior de la distribución *F*. Por ejemplo, con 4 grados de libertad en el numerador, 8 grados de libertad en el denominador y un área de 0.05 en la cola superior, $F_{0.05} = 3.84$.

Grados de libertad en el denominador	Área en la cola superior	Grados de libertad en el numerador																	
		1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	60	100	1000
1	0.10	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	61.22	61.74	62.05	62.26	62.53	62.79	63.01	63.30
	0.05	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	245.95	248.02	249.26	250.10	251.14	252.20	253.04	254.19
	0.025	647.79	799.48	864.15	899.60	921.83	937.11	948.20	956.64	963.28	968.63	984.87	993.08	998.09	1001.40	1005.60	1009.79	1013.16	1017.76
	0.01	4052.18	4999.34	5403.53	5624.26	5763.96	5858.95	5928.33	5980.95	6022.40	6055.93	6156.97	6208.66	6239.86	6260.35	6286.43	6312.97	6333.92	6362.80
2	0.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
	0.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.45	19.46	19.46	19.47	19.48	19.49	19.49
	0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
	0.01	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39	99.40	99.43	99.45	99.46	99.47	99.48	99.48	99.49	99.50
3	0.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.20	5.18	5.17	5.17	5.16	5.15	5.14	5.13
	0.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.63	8.62	8.59	8.57	8.55	8.53
	0.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.25	14.17	14.12	14.08	14.04	13.99	13.96	13.91
	0.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	26.87	26.69	26.58	26.50	26.41	26.32	26.24	26.14
4	0.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
	0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
	0.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.66	8.56	8.50	8.46	8.41	8.36	8.32	8.26
	0.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.20	14.02	13.91	13.84	13.75	13.65	13.58	13.47
5	0.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.324	3.21	3.19	3.17	3.16	3.14	3.13	3.11
	0.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.52	4.50	4.46	4.43	4.41	4.37
	0.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43	6.33	6.27	6.23	6.18	6.12	6.08	6.02
	0.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.72	9.55	9.45	9.38	9.29	9.20	9.13	9.03

TABLA 4 Distribución *F* (*continuación*)

Grados de libertad en el denominador	Área en la cola superior	Grados de libertad en el numerador																	
		1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	60	100	1 000
6	0.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.87	2.84	2.81	2.80	2.78	2.76	2.75	2.72
	0.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.83	3.81	3.77	3.74	3.71	3.67
	0.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.27	5.17	5.11	5.07	5.01	4.96	4.92	4.86
	0.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56	7.40	7.30	7.23	7.14	7.06	6.99	6.89
7	0.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.63	2.59	2.57	2.56	2.54	2.51	2.50	2.47
	0.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.40	3.38	3.34	3.30	3.27	3.23
	0.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.57	4.47	4.40	4.36	4.31	4.25	4.21	4.15
	0.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31	6.16	6.06	5.99	5.91	5.82	5.75	5.66
8	0.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.30
	0.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.11	3.08	3.04	3.01	2.97	2.93
	0.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.10	4.00	3.94	3.89	3.84	3.78	3.74	3.68
	0.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52	5.36	5.26	5.20	5.12	5.03	4.96	4.87
9	0.10	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.34	2.30	2.27	2.25	2.23	2.21	2.19	2.16
	0.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.89	2.86	2.83	2.79	2.76	2.71
	0.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.77	3.67	3.60	3.56	3.51	3.45	3.40	3.34
	0.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96	4.81	4.71	4.65	4.57	4.48	4.41	4.32
10	0.10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.24	2.20	2.17	2.16	2.13	2.11	2.09	2.06
	0.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.73	2.70	2.66	2.62	2.59	2.54
	0.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.52	3.42	3.35	3.31	3.26	3.20	3.15	3.09
	0.01	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56	4.41	4.31	4.25	4.17	4.08	4.01	3.92
11	0.10	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.17	2.12	2.10	2.08	2.05	2.03	2.01	1.98
	0.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.60	2.57	2.53	2.49	2.46	2.41
	0.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.33	3.23	3.16	3.12	3.06	3.00	2.96	2.89
	0.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.25	4.10	4.01	3.94	3.86	3.78	3.71	3.61
12	0.10	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.10	2.06	2.03	2.01	1.99	1.96	1.94	1.91
	0.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.50	2.47	2.43	2.38	2.35	2.30
	0.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.18	3.07	3.01	2.96	2.91	2.85	2.80	2.73
	0.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01	3.86	3.76	3.70	3.62	3.54	3.47	3.37
13	0.10	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
	0.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.41	2.38	2.34	2.30	2.26	2.21
	0.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.05	2.95	2.88	2.84	2.78	2.72	2.67	2.60
	0.01	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.82	3.66	3.57	3.51	3.43	3.34	3.27	3.18
14	0.10	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.01	1.96	1.93	1.99	1.89	1.86	1.83	1.80
	0.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.34	2.31	2.27	2.22	2.19	2.14
	0.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	2.95	2.84	2.78	2.73	2.67	2.61	2.56	2.50
	0.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.66	3.51	3.41	3.35	3.27	3.18	3.11	3.02
15	0.10	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	1.97	1.92	1.89	1.87	1.85	1.82	1.79	1.76
	0.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.28	2.25	2.20	2.16	2.12	2.07
	0.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.86	2.76	2.69	2.64	2.59	2.52	2.47	2.40
	0.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.52	3.37	3.28	3.21	3.13	3.05	2.98	2.88

Grados de libertad en el denominador	Área en la cola superior	Grados de libertad en el numerador																	
		1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	60	100	1 000
16	0.10	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.94	1.89	1.86	1.84	1.81	1.78	1.76	1.72
	0.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.23	2.19	2.15	2.11	2.07	2.02
	0.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.79	2.68	2.61	2.57	2.51	2.45	2.40	2.32
	0.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.41	3.26	3.16	3.10	3.02	2.93	2.86	2.76
17	0.10	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.91	1.86	1.83	1.81	1.78	1.75	1.73	1.69
	0.05	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.18	2.15	2.10	2.06	2.02	1.97
	0.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.72	2.62	2.55	2.50	2.44	2.38	2.33	2.26
	0.01	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.31	3.16	3.07	3.00	2.92	2.83	2.76	2.66
18	0.10	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.89	1.84	1.80	1.78	1.75	1.72	1.70	1.66
	0.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.14	2.11	2.06	2.02	1.98	1.92
	0.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.67	2.56	2.49	2.44	2.38	2.32	2.27	2.20
	0.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.23	3.08	2.98	2.92	2.84	2.75	2.68	2.58
19	0.10	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.86	1.81	1.78	1.76	1.73	1.70	1.67	1.64
	0.05	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.11	2.07	2.03	1.98	1.94	1.88
	0.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.62	2.51	2.44	2.39	2.33	2.27	2.22	2.14
	0.01	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.15	3.00	2.91	2.84	2.76	2.67	2.60	2.50
20	0.10	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.84	1.79	1.76	1.74	1.71	1.68	1.65	1.61
	0.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.07	2.04	1.99	1.95	1.91	1.85
	0.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.57	2.46	2.40	2.35	2.29	2.22	2.17	2.09
	0.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09	2.94	2.84	2.78	2.69	2.61	2.54	2.43
21	0.10	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.83	1.78	1.74	1.72	1.69	1.66	1.63	1.59
	0.05	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.18	2.10	2.05	2.01	1.96	1.92	1.88	1.82
	0.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.53	2.42	2.36	2.31	2.25	2.18	2.13	2.05
	0.01	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.03	2.88	2.79	2.72	2.64	2.55	2.48	2.37
22	0.10	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.81	1.76	1.73	1.70	1.67	1.64	1.61	1.57
	0.05	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.07	2.02	1.98	1.94	1.89	1.85	1.79
	0.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.50	2.39	2.32	2.27	2.21	2.14	2.09	2.01
	0.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	2.98	2.83	2.73	2.67	2.58	2.50	2.42	2.32
23	0.10	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.80	1.74	1.71	1.69	1.66	1.62	1.59	1.55
	0.05	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.13	2.05	2.00	1.96	1.91	1.86	1.82	1.76
	0.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.47	2.36	2.29	2.24	2.18	2.11	2.06	1.98
	0.01	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	2.93	2.78	2.69	2.62	2.54	2.45	2.37	2.27
24	0.10	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.78	1.73	1.70	1.67	1.64	1.61	1.58	1.54
	0.05	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11	2.03	1.97	1.94	1.89	1.84	1.80	1.74
	0.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.44	2.33	2.26	2.21	2.15	2.08	2.02	1.94
	0.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	2.89	2.74	2.64	2.58	2.49	2.40	2.33	2.22

TABLA 4 Distribución *F* (*continuación*)

Grados de libertad en el denominador	Área en la cola superior	Grados de libertad en el numerador																	
		1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	60	100	1 000
25	0.10	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.77	1.72	1.68	1.66	1.63	1.59	1.56	1.52
	0.05	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.09	2.01	1.96	1.92	1.87	1.82	1.78	1.72
	0.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.41	2.30	2.23	2.18	2.12	2.05	2.00	1.91
	0.01	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.85	2.70	2.60	2.54	2.45	2.36	2.29	2.18
26	0.10	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.76	1.71	1.67	1.65	1.61	1.58	1.55	1.51
	0.05	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07	1.99	1.94	1.90	1.85	1.80	1.76	1.70
	0.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.39	2.28	2.21	2.16	2.09	2.03	1.97	1.89
	0.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.81	2.66	2.57	2.50	2.42	2.33	2.25	2.14
27	0.10	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.75	1.70	1.66	1.64	1.60	1.57	1.54	1.50
	0.05	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.06	1.97	1.92	1.88	1.84	1.79	1.74	1.68
	0.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.36	2.25	2.18	2.13	2.07	2.00	1.94	1.86
	0.01	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.78	2.63	2.54	2.47	2.38	2.29	2.22	2.11
28	0.10	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.74	1.69	1.65	1.63	1.59	1.56	1.53	1.48
	0.05	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04	1.96	1.91	1.87	1.82	1.77	1.73	1.66
	0.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.34	2.23	2.16	2.11	2.05	1.98	1.92	1.84
	0.01	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.75	2.60	2.51	2.44	2.35	2.26	2.19	2.08
29	0.10	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.73	1.68	1.64	1.62	1.58	1.55	1.52	1.47
	0.05	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.03	1.94	1.89	1.85	1.81	1.75	1.71	1.65
	0.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.32	2.21	2.14	2.09	2.03	1.96	1.90	1.82
	0.01	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.73	2.57	2.48	2.41	2.33	2.23	2.16	2.05
30	0.10	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.72	1.67	1.63	1.61	1.57	1.54	1.51	1.46
	0.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.88	1.84	1.79	1.74	1.70	1.63
	0.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.31	2.20	2.12	2.07	2.01	1.94	1.88	1.80
	0.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.70	2.55	2.45	2.39	2.30	2.21	2.13	2.02
40	0.10	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.66	1.61	1.57	1.54	1.51	1.47	1.43	1.38
	0.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.78	1.74	1.69	1.64	1.59	1.52
	0.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.18	2.07	1.99	1.94	1.88	1.80	1.74	1.65
	0.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.52	2.37	2.27	2.20	2.11	2.02	1.94	1.82
60	0.10	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.60	1.54	1.50	1.48	1.44	1.40	1.36	1.30
	0.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.84	1.75	1.69	1.65	1.59	1.53	1.48	1.40
	0.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.06	1.94	1.87	1.82	1.74	1.67	1.60	1.49
	0.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.35	2.20	2.10	2.03	1.94	1.84	1.75	1.62
100	0.10	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66	1.56	1.49	1.45	1.42	1.38	1.34	1.29	1.22
	0.05	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77	1.68	1.62	1.57	1.52	1.45	1.39	1.30
	0.025	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	1.97	1.85	1.77	1.71	1.64	1.56	1.48	1.36
	0.01	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.22	2.07	1.97	1.89	1.80	1.69	1.60	1.45
1 000	0.10	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.64	1.61	1.49	1.43	1.38	1.35	1.30	1.25	1.20	1.08
	0.05	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.68	1.58	1.52	1.47	1.41	1.33	1.26	1.11
	0.025	5.04	3.70	3.13	2.80	2.58	2.42	2.30	2.20	2.13	2.06	1.85	1.72	1.64	1.58	1.50	1.41	1.32	1.13
	0.01	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.06	1.90	1.79	1.72	1.61	1.50	1.38	1.16

TABLA 5 Probabilidades binomiales

Las entradas en la tabla proporcionan la probabilidad de x éxitos en n ensayos de un experimento binomial, donde p es la probabilidad de un éxito en un ensayo. Por ejemplo, para seis ensayos y $p = 0.05$, la probabilidad de dos éxitos es de 0.0305.

		p								
n	x	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
2	0	0.9801	0.9604	0.9409	0.9216	0.9025	0.8836	0.8649	0.8464	0.8281
	1	0.0198	0.0392	0.0582	0.0768	0.0950	0.1128	0.1302	0.1472	0.1638
	2	0.0001	0.0004	0.0009	0.0016	0.0025	0.0036	0.0049	0.0064	0.0081
3	0	0.9703	0.9412	0.9127	0.8847	0.8574	0.8306	0.8044	0.7787	0.7536
	1	0.0294	0.0576	0.0847	0.1106	0.1354	0.1590	0.1816	0.2031	0.2236
	2	0.0003	0.0012	0.0026	0.0046	0.0071	0.0102	0.0137	0.0177	0.0221
	3	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0003	0.0005	0.0007
4	0	0.9606	0.9224	0.8853	0.8493	0.8145	0.7807	0.7481	0.7164	0.6857
	1	0.0388	0.0753	0.1095	0.1416	0.1715	0.1993	0.2252	0.2492	0.2713
	2	0.0006	0.0023	0.0051	0.0088	0.0135	0.0191	0.0254	0.0325	0.0402
	3	0.0000	0.0000	0.0001	0.0002	0.0005	0.0008	0.0013	0.0019	0.0027
	4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
5	0	0.9510	0.9039	0.8587	0.8154	0.7738	0.7339	0.6957	0.6591	0.6240
	1	0.0480	0.0922	0.1328	0.1699	0.2036	0.2342	0.2618	0.2866	0.3086
	2	0.0010	0.0038	0.0082	0.0142	0.0214	0.0299	0.0394	0.0498	0.0610
	3	0.0000	0.0001	0.0003	0.0006	0.0011	0.0019	0.0030	0.0043	0.0060
	4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0003
	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	0	0.9415	0.8858	0.8330	0.7828	0.7351	0.6899	0.6470	0.6064	0.5679
	1	0.0571	0.1085	0.1546	0.1957	0.2321	0.2642	0.2922	0.3164	0.3370
	2	0.0014	0.0055	0.0120	0.0204	0.0305	0.0422	0.0550	0.0688	0.0833
	3	0.0000	0.0002	0.0005	0.0011	0.0021	0.0036	0.0055	0.0080	0.0110
	4	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0005	0.0008
	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	0	0.9321	0.8681	0.8080	0.7514	0.6983	0.6485	0.6017	0.5578	0.5168
	1	0.0659	0.1240	0.1749	0.2192	0.2573	0.2897	0.3170	0.3396	0.3578
	2	0.0020	0.0076	0.0162	0.0274	0.0406	0.0555	0.0716	0.0886	0.1061
	3	0.0000	0.0003	0.0008	0.0019	0.0036	0.0059	0.0090	0.0128	0.0175
	4	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0011	0.0017
	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8	0	0.9227	0.8508	0.7837	0.7214	0.6634	0.6096	0.5596	0.5132	0.4703
	1	0.0746	0.1389	0.1939	0.2405	0.2793	0.3113	0.3370	0.3570	0.3721
	2	0.0026	0.0099	0.0210	0.0351	0.0515	0.0695	0.0888	0.1087	0.1288
	3	0.0001	0.0004	0.0013	0.0029	0.0054	0.0089	0.0134	0.0189	0.0255
	4	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0013	0.0021	0.0031
	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

TABLA 5 Probabilidades binomiales (*continuación*)

		p								
n	***x***	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
9	0	0.9135	0.8337	0.7602	0.6925	0.6302	0.5730	0.5204	0.4722	0.4279
	1	0.0830	0.1531	0.2116	0.2597	0.2985	0.3292	0.3525	0.3695	0.3809
	2	0.0034	0.0125	0.0262	0.0433	0.0629	0.0840	0.1061	0.1285	0.1507
	3	0.0001	0.0006	0.0019	0.0042	0.0077	0.0125	0.0186	0.0261	0.0348
	4	0.0000	0.0000	0.0001	0.0003	0.0006	0.0012	0.0021	0.0034	0.0052
	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0005
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
10	0	0.9044	0.8171	0.7374	0.6648	0.5987	0.5386	0.4840	0.4344	0.3894
	1	0.0914	0.1667	0.2281	0.2770	0.3151	0.3438	0.3643	0.3777	0.3851
	2	0.0042	0.0153	0.0317	0.0519	0.0746	0.0988	0.1234	0.1478	0.1714
	3	0.0001	0.0008	0.0026	0.0058	0.0105	0.0168	0.0248	0.0343	0.0452
	4	0.0000	0.0000	0.0001	0.0004	0.0010	0.0019	0.0033	0.0052	0.0078
	5	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0003	0.0005	0.0009
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
12	0	0.8864	0.7847	0.6938	0.6127	0.5404	0.4759	0.4186	0.3677	0.3225
	1	0.1074	0.1922	0.2575	0.3064	0.3413	0.3645	0.3781	0.3837	0.3827
	2	0.0060	0.0216	0.0438	0.0702	0.0988	0.1280	0.1565	0.1835	0.2082
	3	0.0002	0.0015	0.0045	0.0098	0.0173	0.0272	0.0393	0.0532	0.0686
	4	0.0000	0.0001	0.0003	0.0009	0.0021	0.0039	0.0067	0.0104	0.0153
	5	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004	0.0008	0.0014	0.0024
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0003
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
15	0	0.8601	0.7386	0.6333	0.5421	0.4633	0.3953	0.3367	0.2863	0.2430
	1	0.1303	0.2261	0.2938	0.3388	0.3658	0.3785	0.3801	0.3734	0.3605
	2	0.0092	0.0323	0.0636	0.0988	0.1348	0.1691	0.2003	0.2273	0.2496
	3	0.0004	0.0029	0.0085	0.0178	0.0307	0.0468	0.0653	0.0857	0.1070
	4	0.0000	0.0002	0.0008	0.0022	0.0049	0.0090	0.0148	0.0223	0.0317
	5	0.0000	0.0000	0.0001	0.0002	0.0006	0.0013	0.0024	0.0043	0.0069
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0006	0.0011
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

TABLA 5 Probabilidades binomiales (*continuación*)

		p								
n	***x***	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
18	0	0.8345	0.6951	0.5780	0.4796	0.3972	0.3283	0.2708	0.2229	0.1831
	1	0.1517	0.2554	0.3217	0.3597	0.3763	0.3772	0.3669	0.3489	0.3260
	2	0.0130	0.0443	0.0846	0.1274	0.1683	0.2047	0.2348	0.2579	0.2741
	3	0.0007	0.0048	0.0140	0.0283	0.0473	0.0697	0.0942	0.1196	0.1446
	4	0.0000	0.0004	0.0016	0.0044	0.0093	0.0167	0.0266	0.0390	0.0536
	5	0.0000	0.0000	0.0001	0.0005	0.0014	0.0030	0.0056	0.0095	0.0148
	6	0.0000	0.0000	0.0000	0.0000	0.0002	0.0004	0.0009	0.0018	0.0032
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0005
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
20	0	0.8179	0.6676	0.5438	0.4420	0.3585	0.2901	0.2342	0.1887	0.1516
	1	0.1652	0.2725	0.3364	0.3683	0.3774	0.3703	0.3526	0.3282	0.3000
	2	0.0159	0.0528	0.0988	0.1458	0.1887	0.2246	0.2521	0.2711	0.2818
	3	0.0010	0.0065	0.0183	0.0364	0.0596	0.0860	0.1139	0.1414	0.1672
	4	0.0000	0.0006	0.0024	0.0065	0.0133	0.0233	0.0364	0.0523	0.0703
	5	0.0000	0.0000	0.0002	0.0009	0.0022	0.0048	0.0088	0.0145	0.0222
	6	0.0000	0.0000	0.0000	0.0001	0.0003	0.0008	0.0017	0.0032	0.0055
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0005	0.0011
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

TABLA 5 Probabilidades binomiales (*continuación*)

		p								
n	***x***	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
2	0	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500
	1	0.1800	0.2550	0.3200	0.3750	0.4200	0.4550	0.4800	0.4950	0.5000
	2	0.0100	0.0225	0.0400	0.0625	0.0900	0.1225	0.1600	0.2025	0.2500
3	0	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250
	1	0.2430	0.3251	0.3840	0.4219	0.4410	0.4436	0.4320	0.4084	0.3750
	2	0.0270	0.0574	0.0960	0.1406	0.1890	0.2389	0.2880	0.3341	0.3750
	3	0.0010	0.0034	0.0080	0.0156	0.0270	0.0429	0.0640	0.0911	0.1250
4	0	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625
	1	0.2916	0.3685	0.4096	0.4219	0.4116	0.3845	0.3456	0.2995	0.2500
	2	0.0486	0.0975	0.1536	0.2109	0.2646	0.3105	0.3456	0.3675	0.3750
	3	0.0036	0.0115	0.0256	0.0469	0.0756	0.1115	0.1536	0.2005	0.2500
	4	0.0001	0.0005	0.0016	0.0039	0.0081	0.0150	0.0256	0.0410	0.0625
5	0	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0312
	1	0.3280	0.3915	0.4096	0.3955	0.3602	0.3124	0.2592	0.2059	0.1562
	2	0.0729	0.1382	0.2048	0.2637	0.3087	0.3364	0.3456	0.3369	0.3125
	3	0.0081	0.0244	0.0512	0.0879	0.1323	0.1811	0.2304	0.2757	0.3125
	4	0.0004	0.0022	0.0064	0.0146	0.0284	0.0488	0.0768	0.1128	0.1562
	5	0.0000	0.0001	0.0003	0.0010	0.0024	0.0053	0.0102	0.0185	0.0312
6	0	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156
	1	0.3543	0.3993	0.3932	0.3560	0.3025	0.2437	0.1866	0.1359	0.0938
	2	0.0984	0.1762	0.2458	0.2966	0.3241	0.3280	0.3110	0.2780	0.2344
	3	0.0146	0.0415	0.0819	0.1318	0.1852	0.2355	0.2765	0.3032	0.3125
	4	0.0012	0.0055	0.0154	0.0330	0.0595	0.0951	0.1382	0.1861	0.2344
	5	0.0001	0.0004	0.0015	0.0044	0.0102	0.0205	0.0369	0.0609	0.0938
	6	0.0000	0.0000	0.0001	0.0002	0.0007	0.0018	0.0041	0.0083	0.0156
7	0	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078
	1	0.3720	0.3960	0.3670	0.3115	0.2471	0.1848	0.1306	0.0872	0.0547
	2	0.1240	0.2097	0.2753	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641
	3	0.0230	0.0617	0.1147	0.1730	0.2269	0.2679	0.2903	0.2918	0.2734
	4	0.0026	0.0109	0.0287	0.0577	0.0972	0.1442	0.1935	0.2388	0.2734
	5	0.0002	0.0012	0.0043	0.0115	0.0250	0.0466	0.0774	0.1172	0.1641
	6	0.0000	0.0001	0.0004	0.0013	0.0036	0.0084	0.0172	0.0320	0.0547
	7	0.0000	0.0000	0.0000	0.0001	0.0002	0.0006	0.0016	0.0037	0.0078
8	0	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039
	1	0.3826	0.3847	0.3355	0.2670	0.1977	0.1373	0.0896	0.0548	0.0312
	2	0.1488	0.2376	0.2936	0.3115	0.2965	0.2587	0.2090	0.1569	0.1094
	3	0.0331	0.0839	0.1468	0.2076	0.2541	0.2786	0.2787	0.2568	0.2188
	4	0.0046	0.0185	0.0459	0.0865	0.1361	0.1875	0.2322	0.2627	0.2734
	5	0.0004	0.0026	0.0092	0.0231	0.0467	0.0808	0.1239	0.1719	0.2188
	6	0.0000	0.0002	0.0011	0.0038	0.0100	0.0217	0.0413	0.0703	0.1094
	7	0.0000	0.0000	0.0001	0.0004	0.0012	0.0033	0.0079	0.0164	0.0313
	8	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0017	0.0039

TABLA 5 Probabilidades binomiales (*continuación*)

		p								
n	***x***	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
9	0	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020
	1	0.3874	0.3679	0.3020	0.2253	0.1556	0.1004	0.0605	0.0339	0.0176
	2	0.1722	0.2597	0.3020	0.3003	0.2668	0.2162	0.1612	0.1110	0.0703
	3	0.0446	0.1069	0.1762	0.2336	0.2668	0.2716	0.2508	0.2119	0.1641
	4	0.0074	0.0283	0.0661	0.1168	0.1715	0.2194	0.2508	0.2600	0.2461
	5	0.0008	0.0050	0.0165	0.0389	0.0735	0.1181	0.1672	0.2128	0.2461
	6	0.0001	0.0006	0.0028	0.0087	0.0210	0.0424	0.0743	0.1160	0.1641
	7	0.0000	0.0000	0.0003	0.0012	0.0039	0.0098	0.0212	0.0407	0.0703
	8	0.0000	0.0000	0.0000	0.0001	0.0004	0.0013	0.0035	0.0083	0.0176
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0008	0.0020
10	0	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010
	1	0.3874	0.3474	0.2684	0.1877	0.1211	0.0725	0.0403	0.0207	0.0098
	2	0.1937	0.2759	0.3020	0.2816	0.2335	0.1757	0.1209	0.0763	0.0439
	3	0.0574	0.1298	0.2013	0.2503	0.2668	0.2522	0.2150	0.1665	0.1172
	4	0.0112	0.0401	0.0881	0.1460	0.2001	0.2377	0.2508	0.2384	0.2051
	5	0.0015	0.0085	0.0264	0.0584	0.1029	0.1536	0.2007	0.2340	0.2461
	6	0.0001	0.0012	0.0055	0.0162	0.0368	0.0689	0.1115	0.1596	0.2051
	7	0.0000	0.0001	0.0008	0.0031	0.0090	0.0212	0.0425	0.0746	0.1172
	8	0.0000	0.0000	0.0001	0.0004	0.0014	0.0043	0.0106	0.0229	0.0439
	9	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005	0.0016	0.0042	0.0098
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0010
12	0	0.2824	0.1422	0.0687	0.0317	0.0138	0.0057	0.0022	0.0008	0.0002
	1	0.3766	0.3012	0.2062	0.1267	0.0712	0.0368	0.0174	0.0075	0.0029
	2	0.2301	0.2924	0.2835	0.2323	0.1678	0.1088	0.0639	0.0339	0.0161
	3	0.0853	0.1720	0.2362	0.2581	0.2397	0.1954	0.1419	0.0923	0.0537
	4	0.0213	0.0683	0.1329	0.1936	0.2311	0.2367	0.2128	0.1700	0.1208
	5	0.0038	0.0193	0.0532	0.1032	0.1585	0.2039	0.2270	0.2225	0.1934
	6	0.0005	0.0040	0.0155	0.0401	0.0792	0.1281	0.1766	0.2124	0.2256
	7	0.0000	0.0006	0.0033	0.0115	0.0291	0.0591	0.1009	0.1489	0.1934
	8	0.0000	0.0001	0.0005	0.0024	0.0078	0.0199	0.0420	0.0762	0.1208
	9	0.0000	0.0000	0.0001	0.0004	0.0015	0.0048	0.0125	0.0277	0.0537
	10	0.0000	0.0000	0.0000	0.0000	0.0002	0.0008	0.0025	0.0068	0.0161
	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0010	0.0029
	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002
15	0	0.2059	0.0874	0.0352	0.0134	0.0047	0.0016	0.0005	0.0001	0.0000
	1	0.3432	0.2312	0.1319	0.0668	0.0305	0.0126	0.0047	0.0016	0.0005
	2	0.2669	0.2856	0.2309	0.1559	0.0916	0.0476	0.0219	0.0090	0.0032
	3	0.1285	0.2184	0.2501	0.2252	0.1700	0.1110	0.0634	0.0318	0.0139
	4	0.0428	0.1156	0.1876	0.2252	0.2186	0.1792	0.1268	0.0780	0.0417
	5	0.0105	0.0449	0.1032	0.1651	0.2061	0.2123	0.1859	0.1404	0.0916
	6	0.0019	0.0132	0.0430	0.0917	0.1472	0.1906	0.2066	0.1914	0.1527
	7	0.0003	0.0030	0.0138	0.0393	0.0811	0.1319	0.1771	0.2013	0.1964
	8	0.0000	0.0005	0.0035	0.0131	0.0348	0.0710	0.1181	0.1647	0.1964
	9	0.0000	0.0001	0.0007	0.0034	0.0016	0.0298	0.0612	0.1048	0.1527
	10	0.0000	0.0000	0.0001	0.0007	0.0030	0.0096	0.0245	0.0515	0.0916
	11	0.0000	0.0000	0.0000	0.0001	0.0006	0.0024	0.0074	0.0191	0.0417
	12	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0016	0.0052	0.0139
	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0010	0.0032
	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

TABLA 5 Probabilidades binomiales (*continuación*)

		p								
n	***x***	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
18	0	0.1501	0.0536	0.0180	0.0056	0.0016	0.0004	0.0001	0.0000	0.0000
	1	0.3002	0.1704	0.0811	0.0338	0.0126	0.0042	0.0012	0.0003	0.0001
	2	0.2835	0.2556	0.1723	0.0958	0.0458	0.0190	0.0069	0.0022	0.0006
	3	0.1680	0.2406	0.2297	0.1704	0.1046	0.0547	0.0246	0.0095	0.0031
	4	0.0700	0.1592	0.2153	0.2130	0.1681	0.1104	0.0614	0.0291	0.0117
	5	0.0218	0.0787	0.1507	0.1988	0.2017	0.1664	0.1146	0.0666	0.0327
	6	0.0052	0.0301	0.0816	0.1436	0.1873	0.1941	0.1655	0.1181	0.0708
	7	0.0010	0.0091	0.0350	0.0820	0.1376	0.1792	0.1892	0.1657	0.1214
	8	0.0002	0.0022	0.0120	0.0376	0.0811	0.1327	0.1734	0.1864	0.1669
	9	0.0000	0.0004	0.0033	0.0139	0.0386	0.0794	0.1284	0.1694	0.1855
	10	0.0000	0.0001	0.0008	0.0042	0.0149	0.0385	0.0771	0.1248	0.1669
	11	0.0000	0.0000	0.0001	0.0010	0.0046	0.0151	0.0374	0.0742	0.1214
	12	0.0000	0.0000	0.0000	0.0002	0.0012	0.0047	0.0145	0.0354	0.0708
	13	0.0000	0.0000	0.0000	0.0000	0.0002	0.0012	0.0045	0.0134	0.0327
	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0011	0.0039	0.0117
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0009	0.0031
	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006
	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
20	0	0.1216	0.0388	0.0115	0.0032	0.0008	0.0002	0.0000	0.0000	0.0000
	1	0.2702	0.1368	0.0576	0.0211	0.0068	0.0020	0.0005	0.0001	0.0000
	2	0.2852	0.2293	0.1369	0.0669	0.0278	0.0100	0.0031	0.0008	0.0002
	3	0.1901	0.2428	0.2054	0.1339	0.0716	0.0323	0.0123	0.0040	0.0011
	4	0.0898	0.1821	0.2182	0.1897	0.1304	0.0738	0.0350	0.0139	0.0046
	5	0.0319	0.1028	0.1746	0.2023	0.1789	0.1272	0.0746	0.0365	0.0148
	6	0.0089	0.0454	0.1091	0.1686	0.1916	0.1712	0.1244	0.0746	0.0370
	7	0.0020	0.0160	0.0545	0.1124	0.1643	0.1844	0.1659	0.1221	0.0739
	8	0.0004	0.0046	0.0222	0.0609	0.1144	0.1614	0.1797	0.1623	0.1201
	9	0.0001	0.0011	0.0074	0.0271	0.0654	0.1158	0.1597	0.1771	0.1602
	10	0.0000	0.0002	0.0020	0.0099	0.0308	0.0686	0.1171	0.1593	0.1762
	11	0.0000	0.0000	0.0005	0.0030	0.0120	0.0336	0.0710	0.1185	0.1602
	12	0.0000	0.0000	0.0001	0.0008	0.0039	0.0136	0.0355	0.0727	0.1201
	13	0.0000	0.0000	0.0000	0.0002	0.0010	0.0045	0.0146	0.0366	0.0739
	14	0.0000	0.0000	0.0000	0.0000	0.0002	0.0012	0.0049	0.0150	0.0370
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0013	0.0049	0.0148
	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0013	0.0046
	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0011
	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002
	19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

TABLA 5 Probabilidades binomiales (*continuación*)

		p								
n	***x***	**0.55**	**0.60**	**0.65**	**0.70**	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**
2	0	0.2025	0.1600	0.1225	0.0900	0.0625	0.0400	0.0225	0.0100	0.0025
	1	0.4950	0.4800	0.4550	0.4200	0.3750	0.3200	0.2550	0.1800	0.0950
	2	0.3025	0.3600	0.4225	0.4900	0.5625	0.6400	0.7225	0.8100	0.9025
3	0	0.0911	0.0640	0.0429	0.0270	0.0156	0.0080	0.0034	0.0010	0.0001
	1	0.3341	0.2880	0.2389	0.1890	0.1406	0.0960	0.0574	0.0270	0.0071
	2	0.4084	0.4320	0.4436	0.4410	0.4219	0.3840	0.3251	0.2430	0.1354
	3	0.1664	0.2160	0.2746	0.3430	0.4219	0.5120	0.6141	0.7290	0.8574
4	0	0.0410	0.0256	0.0150	0.0081	0.0039	0.0016	0.0005	0.0001	0.0000
	1	0.2005	0.1536	0.1115	0.0756	0.0469	0.0256	0.0115	0.0036	0.0005
	2	0.3675	0.3456	0.3105	0.2646	0.2109	0.1536	0.0975	0.0486	0.0135
	3	0.2995	0.3456	0.3845	0.4116	0.4219	0.4096	0.3685	0.2916	0.1715
	4	0.0915	0.1296	0.1785	0.2401	0.3164	0.4096	0.5220	0.6561	0.8145
5	0	0.0185	0.0102	0.0053	0.0024	0.0010	0.0003	0.0001	0.0000	0.0000
	1	0.1128	0.0768	0.0488	0.0284	0.0146	0.0064	0.0022	0.0005	0.0000
	2	0.2757	0.2304	0.1811	0.1323	0.0879	0.0512	0.0244	0.0081	0.0011
	3	0.3369	0.3456	0.3364	0.3087	0.2637	0.2048	0.1382	0.0729	0.0214
	4	0.2059	0.2592	0.3124	0.3601	0.3955	0.4096	0.3915	0.3281	0.2036
	5	0.0503	0.0778	0.1160	0.1681	0.2373	0.3277	0.4437	0.5905	0.7738
6	0	0.0083	0.0041	0.0018	0.0007	0.0002	0.0001	0.0000	0.0000	0.0000
	1	0.0609	0.0369	0.0205	0.0102	0.0044	0.0015	0.0004	0.0001	0.0000
	2	0.1861	0.1382	0.0951	0.0595	0.0330	0.0154	0.0055	0.0012	0.0001
	3	0.3032	0.2765	0.2355	0.1852	0.1318	0.0819	0.0415	0.0146	0.0021
	4	0.2780	0.3110	0.3280	0.3241	0.2966	0.2458	0.1762	0.0984	0.0305
	5	0.1359	0.1866	0.2437	0.3025	0.3560	0.3932	0.3993	0.3543	0.2321
	6	0.0277	0.0467	0.0754	0.1176	0.1780	0.2621	0.3771	0.5314	0.7351
7	0	0.0037	0.0016	0.0006	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
	1	0.0320	0.0172	0.0084	0.0036	0.0013	0.0004	0.0001	0.0000	0.0000
	2	0.1172	0.0774	0.0466	0.0250	0.0115	0.0043	0.0012	0.0002	0.0000
	3	0.2388	0.1935	0.1442	0.0972	0.0577	0.0287	0.0109	0.0026	0.0002
	4	0.2918	0.2903	0.2679	0.2269	0.1730	0.1147	0.0617	0.0230	0.0036
	5	0.2140	0.2613	0.2985	0.3177	0.3115	0.2753	0.2097	0.1240	0.0406
	6	0.0872	0.1306	0.1848	0.2471	0.3115	0.3670	0.3960	0.3720	0.2573
	7	0.0152	0.0280	0.0490	0.0824	0.1335	0.2097	0.3206	0.4783	0.6983
8	0	0.0017	0.0007	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0164	0.0079	0.0033	0.0012	0.0004	0.0001	0.0000	0.0000	0.0000
	2	0.0703	0.0413	0.0217	0.0100	0.0038	0.0011	0.0002	0.0000	0.0000
	3	0.1719	0.1239	0.0808	0.0467	0.0231	0.0092	0.0026	0.0004	0.0000
	4	0.2627	0.2322	0.1875	0.1361	0.0865	0.0459	0.0185	0.0046	0.0004
	5	0.2568	0.2787	0.2786	0.2541	0.2076	0.1468	0.0839	0.0331	0.0054
	6	0.1569	0.2090	0.2587	0.2965	0.3115	0.2936	0.2376	0.1488	0.0515
	7	0.0548	0.0896	0.1373	0.1977	0.2670	0.3355	0.3847	0.3826	0.2793
	8	0.0084	0.0168	0.0319	0.0576	0.1001	0.1678	0.2725	0.4305	0.6634

TABLA 5 Probabilidades binomiales (*continuación*)

		p								
n	***x***	**0.55**	**0.60**	**0.65**	**0.70**	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**
9	0	0.0008	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0083	0.0035	0.0013	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000
	2	0.0407	0.0212	0.0098	0.0039	0.0012	0.0003	0.0000	0.0000	0.0000
	3	0.1160	0.0743	0.0424	0.0210	0.0087	0.0028	0.0006	0.0001	0.0000
	4	0.2128	0.1672	0.1181	0.0735	0.0389	0.0165	0.0050	0.0008	0.0000
	5	0.2600	0.2508	0.2194	0.1715	0.1168	0.0661	0.0283	0.0074	0.0006
	6	0.2119	0.2508	0.2716	0.2668	0.2336	0.1762	0.1069	0.0446	0.0077
	7	0.1110	0.1612	0.2162	0.2668	0.3003	0.3020	0.2597	0.1722	0.0629
	8	0.0339	0.0605	0.1004	0.1556	0.2253	0.3020	0.3679	0.3874	0.2985
	9	0.0046	0.0101	0.0207	0.0404	0.0751	0.1342	0.2316	0.3874	0.6302
10	0	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0042	0.0016	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0229	0.0106	0.0043	0.0014	0.0004	0.0001	0.0000	0.0000	0.0000
	3	0.0746	0.0425	0.0212	0.0090	0.0031	0.0008	0.0001	0.0000	0.0000
	4	0.1596	0.1115	0.0689	0.0368	0.0162	0.0055	0.0012	0.0001	0.0000
	5	0.2340	0.2007	0.1536	0.1029	0.0584	0.0264	0.0085	0.0015	0.0001
	6	0.2384	0.2508	0.2377	0.2001	0.1460	0.0881	0.0401	0.0112	0.0010
	7	0.1665	0.2150	0.2522	0.2668	0.2503	0.2013	0.1298	0.0574	0.0105
	8	0.0763	0.1209	0.1757	0.2335	0.2816	0.3020	0.2759	0.1937	0.0746
	9	0.0207	0.0403	0.0725	0.1211	0.1877	0.2684	0.3474	0.3874	0.3151
	10	0.0025	0.0060	0.0135	0.0282	0.0563	0.1074	0.1969	0.3487	0.5987
12	0	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0010	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0068	0.0025	0.0008	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0277	0.0125	0.0048	0.0015	0.0004	0.0001	0.0000	0.0000	0.0000
	4	0.0762	0.0420	0.0199	0.0078	0.0024	0.0005	0.0001	0.0000	0.0000
	5	0.1489	0.1009	0.0591	0.0291	0.0115	0.0033	0.0006	0.0000	0.0000
	6	0.2124	0.1766	0.1281	0.0792	0.0401	0.0155	0.0040	0.0005	0.0000
	7	0.2225	0.2270	0.2039	0.1585	0.1032	0.0532	0.0193	0.0038	0.0002
	8	0.1700	0.2128	0.2367	0.2311	0.1936	0.1329	0.0683	0.0213	0.0021
	9	0.0923	0.1419	0.1954	0.2397	0.2581	0.2362	0.1720	0.0852	0.0173
	10	0.0339	0.0639	0.1088	0.1678	0.2323	0.2835	0.2924	0.2301	0.0988
	11	0.0075	0.0174	0.0368	0.0712	0.1267	0.2062	0.3012	0.3766	0.3413
	12	0.0008	0.0022	0.0057	0.0138	0.0317	0.0687	0.1422	0.2824	0.5404
15	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0010	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0052	0.0016	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.0191	0.0074	0.0024	0.0006	0.0001	0.0000	0.0000	0.0000	0.0000
	5	0.0515	0.0245	0.0096	0.0030	0.0007	0.0001	0.0000	0.0000	0.0000
	6	0.1048	0.0612	0.0298	0.0116	0.0034	0.0007	0.0001	0.0000	0.0000
	7	0.1647	0.1181	0.0710	0.0348	0.0131	0.0035	0.0005	0.0000	0.0000
	8	0.2013	0.1771	0.1319	0.0811	0.0393	0.0138	0.0030	0.0003	0.0000
	9	0.1914	0.2066	0.1906	0.1472	0.0917	0.0430	0.0132	0.0019	0.0000
	10	0.1404	0.1859	0.2123	0.2061	0.1651	0.1032	0.0449	0.0105	0.0006
	11	0.0780	0.1268	0.1792	0.2186	0.2252	0.1876	0.1156	0.0428	0.0049

TABLA 5 Probabilidades binomiales (*continuación*)

		p								
n	***x***	**0.55**	**0.60**	**0.65**	**0.70**	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**
	12	0.0318	0.0634	0.1110	0.1700	0.2252	0.2501	0.2184	0.1285	0.0307
	13	0.0090	0.0219	0.0476	0.0916	0.1559	0.2309	0.2856	0.2669	0.1348
	14	0.0016	0.0047	0.0126	0.0305	0.0668	0.1319	0.2312	0.3432	0.3658
	15	0.0001	0.0005	0.0016	0.0047	0.0134	0.0352	0.0874	0.2059	0.4633
18	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0009	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.0039	0.0011	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.0134	0.0045	0.0012	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
	6	0.0354	0.0145	0.0047	0.0012	0.0002	0.0000	0.0000	0.0000	0.0000
	7	0.0742	0.0374	0.0151	0.0046	0.0010	0.0001	0.0000	0.0000	0.0000
	8	0.1248	0.0771	0.0385	0.0149	0.0042	0.0008	0.0001	0.0000	0.0000
	9	0.1694	0.1284	0.0794	0.0386	0.0139	0.0033	0.0004	0.0000	0.0000
	10	0.1864	0.1734	0.1327	0.0811	0.0376	0.0120	0.0022	0.0002	0.0000
	11	0.1657	0.1892	0.1792	0.1376	0.0820	0.0350	0.0091	0.0010	0.0000
	12	0.1181	0.1655	0.1941	0.1873	0.1436	0.0816	0.0301	0.0052	0.0002
	13	0.0666	0.1146	0.1664	0.2017	0.1988	0.1507	0.0787	0.0218	0.0014
	14	0.0291	0.0614	0.1104	0.1681	0.2130	0.2153	0.1592	0.0700	0.0093
	15	0.0095	0.0246	0.0547	0.1046	0.1704	0.2297	0.2406	0.1680	0.0473
	16	0.0022	0.0069	0.0190	0.0458	0.0958	0.1723	0.2556	0.2835	0.1683
	17	0.0003	0.0012	0.0042	0.0126	0.0338	0.0811	0.1704	0.3002	0.3763
	18	0.0000	0.0001	0.0004	0.0016	0.0056	0.0180	0.0536	0.1501	0.3972
20	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.0013	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.0049	0.0013	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	6	0.0150	0.0049	0.0012	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
	7	0.0366	0.0146	0.0045	0.0010	0.0002	0.0000	0.0000	0.0000	0.0000
	8	0.0727	0.0355	0.0136	0.0039	0.0008	0.0001	0.0000	0.0000	0.0000
	9	0.1185	0.0710	0.0336	0.0120	0.0030	0.0005	0.0000	0.0000	0.0000
	10	0.1593	0.1171	0.0686	0.0308	0.0099	0.0020	0.0002	0.0000	0.0000
	11	0.1771	0.1597	0.1158	0.0654	0.0271	0.0074	0.0011	0.0001	0.0000
	12	0.1623	0.1797	0.1614	0.1144	0.0609	0.0222	0.0046	0.0004	0.0000
	13	0.1221	0.1659	0.1844	0.1643	0.1124	0.0545	0.0160	0.0020	0.0000
	14	0.0746	0.1244	0.1712	0.1916	0.1686	0.1091	0.0454	0.0089	0.0003
	15	0.0365	0.0746	0.1272	0.1789	0.2023	0.1746	0.1028	0.0319	0.0022
	16	0.0139	0.0350	0.0738	0.1304	0.1897	0.2182	0.1821	0.0898	0.0133
	17	0.0040	0.0123	0.0323	0.0716	0.1339	0.2054	0.2428	0.1901	0.0596
	18	0.0008	0.0031	0.0100	0.0278	0.0669	0.1369	0.2293	0.2852	0.1887
	19	0.0001	0.0005	0.0020	0.0068	0.0211	0.0576	0.1368	0.2702	0.3774
	20	0.0000	0.0000	0.0002	0.0008	0.0032	0.0115	0.0388	0.1216	0.3585

TABLA 6 Valores de $e^{-\mu}$

μ	$e^{-\mu}$	μ	$e^{-\mu}$	μ	$e^{-\mu}$
0.00	1.0000	2.00	0.1353	4.00	0.0183
0.05	0.9512	2.05	0.1287	4.05	0.0174
0.10	0.9048	2.10	0.1225	4.10	0.0166
0.15	0.8607	2.15	0.1165	4.15	0.0158
0.20	0.8187	2.20	0.1108	4.20	0.0150
0.25	0.7788	2.25	0.1054	4.25	0.0143
0.30	0.7408	2.30	0.1003	4.30	0.0136
0.35	0.7047	2.35	0.0954	4.35	0.0129
0.40	0.6703	2.40	0.0907	4.40	0.0123
0.45	0.6376	2.45	0.0863	4.45	0.0117
0.50	0.6065	2.50	0.0821	4.50	0.0111
0.55	0.5769	2.55	0.0781	4.55	0.0106
0.60	0.5488	2.60	0.0743	4.60	0.0101
0.65	0.5220	2.65	0.0707	4.65	0.0096
0.70	0.4966	2.70	0.0672	4.70	0.0091
0.75	0.4724	2.75	0.0639	4.75	0.0087
0.80	0.4493	2.80	0.0608	4.80	0.0082
0.85	0.4274	2.85	0.0578	4.85	0.0078
0.90	0.4066	2.90	0.0550	4.90	0.0074
0.95	0.3867	2.95	0.0523	4.95	0.0071
1.00	0.3679	3.00	0.0498	5.00	0.0067
1.05	0.3499	3.05	0.0474	6.00	0.0025
1.10	0.3329	3.10	0.0450	7.00	0.0009
1.15	0.3166	3.15	0.0429	8.00	0.000335
1.20	0.3012	3.20	0.0408	9.00	0.000123
				10.00	0.000045
1.25	0.2865	3.25	0.0388		
1.30	0.2725	3.30	0.0369		
1.35	0.2592	3.35	0.0351		
1.40	0.2466	3.40	0.0334		
1.45	0.2346	3.45	0.0317		
1.50	0.2231	3.50	0.0302		
1.55	0.2122	3.55	0.0287		
1.60	0.2019	3.60	0.0273		
1.65	0.1920	3.65	0.0260		
1.70	0.1827	3.70	0.0247		
1.75	0.1738	3.75	0.0235		
1.80	0.1653	3.80	0.0224		
1.85	0.1572	3.85	0.0213		
1.90	0.1496	3.90	0.0202		
1.95	0.1423	3.95	0.0193		

TABLA 7 Probabilidades de Poisson

Las entradas en la tabla proporcionan la probabilidad de que se presenten x eventos u ocurrencias en un proceso de Poisson con una media μ. Por ejemplo, cuando $\mu = 2.5$, la probabilidad de que ocurran 4 eventos es 0.1336.

	μ									
x	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679
1	0.0905	0.1637	0.2222	0.2681	0.3033	0.3293	0.3476	0.3595	0.3659	0.3679
2	0.0045	0.0164	0.0333	0.0536	0.0758	0.0988	0.1217	0.1438	0.1647	0.1839
3	0.0002	0.0011	0.0033	0.0072	0.0126	0.0198	0.0284	0.0383	0.0494	0.0613
4	0.0000	0.0001	0.0002	0.0007	0.0016	0.0030	0.0050	0.0077	0.0111	0.0153
5	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0012	0.0020	0.0031
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0005
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001

	μ									
x	**1.1**	**1.2**	**1.3**	**1.4**	**1.5**	**1.6**	**1.7**	**1.8**	**1.9**	**2.0**
0	0.3329	0.3012	0.2725	0.2466	0.2231	0.2019	0.1827	0.1653	0.1496	0.1353
1	0.3662	0.3614	0.3543	0.3452	0.3347	0.3230	0.3106	0.2975	0.2842	0.2707
2	0.2014	0.2169	0.2303	0.2417	0.2510	0.2584	0.2640	0.2678	0.2700	0.2707
3	0.0738	0.0867	0.0998	0.1128	0.1255	0.1378	0.1496	0.1607	0.1710	0.1804
4	0.0203	0.0260	0.0324	0.0395	0.0471	0.0551	0.0636	0.0723	0.0812	0.0902
5	0.0045	0.0062	0.0084	0.0111	0.0141	0.0176	0.0216	0.0260	0.0309	0.0361
6	0.0008	0.0012	0.0018	0.0026	0.0035	0.0047	0.0061	0.0078	0.0098	0.0120
7	0.0001	0.0002	0.0003	0.0005	0.0008	0.0011	0.0015	0.0020	0.0027	0.0034
8	0.0000	0.0000	0.0001	0.0001	0.0001	0.0002	0.0003	0.0005	0.0006	0.0009
9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0002

	μ									
x	**2.1**	**2.2**	**2.3**	**2.4**	**2.5**	**2.6**	**2.7**	**2.8**	**2.9**	**3.0**
0	0.1225	0.1108	0.1003	0.0907	0.0821	0.0743	0.0672	0.0608	0.0550	0.0498
1	0.2572	0.2438	0.2306	0.2177	0.2052	0.1931	0.1815	0.1703	0.1596	0.1494
2	0.2700	0.2681	0.2652	0.2613	0.2565	0.2510	0.2450	0.2384	0.2314	0.2240
3	0.1890	0.1966	0.2033	0.2090	0.2138	0.2176	0.2205	0.2225	0.2237	0.2240
4	0.0992	0.1082	0.1169	0.1254	0.1336	0.1414	0.1488	0.1557	0.1622	0.1680
5	0.0417	0.0476	0.0538	0.0602	0.0668	0.0735	0.0804	0.0872	0.0940	0.1008
6	0.0146	0.0174	0.0206	0.0241	0.0278	0.0319	0.0362	0.0407	0.0455	0.0504
7	0.0044	0.0055	0.0068	0.0083	0.0099	0.0118	0.0139	0.0163	0.0188	0.0216
8	0.0011	0.0015	0.0019	0.0025	0.0031	0.0038	0.0047	0.0057	0.0068	0.0081
9	0.0003	0.0004	0.0005	0.0007	0.0009	0.0011	0.0014	0.0018	0.0022	0.0027
10	0.0001	0.0001	0.0001	0.0002	0.0002	0.0003	0.0004	0.0005	0.0006	0.0008
11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0002	0.0002
12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001

TABLA 7 Probabilidades de Poisson (*continuación*)

	μ									
x	**3.1**	**3.2**	**3.3**	**3.4**	**3.5**	**3.6**	**3.7**	**3.8**	**3.9**	**4.0**
0	0.0450	0.0408	0.0369	0.0344	0.0302	0.0273	0.0247	0.0224	0.0202	0.0183
1	0.1397	0.1304	0.1217	0.1135	0.1057	0.0984	0.0915	0.0850	0.0789	0.0733
2	0.2165	0.2087	0.2008	0.1929	0.1850	0.1771	0.1692	0.1615	0.1539	0.1465
3	0.2237	0.2226	0.2209	0.2186	0.2158	0.2125	0.2087	0.2046	0.2001	0.1954
4	0.1734	0.1781	0.1823	0.1858	0.1888	0.1912	0.1931	0.1944	0.1951	0.1954
5	0.1075	0.1140	0.1203	0.1264	0.1322	0.1377	0.1429	0.1477	0.1522	0.1563
6	0.0555	0.0608	0.0662	0.0716	0.0771	0.0826	0.0881	0.0936	0.0989	0.1042
7	0.0246	0.0278	0.0312	0.0348	0.0385	0.0425	0.0466	0.0508	0.0551	0.0595
8	0.0095	0.0111	0.0129	0.0148	0.0169	0.0191	0.0215	0.0241	0.0269	0.0298
9	0.0033	0.0040	0.0047	0.0056	0.0066	0.0076	0.0089	0.0102	0.0116	0.0132
10	0.0010	0.0013	0.0016	0.0019	0.0023	0.0028	0.0033	0.0039	0.0045	0.0053
11	0.0003	0.0004	0.0005	0.0006	0.0007	0.0009	0.0011	0.0013	0.0016	0.0019
12	0.0001	0.0001	0.0001	0.0002	0.0002	0.0003	0.0003	0.0004	0.0005	0.0006
13	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002
14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001

	μ									
x	**4.1**	**4.2**	**4.3**	**4.4**	**4.5**	**4.6**	**4.7**	**4.8**	**4.9**	**5.0**
0	0.0166	0.0150	0.0136	0.0123	0.0111	0.0101	0.0091	0.0082	0.0074	0.0067
1	0.0679	0.0630	0.0583	0.0540	0.0500	0.0462	0.0427	0.0395	0.0365	0.0337
2	0.1393	0.1323	0.1254	0.1188	0.1125	0.1063	0.1005	0.0948	0.0894	0.0842
3	0.1904	0.1852	0.1798	0.1743	0.1687	0.1631	0.1574	0.1517	0.1460	0.1404
4	0.1951	0.1944	0.1933	0.1917	0.1898	0.1875	0.1849	0.1820	0.1789	0.1755
5	0.1600	0.1633	0.1662	0.1687	0.1708	0.1725	0.1738	0.1747	0.1753	0.1755
6	0.1093	0.1143	0.1191	0.1237	0.1281	0.1323	0.1362	0.1398	0.1432	0.1462
7	0.0640	0.0686	0.0732	0.0778	0.0824	0.0869	0.0914	0.0959	0.1002	0.1044
8	0.0328	0.0360	0.0393	0.0428	0.0463	0.0500	0.0537	0.0575	0.0614	0.0653
9	0.0150	0.0168	0.0188	0.0209	0.0232	0.0255	0.0280	0.0307	0.0334	0.0363
10	0.0061	0.0071	0.0081	0.0092	0.0104	0.0118	0.0132	0.0147	0.0164	0.0181
11	0.0023	0.0027	0.0032	0.0037	0.0043	0.0049	0.0056	0.0064	0.0073	0.0082
12	0.0008	0.0009	0.0011	0.0014	0.0016	0.0019	0.0022	0.0026	0.0030	0.0034
13	0.0002	0.0003	0.0004	0.0005	0.0006	0.0007	0.0008	0.0009	0.0011	0.0013
14	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0003	0.0003	0.0004	0.0005
15	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002

	μ									
x	**5.1**	**5.2**	**5.3**	**5.4**	**5.5**	**5.6**	**5.7**	**5.8**	**5.9**	**6.0**
0	0.0061	0.0055	0.0050	0.0045	0.0041	0.0037	0.0033	0.0030	0.0027	0.0025
1	0.0311	0.0287	0.0265	0.0244	0.0225	0.0207	0.0191	0.0176	0.0162	0.0149
2	0.0793	0.0746	0.0701	0.0659	0.0618	0.0580	0.0544	0.0509	0.0477	0.0446
3	0.1348	0.1293	0.1239	0.1185	0.1133	0.1082	0.1033	0.0985	0.0938	0.0892
4	0.1719	0.1681	0.1641	0.1600	0.1558	0.1515	0.1472	0.1428	0.1383	0.1339

TABLA 7 Probabilidades de Poisson (*continuación*)

	μ									
x	**5.1**	**5.2**	**5.3**	**5.4**	**5.5**	**5.6**	**5.7**	**5.8**	**5.9**	**6.0**
5	0.1753	0.1748	0.1740	0.1728	0.1714	0.1697	0.1678	0.1656	0.1632	0.1606
6	0.1490	0.1515	0.1537	0.1555	0.1571	0.1587	0.1594	0.1601	0.1605	0.1606
7	0.1086	0.1125	0.1163	0.1200	0.1234	0.1267	0.1298	0.1326	0.1353	0.1377
8	0.0692	0.0731	0.0771	0.0810	0.0849	0.0887	0.0925	0.0962	0.0998	0.1033
9	0.0392	0.0423	0.0454	0.0486	0.0519	0.0552	0.0586	0.0620	0.0654	0.0688
10	0.0200	0.0220	0.0241	0.0262	0.0285	0.0309	0.0334	0.0359	0.0386	0.0413
11	0.0093	0.0104	0.0116	0.0129	0.0143	0.0157	0.0173	0.0190	0.0207	0.0225
12	0.0039	0.0045	0.0051	0.0058	0.0065	0.0073	0.0082	0.0092	0.0102	0.0113
13	0.0015	0.0018	0.0021	0.0024	0.0028	0.0032	0.0036	0.0041	0.0046	0.0052
14	0.0006	0.0007	0.0008	0.0009	0.0011	0.0013	0.0015	0.0017	0.0019	0.0022
15	0.0002	0.0002	0.0003	0.0003	0.0004	0.0005	0.0006	0.0007	0.0008	0.0009
16	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002	0.0003	0.0003
17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001

	μ									
x	**6.1**	**6.2**	**6.3**	**6.4**	**6.5**	**6.6**	**6.7**	**6.8**	**6.9**	**7.0**
0	0.0022	0.0020	0.0018	0.0017	0.0015	0.0014	0.0012	0.0011	0.0010	0.0009
1	0.0137	0.0126	0.0116	0.0106	0.0098	0.0090	0.0082	0.0076	0.0070	0.0064
2	0.0417	0.0390	0.0364	0.0340	0.0318	0.0296	0.0276	0.0258	0.0240	0.0223
3	0.0848	0.0806	0.0765	0.0726	0.0688	0.0652	0.0617	0.0584	0.0552	0.0521
4	0.1294	0.1249	0.1205	0.1162	0.1118	0.1076	0.1034	0.0992	0.0952	0.0912
5	0.1579	0.1549	0.1519	0.1487	0.1454	0.1420	0.1385	0.1349	0.1314	0.1277
6	0.1605	0.1601	0.1595	0.1586	0.1575	0.1562	0.1546	0.1529	0.1511	0.1490
7	0.1399	0.1418	0.1435	0.1450	0.1462	0.1472	0.1480	0.1486	0.1489	0.1490
8	0.1066	0.1099	0.1130	0.1160	0.1188	0.1215	0.1240	0.1263	0.1284	0.1304
9	0.0723	0.0757	0.0791	0.0825	0.0858	0.0891	0.0923	0.0954	0.0985	0.1014
10	0.0441	0.0469	0.0498	0.0528	0.0558	0.0588	0.0618	0.0649	0.0679	0.0710
11	0.0245	0.0265	0.0285	0.0307	0.0330	0.0353	0.0377	0.0401	0.0426	0.0452
12	0.0124	0.0137	0.0150	0.0164	0.0179	0.0194	0.0210	0.0227	0.0245	0.0264
13	0.0058	0.0065	0.0073	0.0081	0.0089	0.0098	0.0108	0.0119	0.0130	0.0142
14	0.0025	0.0029	0.0033	0.0037	0.0041	0.0046	0.0052	0.0058	0.0064	0.0071
15	0.0010	0.0012	0.0014	0.0016	0.0018	0.0020	0.0023	0.0026	0.0029	0.0033
16	0.0004	0.0005	0.0005	0.0006	0.0007	0.0008	0.0010	0.0011	0.0013	0.0014
17	0.0001	0.0002	0.0002	0.0002	0.0003	0.0003	0.0004	0.0004	0.0005	0.0006
18	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002
19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001

	μ									
x	**7.1**	**7.2**	**7.3**	**7.4**	**7.5**	**7.6**	**7.7**	**7.8**	**7.9**	**8.0**
0	0.0008	0.0007	0.0007	0.0006	0.0006	0.0005	0.0005	0.0004	0.0004	0.0003
1	0.0059	0.0054	0.0049	0.0045	0.0041	0.0038	0.0035	0.0032	0.0029	0.0027
2	0.0208	0.0194	0.0180	0.0167	0.0156	0.0145	0.0134	0.0125	0.0116	0.0107
3	0.0492	0.0464	0.0438	0.0413	0.0389	0.0366	0.0345	0.0324	0.0305	0.0286
4	0.0874	0.0836	0.0799	0.0764	0.0729	0.0696	0.0663	0.0632	0.0602	0.0573

TABLA 7 Probabilidades de Poisson (*continuación*)

	μ									
x	**7.1**	**7.2**	**7.3**	**7.4**	**7.5**	**7.6**	**7.7**	**7.8**	**7.9**	**8.0**
5	0.1241	0.1204	0.1167	0.1130	0.1094	0.1057	0.1021	0.0986	0.0951	0.0916
6	0.1468	0.1445	0.1420	0.1394	0.1367	0.1339	0.1311	0.1282	0.1252	0.1221
7	0.1489	0.1486	0.1481	0.1474	0.1465	0.1454	0.1442	0.1428	0.1413	0.1396
8	0.1321	0.1337	0.1351	0.1363	0.1373	0.1382	0.1388	0.1392	0.1395	0.1396
9	0.1042	0.1070	0.1096	0.1121	0.1144	0.1167	0.1187	0.1207	0.1224	0.1241
10	0.0740	0.0770	0.0800	0.0829	0.0858	0.0887	0.0914	0.0941	0.0967	0.0993
11	0.0478	0.0504	0.0531	0.0558	0.0585	0.0613	0.0640	0.0667	0.0695	0.0722
12	0.0283	0.0303	0.0323	0.0344	0.0366	0.0388	0.0411	0.0434	0.0457	0.0481
13	0.0154	0.0168	0.0181	0.0196	0.0211	0.0227	0.0243	0.0260	0.0278	0.0296
14	0.0078	0.0086	0.0095	0.0104	0.0113	0.0123	0.0134	0.0145	0.0157	0.0169
15	0.0037	0.0041	0.0046	0.0051	0.0057	0.0062	0.0069	0.0075	0.0083	0.0090
16	0.0016	0.0019	0.0021	0.0024	0.0026	0.0030	0.0033	0.0037	0.0041	0.0045
17	0.0007	0.0008	0.0009	0.0010	0.0012	0.0013	0.0015	0.0017	0.0019	0.0021
18	0.0003	0.0003	0.0004	0.0004	0.0005	0.0006	0.0006	0.0007	0.0008	0.0009
19	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002	0.0003	0.0003	0.0003	0.0004
20	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002
21	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001

	μ									
x	**8.1**	**8.2**	**8.3**	**8.4**	**8.5**	**8.6**	**8.7**	**8.8**	**8.9**	**9.0**
0	0.0003	0.0003	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001
1	0.0025	0.0023	0.0021	0.0019	0.0017	0.0016	0.0014	0.0013	0.0012	0.0011
2	0.0100	0.0092	0.0086	0.0079	0.0074	0.0068	0.0063	0.0058	0.0054	0.0050
3	0.0269	0.0252	0.0237	0.0222	0.0208	0.0195	0.0183	0.0171	0.0160	0.0150
4	0.0544	0.0517	0.0491	0.0466	0.0443	0.0420	0.0398	0.0377	0.0357	0.0337
5	0.0882	0.0849	0.0816	0.0784	0.0752	0.0722	0.0692	0.0663	0.0635	0.0607
6	0.1191	0.1160	0.1128	0.1097	0.1066	0.1034	0.1003	0.0972	0.0941	0.0911
7	0.1378	0.1358	0.1338	0.1317	0.1294	0.1271	0.1247	0.1222	0.1197	0.1171
8	0.1395	0.1392	0.1388	0.1382	0.1375	0.1366	0.1356	0.1344	0.1332	0.1318
9	0.1256	0.1269	0.1280	0.1290	0.1299	0.1306	0.1311	0.1315	0.1317	0.1318
10	0.1017	0.1040	0.1063	0.1084	0.1104	0.1123	0.1140	0.1157	0.1172	0.1186
11	0.0749	0.0776	0.0802	0.0828	0.0853	0.0878	0.0902	0.0925	0.0948	0.0970
12	0.0505	0.0530	0.0555	0.0579	0.0604	0.0629	0.0654	0.0679	0.0703	0.0728
13	0.0315	0.0334	0.0354	0.0374	0.0395	0.0416	0.0438	0.0459	0.0481	0.0504
14	0.0182	0.0196	0.0210	0.0225	0.0240	0.0256	0.0272	0.0289	0.0306	0.0324
15	0.0098	0.0107	0.0116	0.0126	0.0136	0.0147	0.0158	0.0169	0.0182	0.1094
16	0.0050	0.0055	0.0060	0.0066	0.0072	0.0079	0.0086	0.0093	0.0101	0.0109
17	0.0024	0.0026	0.0029	0.0033	0.0036	0.0040	0.0044	0.0048	0.0053	0.0058
18	0.0011	0.0012	0.0014	0.0015	0.0017	0.0019	0.0021	0.0024	0.0026	0.0029
19	0.0005	0.0005	0.0006	0.0007	0.0008	0.0009	0.0010	0.0011	0.0012	0.0014
20	0.0002	0.0002	0.0002	0.0003	0.0003	0.0004	0.0004	0.0005	0.0005	0.0006
21	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002	0.0002	0.0003
22	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

TABLA 7 Probabilidades de Poisson (*continuación*)

	μ									
x	**9.1**	**9.2**	**9.3**	**9.4**	**9.5**	**9.6**	**9.7**	**9.8**	**9.9**	**10**
0	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000
1	0.0010	0.0009	0.0009	0.0008	0.0007	0.0007	0.0006	0.0005	0.0005	0.0005
2	0.0046	0.0043	0.0040	0.0037	0.0034	0.0031	0.0029	0.0027	0.0025	0.0023
3	0.0140	0.0131	0.0123	0.0115	0.0107	0.0100	0.0093	0.0087	0.0081	0.0076
4	0.0319	0.0302	0.0285	0.0269	0.0254	0.0240	0.0226	0.0213	0.0201	0.0189
5	0.0581	0.0555	0.0530	0.0506	0.0483	0.0460	0.0439	0.0418	0.0398	0.0378
6	0.0881	0.0851	0.0822	0.0793	0.0764	0.0736	0.0709	0.0682	0.0656	0.0631
7	0.1145	0.1118	0.1091	0.1064	0.1037	0.1010	0.0982	0.0955	0.0928	0.0901
8	0.1302	0.1286	0.1269	0.1251	0.1232	0.1212	0.1191	0.1170	0.1148	0.1126
9	0.1317	0.1315	0.1311	0.1306	0.1300	0.1293	0.1284	0.1274	0.1263	0.1251
10	0.1198	0.1210	0.1219	0.1228	0.1235	0.1241	0.1245	0.1249	0.1250	0.1251
11	0.0991	0.1012	0.1031	0.1049	0.1067	0.1083	0.1098	0.1112	0.1125	0.1137
12	0.0752	0.0776	0.0799	0.0822	0.0844	0.0866	0.0888	0.0908	0.0928	0.0948
13	0.0526	0.0549	0.0572	0.0594	0.0617	0.0640	0.0662	0.0685	0.0707	0.0729
14	0.0342	0.0361	0.0380	0.0399	0.0419	0.0439	0.0459	0.0479	0.0500	0.0521
15	0.0208	0.0221	0.0235	0.0250	0.0265	0.0281	0.0297	0.0313	0.0330	0.0347
16	0.0118	0.0127	0.0137	0.0147	0.0157	0.0168	0.0180	0.0192	0.0204	0.0217
17	0.0063	0.0069	0.0075	0.0081	0.0088	0.0095	0.0103	0.0111	0.0119	0.0128
18	0.0032	0.0035	0.0039	0.0042	0.0046	0.0051	0.0055	0.0060	0.0065	0.0071
19	0.0015	0.0017	0.0019	0.0021	0.0023	0.0026	0.0028	0.0031	0.0034	0.0037
20	0.0007	0.0008	0.0009	0.0010	0.0011	0.0012	0.0014	0.0015	0.0017	0.0019
21	0.0003	0.0003	0.0004	0.0004	0.0005	0.0006	0.0006	0.0007	0.0008	0.0009
22	0.0001	0.0001	0.0002	0.0002	0.0002	0.0002	0.0003	0.0003	0.0004	0.0004
23	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002
24	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001

	μ									
x	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0010	0.0004	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.0037	0.0018	0.0008	0.0004	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
4	0.0102	0.0053	0.0027	0.0013	0.0006	0.0003	0.0001	0.0001	0.0000	0.0000
5	0.0224	0.0127	0.0070	0.0037	0.0019	0.0010	0.0005	0.0002	0.0001	0.0001
6	0.0411	0.0255	0.0152	0.0087	0.0048	0.0026	0.0014	0.0007	0.0004	0.0002
7	0.0646	0.0437	0.0281	0.0174	0.0104	0.0060	0.0034	0.0018	0.0010	0.0005
8	0.0888	0.0655	0.0457	0.0304	0.0194	0.0120	0.0072	0.0042	0.0024	0.0013
9	0.1085	0.0874	0.0661	0.0473	0.0324	0.0213	0.0135	0.0083	0.0050	0.0029
10	0.1194	0.1048	0.0859	0.0663	0.0486	0.0341	0.0230	0.0150	0.0095	0.0058
11	0.1194	0.1144	0.1015	0.0844	0.0663	0.0496	0.0355	0.0245	0.0164	0.0106
12	0.1094	0.1144	0.1099	0.0984	0.0829	0.0661	0.0504	0.0368	0.0259	0.0176
13	0.0926	0.1056	0.1099	0.1060	0.0956	0.0814	0.0658	0.0509	0.0378	0.0271
14	0.0728	0.0905	0.1021	0.1060	0.1024	0.0930	0.0800	0.0655	0.0514	0.0387

TABLA 7 Probabilidades de Poisson (*continuación*)

	μ									
x	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
15	0.0534	0.0724	0.0885	0.0989	0.1024	0.0992	0.0906	0.0786	0.0650	0.0516
16	0.0367	0.0543	0.0719	0.0866	0.0960	0.0992	0.0963	0.0884	0.0772	0.0646
17	0.0237	0.0383	0.0550	0.0713	0.0847	0.0934	0.0963	0.0936	0.0863	0.0760
18	0.0145	0.0256	0.0397	0.0554	0.0706	0.0830	0.0909	0.0936	0.0911	0.0844
19	0.0084	0.0161	0.0272	0.0409	0.0557	0.0699	0.0814	0.0887	0.0911	0.0888
20	0.0046	0.0097	0.0177	0.0286	0.0418	0.0559	0.0692	0.0798	0.0866	0.0888
21	0.0024	0.0055	0.0109	0.0191	0.0299	0.0426	0.0560	0.0684	0.0783	0.0846
22	0.0012	0.0030	0.0065	0.0121	0.0204	0.0310	0.0433	0.0560	0.0676	0.0769
23	0.0006	0.0016	0.0037	0.0074	0.0133	0.0216	0.0320	0.0438	0.0559	0.0669
24	0.0003	0.0008	0.0020	0.0043	0.0083	0.0144	0.0226	0.0328	0.0442	0.0557
25	0.0001	0.0004	0.0010	0.0024	0.0050	0.0092	0.0154	0.0237	0.0336	0.0446
26	0.0000	0.0002	0.0005	0.0013	0.0029	0.0057	0.0101	0.0164	0.0246	0.0343
27	0.0000	0.0001	0.0002	0.0007	0.0016	0.0034	0.0063	0.0109	0.0173	0.0254
28	0.0000	0.0000	0.0001	0.0003	0.0009	0.0019	0.0038	0.0070	0.0117	0.0181
29	0.0000	0.0000	0.0001	0.0002	0.0004	0.0011	0.0023	0.0044	0.0077	0.0125
30	0.0000	0.0000	0.0000	0.0001	0.0002	0.0006	0.0013	0.0026	0.0049	0.0083
31	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0007	0.0015	0.0030	0.0054
32	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0004	0.0009	0.0018	0.0034
33	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0005	0.0010	0.0020
34	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0006	0.0012
35	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0007
36	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004
37	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002
38	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
39	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001

Apéndice C Notación de suma

Suma

Definición

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n \tag{C.1}$$

Ejemplo para $x_1 = 5, x_2 = 8, x_3 = 14$

$$\begin{aligned}\sum_{i=1}^{3} x_i &= x_1 + x_2 + x_3 \\ &= 5 + 8 + 14 \\ &= 27\end{aligned}$$

Resultado 1

Para una constante c

$$\sum_{i=1}^{n} c = \underbrace{(c + c + \cdots + c)}_{n \text{ veces}} = nc \tag{C.2}$$

Ejemplo para $c = 5, n = 10$

$$\sum_{i=1}^{10} 5 = 10(5) = 50$$

Ejemplo para $c = \bar{x}$

$$\sum_{i=1}^{n} \bar{x} = n\bar{x}$$

Resultado 2

$$\begin{aligned}\sum_{i=1}^{n} cx_i &= cx_1 + cx_2 + \cdots + cx_n \\ &= c(x_1 + x_2 + \cdots + x_n) = c\sum_{i=1}^{n} x_i\end{aligned} \tag{C.3}$$

Ejemplo para $x_1 = 5, x_2 = 8, x_3 = 14, c = 2$

$$\sum_{i=1}^{3} 2x_i = 2\sum_{i=1}^{3} x_i = 2(27) = 54$$

Resultado 3

$$\sum_{i=1}^{n} (ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i \tag{C.4}$$

Ejemplo para $x_1 = 5, x_2 = 8, x_3 = 14, a = 2, y_1 = 7, y_2 = 3, y_3 = 8, b = 4$

$$\begin{aligned}\sum_{i=1}^{3}(2x_i + 4y_i) &= 2\sum_{i=1}^{3}x_i + 4\sum_{i=1}^{3}y_i \\ &= 2(27) + 4(18) \\ &= 54 + 72 \\ &= 126\end{aligned}$$

Doble suma

Considere los siguientes datos que involucran la variable x_{ij}, donde i es el subíndice que denota la posición en una fila o renglón, y j es el subíndice que denota la posición en la columna.

		Columna		
		1	**2**	**3**
Fila o renglón	**1**	$x_{11} = 10$	$x_{12} = 8$	$x_{13} = 6$
	2	$x_{21} = 7$	$x_{22} = 4$	$x_{23} = 12$

Definición

$$\sum_{i=1}^{n}\sum_{j=1}^{m}x_{ij} = (x_{11} + x_{12} + \cdots + x_{1m}) + (x_{21} + x_{22} + \cdots + x_{2m}) + (x_{31} + x_{32} + \cdots + x_{3m}) + \cdots + (x_{n1} + x_{n2} + \cdots + x_{nm}) \quad \textbf{(C.5)}$$

Ejemplo

$$\begin{aligned}\sum_{i=1}^{2}\sum_{i=1}^{3}x_{ij} &= x_{11} + x_{12} + x_{13} + x_{21} + x_{22} + x_{23} \\ &= 10 + 8 + 6 + 7 + 4 + 12 \\ &= 47\end{aligned}$$

Definición

$$\sum_{i=1}^{n}x_{ij} = x_{1j} + x_{2j} + \cdots + x_{nj} \quad \textbf{(C.6)}$$

Ejemplo

$$\begin{aligned}\sum_{i=1}^{2}x_{i2} &= x_{12} + x_{22} \\ &= 8 + 4 \\ &= 12\end{aligned}$$

Notación abreviada

Algunas veces, cuando una suma es para todos los valores del subíndice, se utilizan las siguientes notaciones abreviadas.

$$\sum_{i=1}^{n}x_i = \sum x_i \quad \textbf{(C.7)}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m}x_{ij} = \sum\sum x_{ij} \quad \textbf{(C.8)}$$

$$\sum_{i=1}^{n}x_{ij} = \sum_{i}x_{ij} \quad \textbf{(C.9)}$$

Apéndice D Soluciones a las autoevaluaciones y respuestas a los ejercicios de números pares

Capítulo 1

2. *a*) 10
b) 5
c) Variables categóricas: el tamaño y el combustible
Variables cuantitativas: cilindros, MPG en la ciudad y MPG en autopista
d)

Variable	Medidas a escala
Tamaño	Ordinal
Cilindros	de razón
MPG en la ciudad	de razón
MPG en autopista	de razón
Combustible	Nominal

3. *a*) Promedio al conducir en la ciudad = 182/10 = 18.2 mpg
b) Promedio al conducir en autopista = 261/10 = 26.1 mpg
En promedio, las millas por galón al conducir en autopista es de 7.9 mpg mayor que al conducir en la ciudad
c) 3 de 10 o 30% tienen motores de cuatro cilindros
d) 6 de 10 o 60% utilizan gasolina regular

4. *a*) 7
b) 5
c) Variables categóricas: estado, campo de juego y la división de la NCAA
d) Variables cuantitativas: activos y solicitudes admitidas

6. *a*) Cuantitativos
b) Categóricos
c) Categóricos
d) Cuantitativos
e) Categóricos

8. *a*) 1 015
b) Categóricos
c) Porcentajes
d) 0.10(1015) = 101.5; 101 o 102 personas encuestadas

10. *a*) Cuantitativo; de razón
b) Categórico; nominal
c) Categórico: ordinal
d) Cuantitativo; de razón
e) Categórico; nominal

12. *a*) Todas las personas que visitan Hawaii
b) Sí
c) Las preguntas primera y cuarta proporcionan datos cuantitativos.
La segunda y tercera preguntas proporcionan datos categóricos.

13. *a*) El gasto federal ($ billones)
b) Cuantitativos
c) Series de tiempo
d) El gasto federal se ha incrementado con el tiempo

14. *a*) Trace una gráfica con una línea de serie de tiempo para cada fabricante
b) Toyota supera a General Motors en 2006 y se convierte en el principal fabricante de automóviles
c) Una gráfica de barras mostraría los datos de corte transversal para 2007; la altura de las barras sería GM 8.8, Ford 7.9, DC 4.6 y Toyota 9.6

18. *a*) 36%
b) 189
c) Categóricos

20. *a*) 43% de los administradores fueron optimistas (a la alza) o muy optimistas (muy a la alza), y 21% de los administradores espera que la asistencia médica sea la industria principal durante los próximos 12 meses
b) El rendimiento promedio esperado a 12 meses se estima que sea de 11.2% para la población de los administradores de inversiones
c) El promedio muestral de 2.5 años es una estimación de cuánto le tomará conseguir a la población de administradores de inversiones un crecimiento sostenido

22. *a*) La población consiste de todos los clientes de la cadena de tiendas en Charlotte, Carolina del Norte
b) Algunas formas que la cadena de tiendas de abarrotes podría utilizar para recabar los datos son:
- Encuestar a clientes que entran o salen de la tienda
- La encuesta podría ser enviada por correo a clientes que tengan una tarjeta de compras del club
- A los clientes se les proporcionaría una encuesta impresa al salir de la tienda
- A los clientes se les entregaría un cupón en el que se les solicitaría completar un breve cuestionario en línea; si lo hacen, recibirán 5% de descuento en su siguiente compra

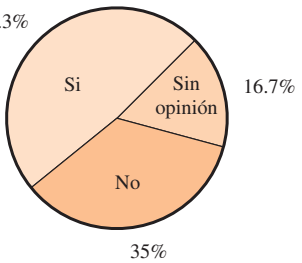
24. *a*) Correcto
b) Incorrecto
c) Correcto
d) Incorrecto
e) Incorrecto

Capítulo 2

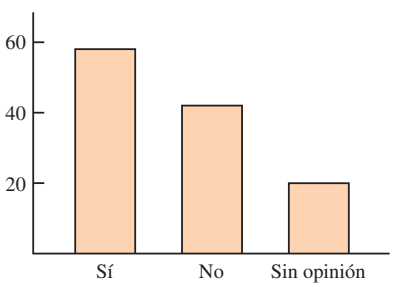
2. *a*) 0.20
b) 40
c)/*d*)

Clase	Frecuencia	Frecuencia porcentual
A	44	22
B	36	18
C	80	40
D	40	20
Total	200	100

3. *a*) $360° \times 58/120 = 174°$
b) $360° \times 42/120 = 126°$
c)



d)



4. *a*) Cualitativos
b)

Programa de televisión	Frecuencia	Frecuencia porcentual
La ley y el orden (LyO)	10	20%
CSI	18	36%
Sin rastro	9	18%
Esposas desesperadas (ED)	13	26%
Total:	50	100%

d) *CSI* tuvo la mayor audiencia; *Esposas desesperadas* ocupó el segundo lugar

6. *a*)

Cadena de televisión	Frecuencia	Frecuencia porcentual
ABC	15	30
CBS	17	34
FOX	1	2
NBC	17	34

b) CBS y NBC empataron en primer lugar; ABC quedó cerca con 15

7.

Calificación	Frecuencia	Frecuencia relativa
Sobresaliente	19	0.38
Muy bueno	13	0.26
Bueno	10	0.20
Regular	6	0.12
Malo	2	0.04

La gerencia debe estar satisfecha con estos resultados: 64% de las evaluaciones son de muy bueno a sobresaliente, y 84% son bueno o mejor; comparando estas calificaciones con los resultados anteriores, se demostrará si el restaurante está mejorando en las calificaciones de sus clientes con respecto a la calidad de sus alimentos

8. *a*)

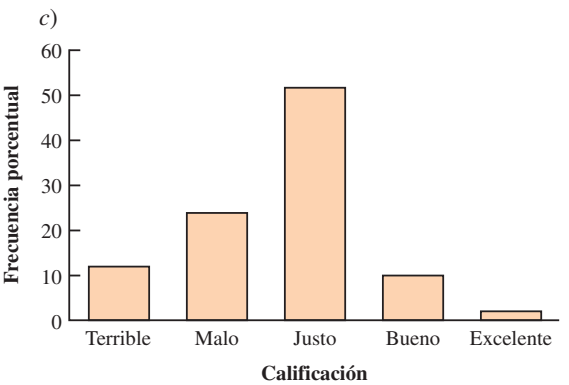
Posición	Frecuencia	Frecuencia relativa
P	17	0.309
H	4	0.073
1	5	0.091
2	4	0.073
3	2	0.036
S	5	0.091
L	6	0.109
C	5	0.091
R	7	0.127
Totales	55	1.000

b) Pitcher
c) 3a base
d) Jardinero derecho
e) 16 jugadores de cuadro (infielders) en comparación con 18 jugadores de campo (outfielders)

10. *a*)/*b*)

Calificación	Frecuencia	Frecuencia porcentual
Excelente	20	2
Bueno	101	10
Justo	528	52
Malo	244	24
Terrible	122	12
Total	1015	100

c)



d) 36% un desempeño malo o terrible
12% un buen o excelente desempeño
e) 50% un malo o terrible desempeño
4% de bueno a excelente desempeño
Más pesimismo en España

12.

Clase	Frecuencia acumulada	Frecuencia relativa acumulada
≤19	10	0.20
≤29	24	0.48
≤39	41	0.82
≤49	48	0.96
≤59	50	1.00

14. *b*)/*c*)

Clase	Frecuencia	Frecuencia porcentual
6.0–7.9	4	20
8.0–9.9	2	10
10.0–11.9	8	40
12.0–13.9	3	15
14.0–15.9	3	15
Totales	20	100

15. *a*)/*b*)

Tiempo de espera	Frecuencia	Frecuencia relativa
0–4	4	0.20
5–9	8	0.40
10–14	5	0.25
15–19	2	0.10
20–24	1	0.05
Totales	20	1.00

c)/*d*)

Tiempo de espera	Frecuencia acumulada	Frecuencia relativa acumulada
≤4	4	0.20
≤9	12	0.60
≤14	17	0.85
≤19	19	0.95
≤24	20	1.00

e) 12/20 = 0.60

16. *a*)

Sueldo	Frecuencia
150–159	1
160–169	3
170–179	7
180–189	5
190–199	1
200–209	2
210–219	1
Total	20

b)

Sueldo	Frecuencia porcentual
150–159	5
160–169	15
170–179	35
180–189	25
190–199	5
200–209	10
210–219	5
Total	100

c)

Sueldo	Frecuencia porcentual acumulada
Menor o igual que 159	5
Menor o igual que 169	20
Menor o igual que 179	55
Menor o igual que 189	80
Menor o igual que 199	85
Menor o igual que 209	95
Menor o igual que 219	100
Total	100

e) Existe un sesgo hacia la derecha

f) 15%

18. *a*) Más bajo de $180; más alto de $2 050

b)

Gasto	Frecuencia	Frecuencia porcentual
$ 0–249	3	12
250–499	6	24
500–749	5	20
750–999	5	20
1 000–1 249	3	12
1 250–1 499	1	4
1 500–1 749	0	0
1 750–1 999	1	4
2 000–2 249	1	4
Total	25	100

c) La distribución muestra un sesgo positivo

d) La mayoría de los consumidores (64%) gasta entre $250 y $1000; el valor intermedio es de aproximadamente $750; y dos personas gastaron más de $1 750

20. *a*)

Ingreso fuera de campo ($1 000s)	Frecuencia	Frecuencia porcentual
0–4 999	30	60
5 000–9 999	9	18
10 000–14 999	4	8
15 000–19 999	0	0
20 000–24 999	3	6
25 000–29 999	2	4
30 000–34 999	0	0
35 000–39 999	0	0
40 000–44 999	1	2
45 000–49 999	0	0
Más de 50 000	1	2
Total	50	100

c) El ingreso fuera de campo tiene un sesgo hacia la derecha; únicamente Tiger Woods gana más de $50 millones

d) La mayoría (60%) gana menos de $5 millones, 78% gana menos de $10 millones y 5 golfistas (10%) obtienen entre $20 y $30 millones; únicamente Tiger Woods y Phil Mickelson ganan más de $40 millones

22.

5	7 8
6	4 5 8
7	0 2 2 5 5 6 8
8	0 2 3 5

23. Hoja unitaria = 0.1

6	3
7	5 5 7
8	1 3 4 8
9	3 6
10	0 4 5
11	3

24. Hoja unitaria = 10

11	6
12	0 2
13	0 6 7
14	2 2 7
15	5
16	0 2 8
17	0 2 3

25.

9	8 9
10	2 4 6 6
11	4 5 7 8 8 9
12	2 4 5 7
13	1 2
14	4
15	1

26. *a*)

1	0 3 7 7
2	4 5 5
3	0 0 5 5 9
4	0 0 0 5 5 8
5	0 0 0 4 5 5

b)

0	5 7
1	0 1 1 3 4
1	5 5 5 8
2	0 0 0 0 0 0
2	5 5
3	0 0 0
3	6
4	
4	
5	
5	
6	3

28. *a*)

2	14
2	67
3	011123
3	5677
4	003333344
4	6679
5	00022
5	5679
6	14
6	6
7	2

b) 40–44, con 9

c) 43, con 5

d) 10%; una participación relativamente baja en la carrera

29. *a*)

		y = 1	y = 2	Total
x	A	5	0	5
	B	11	2	13
	C	2	10	12
	Total	18	12	30

b)

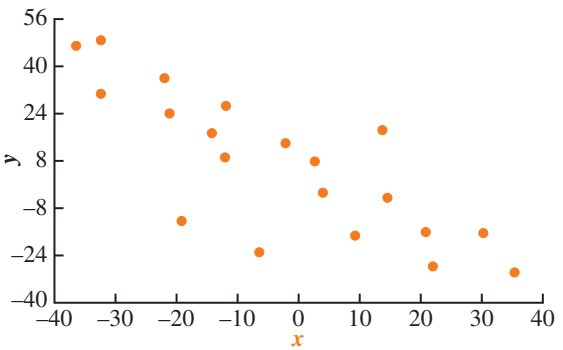
		y = 1	y = 2	Total
x	A	100.0	0.0	100.0
	B	84.6	15.4	100.0
	C	16.7	83.3	100.0

c)

		y = 1	y = 2
x	A	27.8	0.0
	B	61.1	16.7
	C	11.1	83.3
	Total	100.0	100.0

d) Los valores *A* siempre están en $y = 1$
Los valores *B* están con más frecuencia en $y = 1$
Los valores *C* están con más frecuencia en $y = 2$

30. *a*)



b) Una relación negativa entre x y y; y decrece a medida que x crece

32. *a*)

Ingresos familiares ($1 000)

Nivel de educación	Menos de 25	25.0–49.9	50.0–74.9	75.0–99.9	100 o más	Total
Sin educación media	32.10	18.71	9.13	5.26	2.20	13.51
Educación media	37.52	37.05	33.04	25.73	16.00	29.97
Educación superior inconclusa	21.42	28.44	30.74	31.71	24.43	27.21
Educación superior	6.75	11.33	18.72	25.19	32.26	18.70
Maestría o doctorado	2.21	4.48	8.37	12.11	25.11	10.61
Total	100.00	100.00	100.00	100.00	100.00	100.00

13.51% de los jefes de familia no se graduó del bachillerato

b) 25.11%, 53.54%

c) Una relación positiva entre el ingreso y el nivel educación

34. *a*)

Rendimiento promedio a 5 años

Tipo de fondo	0–9.99	10–19.99	20–29.99	30–39.99	40–49.99	50–59.99	Total
DE	1	25	1	0	0	0	27
FI	9	1	0	0	0	0	10
IE	0	2	3	2	0	1	8
Total	10	28	4	2	0	1	45

b)

Rendimiento promedio a 5 años	Frecuencia
0–9.99	10
10–19.99	28
20–29.99	4
30–39.99	2
40–49.99	0
50–59.99	1
Total	45

c)

Tipo de fondo	Frecuencia
DE	27
FI	10
IE	8
Total	45

d) El margen de la tabulación cruzada muestra estas distribuciones de frecuencia

e) Los mayores rendimientos – fondos de capital variable internacional (IE)
Los rendimientos más bajos – fondos de renta fija (FI)

36. *b*) Los rendimientos a 5 años están asociados con el mayor valor de los activos netos

38. *a*)

MPG autopista

Desplazamiento	15–19	20–24	25–29	30–34	35–39	Total
1.0–2.9	0	6	72	46	4	128
3.0–4.9	3	56	86	0	0	145
5.0–6.9	23	14	1	0	0	38
Total	26	76	159	46	4	311

b) Las mayores eficiencias de combustible se asocian con un motor de menor cilindrada
Las menores eficiencias de combustible se asocian con un motor de mayor cilindrada

d) Las menores eficiencias de combustible se asocian con los motores de mayor cilindrada

e) Diagrama de dispersión

40. *a*)

División	Frecuencia	Porcentaje
Buick	10	5
Cadillac	10	5
Chevrolet	122	61
GMC	24	12
Hummer	2	1
Pontiac	18	9
Saab	2	1
Saturn	12	6
Total	200	100

b) Chevrolet, 61%

c) Hummer y Saab, los dos únicamente con 1%
Chevrolet y GMC se mantienen

42. *a*)

Calificación en el SAT	Frecuencia
800–999	1
1000–1199	3
1200–1399	6
1400–1599	10
1600–1799	7
1800–1999	2
2000–2199	1
Total	30

b) Casi simétrica

c) El 33% de las calificaciones cayó entre 1 400 y 1 599. Las puntuaciones por debajo de 800 o arriba de 2 200 son inusuales. El promedio está cercano o ligeramente arriba de 1 500

44. *a*)

Population (Población)	Frequency (Frecuencia)	Percent Frequency (Frecuencia porcentual)
0.0–2.4	17	34
2.5–4.9	12	24
5.0–7.4	9	18
7.5–9.9	4	8
10.0–12.4	3	6
12.5–14.9	1	2
15.0–17.4	1	2
17.5–19.9	1	2
20.0–22.4	0	0
22.5–24.9	1	2
25.0–27.4	0	0
27.5–29.9	0	0
30.0–32.4	0	0
32.5–34.9	0	0
35.0–37.4	1	2
Total	50	100

c) Un sesgo positivo alto

d) 17 (34%) cuenta con una población menor de 2.5 millones
29 (58%) tiene una población menor de 5 millones
8 (16%) cuenta con una población mayor de 10 millones
El mayor tiene más de 35.9 millones (California)
El menor tiene menos de 0.5 millones (Wyoming)

46. *a*) **Temperaturas altas**

1	
2	
3	0
4	1 2 2 5
5	2 4 5
6	0 0 0 1 2 2 5 6 8
7	0 7
8	4

b) **Temperaturas bajas**

1	1
2	1 2 6 7 9
3	1 5 6 8 9
4	0 3 3 6 7
5	0 0 4
6	5
7	
8	

c) El intervalo de frecuencia más alto está en la década de 1960 (9 de 20) y sólo hay una temperatura menor de 54. La mayor parte de las temperaturas altas está entre 41. y 68, mientras que la mayor parte de las temperaturas bajas varía de 21 a 47
La baja fue de 11 y la alta de 84

d)

Temperatura alta	Frecuencia	Temperatura baja	Frecuencia
10–19	0	10–19	1
20–29	0	20–29	5
30–39	1	30–39	5
40–49	4	40–49	5
50–59	3	50–59	3
60–69	9	60–69	1
70–79	2	70–79	0
80–89	1	80–89	0
Total	20	Total	20

48. *a*)

Nivel de apoyo	Frecuencia porcentual
Totalmente a favor	30.10
Más a favor que en contra	34.83
Más en contra que a favor	21.13
Totalmente en contra	13.94
Total	100.00

En general a favor de impuestos más altos
= 30.10% + 34.83%
= 64.93%

b) 20.2, 19.5, 20.6, 20.7, 19.0
Aproximadamente 20% por país

c) Tabulación cruzada con columna de porcentajes:

	País				
Apoyo	Gran Bretaña	Italia	España	Alemania	Estados Unidos
Totalmente a favor	31.00	31.96	45.99	19.98	20.98
Más a favor que en contra	34.04	39.04	32.01	36.99	32.06
Más en contra que a favor	23.00	17.99	13.98	24.03	26.96
Totalmente en contra	11.96	11.01	8.03	18.99	20.00
Total	100.00	100.00	100.00	100.00	100.00

Los porcentajes de encuestados que expresaron estar a favor de un impuesto más alto, ya sea diciendo "totalmente a favor" o a "más a favor que en contra" son 65.04%, 71.00%, 78.00%, 56.97%, y 53.04 para los cinco países; todos muestran más de 50% de apoyo, pero los europeos manifiestan más apoyo para el impuesto que Estados Unidos; Italia y España demuestran el mayor nivel de apoyo.

50. *a*) Total de filas: 247; 54; 82; 121
Total de columnas: 149; 317; 17; 7; 14

b)

Año	Frecuencia	Combustible	Frecuencia
1973 o antes	247	Elect.	149
1974–79	54	Gas Nat.	317
1980–86	82	Petróleo	17
1987–91	121	Propano	7
Total	504	Otros	14
		Total	504

c) Tabulación cruzada de porcentajes de columnas

Año de construcción	Tipo de combustible				
	Elect.	Gas Nat.	Petróleo	Propano	Otro
1973 o antes	26.9	57.7	70.5	71.4	50.0
1974–1979	16.1	8.2	11.8	28.6	0.0
1980–1986	24.8	12.0	5.9	0.0	42.9
1987–1991	32.2	22.1	11.8	0.0	7.1
Total	100.0	100.0	100.0	100.0	100.0

d) Tabulación cruzada de los porcentajes de fila

Año de construcción	Tipo de combustible					
	Elect.	Gas Nat.	Petróleo	Propano	Otro	Total
1973 o antes	16.2	74.1	4.9	2.0	2.8	100.0
1974–1979	44.5	48.1	3.7	3.7	0.0	100.0
1980–1986	45.1	46.4	1.2	0.0	7.3	100.0
1987–1991	39.7	57.8	1.7	0.0	0.8	100.0

52. *a*) Tabulación cruzada del valor de mercado y las unidades

Valor de mercado ($1 000s)	Utilidades ($1 000s)				
	0–300	300–600	600–900	900–1200	Total
0–8 000	23	4			27
8 000–16 000	4	4	2	2	12
16 000–24 000		2	1	1	4
24 000–32 000		1	2	1	4
32 000–40 000		2	1		3
Total	27	13	6	4	50

b) Tabulación cruzada de los porcentajes de fila

Valor de mercado ($1 000s)	Utilidades ($1 000s)				
	0–300	300–600	600–900	900–1200	Total
0–8 000	85.19	14.81	0.00	0.00	100
8000–16 000	33.33	33.33	16.67	16.67	100
16 000–24 000	0.00	50.00	25.00	25.00	100
24 000–32 000	0.00	25.00	50.00	25.00	100
32 000–40 000	0.00	66.67	33.33	0.00	100

c) Una relación positiva está indicada entre las utilidades y el valor de mercado; conforme las utilidades se incrementan, el valor de mercado aumenta

54. *b*) Se demuestra una relación positiva entre el valor de mercado y el capital de los accionistas

Capítulo 3

2. 16, 16.5

3. Se ordenan los datos de menor a mayor: 15, 20, 25, 25, 27, 28, 30, 34

$i = \frac{20}{100}(8) = 1.6$; redondee hacia arriba hasta la posición 2

Percentil 20 = 20

$i = \frac{25}{100}(8) = 2$; utilice las posiciones 2 y 3

Percentil 25 $= \frac{20 + 25}{2} = 22.5$

$i = \frac{65}{100}(8) = 5.2$; redondee a la posición 6

Percentil 75 = 28

$i = \frac{75}{100}(8) = 6$; utilice las posiciones 6 y 7

Percentil 75 $= \frac{28 + 30}{2} = 29$

4. 59.73, 57, 53

6. *a*) 18.42
b) 6.32
c) 34.3%
d) Las reducciones de sólo 0.65 disparos y 0.9% tiros realizados por juego.
Sí, de acuerdo pero no de manera espectacular

8. *a*) $\bar{x} = \frac{\Sigma x_i}{n} = \frac{3200}{20} = 160$

Orden de los datos desde el bajo 100 hasta el alto 360

Mediana: $i = \left(\frac{50}{100}\right)20 = 10$; ocupa la 10a. y 11a posiciones

Mediana $= \left(\frac{130 + 140}{2}\right) = 135$

Moda = 120 (se presentó 3 veces)

b) $i = \left(\frac{25}{100}\right)20 = 5$; ocupa la 5a. y 6a. posiciones

$Q_1 = \left(\frac{115 + 115}{2}\right) = 115$

$i = \left(\frac{75}{100}\right)20 = 15$; ocupa la 15a. y 16a. posiciones

$Q_3 = \left(\frac{180 + 195}{2}\right) = 187.5$

c) $i = \left(\frac{90}{100}\right)20 = 18$; ocupa la 18a. y 19a. posiciones

Percentil 90 $= \left(\frac{235 + 255}{2}\right) = 245$

90% del costo de la declaración de impuestos es de $245 o menos

10. *a*) 0.4%, 3.5%
b) 2.3%, 2.5%, 2.7%
c) 2.0%, 2.8%
d) Optimista

12. Disney: 3321, 255.5, 253, 169, 325
Pixar: 3231, 538.5, 505, 363, 631
Las películas de Pixar generan ingresos de taquilla de aproximadamente el doble por película

14. 16, 4

15. Rango = 34 − 15 = 19
Ordene los datos de menor a mayor: 15, 20, 25, 25, 27, 28, 30, 34

$i = \frac{25}{100}(8) = 2$; $Q_1 = \frac{20 + 25}{2} = 22.5$

$$i = \frac{75}{100}(8) = 6;\ Q_3 = \frac{28 + 30}{2} = 29$$

$$\text{RIC} = Q_3 - Q_1 = 29 - 22.5 = 6.5$$

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{204}{8} = 25.5$$

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
27	1.5	2.25
25	−0.5	0.25
20	−5.5	30.25
15	−10.5	110.25
30	4.5	20.25
34	8.5	72.25
28	2.5	6.25
25	−0.5	0.25
		242.00

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{242}{8 - 1} = 34.57$$

$$s = \sqrt{34.57} = 5.88$$

16. *a*) Rango = 190 − 168 = 22

b) $\bar{x} = \dfrac{\Sigma x_i}{n} = \dfrac{1\,068}{6} = 178$

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{4^2 + (-10)^2 + 6^2 + 12^2 + (-8)^2 + (-4)^2}{6 - 1} = \frac{376}{5} = 75.2$$

c) $s = \sqrt{75.2} = 8.67$

d) $\dfrac{s}{\bar{x}}(100) = \dfrac{8.67}{178}(100\%) = 4.87\%$

18. *a*) 38, 97, 9.85

b) El este muestra más variación

20. *Dawson:* rango = 2; $s = 0.67$

Clark: rango = 8; $s = 2.58$

22. *a*) 1285, 433

Los de primer año más

b) 1 720, 352

c) 404, 131.5

d) 367.04, 96.96

e) Los de primer año muestran más variación

24. Tiempos de cuarto de milla*:* $s = 0.0564$; coeficiente de variación = 5.8%

Tiempos de milla*:* $s = 0.1295$; coeficiente de variación = 2.9%

26. 0.20, 1.50, 0, −0.50, −2.20

27. Teorema de Chebyshev: *por lo menos* $(1 - 1/z^2)$

a) $z = \dfrac{40 - 30}{5} = 2;\ 1 - \dfrac{1}{(2)^2} = 0.75$

b) $z = \dfrac{45 - 30}{5} = 3;\ 1 - \dfrac{1}{(3)^2} = 0.89$

c) $z = \dfrac{38 - 30}{5} = 1.6;\ 1 - \dfrac{1}{(1.6)^2} = 0.61$

d) $z = \dfrac{42 - 30}{5} = 2.4;\ 1 - \dfrac{1}{(2.4)^2} = 0.83$

e) $z = \dfrac{48 - 30}{5} = 3.6;\ 1 - \dfrac{1}{(3.6)^2} = 0.92$

28. *a*) 95%

b) Casi todos

c) 68%

29. *a*) $z = 2$ desviaciones estándar

$1 - \dfrac{1}{z^2} = 1 - \dfrac{1}{2^2} = \dfrac{3}{4}$; por lo menos 75%

b) $z = 2.5$ desviaciones estándar

$1 - \dfrac{1}{z^2} = 1 - \dfrac{1}{2.5^2} = 0.84$; por lo menos 84%

c) $z = 2$ desviaciones estándar

Regla empírica: 95%

30. *a*) 68%

b) 81.5%

c) 2.5%

32. *a*) −0.67

b) 1.50

c) Ninguna observación atípica

d) Sí; $z = 8.25$

34. *a*) 76.5, 7

b) 16%, 2.5%

c) 12.2, 7.89; no

36. 15, 22.5, 26, 29, 34

38. Datos ordenados de menor a mayor: 5, 6, 8, 10, 10, 12, 15, 16, 18

$i = \dfrac{25}{100}(9) = 2.25$; redondeo hacia arriba a la posición 3

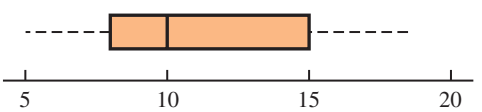
$Q_1 = 8$

Mediana (5a) posición) = 10

$i = \dfrac{75}{100}(9) = 6.75$; redondeo hacia arriba a la posición 7

$Q_3 = 15$

Resumen de cinco números: 5, 8, 10, 15, 18



40. *a*) El 1er lugar de los hombres 43.73 minutos más rápido

b) Medianas: 109.64, 131.67

El tiempo mediano de los hombres, 22.03 minutos más rápido

c) 65.30, 87.18, 109.64, 128.40, 148.70

109.03, 122.08, 131.67, 147.18, 189.28

d) Límites varoniles: 25.35 hasta 190.23; sin observaciones atípicas

Límites femeniles: 84.43 hasta 184.83; 2 observaciones atípicas

e) Las corredoras mostraron menor variación

41. *a*) Organice los datos de menor a mayor

$i = \dfrac{25}{100}(21) = 5.25$; redondee hacia arriba hasta la posición 6

$Q_1 = 1\,872$
Mediana (11a. posición) = 4 019
$i = \frac{75}{100}(21) = 15.75$; redondee hacia arriba hasta la posición 16

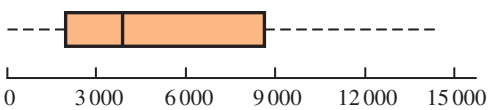
$Q_3 = 8\,305$
Resumen de cinco números: 608, 1 872, 4 019, 8 305, 14 138

b) RIC $= Q_3 - Q_1 = 8\,305 - 1\,872 = 6\,433$
Límite inferior: $1872 - 1.5(6\,433) = -7\,777.5$
Límite superior: $8\,305 + 1.5(6\,433) = 17\,955$

c) No: los datos están dentro de los límites

d) 41 138 > 27 604; 41 138 sería una observación atípica; el valor de los datos será revisado y corregido

e)



42. *a*) 73.5
b) 68, 71.5, 73.5, 74.5, 77
c) Límites: 67 y 79; ninguna observación atípica
d) 66, 68, 71, 73, 75; 60.5 y 80.5
63, 65, 66, 67.6, 69; 61.25 y 71.25
75, 77, 78.5, 79.5, 81; 73.25 y 83.25
Sin observaciones atípicas para cualesquiera servicios
e) Verizon es el mejor
Sprint tiene la calificación más baja

44. *a*) 18.2, 15.35
b) 11.7, 23.5
c) 3.4, 11.7, 15.35, 23.5, 41.3
d) Sí; Alger SmallCap, 41.3

45. *b*) Parece haber una relación lineal negativa entre x y y

c)

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
4	50	−4	4	−16
6	50	−2	4	−8
11	40	3	−6	−18
3	60	−5	14	−70
16	30	8	−16	−128
40	230	0	0	−240

$\bar{x} = 8; \bar{y} = 46$

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{-240}{4} = -60$$

La covarianza muestral indica una asociación lineal negativa entre x y y

d) $r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{-60}{(5.43)(11.40)} = -0.969$

El coeficiente de correlación muestral −0.969 indica una fuerte relación lineal negativa

46. *b*) Parece haber una relación lineal positiva entre x y y
c) $s_{xy} = 26.5$
d) $r_{xy} = 0.693$

48. −0.91; relación negativa

50. *b*) 0.910
c) Una fuerte relación lineal positiva; no

52. *a*) 3.69
b) 3.175

53. *a*)

f_i	M_i	$f_i M_i$
4	5	20
7	10	70
9	15	135
5	20	100
25		325

$$\bar{x} = \frac{\Sigma f_i M_i}{n} = \frac{325}{25} = 13$$

b)

f_i	M_i	$(M_i - \bar{x})$	$(M_i - \bar{x})^2$	$f_i(M_i - \bar{x})^2$
4	5	−8	64	256
7	10	−3	9	63
9	15	2	4	36
5	20	7	49	245
25				600

$$s^2 = \frac{\Sigma f_i(M_i - \bar{x})^2}{n - 1} = \frac{600}{25 - 1} = 25$$
$$s = \sqrt{25} = 5$$

54. *a*)

Calificación x_i	Ponderación w_i
4 (A)	9
3 (B)	15
2 (C)	33
1 (D)	3
0 (F)	0
	60 horas de crédito

$$\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i} = \frac{9(4) + 15(3) + 33(2) + 3(1)}{9 + 15 + 33 + 3} = \frac{150}{60} = 2.5$$

b) Sí

56. 3.8, 3.7

58. *a*) 1 800, 1 351
b) 387, 1 710
c) 7 280, 1 323
d) 3 675 303, 1 917
e) Sesgo positivo alto
f) Utilizando un diagrama de caja: 4 135 y 7 450 son observaciones atípicas

60. *a*) 2.3, 1.85
b) 1.90, 1.38
c) Group Altria 5%
d) −0.51, menor que la media
e) 1.02, mayor que la media
f) No

62. *a*) $670
b) $456
c) $z = 3$; sí
d) Ahorra tiempo y previene costos de penalización

64. *a*) 215.9
b) 55%
c) 175.0, 628.3
d) 48.8, 175.0, 215.9, 628.3, 2 325.0
e) Sí, cualquier precio mayor de 1308.25
f) 482.1

66. *a*) 364 habitaciones
b) $457
c) −0.293; una ligera correlación negativa
Un costo más alto por noche tiende a estar relacionado con hoteles pequeños

68. *a*) 0.268, baja o una correlación positiva débil
b) Muy pobre pronóstico; el entrenamiento en primavera es práctico y no se toma en cuenta hacia la clasificación o *playoffs*

70. *a*) 60.68
b) $s^2 = 31.23$; $s = 5.59$

Capítulo 4

2. $\binom{6}{3} = \frac{6!}{3!3!} = \frac{6\cdot5\cdot4\cdot3\cdot2\cdot1}{(3\cdot2\cdot1)(3\cdot2\cdot1)} = 20$

ABC	ACE	BCD	BEF
ABD	ACF	BCE	CDE
ABE	ADE	BCF	CDF
ABF	ADF	BDE	CEF
ACD	AEF	BDF	DEF

4. *b*) (H,H,H), (H,H,T), (H,T,H), (H,T,T), (T,H,H), (T,H,T), (T,T,H), (T,T,T)
c) 1/8

6. $P(E_1) = 0.40$, $P(E_2) = 0.26$, $P(E_3) = 0.34$
Método de frecuencia relativa

8. *a*) 4: Comisión, positiva—aprueba el Consejo
Comisión, positiva—desaprueba el Consejo
Comisión, negativa—aprueba el Consejo
Comisión, negativa—desaprueba el Consejo

9. $\binom{50}{4} = \frac{50!}{4!46!} = \frac{50\cdot49\cdot48\cdot47}{4\cdot3\cdot2\cdot1} = 230\,300$

10. *a*) Utilizando la tabla, P(deuda) = 0.94
b) Cinco de las ocho instituciones, P(más de 60%) = 5/8 = 0.625
c) Dos de las ocho instituciones, P(más de $30 000) = 2/8 = 0.25
d) P(sin deuda) = 1 − P(deuda) = 1 − 0.72 = 0.28
e) Un promedio ponderado con 72% tiene una deuda promedio de $32 980 y 28% no tiene deuda

$$\text{Deuda promedio por graduado} = \frac{0.72(\$32\,980) + 0.28(\$0)}{0.72 + 0.28} = \$23\,746$$

12. *a*) 3,478,761
b) 1/3,478,761
c) 1/146,107,962

14. *a*) 1/4
b) 1/2
c) 3/4

15. *a*) S = {as de bastos, as de diamantes, as de corazones, as de espadas}
b) S = {2 de bastos, 3 de bastos, . . . , 10 de bastos J de bastos, Q de bastos, K de bastos, A de bastos}
c) Hay 12; jota, reina, o rey en cada uno de los cuatro turnos
d) Para *a*): 4/52 = 1/13 = 0.08
Para *b*): 13/52 = 1/4 = 0.25
Para *c*): 12/52 = 0.23

16. *a*) 36
c) 1/6
d) 5/18
e) No; P(impar) = P(par) = 1/2
f) Clásico

17. *a*) (4, 6), (4, 7), (4, 8)
b) 0.05 + 0.10 + 0.15 = 0.30
c) (2, 8), (3, 8), (4, 8)
d) 0.05 + 0.05 + 0.15 = 0.25
e) 0.15

18. *a*) 0.0222
b) 0.8226
c) 0.1048

20. *a*) 0.108
b) 0.096
c) 0.434

22. *a*) 0.40, 0.40, 0.60
b) 0.80; sí
c) $A^c = \{E_3, E_4, E_5\}$; $C^c = \{E_1, E_4\}$; $P(A^c) = 0.60$; $P(C^c) = 0.40$
d) (E_1, E_2, E_5); 0.60
e) 0.80

23. *a*) $P(A) = P(E_1) + P(E_4) + P(E_6)$
$= 0.05 + 0.25 + 0.10 = 0.40$
$P(B) = P(E_2) + P(E_4) + P(E_7)$
$= 0.20 + 0.25 + 0.05 = 0.50$
$P(C) = P(E_2) + P(E_3) + P(E_5) + P(E_7)$
$= 0.20 + 0.20 + 0.15 + 0.05 = 0.60$
b) $A \cup B = \{E_1, E_2, E_4, E_6, E_7\}$;
$P(A \cup B) = P(E_1) + P(E_2) + P(E_4) + P(E_6) + P(E_7)$
$= 0.05 + 0.20 + 0.25 + 0.10 + 0.05$
$= 0.65$
c) $A \cap B = \{E_4\}$; $P(A \cap B) = P(E_4) = 0.25$
d) Sí, los dos son mutuamente excluyentes
e) $B^c = \{E_1, E_3, E_5, E_6\}$;
$P(B^c) = P(E_1) + P(E_3) + P(E_5) + P(E_6)$
$= 0.05 + 0.20 + 0.15 + 0.10$
$= 0.50$

24. *a*) 0.05
b) 0.70

26. *a*) 0.64
b) 0.48
c) 0.36
d) 0.76

28. Sea B = rentar un automóvil por razones de trabajo
P = rentar un automóvil por razones personales
a) $P(B \cup P) = P(B) + P(P) - P(B \cap P)$
$= 0.540 + 0.458 - 0.300$
$= 0.698$
b) P(por ninguna de las dos) $= 1 - 0.698 = 0.302$

30. *a*) $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{0.40}{0.60} = 0.6667$
b) $P(B \mid A) = \dfrac{P(A \cap B)}{P(A)} = \dfrac{0.40}{0.50} = 0.80$
c) No, debido a que $P(A \mid B) \neq P(A)$

32. *a*)

	Automóvil	Camión ligero	Total
Estadounidense	0.1330	0.2939	0.4269
No estadounidense	0.3478	0.2253	0.5731
Total	0.4808	0.5192	1.0000

b) 0.4269, 0.5731, marca no estadounidense, más altas 0.4808, 0.5192, camión ligero un poco más alta
c) 0.3115, 0.6885 camión ligero más alta
d) 0.6909, 0.3931 automóvil más alta
e) 0.5661, más alta para los camiones ligeros fabricados en Estados Unidos

33. *a*)

Razón de solicitud

	Calidad	Costo/ Conveniencia	Otras	Total
Tiempo completo	0.218	0.204	0.039	0.461
Tiempo parcial	0.208	0.307	0.024	0.539
Total	0.426	0.511	0.063	1.000

b) Un estudiante menciona la mayoría de las veces el costo o la conveniencia como la primera razón (probabilidad = 0.511); la calidad de la escuela es la segunda razón mencionada por la mayoría de los estudiantes (probabilidad = 0.426)
c) P(calidad | tiempo completo) $= 0.218/0.461 = 0.473$
d) P(calidad | tiempo parcial) $= 0.208/0.539 = 0.386$
e) Para la independencia se debe tener $P(A)P(B) = P(A \cap B)$; a partir de la tabla
$P(A \cap B) = 0.218$, $P(A) = 0.461$, $P(B) = 0.426$
$P(A)P(B) = (0.461)(0.426) = 0.196$
Debido a que $P(A)P(B) \neq P(A \cap B)$, los eventos no son independientes

34. *a*)

	A tiempo	Tarde	Total
Southwest	0.3336	0.0664	0.40
US Airways	0.2629	0.0871	0.35
JetBlue	0.1753	0.0747	0.25
Total	0.7718	0.2282	1.00

b) Southwest (0.40)
c) 0.7718
d) US Airways (0.3817); Southwest (0.2910)

36. *a*) 0.7921
b) 0.9879
c) 0.0121
d) 0.3364, 0.8236, 0.1764
No le cometen falta a Jerry Stackhouse

38. *a*) 0.70
b) 0.30
c) 0.67, 0.33
d) 0.20, 0.10
e) 0.40
f) 0.20
g) No; $P(S \mid M) \neq P(S)$

39. *a*) Sí, debido a que $P(A_1 \cap A_2) = 0$
b) $P(A_1 \cap B) = P(A_1)P(B \mid A_1) = 0.40(0.20) = 0.08$
$P(A_2 \cap B) = P(A_2)P(B \mid A_2) = 0.60(0.05) = 0.03$
c) $P(B) = P(A_1 \cap B) + P(A_2 \cap B) = 0.08 + 0.03 = 0.11$
d) $P(A_1 \mid B) = \dfrac{0.08}{0.11} = 0.7273$
$P(A_2 \mid B) = \dfrac{0.03}{0.11} = 0.2727$

40. *a*) 0.10, 0.20, 0.09
b) 0.51
c) 0.26, 0.51, 0.23

42. M = pagos atrasados
D_1 = incumplimiento de los clientes
D_2 = los clientes que no dejan de pagar
$P(D_1) = 0.05$, $P(D_2) = 0.95$, $P(M \mid D_2) = 0.2$, $P(M \mid D_1) = 1$
a) $P(D_1 \mid M) = \dfrac{P(D_1)P(M \mid D_1)}{P(D_1)P(M \mid D_1) + P(D_2)\,P(M \mid D_2)}$
$= \dfrac{(0.05)(1)}{(0.05)(1) + (0.95)(0.2)}$
$= \dfrac{0.05}{0.24} = 0.21$
b) Sí, la probabilidad de que el cliente incumpla es mayor de 0.20

44. *a*) 0.47, 0.53, 0.50, 0.45
b) 0.4963
c) 0.4463
d) 47%, 53%

46. *a*) 0.60
b) 0.26
c) 0.40
d) 0.74

48. *a*) 315
b) 0.29
c) No
d) Los republicanos

50. *a*) 0.76
b) 0.24

52. *b*) 0.2022
c) 0.4618
d) 0.4005

54. *a*) 0.49
b) 0.44
c) 0.54
d) No
e) Sí

56. *a*) 0.25
b) 0.125
c) 0.0125
d) 0.10
e) No

58. *a*)

	Adultos jóvenes	Adultos mayores	Total
Blogger	0.0432	0.0368	0.08
No blogger	0.2208	0.6992	0.92
Total	0.2640	0.7360	1.00

b) 0.2640
c) 0.0432
d) 0.1636

60. *a*) 0.40
b) 0.67

Capítulo 5

1. *a*) Cara, cara (H, H)
Cara, cruz (H, T)
Cruz, cara (T, H)
Cruz, Cruz (T, T)
b) x = número de caras en el segundo lanzamiento de la moneda
c)

Resultado	Valores de x
(H, H)	2
(H, T)	1
(T, H)	1
(T, T)	0

d) Discreta; 0, 1, y 2

2. *a*) x = tiempo requerido en minutos para ensamblar un producto
b) Cualquier valor positivo: $x > 0$
c) Continua

3. Sea Y = oferta de empleo
N = ninguna oferta de empleo
a) $S = \{(Y, Y, Y), (Y, Y, N), (Y, N, Y), (Y, N, N), (N, Y, Y), (N, Y, N), (N, N, Y), (N, N, N)\}$
b) Sea N = número de ofertas formuladas; N es una variable aleatoria discreta
c)

Resultado experimental	(Y, Y, Y)	(Y, Y, N)	(Y, N, Y)	(Y, N, N)	(N, Y, Y)	(N, Y, N)	(N, N, Y)	(N, N, N)
Valor de N	3	2	2	1	2	1	1	0

4. $x = 0, 1, 2, \ldots, 9$

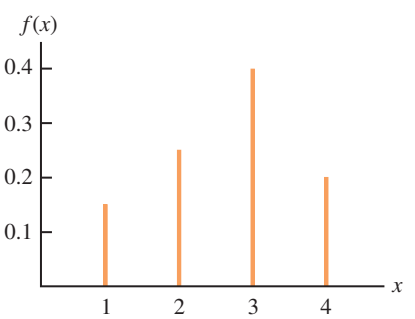
6. *a*) 0, 1, 2, . . . , 20; discreta
b) 0, 1, 2, . . . ; discreta
c) 0, 1, 2, . . . , 50; discreta
d) $0 \leq x \leq 8$; continua
e) $x > 0$; continua

7. *a*) $f(x) \geq 0$ para todos los valores de x
$\Sigma f(x) = 1$; por tanto, esta es una distribución de probabilidad válida
b) Probabilidad de que $x = 30$ es $f(30) = 0.25$
c) La probabilidad de que $x \leq 25$ es $f(20) + f(25) = 0.20 + 0.15 = 0.35$
d) Probabilidad de que $x > 30$ es $f(35) = 0.40$

8. *a*)

x	$f(x)$
1	3/20 = 0.15
2	5/20 = 0.25
3	8/20 = 0.40
4	4/20 = 0.20
	Total 1.00

b)



c) $f(x) \geq 0$ para $x = 1, 2, 3, 4$
$\Sigma f(x) = 1$

10. *a*)

x	1	2	3	4	5
$f(x)$	0.05	0.09	0.03	0.42	0.41

b)

x	1	2	3	4	5
$f(x)$	0.04	0.10	0.12	0.46	0.28

c) 0.83
d) 0.28
e) Los altos directivos están más satisfechos

12. *a*) Sí
b) 0.15
c) 0.10

14. *a*) 0.05
b) 0.70
c) 0.40

16. *a*)

y	$f(y)$	$yf(y)$
2	0.20	0.4
4	0.30	1.2
7	0.40	2.8
8	0.10	0.8
Totales	1.00	5.2

$E(y) = \mu = 5.2$

b)

y	$y - \mu$	$(y - \mu)^2$	$f(y)$	$(y - \mu)^2 f(y)$
2	−3.20	10.24	0.20	2.048
4	−1.20	1.44	0.30	0.432
7	1.80	3.24	0.40	1.296
8	2.80	7.84	0.10	0.784
			Total	4.560

$\text{Var}(y) = 4.56$

$\sigma = \sqrt{4.56} = 2.14$

18. *a*)/*b*)

x	$f(x)$	$xf(x)$	$x - \mu$	$(x - \mu)^2$	$(x - \mu)^2 f(x)$
0	0.04	0.00	−1.84	3.39	0.12
1	0.34	0.34	−0.84	0.71	0.24
2	0.41	0.82	0.16	0.02	0.01
3	0.18	0.53	1.16	1.34	0.24
4	0.04	0.15	2.16	4.66	0.17
Total	1.00	1.84			0.79
		↑			↑
		$E(x)$			$\text{Var}(x)$

c)/*d*)

y	$f(y)$	$yf(y)$	$y - \mu$	$(y - \mu)^2$	$(y - \mu)^2 f(y)$
0	0.00	0.00	−2.93	8.58	0.01
1	0.03	0.03	−1.93	3.72	0.12
2	0.23	0.45	−0.93	0.86	0.20
3	0.52	1.55	0.07	0.01	0.00
4	0.22	0.90	1.07	1.15	0.26
Total	1.00	2.93			0.59
		↑			↑
		$E(y)$			$\text{Var}(y)$

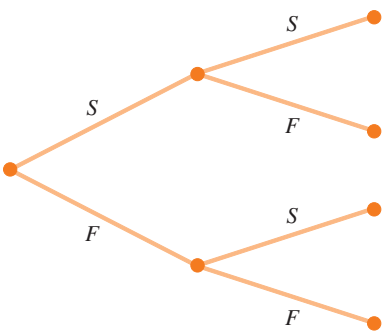
e) El número de recámaras en casas ocupadas por los propietarios es mayor que en rentadas las casas; el número esperado de recámaras es de 2.93 − 1.84 = 1.09 mayor, y la variabilidad en su número es menor para las casas propias

20. *a*) 430
b) −90; porque busca protegerse a toda costa contra el gasto de una gran pérdida

22. *a*) 445
b) Perderá $1250

24. *a*) Mediana: 145; alta: 140
b) Mediana: 2 725; alta: 12 400

25. *a*)



b) $f(1) = \binom{2}{1}(0.4)^1(0.6)^1 = \frac{2!}{1!1!}(0.4)(0.6) = 0.48$

c) $f(0) = \binom{2}{0}(0.4)^0(0.6)^2 = \frac{2!}{0!2!}(1)(0.36) = 0.36$

d) $f(2) = \binom{2}{2}(0.4)^2(0.6)^0 = \frac{2!}{2!0!}(0.16)(0.1) = 0.16$

e) $P(x \geq 1) = f(1) + f(2) = 0.48 + 0.16 = 0.64$

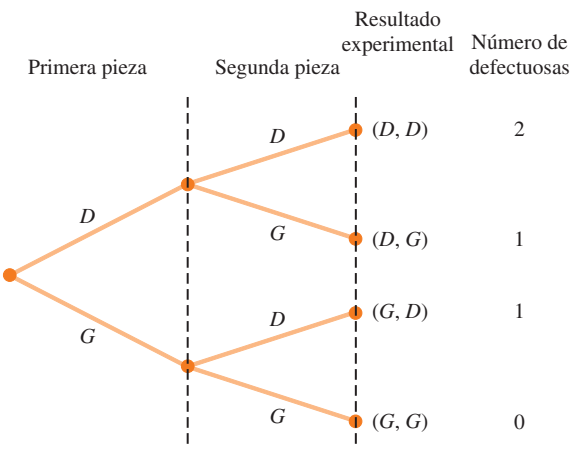
f) $E(x) = np = 2(0.4) = 0.8$
$\text{Var}(x) = np(1 - p) = 2(0.4)(0.6) = 0.48$
$\sigma = \sqrt{0.48} = 0.6928$

26. *a*) 0.3487
b) 0.1937
c) 0.9298
d) 0.6513
e) 1
f) 0.9, 0.95

28. *a*) 0.2789
b) 0.4181
c) 0.0733

30. *a*) La probabilidad de que un artículo producido esté defectuoso debe ser de 0.03 por cada pieza seleccionada; las partes deben ser seleccionadas de forma independiente

b) Sea D = defectuosa
G = sin defecto



c) Los dos resultados presentaron exactamente un defecto

d) $P(\text{sin defectos}) = (0.97)(0.97) = 0.9409$
$P(1 \text{ defecto}) = 2(0.03)(0.97) = 0.0582$
$P(2 \text{ defectos}) = (0.03)(0.03) = 0.0009$

32. *a*) 0.90
b) 0.99
c) 0.999
d) Sí

34. *a*) 0.2262
b) 0.8355

36. *a*) 0.1897
b) 0.9757
c) $f(12) = 0.0008$; sí
d) 5

38. *a*) $f(x) = \dfrac{3^x e^{-3}}{x!}$
b) 0.2241
c) 0.1494
d) 0.8008

39. *a*) $f(x) = \dfrac{2^x e^{-2}}{x!}$
b) $\mu = 6$ para 3 periodos
c) $f(x) = \dfrac{6^x e^{-6}}{x!}$
d) $f(2) = \dfrac{2^2 e^{-2}}{2!} = \dfrac{4(0.1353)}{2} = 0.2706$
e) $f(6) = \dfrac{6^6 e^{-6}}{6!} = 0.1606$
f) $f(5) = \dfrac{4^5 e^{-4}}{5!} = 0.1563$

40. *a*) 0.1952
b) 0.1048
c) 0.0183
d) 0.0907

42. *a*) $f(0) = \dfrac{7^0 e^{-7}}{0!} = e^{-7} = 0.0009$
b) Probabilidad $= 1 - [f(0) + f(1)]$
$f(1) = \dfrac{7^1 e^{-7}}{1!} = 7e^{-7} = 0.0064$
Probabilidad $= 1 - [0.0009 + 0.0064] = 0.9927$
c) $\mu = 3.5$
$f(0) = \dfrac{3.5^0 e^{-3.5}}{0!} = e^{-3.5} = 0.0302$
Probabilidad $= 1 - f(0) = 1 - 0.0302 = 0.9698$
d)
$$\begin{aligned}\text{Probabilidad} &= 1 - [f(0) + f(1) + f(2) + f(3) + f(4)] \\ &= 1 - [0.0009 + 0.0064 + 0.0223 + 0.0521 \\ &\quad + 0.0912] \\ &= 0.8271\end{aligned}$$

44. *a*) $\mu = 1.25$
b) 0.2865
c) 0.3581
d) 0.3554

46. *a*) $f(1) = \dfrac{\binom{3}{1}\binom{10-3}{4-1}}{\binom{10}{4}} = \dfrac{\left(\frac{3!}{1!2!}\right)\left(\frac{7!}{3!4!}\right)}{\frac{10!}{4!6!}}$
$= \dfrac{(3)(35)}{210} = 0.50$
b) $f(2) = \dfrac{\binom{3}{2}\binom{10-3}{2-2}}{\binom{10}{2}} = \dfrac{(3)(1)}{45} = 0.067$
c) $f(0) = \dfrac{\binom{3}{0}\binom{10-3}{2-0}}{\binom{10}{2}} = \dfrac{(1)(21)}{45} = 0.4667$
d) $f(2) = \dfrac{\binom{3}{2}\binom{10-3}{4-2}}{\binom{10}{4}} = \dfrac{(3)(21)}{210} = 0.30$
e) $x = 4$ es *mayor que* $r = 3$; por tanto, $f(4) = 0$

48. *a*) 0.5250
b) 0.8167

50. $N = 60, n = 10$
a) $r = 20, x = 0$
$$\begin{aligned}f(0) &= \frac{\binom{20}{0}\binom{40}{10}}{\binom{60}{10}} = \frac{(1)\left(\frac{40!}{10!30!}\right)}{\frac{60!}{10!50!}} \\ &= \left(\frac{40!}{10!30!}\right)\left(\frac{10!50!}{60!}\right) \\ &= \frac{40\cdot 39\cdot 38\cdot 37\cdot 36\cdot 35\cdot 34\cdot 33\cdot 32\cdot 31}{60\cdot 59\cdot 58\cdot 57\cdot 56\cdot 55\cdot 54\cdot 53\cdot 52\cdot 51} \\ &= 0.0112\end{aligned}$$
b) $r = 20, x = 1$
$$\begin{aligned}f(1) &= \frac{\binom{20}{1}\binom{40}{9}}{\binom{60}{10}} = 20\left(\frac{40!}{9!31!}\right)\left(\frac{10!50!}{60!}\right) \\ &= 0.0725\end{aligned}$$
c) $1 - f(0) - f(1) = 1 - 0.0112 - 0.0725 = 0.9163$
d) La misma que la probabilidad de que uno trabaje en Hawai; 0.0725

52. *a*) 0.2917
b) 0.0083
c) 0.5250, 0.1750; un banco
d) 0.7083
e) 0.90, 0.49, 0.70

54. *a*)

x	1	2	3	4	5
$f(x)$	0.24	0.21	0.10	0.21	0.24

b) 3.00, 2.34
c) Bonos $E(x) = 1.36$, $\text{Var}(x) = 0.23$
Acciones: $E(x) = 4$, $\text{Var}(x) = 1$

56. *a*) 0.0596
b) 0.3585

c) 100
d) 95, 9.75

58. *a*) 0.9510
b) 0.0480
c) 0.0490

60. *a*) 240
b) 12.96
c) 12.96

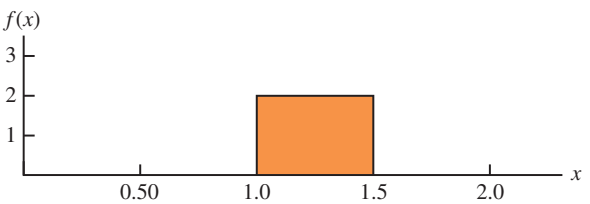
62. 0.1912

64. *a*) 0.2240
b) 0.5767

66. *a*) 0.4667
b) 0.4667
c) 0.0667

Capítulo 6

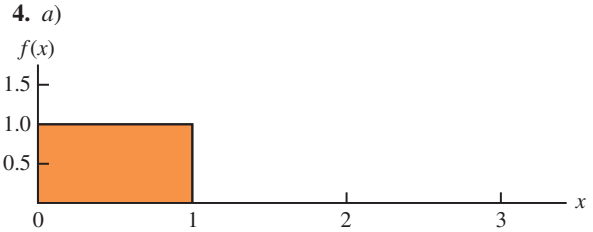
1. *a*)



b) $P(x = 1.25) = 0$; la probabilidad de cualquier punto es cero, debido a que el área bajo la curva sobre cualquier punto es cero
c) $P(1.0 \le x \le 1.25) = 2(0.25) = 0.50$
d) $P(1.20 < x < 1.5) = 2(0.30) = 0.60$

2. *b*) 0.50
c) 0.60
d) 15
e) 8.33

4. *a*)



b) $P(0.25 < x < 0.75) = 1(0.50) = 0.50$
c) $P(x \le 0.30) = 1(0.30) = 0.30$
d) $P(x > 0.60) = 1(0.40) = 0.40$

6. *a*) 0.125
b) 0.50
c) 0.25

10. *a*) 0.9332
b) 0.8413
c) 0.0919
d) 0.4938

12. *a*) 0.2967
b) 0.4418
c) 0.3300
d) 0.5910
e) 0.8849
f) 0.2389

13. *a*) $P(-1.98 \le z \le 0.49) = P(z \le 0.49) - P(z < -1.98) = 0.6879 - 0.0239 = 0.6640$
b) $P(0.52 \le z \le 1.22) = P(z \le 1.22) - P(z < 0.52) = 0.8888 - 0.6985 = 0.1903$
c) $P(-1.75 \le z \le -1.04) = P(z \le -1.04) - P(z < -1.75) = 0.1492 - 0.0401 = 0.1091$

14. *a*) $z = 1.96$
b) $z = 1.96$
c) $z = 0.61$
d) $z = 1.12$
e) $z = 0.44$
f) $z = 0.44$

15. *a*) El valor z que corresponde a una probabilidad acumulada de 0.2119 es $z = -0.80$
b) Calcule 0.9030/2 = 0.4515; la probabilidad acumulada de 0.5000 + 0.4515 = 0.9515 corresponde a $z = 1.66$
c) Calcule 0.2052/2 = 0.1026; z corresponde a la probabilidad acumulada de 0.5000 + 0.1026 = 0.6026, así que $z = 0.26$
d) El valor z correspondiente a una probabilidad acumulada de 0.9948 es $z = 2.56$
e) El área a la izquierda de z es 1 − 0.6915 = 0.3085, por tanto $z = -0.50$

16. *a*) $z = 2.33$
b) $z = 1.96$
c) $z = 1.645$
d) $z = 1.28$

18. $\mu = 30$ y $\sigma = 8.2$

a) Para $x = 40$, $z = \dfrac{40 - 30}{8.2} = 1.22$

$P(z \le 1.22) = 0.8888$
$P(x \ge 40) = 1.000 - 0.8888 = 0.1112$

b) Para $x = 20$, $z = \dfrac{20 - 30}{8.2} = -1.22$

$P(z \le -1.22) = 0.1112$
$P(x \le 20) = 0.1112$

c) El valor z de 1.28 corta un área de aproximadamente 10% en la cola superior
$x = 30 + 8.2(1.28)$
$= 40.50$
El precio de una acción de \$40.50 o más alto colocará a la empresa en el 10% de las principales

20. *a*) 0.0885
b) 12.51%
c) 93.8 horas o más

22. *a*) 0.7193
b) \$35.59
c) 0.0233

24. *a*) 200, 26.04
b) 0.2206

c) 0.1251
d) 242.84 millones

26. *a*) $\mu = np = 100(0.20) = 20$
$\sigma^2 = np(1-p) = 100(0.20)(0.80) = 16$
$\sigma = \sqrt{16} = 4$
b) Sí, debido a que $np = 20$ y $n(1-p) = 80$
c) $P(23.5 \le x \le 24.5)$

$$z = \frac{24.5 - 20}{4} = 1.13 \qquad P(z \le 1.13) = 0.8708$$

$$z = \frac{23.5 - 20}{4} = 0.88 \qquad P(z \le 0.88) = 0.8106$$

$P(23.5 \le x \le 24.5) = P(0.88 \le z \le 1.13)$
$= 0.8708 - 0.8106 = 0.0602$

d) $P(17.5 \le x \le 22.5)$

$$z = \frac{22.5 - 20}{4} = 0.63 \qquad P(z \le 0.63) = 0.7357$$

$$z = \frac{17.5 - 20}{4} = -0.63 \qquad P(z \le -0.63) = 0.2643$$

$P(17.5 \le x \le 22.5) = P(-0.63 \le z \le 0.63)$
$= 0.7357 - 0.2643 = 0.4714$

e) $P(x \le 15.5)$

$$z = \frac{15.5 - 20}{4} = -1.13 \qquad P(z \le -1.13) = 0.1292$$

$P(x \le 15.5) = P(z \le -1.13) = 0.1292$

28. *a*) $\mu = np = 250(0.20) = 50$
b) $\sigma^2 = np(1-p) = 250(0.20)(1 - 20) = 40$
$\sigma = \sqrt{40} = 6.3246$
$P(x < 40) = P(x \le 39.5)$

$$z = \frac{x - \mu}{\sigma} = \frac{39.5 - 50}{6.3246} = -1.66 \quad \text{Área} = 0.0485$$

$P(x \le 39.5) = 0.0485$
c) $P(55 \le x \le 60) = P(54.5 \le x \le 60.5)$

$$z = \frac{x - \mu}{\sigma} = \frac{54.5 - 50}{6.3246} = 0.71 \qquad \text{Área} = 0.7611$$

$$z = \frac{x - \mu}{\sigma} = \frac{60.5 - 50}{6.3246} = 1.66 \qquad \text{Área} = 0.9515$$

$P(54.5 \le x \le 60.5) = 0.9515 - 0.7611 = 0.1904$
d) $P(x \ge 70) = P(x \ge 69.5)$

$$z = \frac{x - \mu}{\sigma} = \frac{69.5 - 50}{6.3246} = 3.08 \qquad \text{Área} = 0.9990$$

$P(x \ge 69.5) = 1 - 0.9990 = 0.0010$

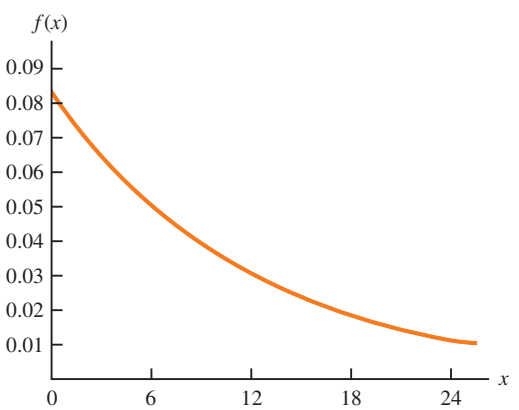
30. *a*) 220
b) 0.0392
c) 0.8962

32. *a*) 0.5276
b) 0.3935
c) 0.4724
d) 0.1341

33. *a*) $P(x \le x_0) = 1 - e^{-x_0/3}$
b) $P(x \le 2) = 1 - e^{-2/3} = 1 - 0.5134 = 0.4866$
c) $P(x \ge 3) = 1 - P(x \le 3) = 1 - (1 - e^{-3/3})$
$= e^{-1} = 0.3679$
d) $P(x \le 5) = 1 - e^{-5/3} = 1 - 0.1889 = 0.8111$
e) $P(2 \le x \le 5) = P(x \le 5) - P(x \le 2)$
$= 0.8111 - 0.4866 = 0.3245$

34. *a*) 0.5624
b) 0.1915
c) 0.2461
d) 0.2259

35. *a*)



b) $P(x \le 12) = 1 - e^{-12/12} = 1 - 0.3679 = 0.6321$
c) $P(x \le 6) = 1 - e^{-6/12} = 1 - 0.6065 = 0.3935$
d) $P(x \ge 30) = 1 - P(x < 30)$
$= 1 - (1 - e^{-30/12})$
$= 0.0821$

36. *a*) 0.3935
b) 0.2386
c) 0.1353

38. *a*) $f(x) = 5.5e^{-5.5x}$
b) 0.2528
c) 0.6002

40. *a*) \$3 780 o menos
b) 19.22%
c) \$8 167.50

42. *a*) 3 229
b) 0.2244
c) \$12 382 o más

44. *a*) 0.0228
b) \$50

46. *a*) 38.3%
b) 3.59% mejor; 96.41% peor
c) 38.21%

48. $\mu = 19.23$ onzas

50. *a*) Una pérdida de \$240
b) 0.1788
c) 0.3557
d) 0.0594

52. *a*) 1/7 de minuto
b) $7e^{-7x}$
c) 0.0009
d) 0.2466

54. *a*) 2 minutos
b) 0.2212
c) 0.3935
d) 0.0821

Capítulo 7

1. *a*) AB, AC, AD, AE, BC, BD, BE, CD, CE, DE
b) Con 10 muestras, cada una tiene una $\frac{1}{10}$ de probabilidad
c) E y C, debido a que 8 y 0 no se usan; 5 se omite porque E está ya dentro de la muestra; 7 no se utiliza; 3 se identifica con C; el 2 ya no es necesario debido a que ya se tiene la muestra de tamaño 2

2. 22, 147, 229, 289

3. 459, 147, 385, 113, 340, 401, 215, 2, 33, 348

4. *a*) Bell South, LSI Logic, General Electric
b) 120

6. 2782, 493, 825, 1 807, 289

8. ExxonMobil, Chevron, Travelers, Microsoft, Pfizer e Intel

10. *a*) finito; *b*) infinito; *c*) infinito; *d*) finito; *e*) infinito

11. *a*) $\bar{x} = \dfrac{\Sigma x_i}{n} = \dfrac{54}{6} = 9$

b) $s = \sqrt{\dfrac{\Sigma(x_i - \bar{x})^2}{n-1}}$

$\Sigma(x_i - \bar{x})^2 = (-4)^2 + (-1)^2 + 1^2 + (-2)^2 + 1^2 + 5^2 = 48$

$s = \sqrt{\dfrac{48}{6-1}} = 3.1$

12. *a*) 0.50
b) 0.3667

13. *a*) $\bar{x} = \dfrac{\Sigma x_i}{n} = \dfrac{465}{5} = 93$

b)

	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
	94	+1	1
	100	+7	49
	85	−8	64
	94	+1	1
	92	−1	1
Totales	465	0	116

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{116}{4}} = 5.39$$

14. *a*) 0.45
b) 0.15
c) 0.45

16. *a*) 0.10
b) 20
c) 0.72

18. *a*) 200
b) 5
c) Normal con $E(\bar{x}) = 200$ y $\sigma_{\bar{x}} = 5$
d) La distribución de probabilidad de $\bar{x}$

19. *a*) La distribución de muestreo es normal con
$E(\bar{x}) = \mu = 200$
$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 50/\sqrt{100} = 5$
Para ± 5, $195 \le \bar{x} \le 205$
Al utilizar la tabla de probabilidad normal estándar:

En $\bar{x} = 205$, $z = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \dfrac{5}{5} = 1$

$P(z \le 1) = 0.8413$

En $\bar{x} = 195$, $z = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \dfrac{-5}{5} = -1$

$P(z < -1) = 0.1587$
$P(195 \le \bar{x} \le 205) = 0.8413 - 0.1587 = 0.6826$

b) Para ± 10, $190 \le \bar{x} \le 210$
Al utilizar la tabla de probabilidad normal estándar:

En $\bar{x} = 210$, $z = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \dfrac{10}{5} = 2$

$P(z \le 2) = 0.9772$

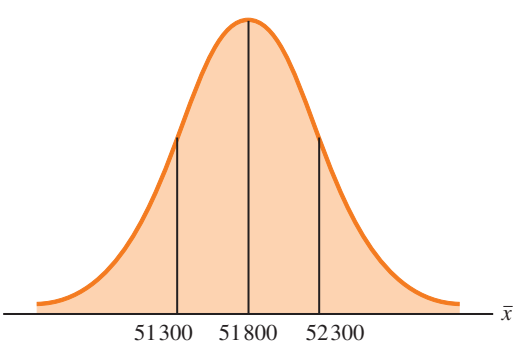
En $\bar{x} = 190$, $z = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \dfrac{-10}{5} = -2$

$P(z < -2) = 0.0228$
$P(190 \le \bar{x} \le 210) = 0.9722 - 0.0228 = 0.9544$

20. 3.54, 2.50, 2.04, 1.77
$\sigma_{\bar{x}}$ disminuye conforme n aumenta

22. *a*) Normal con $E(\bar{x}) = 51\,800$ y $\sigma_{\bar{x}} = 516.40$
b) $\sigma_{\bar{x}}$ disminuye a 365.15
c) $\sigma_{\bar{x}}$ disminuye conforme n aumenta

23. *a*)



$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4\,000}{\sqrt{60}} = 516.40$$

En $\bar{x} = 52\,300$, $z = \dfrac{52\,300 - 51\,800}{516.40} = 0.97$

$$P(\bar{x} \le 52\,300) = P(z \le 0.97) = 0.8340$$

$$\text{En } \bar{x} = 51\,300,\ z = \frac{51\,300 - 51\,800}{516.40} = -0.97$$

$$P(\bar{x} < 51\,300) = P(z < -0.97) = 0.1660$$

$$P(51\,300 \le \bar{x} \le 52\,300) = 0.8340 - 0.1660 = 0.6680$$

b) $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{4\,000}{\sqrt{120}} = 365.15$

$$\text{En } \bar{x} = 52\,300,\ z = \frac{52\,300 - 51\,800}{365.15} = 1.37$$

$$P(\bar{x} \le 52\,300) = P(z \le 1.37) = 0.9147$$

$$\text{En } \bar{x} = 51\,300,\ z = \frac{51\,300 - 51\,800}{365.15} = -1.37$$

$$P(\bar{x} < 51\,300) = P(z < -1.37) = 0.0853$$

$$P(51\,300 \le \bar{x} \le 52\,300) = 0.9147 - 0.0853 = 0.8294$$

24. *a)* Normal con $E(\bar{x}) = 17.5$ y $\sigma_{\bar{x}} = 0.57$
b) 0.9198
c) 0.6212

26. *a)* 0.4246, 0.5284, 0.6922, 0.9586
b) Mayor probabilidad de que la media muestral se acerque a la media poblacional

28. *a)* Normal con $E(\bar{x}) = 95$ y $\sigma_{\bar{x}} = 2.56$
b) 0.7580
c) 0.8502
d) Inciso *c)*, por el tamaño de la muestra mayor

30. *a)* $n/N = 0.01$; no
b) 1.29, 1.30; poca diferencia
c) 0.8764

32. *a)* $E(\bar{p}) = 0.40$

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.40)(0.60)}{200}} = 0.0346$$

Como ± 0.03 significa que $0.37 \le \bar{p} \le 0.43$

$$z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0.03}{0.0346} = 0.87$$

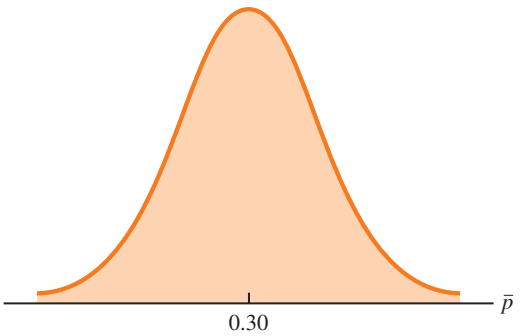
$$P(0.37 \le \bar{p} \le 0.43) = P(-0.87 \le z \le 0.87) = 0.8078 - 0.1922 = 0.6156$$

b) $z = \dfrac{\bar{p} - p}{\sigma_{\bar{p}}} = \dfrac{0.05}{0.0346} = 1.44$

$$P(0.35 \le \bar{p} \le 0.45) = P(-1.44 \le z \le 1.44) = 0.9251 - 0.0749 = 0.8502$$

34. *a)* 0.6156
b) 0.7814
c) 0.9488
d) 0.9942
e) Alta probabilidad con n mayor

35. *a)*



$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.30(0.70)}{100}} = 0.0458$$

La distribución normal es apropiada debido a que $np = 100(0.30) = 30$ y $n(1-p) = 100(0.70) = 70$ son mayores de 5

b) $P(0.20 \le \bar{p} \le 0.40) = ?$

$$z = \frac{0.40 - 0.30}{0.0458} = 2.18$$

$$P(0.20 \le \bar{p} \le 0.40) = P(-2.18 \le z \le 2.18) = 0.9854 - 0.0146 = 0.9708$$

c) $P(0.25 \le \bar{p} \le 0.35) = ?$

$$z = \frac{0.35 - 0.30}{0.0458} = 1.09$$

$$P(0.25 \le \bar{p} \le 0.35) = P(-1.09 \le z \le 1.09) = 0.8621 - 0.1379 = 0.7242$$

36. *a)* Normal con $E(\bar{p}) = 0.66$ y $\sigma_{\bar{p}} = 0.0273$
b) 0.8584
c) 0.9606
d) Sí, el error estándar es menor en el inciso *c)*
e) 0.9616; la probabilidad es mayor debido a que el aumento del tamaño de la muestra reduce el error estándar

38. *a)* Normal con $E(\bar{p}) = 0.56$ y $\sigma_{\bar{p}} = 0.0248$
b) 0.5820
c) 0.8926

40. *a)* Normal con $E(\bar{p}) = 0.76$ y $\sigma_{\bar{p}} = 0.0214$
b) 0.8384
c) 0.9452

42. 122, 99, 25, 55, 115, 102, 61

44. *a)* Normal con $E(\bar{x}) = 115.50$ y $\sigma_{\bar{x}} = 5.53$
b) 0.9298
c) $z = -2.80$, 0.0026

46. *a)* 955
b) 0.50
c) 0.7062
d) 0.8230

48. *a)* 625
b) 0.7888

50. *a*) Normal con $E(\bar{p}) = 0.28$ y $\sigma_{\bar{p}} = 0.0290$
b) 0.8324
c) 0.5098

52. *a*) 0.8882
b) 0.0233

54. *a*) 48
b) Normal, $E(\bar{p}) = 0.25$, $\sigma_{\bar{p}} = 0.0625$
c) 0.2119

Capítulo 8

2. Use $\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$
a) $32 \pm 1.645(6/\sqrt{50})$
32 ± 1.4; de 30.6 a 33.4
b) $32 \pm 1.96(6/\sqrt{50})$
32 ± 1.66; de 30.34 a 33.66
c) $32 \pm 2.576(6/\sqrt{50})$
32 ± 2.19; de 29.81 a 34.19

4. 54

5. *a*) $1.96\sigma/\sqrt{n} = 1.96(5/\sqrt{49}) = 1.40$
b) 24.80 ± 1.40; de 23.40 a 26.20

6. De 8.1 a 8.9

8. *a*) Que la población es aproximadamente normal
b) 3.1
c) 4.1

10. *a*) \$113 638 a \$124 672
b) \$112 581 a \$125 729
c) \$110 515 a \$127 795
d) La amplitud aumenta a medida que se incrementa el nivel de confianza

12. *a*) 2.179
b) −1.676
c) 2.457
d) −1.708 y 1.708
e) −2.014 y 2.014

13. *a*) $\bar{x} = \dfrac{\Sigma x_i}{n} = \dfrac{80}{8} = 10$

b) $s = \sqrt{\dfrac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\dfrac{84}{7}} = 3.464$

c) $t_{0.025}\left(\dfrac{s}{\sqrt{n}}\right) = 2.365\left(\dfrac{3.46}{\sqrt{8}}\right) = 2.9$

d) $\bar{x} \pm t_{0.025}\left(\dfrac{s}{\sqrt{n}}\right)$
10 ± 2.9 (7.1 hasta 12.9)

14. *a*) 21.5 a 23.5
b) 21.3 a 23.7
c) 20.9 a 24.1
d) Un margen mayor de error y un intervalo más amplio

15. $\bar{x} \pm t_{\alpha/2}(s/\sqrt{n})$
95% de confianza: $gl = 64$ y $t_{0.05} = 1.669$
$19.5 \pm 1.669\left(\dfrac{5.2}{\sqrt{65}}\right)$
19.5 ± 1.08 o (18.42 hasta 20.58)

95% de confianza: $gl = 64$ y $t_{0.025} = 1.998$
$19.5 \pm 1.998\left(\dfrac{5.2}{\sqrt{65}}\right)$
19.5 ± 1.29 o (18.21 hasta 20.79)

16. *a*) 1.69
b) 47.31 a 50.69
c) Pocas horas y costo más alto para United

18. *a*) 22 semanas
b) 3.8020
c) 18.20 a 25.80
d) *n* mayor la próxima vez

20. $\bar{x} = 22$; 21.48 a 22.52

22. *a*) \$9 269 a \$12 541
b) 1 523
c) 4 748 714, \$34 millones

24. *a*) Valor planeado $\sigma = \dfrac{\text{rango}}{4} = \dfrac{36}{4} = 9$

b) $n = \dfrac{z^2_{0.025}\,\sigma^2}{E^2} = \dfrac{(1.96)^2(9)^2}{(3)^2} = 34.57$; utilizar $n = 35$

c) $n = \dfrac{(1.96)^2(9)^2}{(2)^2} = 77.79$; utilizar $n = 78$

25. *a*) Utilizar $n = \dfrac{z^2_{\alpha/2}\sigma^2}{E^2}$

$n = \dfrac{(1.96)^2(6.84)^2}{(1.5)^2} = 79.88$; utilizar $n = 80$

b) $n = \dfrac{(1.645)^2(6.84)^2}{(2)^2} = 31.65$; utilizar $n = 32$

26. *a*) 18
b) 35
c) 97

28. *a*) 328
b) 465
c) 803
d) *n* se hace más grande; no lo recomendaría

30. 81

31. *a*) $\bar{p} = \dfrac{100}{400} = 0.25$

b) $\sqrt{\dfrac{\bar{p}(1-\bar{p})}{n}} = \sqrt{\dfrac{0.25(0.75)}{400}} = 0.0217$

c) $\bar{p} \pm z_{0.025}\sqrt{\dfrac{\bar{p}(1-\bar{p})}{n}}$
$0.25 \pm 1.96(0.0217)$
0.25 ± 0.0424; 0.2076 a 0.2924

32. *a*) 0.6733 a 0.7267
b) 0.6682 a 0.7318

34. 1 068

35. *a*) $\bar{p} = \dfrac{1\,760}{2\,000} = 0.88$

b) Margen de error

$z_{0.05} = \sqrt{\dfrac{\bar{p}(1-\bar{p})}{n}} = 1.645\sqrt{\dfrac{0.88(1-0.88)}{2\,000}} = 0.0120$

c) Intervalo de confianza
0.88 ± 0.0120 o 0.868 a 0.892

d) Margen de error

$$z_{0.05}\sqrt{\frac{p(1-p)}{n}} = 1.96\sqrt{\frac{0.88(1-0.88)}{2000}} = 0.0142$$

95% intervalo de confianza
0.88 ± 0.0142 o 0.8658 a 0.8942

36. *a*) 0.23
b) 0.1716 a 0.2884

38. *a*) 0.1790
b) 0.0738; 0.5682 a 0.7158
c) 354

39. *a*) $n = \dfrac{z^2_{0.025}p^*(1-p^*)}{E^2} = \dfrac{(1.96)^2(0.156)(1-0.156)}{(0.03)^2}$
$= 562$

b) $n = \dfrac{z^2_{0.005}p^*(1-p^*)}{E^2} = \dfrac{(2.576)^2(0.156)(1-0.156)}{(0.03)^2}$
$= 970.77$; utilice 971

40. 0.0346 (0.4854 a 0.5546)

42. *a*) 0.0442
b) 601, 1 068, 2 401, 9 604

44. *a*) 4.00
b) $29.77 a $37.77

46. *a*) 122
b) $1751 a $1995
c) $172, 316 millones
d) Menos de $1873

48. *a*) 14 minutos
b) 13.38 a 14.62
c) 32 por día
d) Reduciendo personal

50. 37

52. 176

54. *a*) 0.5420
b) 0.0508
c) 0.4912 a 0.5928

56. *a*) 0.8273
b) 0.7957 a 0.8589

58. *a*) 1 267
b) 1 509

60. *a*) 0.3101
b) 0.2898 a 0.3304
c) 8 219; no, este tamaño de muestra es innecesariamente grande

Capítulo 9

2. *a*) H_0: $\mu \leq 14$
H_a: $\mu > 14$
b) No hay evidencia de que el nuevo plan incremente las ventas
c) La hipótesis de investigación $\mu > 14$ es apoyada, el nuevo plan incrementa las ventas

4. *a*) H_0: $\mu \geq 220$
H_a: $\mu < 220$

5. *a*) Rechazar H_0: $\mu \leq 56.2$ cuando ésta es verdadera
b) Aceptar H_0: $\mu \leq 56.2$ cuando ésta es falsa

6. *a*) H_0: $\mu \leq 1$
H_a: $\mu > 1$
b) Reclamar $\mu > 1$ cuando ésta no es verdadera
c) Aceptando $\mu \leq 1$ cuando ésta no es verdadera

8. *a*) H_0: $\mu \geq 220$
H_a: $\mu < 220$
b) Reclamar $\mu < 220$ cuando esto no es verdadero
c) Reclamar $\mu \geq 220$ cuando esto no es verdadero

10. *a*) $z = \dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = \dfrac{26.4-25}{6/\sqrt{40}} = 1.48$
b) Usando la tabla normal $z = 1.48$: valor-p = 1.0000 − 0.9306 = 0.0694
c) El valor-$p > 0.01$; H_0 no es rechazada
d) H_0 es rechazada si $z \geq 2.33$
$1.48 < 2.33$, H_0 no es rechazada

11. *a*) $z = \dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = \dfrac{14.15-15}{3/\sqrt{50}} = -2.00$
b) El valor-$p = 2(0.0228) = 0.0456$
c) El valor-$p \leq 0.05$; H_0 es rechazada
d) H_0 es rechazada si $z \leq -1.96$ o $z \geq 1.96$
$-2.00 \leq -1.96$, se rechaza H_0

12. *a*) 0.1056; H_0 no es rechazada
b) 0.0062; H_0 es rechazada
c) ≈ 0; H_0 es rechazada
d) 0.7967; H_0 no es rechazada

14. *a*) 0.3844; H_0 no es rechazada
b) 0.0074; H_0 es rechazada
c) 0.0836; H_0 no es rechazada

15. *a*) H_0: $\mu \geq 1056$
H_a: $\mu < 1056$
b) $z = \dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = \dfrac{910-1056}{1600/\sqrt{400}} = -1.83$
El valor-$p = 0.0336$
c) El valor-$p \leq 0.05$; H_0 es rechazada; el reembolso medio de los contribuyentes de "última hora" es menor de $1 056
d) H_0 es rechazada si $z \leq -1.645$
$-1.83 \leq -1.645$; H_0 es rechazada

16. *a*) H_0 es rechazada: $\mu \leq 3173$
H_a: $\mu > 3173$
b) 0.0207
c) H_0 es rechazada; se concluye que el saldo medio de la tarjeta de crédito de los estudiantes universitarios ha aumentado

18. *a*) H_0: $\mu = 4.1$
H_a: $\mu \neq 4.1$
b) −2.21, 0.0272
c) H_0 es rechazada; el rendimiento para los Mid-Cap Growth Funds difiere de los U.S. Diversified Equity

20. *a*) H_0: $\mu \geq 32.79$
H_a: $\mu < 32.79$

b) -2.73
c) 0.0032
d) H_0 es rechaza; se concluye que el promedio mensual facturado en Internet es menor en los estados del sur

22. *a*) H_0: $\mu = 8$
H_a: $\mu \neq 8$
b) 0.1706
c) H_0 no es rechazada; no se puede concluir que el tiempo promedio de espera difiera de 8 minutos
d) 7.83 hasta 8.97; sí

24. *a*) $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{17 - 18}{4.5/\sqrt{48}} = -1.54$
b) Grados de libertad $= n - 1 = 47$
El área de la cola inferior está entre 0.05 y 0.10
El valor-p (para las dos colas) está entre 0.10 y 0.20
El valor-p exacto = 0.1303
c) El valor-$p > 0.05$; H_0 no es rechazada
d) Con $gl = 47$, $t_{0.025} = 2.012$
H_0 es rechazada si $t \leq -2.012$ o $t \geq 2.012$
$t = -1.54$; H_0 no es rechazada

26. *a*) Entre 0.02 y 0.05; el valor-p exacto = 0.0397; H_0 es rechazada
b) Entre 0.01 y 0.02; el valor-p exacto = 0.0125; H_0 es rechazada
c) Entre 0.10 y 0.20; el valor-p exacto = 0.1285; H_0 no es rechazada

27. *a*) H_0: $\mu \geq 238$
H_a: $\mu < 238$
b) $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{231 - 238}{80/\sqrt{100}} = -0.88$
Grados de libertad $= n - 1 = 99$
El valor-p está entre 0.10 y 0.20
El valor-p exacto = 0.1905
c) El valor-$p > 0.05$; H_0 no es rechazada
No se puede concluir que la prestación media semanal en Virginia sea menor a la media nacional
d) $gl = 99$, $t_{0.05} = -1.66$
H_0 es rechazada si $t \leq -1.66$
$-0.88 > -1.66$; H_0 no es rechazada

28. *a*) H_0: $\mu \geq 9$
H_a: $\mu < 9$
b) Entre 0.005 y 0.01
El valor-p exacto = 0.0072
c) H_0 es rechazada; la permanencia media de un CEO es menor de 9 años

30. *a*) H_0: $\mu = 600$
H_a: $\mu \neq 600$
b) Entre 0.20 y 0.40
El valor-p exacto = 0.2491
c) H_0 no es rechazada; se concluye que no se registran cambios en el promedio de audiencia de CNN
d) Una muestra de mayor tamaño

32. *a*) H_0: $\mu = 10\,192$
H_a: $\mu \neq 10\,192$
b) Entre 0.02 y 0.05
El valor-p exacto = 0.0304
c) H_0 es rechazada; el precio promedio del concesionario difiere del precio promedio nacional

34. *a*) H_0: $\mu = 2$
H_a: $\mu \neq 2$
b) 2.2
c) 0.52
d) Entre 0.20 y 0.40
El valor-p exacto = 0.2535
e) H_0 no es rechazada; no existe razón para modificar el criterio de 2 horas en la estimación de costos

36. *a*) $z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} = \dfrac{0.68 - 0.75}{\sqrt{\dfrac{0.75(1 - 0.75)}{300}}} = -2.80$
El valor-p = 0.0026
El valor-$p \leq 0.05$; H_0 es rechazada
b) $z = \dfrac{0.72 - 0.75}{\sqrt{\dfrac{0.75(1 - 0.75)}{300}}} = -1.20$
El valor-p = 0.1151
El valor-$p > 0.05$; H_0 no es rechazada
c) $z = \dfrac{0.70 - 0.75}{\sqrt{\dfrac{0.75(1 - 0.75)}{300}}} = -2.00$
El valor-p = 0.0228
El valor-$p \leq 0.05$; H_0 es rechazada
d) $z = \dfrac{0.77 - 0.75}{\sqrt{\dfrac{0.75(1 - 0.75)}{300}}} = 0.80$
El valor-p = 0.7881
El valor-$p > 0.05$; H_0 no es rechazada

38. *a*) H_0: $p = 0.64$
H_a: $p \neq 0.64$
b) $\bar{p} = 52/100 = 0.52$
$z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} = \dfrac{0.52 - 0.64}{\sqrt{\dfrac{0.64(1 - 0.64)}{100}}} = -2.50$
El valor-$p = 2(0.0062) = 0.0124$
c) El valor-$p \leq 0.05$; H_0 es rechazada
La proporción difiere del 0.64 reportado
d) Sí, porque $\bar{p} = 0.52$ indica que muy pocos creen que la marca de supermercados sea tan buena como la marca nacional

40. *a*) 0.2702
b) H_0: $p \leq 0.22$
H_a: $p > 0.22$
El valor-$p \approx 0$; H_0 es rechazada; existe un incremento significativo después de ver los comerciales
c) Ayuda a evaluar la eficacia de los anuncios

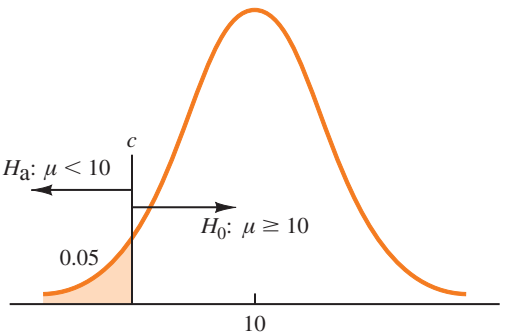
42. *a*) $\bar{p} = 0.15$
b) 0.0718 hasta 0.2282
c) La tasa de cambio de la tienda de Houston es diferente del promedio nacional

44. *a*) H_0: $p \leq 0.51$
H_a: $p > 0.51$

b) $\bar{p} = 0.58$, el valor-$p = 0.0026$

c) H_0 es rechazada; las personas que trabajan en el turno de noche manejan somnolientos con mayor frecuencia

46.



$$c = 10 - 1.645(5/\sqrt{120}) = 9.25$$

H_0 es rechazada si $\bar{x} \leq 9.25$

a) Cuando $\mu = 9$,

$$z = \frac{9.25 - 9}{5/\sqrt{120}} = 0.55$$

$P(H_0 \text{ es rechazada}) = (1.0000 - 0.7088) = 0.2912$

b) Error tipo II

c) Cuando $\mu = 8$,

$$z = \frac{9.25 - 8}{5/\sqrt{120}} = 2.74$$

$$\beta = (1.0000 - 0.9969) = 0.0031$$

48. *a*) Se concluye que $\mu \leq 15$ cuando en realidad no es así

b) 0.2676

c) 0.0179

49. *a*) H_0: $\mu \geq 25$
H_a: $\mu < 25$
H_0 es rechazada si $z \leq -2.05$

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 25}{3/\sqrt{30}} = -2.05$$

Se resuelve por $\bar{x} = 23.88$
Regla de decisión: Aceptar H_0 si $\bar{x} > 23.88$
Rechazar H_0 si $\bar{x} \leq 23.88$

b) Para $\mu = 23$,

$$z = \frac{23.88 - 23}{3/\sqrt{30}} = 1.61$$

$$\beta = 1.0000 - 0.9463 = 0.0537$$

c) Para $\mu = 24$,

$$z = \frac{23.88 - 24}{3/\sqrt{30}} = -0.22$$

$$\beta = 1.0000 - 0.4129 = 0.5871$$

d) En este caso no se puede cometer el error tipo II. Observe que cuando $\mu = 25.5$, H_0 es verdadera; el error tipo II sólo se puede cometer cuando H_0 es falsa

50. *a*) Se concluye que $\mu = 28$ cuando esto no es verdad

b) 0.0853, 0.6179, 0.6179, 0.0853

c) 0.9147

52. 0.1151, 0.0015
Al aumentar n disminuye β

54. $$n = \frac{(z_\alpha + z_\beta)^2\sigma^2}{(\mu_0 - \mu_a)^2} = \frac{(1.645 + 1.28)^2(5)^2}{(10 - 9)^2} = 214$$

56. 109

57. En $\mu_0 = 400$, $\alpha = 0.02$; $z_{0.02} = 2.05$
En $\mu_a = 385$, $\beta = 0.10$; $z_{0.10} = 1.28$
Con $\sigma = 30$,

$$n = \frac{(z_\alpha + z_\beta)^2\sigma^2}{(\mu_0 - \mu_a)^2} = \frac{(2.05 + 1.28)^2(30)^2}{(400 - 385)^2} = 44.4 \text{ o } 45$$

58. 324

60. *a*) H_0: $\mu = 16$
H_a: $\mu \neq 16$

b) 0.0286; H_0 es rechazada
Reajuste de la línea

c) 0.2186; H_0 no es rechazada
Continuar la operación

d) $z = 2.19$; H_0 es rechazada
$z = -1.23$; H_0 es rechazada
Sí, es la misma conclusión

62. *a*) H_0: $\mu \leq 119\,155$
H_a: $\mu > 119\,155$

b) 0.0047

c) H_0 es rechazada; el ingreso anual promedio de los aficionados al teatro en el área de la bahía es mayor

64. $t = -1.05$
El valor-p está entre 0.20 y 0.40
El valor-p exacto = 0.2999
H_0 no es rechazada; no existe evidencia para concluir que la edad en que las mujeres tenían a su primer hijo haya cambiado

66. $t = 2.26$
El valor-p está entre 0.01 y 0.025
El valor-p exacto = 0.0155
H_0 es rechazada; el costo promedio es mayor a $125 000

68. *a*) H_0: $p \leq 0.50$
H_a: $p > 0.50$

b) 0.64

c) 0.0026; H_0 es rechazada; los egresados universitarios dejan de fumar con buena tasa de éxito

70. *a*) H_0: $p \leq 0.80$
H_a: $p > 0.80$

b) 0.84

c) 0.0418

d) H_0 es rechazada; más de 80% de los clientes está satisfecho con el servicio de los agentes en casa

72. H_0: $p \geq 0.90$
H_a: $p < 0.90$
El valor-$p = 0.0808$
H_0 no es rechazada; la demanda de por lo menos 90% no puede ser rechazada

74. *a*) H_0: $\mu \leq 72$
H_a: $\mu > 72$

b) 0.2912

c) 0.7939

d) 0, debido a que H_0 es verdadera

76. *a*) 45

b) 0.0192, 0.2358, 0.7291, 0.7291, 0.2358, 0.0192

Capítulo 10

1. *a)* $\bar{x}_1 - \bar{x}_2 = 13.6 - 11.6 = 2$

b) $z_{\alpha/2} = z_{.05} = 1.645$

$$\bar{x}_1 - \bar{x}_2 \pm 1.645\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$2 \pm 1.645\sqrt{\frac{(2.2)^2}{50} + \frac{(3)^2}{35}}$$

2 ± 0.98 (1.02 hasta 2.98)

c) $z_{\alpha/2} = z_{0.05} = 1.96$

$$2 \pm 1.96\sqrt{\frac{(2.2)^2}{50} + \frac{(3)^2}{35}}$$

2 ± 1.17 (0.83 a 3.17)

2. *a)* $$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(25.2 - 22.8) - 0}{\sqrt{\frac{(5.2)^2}{40} + \frac{(6)^2}{50}}} = 2.03$$

b) El valor-p = 1.0000 − 0.9788 = 0.0212

c) El valor-$p \leq 0.05$; H_0 es rechaza

4. *a)* $\bar{x}_1 - \bar{x}_2 = 85.36 - 81.40 = 3.96$

b) $$z_{0.025}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 1.96\sqrt{\frac{(4.55)^2}{37} + \frac{(3.97)^2}{44}}\ 1.88$$

c) 3.96 ± 1.88 (2.08 a 5.84)

6. El valor-p = 0.0351

H_0 es rechazada; el precio promedio en Atlanta es menor que el precio promedio en Houston

8. *a)* H_0 es rechazada; el servicio al cliente ha mejorado para Rite Aid

b) H_0 no es rechazada; la diferencia estadística no es significativa

c) El valor-p = 0.0336; H_0 es rechazada; el servicio al cliente ha mejorado para Expedia

d) 1.80

e) El incremento de J.C) Penney no es estadísticamente significativo

9. *a)* $\bar{x}_1 - \bar{x}_2 = 22.5 - 20.1 = 2.4$

b) $$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

$$= \frac{\left(\frac{2.5^2}{20} + \frac{4.8^2}{30}\right)^2}{\frac{1}{19}\left(\frac{2.5^2}{20}\right)^2 + \frac{1}{29}\left(\frac{4.8^2}{30}\right)^2} = 45.8$$

c) $gl = 45$, $t_{0.025} = 2.014$

$$t_{0.025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 2.014\sqrt{\frac{2.5^2}{20} + \frac{4.8^2}{30}} = 2.1$$

d) 2.4 ± 2.1 (0.3 a 4.5)

10. *a)* $$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(13.6 - 10.1) - 0}{\sqrt{\frac{5.2^2}{35} + \frac{8.5^2}{40}}} = 2.18$$

b) $$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

$$= \frac{\left(\frac{5.2^2}{35} + \frac{8.5^2}{40}\right)^2}{\frac{1}{34}\left(\frac{5.2^2}{35}\right)^2 + \frac{1}{39}\left(\frac{8.5^2}{40}\right)^2} = 65.7$$

Utilice $gl = 65$

c) $gl = 65$, el área en la cola está entre 0.01 y 0.025; el valor-p para las dos colas se sitúa entre 0.02 y 0.05
Valor-p exacto = 0.0329

d) Valor-$p \leq 0.05$; H_0 es rechazada

12. *a)* $\bar{x}_1 - \bar{x}_2 = 22.5 - 18.6 = 3.9$ millas

b) $$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

$$= \frac{\left(\frac{8.4^2}{50} + \frac{7.4^2}{40}\right)^2}{\frac{1}{49}\left(\frac{8.4^2}{50}\right)^2 + \frac{1}{39}\left(\frac{7.4^2}{40}\right)^2} = 87.1$$

Utilice $gl = 87$, $t_{0.025} = 1.988$

$$3.9 \pm 1.988\sqrt{\frac{8.4^2}{50} + \frac{7.4^2}{40}}$$

3.9 ± 3.3 (0.6 a 7.2)

14. *a)* H_0: $\mu_1 - \mu_2 \geq 0$
H_a: $\mu_1 - \mu_2 < 0$

b) −2.41

c) Al utilizar la tabla t, el valor-p está entre 0.005 y 0.01
Valor-p exacto = 0.009

d) H_0 es rechazada; los sueldos más bajos de enfermería están en Tampa

16. *a)* H_0: $\mu_1 - \mu_2 \leq 0$
H_a: $\mu_1 - \mu_2 > 0$

b) 38

c) $t = 1.80$, $df = 25$
Al utilizar la tabla t, el valor-p está entre 0.025 y 0.05
Valor-p exacto = 0.0420

d) H_0 es rechazada; se concluye que se obtienen las mejores puntuaciones si se tiene grado de licenciatura

18. *a)* H_0: $\mu_1 - \mu_2 \geq 120$
H_a: $\mu_1 - \mu_2 < 120$

b) −2.10
Al utilizar la tabla t, el valor-p está entre 0.01 y 0.025
Valor-p exacto = 0.0195

c) 32 hasta 118

d) La muestra de mayor tamaño

19. *a)* 1, 2, 0, 0, 2

b) $\bar{d} = \Sigma d_i / n = 5/5 = 1$

c) $$s_d = \sqrt{\frac{\Sigma(d_i - \bar{d})^2}{n - 1}} = \sqrt{\frac{4}{5 - 1}} = 1$$

d) $t = \dfrac{\bar{d} - \mu}{s_d/\sqrt{n}} = \dfrac{1 - 0}{1/\sqrt{5}} = 2.24$

$gl = n - 1 = 4$

Utilizando la tabla *t*, el valor-*p* está entre 0.025 y 0.05

Valor-*p* exacto = 0.0443

Valor-*p* ≤ 0.05; H_0 es rechazada

20. *a)* 3, −1, 3, 5, 3, 0, 1

b) 2

c) 2.08

d) 2

e) 0.07 a 3.93

21. H_0: $\mu_d \leq 0$

H_a: $\mu_d > 0$

$\bar{d} = 0.625$

$s_d = 1.30$

$t = \dfrac{\bar{d} - \mu_d}{s_d/\sqrt{n}} = \dfrac{0.625 - 0}{1.30/\sqrt{8}} = 1.36$

$gl = n - 1 = 7$

Utilizando la tabla *t*, el valor-*p* está entre 0.10 y 0.20

Valor-*p* exacto = 0.1080

Valor-*p* > 0.05; H_0 no es rechazada; no se puede concluir que el comercial mejora la media del potencial de comprar

22. $0.10 a $.32; las utilidades se han incrementado

24. $t = 1.32$

Utilizando la taba *t*, el valor-*p* es mayor de 0.10

Valor-*p* exacto = 0.1142

H_0 no es rechazada; no se puede concluir que las tarifas de Dayton sen más altas

26. *a)* $t = -1.42$

Utilizando la tabla *t*, el valor-*p* está entre 0.10 y 0.20

Valor-*p* exacto = 0.1718

H_0 no es rechazada; no hay diferencia media en las puntuaciones

b) −1.05

c) 1.28; sí

28. *a)* $\bar{p}_1 - \bar{p}_2 = 0.48 - 0.36 = 0.12$

b) $\bar{p}_1 - \bar{p}_2 \pm z_{0.05}\sqrt{\dfrac{\bar{p}_1(1 - \bar{p}_1)}{n_1} + \dfrac{\bar{p}_2(1 - \bar{p}_2)}{n_2}}$

$0.12 \pm 1.645\sqrt{\dfrac{0.48(1 - 0.48)}{400} + \dfrac{0.36(1 - 0.36)}{300}}$

0.12 ± 0.0614 (0.0586 hasta 0.1814)

c) $0.12 \pm 1.96\sqrt{\dfrac{0.48(1 - 0.48)}{400} + \dfrac{0.36(1 - 0.36)}{300}}$

0.12 ± 0.0731 (0.0469 a 0.1931)

29. *a)* $\bar{p} = \dfrac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} = \dfrac{200(0.22) + 300(0.16)}{200 + 300} = 0.1840$

$z = \dfrac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1 - \bar{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

$= \dfrac{0.22 - 0.16}{\sqrt{0.1840(1 - 0.1840)\left(\dfrac{1}{200} + \dfrac{1}{300}\right)}} = 1.70$

Valor-*p* = 1.0000 − 0.9554 = 0.0446

b) Valor-*p* ≤ 0.05; H_0 es rechazada

30. $\bar{p}_1 = 0.55, \quad \bar{p}_2 = 0.48$

0.07 ± 0.0691

32. *a)* H_0: $p_w \leq p_m$

H_a: $p_w > p_m$

b) $\bar{p}_w = 0.3699$

c) $\bar{p}_m = 0.3400$

d) Valor-*p* = 0.1093

H_0 no es rechazada; no se puede concluir que las mujeres sean más propensas a preguntar por un domicilio

34. *a)* 0.64

b) 0.45

c) 0.19 ± 0.0813 (0.1087 a 0.2713)

36. *a)* H_0: $p_1 - p_2 = 0$

H_a: $p_1 - p_2 \neq 0$

b) 0.13

c) Valor-*p* = 0.0404

d) H_0 es rechazada; existe una diferencia significativa entre los grupos de edad de jóvenes y adultos

38. *a)* H_0: $\mu_1 - \mu_2 = 0$

H_a: $\mu_1 - \mu_2 \neq 0$

$z = 2.79$

Valor-*p* = 0.0052

H_0 es rechazada; existe una diferencia significativa entre los sistemas existentes

40. *a)* H_0: $\mu_1 - \mu_2 \leq 0$

H_a: $\mu_1 - \mu_2 > 0$

b) $t = 0.60$, $gl = 57$

Utilizando la tabla *t*, el valor-*p* es mayor de 0.20

Valor-*p* exacto = 0.2754

H_0 no es rechazada; no se puede concluir que los fondos de los préstamos tienen una mayor tasa promedio de rendimiento

42. *a)* Una disminución de $2.45

b) 2.45 ± 2.15 ($0.30 hasta 4.60)

c) Disminuyó 8%

d) $23.93

44. *a)* El valor-*p* ≈ 0; H_0 es rechazada

b) 0.0468 hasta 0.1332

46. *a)* 0.35 y 0.47

b) 0.12 ± 0.1037 (0.0163 hasta 0.2237)

c) Sí, se esperaría que las tasas de ocupación sean mayores

Capítulo 11

2. $s^2 = 25$

a) Con 19 grados de libertad, $\chi^2_{0.05} = 30.144$ y $\chi^2_{0.95} = 10.117$

$$\frac{19(25)}{30.144} \leq \sigma^2 \leq \frac{19(25)}{10.117}$$

$$15.76 \leq \sigma^2 \leq 46.95$$

b) Con 19 grados de libertad, $\chi^2_{0.025} = 32.852$ y $\chi^2_{0.975} = 8.907$

$$\frac{19(25)}{32.852} \leq \sigma^2 \leq \frac{19(25)}{8.907}$$

$$14.46 \leq \sigma^2 \leq 53.33$$

c) $3.8 \leq \sigma \leq 7.3$

4. *a*) 0.22 a 0.71
b) 0.47 a 0.84

6. *a*) 0.2205, 47.95, 6.92
b) 5.27 a 10.11

8. *a*) 0.4748
b) 0.6891
c) 0.2383 a 1.3687
0.4882 a 1.1699

9. H_0: $\sigma^2 \leq 0.0004$
H_a: $\sigma^2 > 0.0004$

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(30-1)(0.0005)}{0.0004} = 36.25$$

De la tabla con 29 grados de libertad, el valor-*p* es mayor de 0.10
El valor-$p > 0.05$; H_0 no es rechazada
La especificación del producto no parece violarse

10. H_0: $\sigma^2 \leq 331.24$
H_a: $\sigma^2 > 331.24$
$\chi^2 = 52.07$, $gl = 35$
El valor-*p* está entre 0.025 y 0.05
H_0 es rechazada; la desviación estándar de Vanguard es mayor

12. *a*) 0.8106
b) $\chi^2 = 9.49$
El valor-*p* es mayor que 0.20
H_0 no es rechazada; no se puede concluir que la varianza para la otra revista sea diferente

14. *a*) $F = 2.4$
El valor-*p* está entre 0.025 y 0.05
H_0 es rechazada
b) $F_{0.05} = 2.2$; H_0 es rechazada

15. *a*) La varianza muestral mayores s_1^2

$$F = \frac{s_1^2}{s_2^2} = \frac{8.2}{4} = 2.05$$

Grados de libertad: 20, 25
En la tabla, el área en la cola está entre 0.025 y 0.05
El valor-*p* para las dos colas está situado entre 0.05 y 0.10
El valor-$p > 0.05$; H_0 no es rechazada
b) Para una prueba de dos colas:
$F_{\alpha/2} = F_{0.025} = 2.30$
H_0 es rechazada si $F \geq 2.30$
$2.05 < 2.30$; H_0 no es rechazada

16. $F = 1.59$
El valor-*p* es menor de 0.05
H_0 es rechazada; el fondo de Fidelity tiene mayor varianza

17. *a*) La población 1 de los automóviles de 4 años de antigüedad es
H_0: $\sigma_1^2 \leq \sigma_2^2$
H_a: $\sigma_1^2 > \sigma_2^2$
b) $F = \frac{s_1^2}{s_2^2} = \frac{170^2}{100^2} = 2.89$
Grados de libertad: 25, 24
En las tablas, el valor-*p* es menor de 0.01
El valor-$p \leq 0.01$; H_0 es rechazada
Se concluye que los automóviles de 4 años de antigüedad tienen una mayor varianza en el costo de reparación anual en relación con los de 2 años de antigüedad, lo cual es previsible debido a que los automóviles más viejos son más propensos a las reparaciones más caras que llevan a una mayor variación en los costos de reparación anual

18. $F = 1.44$
El valor-*p* es mayor de 0.20
H_0 no es rechazada; la diferencia entre las varianzas no es estadísticamente significativa

20. $F = 5.29$
El valor-$p \approx 0$
H_0 es rechazada; las varianzas de la población no son iguales para los ejecutivos y los gerentes

22. *a*) $F = 4$
El valor-*p* es menor de 0.01
H_0 es rechazada; se presenta mayor variabilidad en la distancia de frenado en pavimento mojado

24. 10.72 a 24.68

26. *a*) $\chi^2 = 27.44$
El valor-*p* entre 0.01 y 0.025
H_0 es rechazada; la varianza excede los requerimientos máximos
b) 0.00012 a 0.00042

28. $\chi^2 = 31.50$
El valor-*p* está entre 0.05 y 0.10
H_0 es rechazada; se concluye que la varianza de la población es mayor de 1

30. *a*) $n = 15$
b) 6.25 a 11.13

32. $F = 1.39$
H_0 no es rechazada; no se puede concluir que las varianzas del promedio de calificaciones son diferentes

34. $F = 2.08$
El valor-*p* está entre 0.05 y 0.10
H_0 es rechazada; se concluye que las varianzas de la población no son iguales

Capítulo 12

1. *a*) Frecuencias esperadas: $e_1 = 200(0.40) = 80$
$e_2 = 200(0.40) = 80$
$e_3 = 200(0.20) = 40$
Frecuencias observadas: $f_1 = 60, f_2 = 120, f_3 = 20$

$$\chi^2 = \frac{(60-80)^2}{80} + \frac{(120-80)^2}{80} + \frac{(20-40)^2}{40}$$
$$= \frac{400}{80} + \frac{1600}{80} + \frac{400}{40}$$
$$= 5 + 20 + 10 = 35$$

Grados de libertad: $k - 1 = 2$
$\chi^2 = 35$ demuestra que el valor-p es menor que 0.005
El valor-$p \leq 0.01$; se rechaza H_0; las proporciones no son de 0.40, 0.40 y 0.20

b) H_0 es rechazada si $\chi^2 \geq 9.210$
$\chi^2 = 35$; H_0 es rechazada

2. $\chi^2 = 15.33$, $gl = 3$
El valor-p es menor de 0.005
H_0 es rechazada; las proporciones no son todas 0.25

3. H_0: $p_{ABC} = 0.29$, $p_{CBS} = 0.28$, $p_{NBC} = 0.25$, $p_{IND} = 0.18$
H_a: las proporciones no son
$p_{ABC} = 0.29$, $p_{CBS} = 0.28$, $p_{NBC} = 0.25$, $p_{IND} = 0.18$
Frecuencias esperadas: $300(0.29) = 87$, $300(0.28) = 84$
$300(0.25) = 75$, $300(0.18) = 54$
$e_1 = 87$, $e_2 = 84$, $e_3 = 75$, $e_4 = 54$
Frecuencias observadas: $f_1 = 95$, $f_2 = 70$, $f_3 = 89$, $f_4 = 46$

$$\chi^2 = \frac{(95-87)^2}{87} + \frac{(70-84)^2}{84} + \frac{(89-75)^2}{75} + \frac{(46-54)^2}{54} = 6.87$$

Grados de libertad: $k - 1 = 3$
$\chi^2 = 6.87$, el valor-p está entre 0.05 y 0.10
H_0 no es rechazada; no se puede concluir que las proporciones de audiencia hayan cambiado

4. $\chi^2 = 29.51$, $gl = 5$
El valor-p es menor que 0.005
H_0 es rechazada; los porcentajes difieren de aquellos indicados por la empresa

6. *a*) $\chi^2 = 12.21$, $gl = 3$
El valor-p está entre 0.005 y 0.01
Se concluye la diferencia para 2003

b) 21%, 30%, 15%, 34%
Aumento en el uso de la tarjeta de crédito

c) 51%

8. $\chi^2 = 16.31$, $gl = 3$
El valor-p es menor que 0.005
H_0 es rechazada; las evaluaciones difieren, con un servicio telefónico levemente mejor

9. H_0: la columna de las variables es independiente de las filas de las variables

H_a: la columna de las variables no es independiente de la fila de las variables

Las frecuencias esperadas:

	A	B	C
P	28.5	39.9	45.6
Q	21.5	30.1	34.4

$$\chi^2 = \frac{(20-28.5)^2}{28.5} + \frac{(44-39.9)^2}{39.9} + \frac{(50-45.6)^2}{45.6} + \frac{(30-21.5)^2}{21.5} + \frac{(26-30.1)^2}{30.1} + \frac{(30-34.4)^2}{34.4}$$
$$= 7.86$$

Grados de libertad: $(2-1)(3-1) = 2$
$\chi^2 = 7.86$, el valor-p está entre 0.01 y 0.025
Se rechaza H_0; las columnas de variables y las filas de las variables no son independientes

10. $\chi^2 = 19.77$, $gl = 4$
El valor-p es menor que 0.005
H_0 es rechazada; las columnas de las variables y las filas de las variables no son independientes

11. H_0: el tipo del boleto comprado es independiente del tipo de vuelo

H_a: el tipo de boleto comprado no es independiente del tipo de vuelo

Frecuencias esperadas:

$e_{11} = 35.59$ $\quad e_{12} = 15.41$
$e_{21} = 150.73$ $\quad e_{22} = 65.27$
$e_{31} = 455.68$ $\quad e_{32} = 197.32$

Boleto	Vuelo	Frecuencia observada (f_i)	Frecuencia esperada (e_i)	$(f_i - e_i)^2/e_i$
Primera	Nacional	29	35.59	1.22
Primera	International	22	15.41	2.82
Clase de negocios	Nacional	95	150.73	20.61
Clase de negocios	International	121	65.27	47.59
Tarifa completa	Nacional	518	455.68	8.52
Tarifa completa	International	135	197.32	19.68
Totales		920		$\chi^2 = 100.43$

Grados de libertad: $(3-1)(2-1) = 2$
$\chi^2 = 100.43$; el valor-p es menor de 0.005
H_0 es rechazada; el tipo del boleto no es independiente del tipo de vuelo

12. a) $\chi^2 = 7.95$, $gl = 3$
El valor-p está entre 0.025 y 0.05
H_0 es rechazada; la forma de pago no es independiente de la edad del grupo

b) 18 a 24; la mayoría lo utiliza

14. a) $\chi^2 = 8.47$; el valor-p está entre 0.025 y 0.05
H_0 es rechazada; la intención de comprar nuevamente no es independiente del automóvil

b) Accord 77, Camry 71, Taurus 62, Impala 57

c) Impala y Taurus están abajo, Accord y Camry son los primeros; Accord y Camry dan mayor satisfacción a los propietarios, lo que puede ayudar a una futura participación de mercado

16. a) 6446

b) $\chi^2 = 425.4$; el valor-$p = 0$
H_0 es rechazada; la actitud hacia una potencia nuclear no es independiente del país

c) Italia (58%), España (32%)

18. $\chi^2 = 3.01$, $gl = 2$
El valor-p es mayor de 0.10
H_0 no es rechazada; las parejas que trabajan es independiente de la ubicación: 63.3%

20. Primero se estima μ a partir de los datos muestrales (tamaño de la muestra = 120)

$$\mu = \frac{0(39) + 1(30) + 2(30) + 3(18) + 4(3)}{120}$$
$$= \frac{156}{120} = 1.3$$

Por tanto, usamos las probabilidades de Poisson con $\mu = 1.3$ para calcular las frecuencias esperadas

x	Frecuencias observadas	Probabilidad de Poisson	Frecuencias esperada	Diferencia $(f_i - e_i)$
0	39	0.2725	32.70	6.30
1	30	0.3543	42.51	−12.51
2	30	0.2303	27.63	2.37
3	18	0.0998	11.98	6.02
4 o más	3	0.0431	5.16	−2.17

$$\chi^2 = \frac{(6.30)^2}{32.70} + \frac{(-12.51)^2}{42.51} + \frac{(2.37)^2}{27.63} + \frac{(6.02)^2}{11.98} + \frac{(-2.17)^2}{5.16} = 9.04$$

Grados de libertad: $5 - 1 - 1 = 3$
$\chi^2 = 9.04$, el valor-p está entre 0.025 y 0.05
H_0 es rechazada; ninguna distribución de Poisson

21. Con $n = 30$ se utilizarán 6 clases, cada una con 0.1667 de probabilidad relacionada con cada clase

$$\bar{x} = 22.8,\ s = 6.27$$

Los valores z que crean 6 intervalos, cada uno con una probabilidad de 0.1667, son −0.98, −0.43, 0, 0.43, 0.98

z	Valor x
−0.98	22.8 − 0.98(6.27) = 16.66
−0.43	22.8 − 0.43(6.27) = 20.11
0	22.8 + 0.00(6.27) = 22.80
0.43	22.8 + 0.43(6.27) = 25.49
0.98	22.8 + 0.98(6.27) = 28.94

Intervalo	Frecuencia observada	Frecuencia esperada	Diferencia
menor que 16.66	3	5	−2
16.66–20.11	7	5	2
20.11–22.80	5	5	0
22.80–25.49	7	5	2
25.49–28.94	3	5	−2
28.94 y más	5	5	0

$$\chi^2 = \frac{(-2)^2}{5} + \frac{(2)^2}{5} + \frac{(0)^2}{5} + \frac{(2)^2}{5} + \frac{(-2)^2}{5} + \frac{(0)^2}{5}$$
$$= \frac{16}{5} = 3.20$$

Grados de libertad: $6 - 2 - 1 = 3$
$\chi^2 = 3.20$, el valor-p es mayor de 0.10
H_0 no es rechazada
No se rechaza el supuesto de una distribución normal

22. $\chi^2 = 4.30$; $gl = 2$
El valor-p: es mayor de 0.10
H_0 no es rechazada; no se rechaza el supuesto de una distribución Poisson

24. $\chi^2 = 2.8$; $gl = 3$
El valor-p es mayor de 0.10
H_0 no es rechazada; no se rechaza el supuesto de una distribución normal

26. $\chi^2 = 8.04$, $gl = 3$
El valor-p entre 0.025 y 0.05
H_0 es rechazada; el potencial no es el mismo para cada zona de ventas

28. $\chi^2 = 4.64$; $gl = 2$
El valor-p está entre 0.05 y 0.10
H_0 no es rechazada; no se puede concluir que la participación de mercado haya cambiado

30. $\chi^2 = 42.53$; $gl = 4$
El valor-p es menor de 0.005
H_0 es rechazada; se concluye que la satisfacción laboral difiera

32. $\chi^2 = 23.37$; $gl = 3$
El valor-p es menor de 0.005
H_0 es rechazada; la situación laboral no es independiente de la región

34. a) 71%, 22%; menor preferencia
b) $\chi^2 = 2.99$; $gl = 2$
El valor-p es mayor de 0.10
H_0 no es rechazada; no se puede concluir que los hombres y las mujeres difieran en la preferencia

36. $\chi^2 = 6.17$; $gl = 6$
El valor-p es mayor de 0.10
H_0 no es rechazada; el supuesto de que el condado y el día de la semana son independientes no se puede rechazar

38. $\chi^2 = 7.75$; $gl = 3$
El valor-p está entre 0.05 y 0.10
H_0 no es rechazada; no se puede concluir que las oficinas vacantes difieran por área metropolitana

Capítulo 13

1. *a)* $\bar{\bar{x}} = (156 + 142 + 134)/3 = 144$

$$\text{SCTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2$$
$$= 6(156 - 144)^2 + 6(142 - 144)^2 + 6(134 - 144)^2$$
$$= 1\,488$$

b) $\text{CMTR} = \dfrac{\text{SCTR}}{k-1} = \dfrac{1\,488}{2} = 744$

c) $s_1^2 = 164.4$, $s_2^2 = 131.2$, $s_3^2 = 110.4$

$$\text{SCE} = \sum_{j=1}^{k} (n_j - 1)s_j^2$$

$$= 5(164.4) + 5(131.2) + 5(110.4)$$
$$= 2030$$

d) $$\text{CME} = \frac{\text{SCE}}{n_T - k} = \frac{2030}{18 - 3} = 135.3$$

e)

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	*F*	valor-*p*
Tratamientos	1488	2	744	5.50	0.0162
Error	2030	15	135.3		
Total	3518	17			

f) $$F = \frac{\text{CMTR}}{\text{CME}} = \frac{744}{135.3} = 5.50$$

De la tabla *F* (2 grados de libertad en el numerador y 15 grados de libertad en el denominador), el valor-*p* está entre 0.01 y 0.025

Utilizando Excel o Minitab, el valor-*p* que corresponde a $F = 5.50$ es 0.0162

Debido a que el valor-$p \leq \alpha = 0.05$, la hipótesis de que las medidas de los tres tratamientos son iguales es rechazada

2.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	*F*	Valor-*p*
Tratamientos	300	4	75	14.07	0.0000
Error	160	30	5.33		
Total	460	34			

4.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	*F*	Valor-*p*
Tratamientos	150	2	75	4.80	0.0233
Error	250	16	15.63		
Total	400	18			

H_0 es rechazada debido a que el valor-$p \leq \alpha = 0.05$

6. Debido a que el valor-$p = 0.0082$ es menor que $\alpha = 0.05$, la hipótesis nula de que las medias de los tres tratamientos son iguales es rechazada

8. $\bar{\bar{x}} = (79 + 74 + 66)/3 = 73$

$$\text{SCTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 = 6(79 - 73)^2 + 6(74 - 73)^2 + 6(66 - 73)^2 = 516$$

$$\text{CMTR} = \frac{\text{SCTR}}{k - 1} = \frac{516}{2} = 258$$

$$s_1^2 = 34 \qquad s_2^2 = 20 \qquad s_3^2 = 32$$

$$\text{SCE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 5(34) + 5(20) + 5(32) = 430$$

$$\text{CME} = \frac{\text{SCE}}{n_T - k} = \frac{430}{18 - 3} = 28.67$$

$$F = \frac{\text{CMTR}}{\text{CME}} = \frac{258}{28.67} = 9.00$$

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	*F*	valor-*p*
Tratamientos	516	2	258	9.00	0.003
Error	430	15	28.67		
Total	946	17			

Utilizando la tabla *F* (2 grados de libertad en el numerador y 15 en el denominador) el valor-*p* es menor de 0.01

Utilizando Excel o Minitab, el valor-*p* que corresponde a $F = 9.00$ es 0.003

Debido a que el valor-$p \leq \alpha = 0.05$, la hipótesis nula de que las medias para las tres fábricas sean iguales es rechazada; es decir, el análisis de varianza apoya la conclusión de que las puntuaciones de examen de la media poblacional en las tres fábricas NCP no son iguales

10. El valor-$p = 0.0000$

Debido a que el valor-$p \leq \alpha = 0.05$, la hipótesis nula de que las medias para los tres grupos sean iguales es rechazada

12. El valor-$p = 0.0038$

Debido a que el valor-$p \leq \alpha = 0.05$, la hipótesis nula de que los precios promedio de la comida sean los mismos para los tres tipos de restaurantes es rechazada

13. *a)* $\bar{\bar{x}} = (30 + 45 + 36)/3 = 37$

$$\text{SCTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 = 5(30 - 37)^2 + 5(45 - 37)^2 + 5(36 - 37)^2 = 570$$

$$\text{CMTR} = \frac{\text{SCTR}}{k - 1} = \frac{570}{2} = 285$$

$$\text{SCE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 4(6) + 4(4) + 4(6.5) = 66$$

$$\text{CME} = \frac{\text{SCE}}{n_T - k} = \frac{66}{15 - 3} = 5.5$$

$$F = \frac{\text{CMTR}}{\text{CME}} = \frac{285}{5.5} = 51.82$$

Utilizando la tabla *F* (2 grados de libertad en el numerador y 12 en el denominador), el valor-*p* es menor de 0.01

Utilizando Excel o Minitab, el valor-*p* correspondiente a $F = 51.82$ es 0.0000

Debido a que el valor-$p \leq \alpha = 0.05$, la hipótesis nula de que las medias de las tres poblaciones sean iguales es rechazada

b) $$\text{LSD} = t_{\alpha/2}\sqrt{\text{CME}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$
$$= t_{0.025}\sqrt{5.5\left(\frac{1}{5} + \frac{1}{5}\right)}$$
$$= 2.179\sqrt{2.2} = 3.23$$

$|\bar{x}_1 - \bar{x}_2| = |30 - 45| = 15 >$ LSD; diferencia significativa
$|\bar{x}_1 - \bar{x}_3| = |30 - 36| = 6 >$ LSD; diferencia significativa
$|\bar{x}_2 - \bar{x}_3| = |45 - 36| = 9 >$ LSD; diferencia significativa

c) $\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}\sqrt{\text{CME}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

$(30 - 45) \pm 2.179\sqrt{5.5\left(\frac{1}{5} + \frac{1}{5}\right)}$

$-15 \pm 3.23 = -18.23$ a -11.77

14. *a*) Significativa; el valor-$p = 0.0106$

b) LSD = 15.34

1 y 2; significativa

1 y 3; no significativa

2 y 3; significativa

15. *a*)

	Fabricante 1	Fabricante 2	Fabricante 3
Media muestral	23	28	21
Varianza muestral	6.67	4.67	3.33

$$\bar{\bar{x}} = (23 + 28 + 21)/3 = 24$$

$$\text{SCTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2$$
$$= 4(23 - 24)^2 + 4(28 - 24)^2 + 4(21 - 24)^2$$
$$= 104$$

$$\text{CMTR} = \frac{\text{SCTR}}{k - 1} = \frac{104}{2} = 52$$

$$\text{SCE} = \sum_{j=1}^{k} (n_j - 1)s_j^2$$
$$= 3(6.67) + 3(4.67) + 3(3.33) = 44.01$$

$$\text{CME} = \frac{\text{SCE}}{n_T - k} = \frac{44.01}{12 - 3} = 4.89$$

$$F = \frac{\text{CMTR}}{\text{CME}} = \frac{52}{4.89} = 10.63$$

Al utilizar la tabla *F* (2 grados de libertad en el numerador y 9 en el denominador), el valor-*p* es menor de 0.01

Utilizando Excel o Minitab, el valor-*p* correspondiente a $F = 10.63$ es 0.0043

Debido a que el valor-$p \leq \alpha = 0.05$, es rechazada la hipótesis nula de que la media de tiempo necesario para mezclar un lote de material sea el mismo para cada fabricante

b) $\text{LSD} = t_{\alpha/2}\sqrt{\text{CME}\left(\frac{1}{n_1} + \frac{1}{n_3}\right)}$

$= t_{0.025}\sqrt{4.89\left(\frac{1}{4} + \frac{1}{4}\right)}$

$= 2.262\sqrt{2.45} = 3.54$

Puesto que $|\bar{x}_1 - \bar{x}_3| = |23 - 21| = 2 < 3.54$, no parece haber una diferencia significativa entre las medias del fabricante 1 y el fabricante 3

16. $\bar{x}_1 - \bar{x}_2 \pm \text{LSD}$

$23 - 28 \pm 3.54$

$-5 \pm 3.54 = -8.54$ a -1.46

18. *a*) Significativa; el valor-$p = 0.0000$

b) Significativa; $2.3 > \text{LSD} = 1.19$

20. a) Significativa; el valor-$p = 0.011$

b) Comparación entre norte y sur

$|7\,702 - 5\,566| = 2\,136 > \text{LSD} = 1\,620.76$

diferencia significativa

Comparación entre norte y oeste

$|7\,702 - 8\,430| = 728 > \text{LSD} = 1\,620.76$

diferencia no significativa

Comparación entre sur y oeste

$|5\,566 - 8\,430| = 2\,864 > \text{LSD} = 1\,775.45$

diferencia significativa

21. *Medias de tratamiento*

$\bar{x}_{\cdot 1} = 13.6, \quad \bar{x}_{\cdot 2} = 11.0, \quad \bar{x}_{\cdot 3} = 10.6$

Medias del bloque

$\bar{x}_{1\cdot} = 9, \ \bar{x}_{2\cdot} = 7.67, \ \bar{x}_{3\cdot} = 15.67, \ \bar{x}_{4\cdot} = 18.67, \ \bar{x}_{5\cdot} = 7.67$

Otras medias

$\bar{\bar{x}} = 176/15 = 11.73$

Paso 1

$$\text{STC} = \sum_i \sum_j (x_{ij} - \bar{\bar{x}})^2$$
$$= (10 - 11.73)^2 + (9 - 11.73)^2 + \cdots + (8 - 11.73)^2$$
$$= 354.93$$

Paso 2

$$\text{SCTR} = b\sum_j (\bar{x}_{\cdot j} - \bar{\bar{x}})^2$$
$$= 5[(13.6 - 11.73)^2 + (11.0 - 11.73)^2 + (10.6 - 11.73)^2] = 26.53$$

Paso 3

$$\text{SCBL} = k\sum_j (\bar{x}_{\cdot j} - \bar{\bar{x}})^2$$
$$= 3[(9 - 11.73)^2 + (7.67 - 11.73)^2 + (15.67 - 11.73)^2 + (18.67 - 11.73)^2 + (7.67 - 11.73)^2] = 312.32$$

Paso 4

$$\text{SCE} = \text{STC} - \text{SCTR} - \text{SCBL}$$
$$= 354.93 - 26.53 - 312.32 = 16.08$$

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	*F*	Valor-*p*
Tratamientos	26.53	2	13.27	6.60	0.0203
Bloques	312.32	4	78.08		
Error	16.08	8	2.01		
Total	354.93	14			

De la tabla *F* (2 grados de libertad en el numerador y 8 en el denominador), el valor-*p* está entre 0.01 y 0.025

El valor-*p* real = 0.0203

Debido a que el valor-$p \leq \alpha = 0.05$, la hipótesis nula de que las medias de los tres tratamientos son iguales es rechazada

22.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	valor-p
Tratamientos	310	4	77.5	17.69	0.0005
Bloques	85	2	42.5		
Error	35	8	4.38		
Total	430	14			

Significativa; el valor-$p \leq \alpha = 0.05$

24. El valor-$p = 0.0453$

Debido a que el valor-$p \leq \alpha = 0.05$, la hipótesis nula de que el tiempo promedio para afinar un motor sea el mismo para los dos analizadores es rechazada

26. *a*) Significativa: el valor-$p = 0.0231$

b) En la sección de redacción

28. *Paso 1*

$$\text{STC} = \sum_i \sum_j \sum_k (x_{ijk} - \bar{\bar{x}})^2$$
$$= (135 - 111)^2 + (165 - 111)^2 + \cdots + (136 - 111)^2 = 9028$$

Paso 2

$$\text{SCA} = br \sum_i (\bar{x}_{i\cdot} - \bar{\bar{x}})^2$$
$$= 3(2)[(104 - 111)^2 + (118 - 111)^2] = 588$$

Paso 3

$$\text{SCB} = ar \sum_j (\bar{x}_{\cdot j} - \bar{\bar{x}})^2$$
$$= 2(2)[(130 - 111)^2 + (97 - 111)^2 + (106 - 111)^2] = 2328$$

Paso 4

$$\text{SCAB} = r \sum_i \sum_j (\bar{x}_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})^2$$
$$= 2[(150 - 104 - 130 + 111)^2 + (78 - 104 - 97 + 111)^2 + \cdots + (128 - 118 - 106 + 111)^2] = 4392$$

Paso 5

$$\text{SCE} = \text{STC} - \text{SCA} - \text{SCB} - \text{SCAB}$$
$$= 9028 - 588 - 2328 - 4392 = 1720$$

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	valor-p
Factor A	588	1	588	2.05	0.2022
Factor B	2328	2	1164	4.06	0.0767
Interacción	4392	2	2196	7.66	0.0223
Error	1720	6	286.67		
Total	9028	11			

El factor A: $F = 2.05$

Utilizando la tabla F (1 grado de libertad en el numerador y 6 en el denominador), el valor-p es mayor de 0.10

Utilizando Excel o Minitab, el valor-p correspondiente a $F = 2.05$ es 0.2022

Debido a que el valor-$p > \alpha = 0.05$, el factor A no es significativo

El factor B: $F = 4.06$

Utilizando la tabla F (2 grados de libertad en el numerador y 6 en el denominador), el valor-p está entre 0.05 y 0.10

Utilizando Excel o Minitab, el valor-p correspondiente a $F = 4.06$ es 0.0767

Debido a que el valor-$p > \alpha = 0.05$, el factor B no es significativo

La interacción: $F = 7.66$

Utilizando la tabla F (2 grados de libertad en el numerador y 6 en el denominador), el valor-p está entre 0.01 y 0.025

Utilizando Excel o Minitab, el valor-p correspondiente a $F = 7.66$ es 0.0223

Debido a que el valor-$p \leq \alpha = 0.05$, la interacción es significativa

30. Diseño: el valor-$p = 0.0104$; significativo
Tamaño: el valor-$p = 0.1340$; no es significativo
Interacción: el valor-$p = 0.2519$; no es significativo

32. Clase: el valor-$p = 0.0002$; significativo
Tipo: el valor-$p = 0.0006$; significativo
Interacción: el valor-$p = 0.4229$; no es significativo

34. Significativa; el valor-$p = 0.0134$

36. Significativa; el valor-$p = 0.046$

38. No es significativo; el valor-$p = 0.2455$

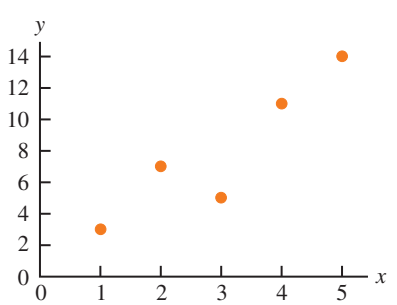
40. *a*) Significativa; el valor-$p = 0.0175$

42. Significativa; el valor-$p = 0.004$

44. El tipo de máquina (el valor-$p = 0.0226$) es significativo; el tipo de suministro (el valor-$p = 0.7913$) y la interacción (el valor-$p = 0.0671$) no son significativos

Capítulo 14

1. *a*)



b) Parece haber una relación lineal positiva entre x y y

c) Se pueden trazar muchas y distintas rectas a efecto de proporcionar una aproximación lineal para la relación entre x y y; en el inciso *d*) se determinará la ecuación de la recta que "mejor" represente la relación según el criterio de los mínimos cuadrados

d) Las sumas necesarias para calcular la pendiente y la intersección de y:

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{15}{5} = 3, \quad \bar{y} = \frac{\Sigma y_i}{n} = \frac{40}{5} = 8,$$
$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 26, \quad \Sigma(x_i - \bar{x})^2 = 10$$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{26}{10} = 2.6$$

$$b_0 = \bar{y} - b_1\bar{x} = 8 - (2.6)(3) = 0.2$$

$$\hat{y} = 0.2 + 2.6x$$

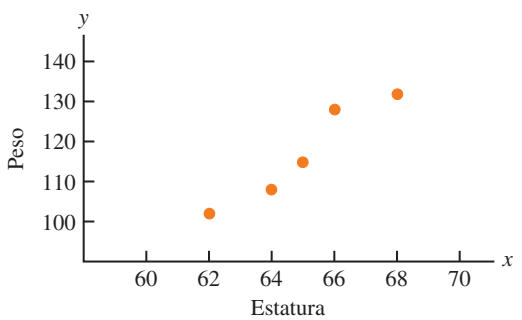
e) $\hat{y} = 0.2 + 2.6x = 0.2 + 2.6(4) = 10.6$

2. *b*) Parece existir una relación lineal negativa entre x y y

d) $\hat{y} = 68 - 3x$

e) 38

4. *a*)



b) Parece haber una relación lineal positiva entre x = estatura y y = peso

c) Muchas y distintas rectas pueden dibujarse para proporcionar una aproximación lineal de la relación entre estatura y peso; en el inciso *d*) se determinará la ecuación para la recta que "mejor" represente la relación con base en el criterio de los mínimos cuadrados

d) Las sumas necesarias para calcular la pendiente y la intersección de y:

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{325}{5} = 65, \quad \bar{y} = \frac{\Sigma y_i}{n} = \frac{585}{5} = 117,$$

$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 110, \quad \Sigma(x_i - \bar{x})^2 = 20$$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{110}{20} = 5.5$$

$$b_0 = \bar{y} - b_1\bar{x} = 117 - (5.5)(65) = -240.5$$

$$\hat{y} = -240.5 + 5.5x$$

e) $\hat{y} = -240.5 + 5.5(63) = 106$

El peso estimado es 106 libras

6. *c*) $\hat{y} = 8.9412 - 0.02633x$

e) 6.3 o aproximadamente \$6 300

8. *c*) $\hat{y} = 359.2668 - 5.2772x$

d) \$254

10. *c*) $\hat{y} = -6\,745.44 + 149.29x$

d) 4 003 o \$4 003 000

12. *c*) $\hat{y} = -8129.4439 + 22.4443x$

d) \$8 704

14. *c*) $\hat{y} = 37.1217 + 0.51758x$

d) 73

15. *a*) $\hat{y}_i = 0.2 + 2.6x_i$ y $\bar{y} = 8$

x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	3	2.8	0.2	0.04	−5	25
2	7	5.4	1.6	2.56	−1	1
3	5	8.0	−3.0	9.00	−3	9
4	11	10.6	0.4	0.16	3	9
5	14	13.2	0.8	0.64	6	36
				SCE = 12.40		STC = 80

SCR = STC − SCE = 80 − 12.4 = 67.6

b) $r^2 = \frac{\text{SCR}}{\text{STC}} = \frac{67.6}{80} = 0.845$

La recta de mínimos cuadrados proporciona un buen ajuste; 84.5% de la variabilidad en y ha sido explicado por la recta de mínimos cuadrados

c) $r_{xy} = \sqrt{0.845} = +0.9192$

16. *a*) SCE = 230, STC = 1 850, SCR = 1 620

b) $r^2 = 0.876$

c) $r_{xy} = -0.936$

18. *a*) Ecuación de regresión estimada y media para la variable dependiente:

$\hat{y} = 1\,790.5 + 581.1x, \quad \bar{y} = 3\,650$

La suma de cuadrados debido al error y la suma total de cuadrados:

$\text{SCE} = \Sigma(y_i - \hat{y}_i)^2 = 85\,135.14$

$\text{STC} = \Sigma(y_i - \bar{y})^2 = 335\,000$

Por tanto, SCR = STC − SCE

= 335 000 − 85 135.14 = 249 864.86

b) $r^2 = \frac{\text{SCR}}{\text{STC}} = \frac{249\,864.86}{335{,}000} = 0.746$

La recta de mínimos cuadrados representa 74.6% de la suma de cuadrados total

c) $r_{xy} = \sqrt{0.746} = +0.8637$

20. *a*) $\hat{y} = 12.0169 + 0.0127x$

b) $r^2 = 0.4503$

c) 53

22. *a*) 0.77

b) Sí

c) $r_{xy} = +0.88$, alta

23. *a*) $s^2 = \text{CME} = \frac{\text{SCE}}{n-2} = \frac{12.4}{3} = 4.133$

b) $s = \sqrt{\text{CME}} = \sqrt{4.133} = 2.033$

c) $\Sigma(x_i - \bar{x})^2 = 10$

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} = \frac{2.033}{\sqrt{10}} = 0.643$$

d) $t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{2.6 - 0}{0.643} = 4.044$

En la tabla t (3 grados de libertad), el área en la cola está entre 0.01 y 0.025

El valor-p está entre 0.02 y 0.05

Usando Excel o Minitab, el valor-p que corresponde a $t = 4.04$ es 0.0272

Debido a que el valor-$p \leq \alpha$, H_0 es rechazada: $\beta_1 = 0$

e) $\text{CMR} = \dfrac{\text{SCR}}{1} = 67.6$

$$F = \frac{\text{CMR}}{\text{CME}} = \frac{67.6}{4.133} = 16.36$$

De la tabla F (1 grado de libertad en el numerador y 3 en el denominador), el valor-p está entre 0.025 y 0.05

Utilizando Excel o Minitab, el valor-p correspondiente a $F = 16.36$ es .0272

Debido a que el valor-$p \le \alpha$, H_0 es rechazada: $\beta_1 = 0$

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	valor-p
Regresión	67.6	1	67.6	16.36	0.0272
Error	12.4	3	4.133		
Total	80	4			

24. *a*) 76.6667
b) 8.7560
c) 0.6526
d) Significativa: el valor-$p = 0.0193$
e) Significativa: el valor-$p = 0.0193$

26. *a*) $s^2 = \text{CME} = \dfrac{\text{SCE}}{n-2} = \dfrac{85{,}135.14}{4} = 21\,283.79$

$s = \sqrt{\text{CME}} = \sqrt{21\,283.79} = 145.89$

$\Sigma(x_i - \bar{x})^2 = 0.74$

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} = \frac{145.89}{\sqrt{0.74}} = 169.59$$

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{581.08 - 0}{169.59} = 3.43$$

De la tabla t (4 grados de libertad), el área en la cola está entre 0.01 y 0.025

El valor-p está entre 0.02 y 0.05

Utilizando Excel o Minitab, el valor-p correspondiente a $t = 3.43$ es 0.0266

Debido a que el valor-$p \le \alpha$, H_0 es rechazada: $\beta_1 = 0$

b) $\text{CMR} = \dfrac{\text{SCR}}{1} = \dfrac{249\,864.86}{1} = 249\,864.86$

$$F = \frac{\text{CMR}}{\text{CME}} = \frac{249\,864.86}{21\,283.79} = 11.74$$

De la tabla F (1 grado de libertad en el numerador y 4 en el denominador), el valor-p está entre 0.025 y 0.05

Utilizando Excel o Minitab, el valor-p correspondiente a $F = 11.74$ es 0.0266

Debido a que el valor-$p \le \alpha$, H_0 es rechazada: $\beta_1 = 0$

c)

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	valor-p
Regresión	249 864.86	1	249 864.86	11.74	0.0266
Error	85 135.14	4	21 283.79		
Total	335 000	5			

28. Están relacionados; el valor-$p = 0.000$

30. Significativa; el valor-$p = 0.0042$

32. *a*) $s = 2.033$

$\bar{x} = 3$, $\Sigma(x_i - \bar{x})^2 = 10$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

$$= 2.033\sqrt{\frac{1}{5} + \frac{(4-3)^2}{10}} = 1.11$$

b) $\hat{y} = 0.2 + 2.6x = 0.2 + 2.6(4) = 10.6$

$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p}$

$10.6 \pm 3.182(1.11)$

10.6 ± 3.53, o 7.07 a 14.13

c) $$s_{\text{ind}} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

$$= 2.033\sqrt{1 + \frac{1}{5} + \frac{(4-3)^2}{10}} = 2.32$$

d) $\hat{y}_p \pm t_{\alpha/2} s_{\text{ind}}$

$10.6 \pm 3.182(2.32)$

10.6 ± 7.38, o 3.22 a 17.98

34. El intervalo de confianza: 8.65 a 21.15
Intervalo de pronóstico: −4.50 a 41.30

35. *a*) $s = 145.89$, $\bar{x} = 3.2$, $\Sigma(x_i - \bar{x})^2 = 0.74$

$\hat{y} = 1\,790.5 + 581.1x = 1\,790.5 + 581.1(3)$
$= 3\,533.8$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

$$= 145.89\sqrt{\frac{1}{6} + \frac{(3 - 3.2)^2}{0.74}} = 68.54$$

$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p}$

$3\,533.8 \pm 2.776(68.54)$

$3\,533.8 \pm 190.27$ o \$3 343.53 a \$3 724.07

b) $$s_{\text{ind}} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

$$= 145.89\sqrt{1 + \frac{1}{6} + \frac{(3 - 3.2)^2}{0.74}} = 161.19$$

$\hat{y}_p \pm t_{\alpha/2} s_{\text{ind}}$

$3\,533.8 \pm 2.776(161.19)$

$3\,533.8 \pm 447.46$ o \$3 086.34 a \$3 981.26

36. *a*) \$201
b) 167.25 a 234.65
c) 108.75 a 293.15

38. *a*) \$5 046.67
b) \$3 815.10 a \$6278.24
c) No está fuera de línea

40. *a*) 9
b) $\hat{y} = 20.0 + 7.21x$
c) 1.3626
d) $\text{SCE} = \text{STC} - \text{SCR} = 51\,984.1 - 41\,587.3 = 10\,396.8$

$\text{CME} = 10\,396.8/7 = 1\,485.3$

$$F = \frac{\text{CMR}}{\text{CME}} = \frac{41\,587.3}{1\,485.3} = 28.0$$

De la tabla F (1 grado de libertad en el numerador y 7 en el denominador), el valor-p es menor de 0.01

Utilizando Excel y Minitab, el valor-p que corresponde a $F = 28.0$ es 0.0011

Debido a que el valor-$p \leq \alpha = 0.05$, H_0: es rechazada $\beta_1 = 0$

e) $\hat{y} = 20.0 + 7.21(50) = 380.5$, o \$380 500

42. *a*) $\hat{y} = 80.0 + 50.0x$

b) 30

c) Significativo: el valor-$p = 0.000$

d) \$680 000

44. *b*) Sí

c) $\hat{y} = 2044.38 - 28.35$ peso

d) Significativo; el valor-$p = 0.000$

e) 0.774; un buen ajuste

45. *a*) $\bar{x} = \dfrac{\Sigma x_i}{n} = \dfrac{70}{5} = 14,\ \bar{y} = \dfrac{\Sigma y_i}{n} = \dfrac{76}{5} = 15.2,$

$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 200,\ \Sigma(x_i - \bar{x})^2 = 126$

$b_1 = \dfrac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \dfrac{200}{126} = 1.5873$

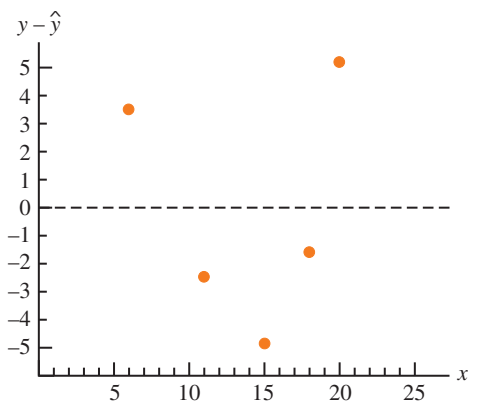
$b_0 = \bar{y} - b_1\bar{x} = 152 - (1.5873)(14) = -7.0222$

$\hat{y} = -7.02 + 1.59x$

b)

x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
6	6	2.52	3.48
11	8	10.47	−2.47
15	12	16.83	−4.83
18	20	21.60	−1.60
20	30	24.78	5.22

c)



Con sólo cinco observaciones es difícil determinar si los supuestos son satisfechos sin embargo, la gráfica sugiere una curvatura en los residuales, lo que indicaría que el supuesto término del error no está satisfecho; el diagrama de dispersión para estos datos también indica que la relación subyacente entre x y y puede ser curvilínea

d) $s^2 = 23.78$

$h_i = \dfrac{1}{n} + \dfrac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}$

$= \dfrac{1}{5} + \dfrac{(x_i - 14)^2}{126}$

x_i	h_i	$s_{y_i - \hat{y}_i}$	$y_i - \hat{y}_i$	**Residuales estandarizados**
6	0.7079	2.64	3.48	1.32
11	0.2714	4.16	−2.47	−0.59
15	0.2079	4.34	−4.83	−1.11
18	0.3270	4.00	−1.60	−0.40
20	0.4857	3.50	5.22	1.49

e) La gráfica de residuales estandarizados contra $\hat{y}$ tiene la misma forma de la gráfica de residuales original: como se indica en el inciso *c*), la curvatura observada muestra que los supuestos sobre el término del error no puede ser satisfecha

46. *a*) $\hat{y} = 2.32 + 0.64x$

b) No; la varianza parece aumentar para los valores más grandes de x

47. *a*) Sea $x =$ los gastos de publicidad y $y =$ los ingresos

$\hat{y} = 29.4 + 1.55x$

b) STC = 1002, SCE = 310.28, SCR = 691.72

$\text{CMR} = \dfrac{\text{SCR}}{1} = 691.72$

$\text{CME} = \dfrac{\text{SCE}}{n - 2} = \dfrac{310.28}{5} = 62.0554$

$F = \dfrac{\text{CMR}}{\text{CME}} = \dfrac{691.72}{62.0554} = 11.15$

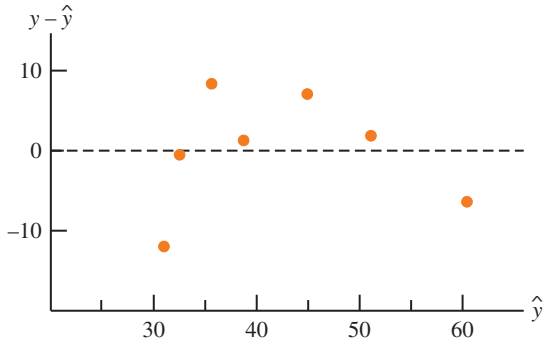
De la tabla F (1 grado de libertad en el numerador y 5 en el denominador), el valor-p está entre 0.01 y 0.025

Utilizando Excel o Minitab, el valor-$p = 0.0206$

Debido a que el valor-$p \leq \alpha = 0.05$, se concluye que las dos variables están relacionadas

c)

x_i	y_i	$\hat{y}_i = 29.40 + 1.55x_i$	$y_i - \hat{y}_i$
1	19	30.95	−11.95
2	32	32.50	−0.50
4	44	35.60	8.40
6	40	38.70	1.30
10	52	44.90	7.10
14	53	51.10	1.90
20	54	60.40	−6.40



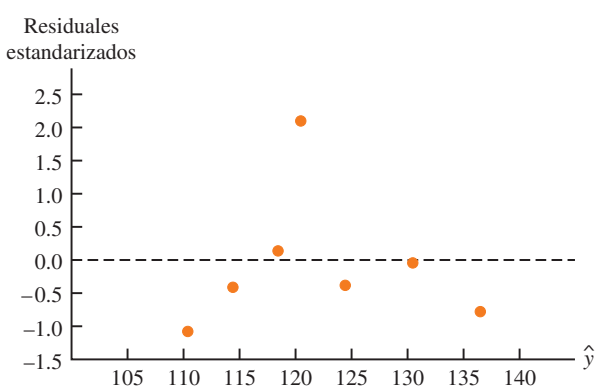
d) La gráfica de residuales lleva a cuestionar la hipótesis de una relación lineal entre x y y; a pesar de que la relación es significativa en el nivel $\alpha = 0.05$, sería extremadamente riesgoso extrapolarla más allá del alcance de los datos

48. *b*) Sí

50. *a*) Utilizando Minitab, se obtuvo la ecuación de regresión estimada $\hat{y} = 66.1 + 0.402x$; una parte de los resultados de Minitab se muestra en la figura D14.50; los valores ajustados y los residuales estandarizados son los siguientes:

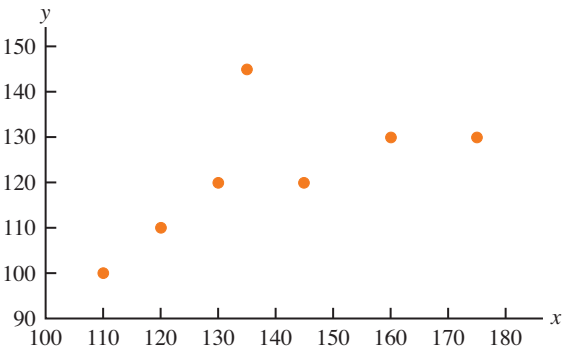
x_i	y_i	$\hat{y}_i$	Residuales estandarizados
135	145	120.41	2.11
110	100	110.35	−1.08
130	120	118.40	0.14
145	120	124.43	−0.38
175	130	136.50	−0.78
160	130	130.47	−0.04
120	110	114.38	−0.41

b)



La gráfica de los residuales estandarizados indica que la observación $x = 135$, $y = 145$ puede ser de un valor atípico; note que esta observación tiene un residual estandarizado de 2.11

c) El diagrama de dispersión es el siguiente:



El diagrama de dispersión también indica que la observación $x = 135$, $y = 145$ puede ser un valor atípico; la implicación es que se pueda identificar para los valores extremos de la regresión lineal simple observando el diagrama

52. *a*) Una parte del resultado de Minitab se muestra en la figura D14.52
b) Minitab identifica la observación 1 al tener un residual estandarizado mayor; por tanto, se considerará la observación 1 como valor atípico

54. *b*) El valor = −252 + ingreso de 5.83
c) Existen cinco observaciones inusuales (9, 19, 21, 22 y 32).

58. *a*) $\hat{y} = 9.26 + 0.711x$
b) Significativo; el valor-$p = 0.001$
c) $r^2 = 0.744$; un buen ajuste
d) \$13.53

60. *b*) GR(%) = 25.4 + 0.285 RR(%)
c) Significativo: el valor-$p = 0.000$
d) No; $r^2 = 0.449$
e) Sí
f) Sí

62. *a*) $\hat{y} = 22.2 - 0.148x$
b) Relación significativa: el valor-$p = 0.028$
c) Buen ajuste; $r^2 = 0.739$
d) 12.294 a 17.271

64. *a*) $\hat{y} = 220 + 132x$
b) Significativo: el valor-$p = 0.000$
c) $r^2 = 0.873$; muy buen ajuste
d) \$559.50 a \$933.90

66. *a*) El mercado beta = 0.95
b) Significativo: el valor-$p = 0.029$
c) $r^2 = 0.470$; sin buen ajuste
d) Xerox tiene un mayor riesgo

68. *b*) Parece existir una relación lineal positiva entre las dos variables
c) $\hat{y} = 9.37 + 1.2875$ cinco mejores (%)
d) Significativo: el valor-$p = 0.000$
e) $r^2 = 0.741$; un buen ajuste
f) $r_{xy} = 0.86$

FIGURA D14.50

```
The regression equation is
Y = 66.1 + 0.402 X

Predictor      Coef    SE Coef       T        p
Constant      66.10      32.06    2.06    0.094
X            0.4023     0.2276    1.77    0.137

S = 12.62     R-sq = 38.5%     R-sq(adj) = 26.1%

Analysis of Variance

SOURCE            DF        SS       MS       F        p
Regression         1     497.2    497.2    3.12    0.137
Residual Error     5     795.7    159.1
Total              6    1292.9

Unusual Observations
Obs      X         Y       Fit    SE Fit    Residual    St Resid
  1    135    145.00    120.42      4.87       24.58        2.11R

  R denotes an observation with a large standardized residual
```

FIGURA D14.52

```
The regression equation is
Shipment = 4.09 + 0.196 Media$

Predictor       Coef    SE Coef       T        p
Constant       4.089      2.168    1.89    0.096
Media$       0.19552    0.03635    5.38    0.000

S = 5.044     R-Sq = 78.3%     R-Sq(adj) = 75.6%

Analysis of Variance
Source            DF        SS        MS        F        p
Regression         1    735.84    735.84    28.93    0.000
Residual Error     8    203.51     25.44
Total              9    939.35

Unusual Observations
Obs    Media$    Shipment      Fit    SE Fit    Residual    St Resid
  1       120       36.30    27.55      3.30        8.75        2.30R

R denotes an observation with a large standardized residual
```

Capítulo 15

2. *a*) La ecuación de regresión estimada es
$\hat{y} = 45.06 + 1.94x_1$
La estimación de y cuando $x_1 = 45$ es
$\hat{y} = 45.06 + 1.94(45) = 132.36$

b) La ecuación de regresión estimada es
$\hat{y} = 85.22 + 4.32x_2$
La estimación de y cuando $x_2 = 15$ es
$\hat{y} = 85.22 + 4.32(15) = 150.02$

c) La ecuación de regresión estimada es
$\hat{y} = -18.37 + 2.01x_1 + 4.74x_2$
La estimación de y cuando $x_1 = 45$ y $x_2 = 15$ es
$\hat{y} = -18.37 + 2.01(45) + 4.74(15) = 143.18$

4. *a*) \$255 000

5. *a*) El resultado de Minitab se muestra en la figura D15.5*a*
b) El resultado de Minitab se muestra en la figura D15.5*b*
c) Es de 1.60 en el inciso *a*) y de 2.29 en el inciso *b*). En el inciso *a*) el coeficiente es una estimación de la

FIGURA D15.5*a*

```
The regression equation is
Revenue = 88.6 + 1.60 TVAdv

Predictor       Coef    SE Coef        T        p
Constant      88.638      1.582    56.02    0.000
TVAdv         1.6039     0.4778     3.36    0.015

S = 1.215     R-sq = 65.3%     R-sq(adj) = 59.5%

Analysis of Variance

SOURCE            DF        SS        MS        F        p
Regression         1    16.640    16.640    11.27    0.015
Residual Error     6     8.860     1.477
Total              7    25.500
```

FIGURA D15.5*b*

```
The regression equation is
Revenue = 83.2 + 2.29 TVAdv + 1.30 NewsAdv

Predictor       Coef    SE Coef        T        p
Constant      83.230      1.574    52.88    0.000
TVAdv         2.2902     0.3041     7.53    0.001
NewsAdv       1.3010     0.3207     4.06    0.010

S = 0.6426     R-sq = 91.9%     R-sq(adj) = 88.7%

Analysis of Variance

SOURCE            DF        SS        MS        F        p
Regression         2    23.435    11.718    28.38    0.002
Residual Error     5     2.065     0.413
Total              7    25.500
```

variación de los ingresos debido a una variación de una unidad en los gastos de publicidad en televisión, y en el inciso *b*) ésta representa una estimación de la variación de los ingresos debido a una variación de una unidad en los gastos de publicidad por televisión, cuando la cantidad de publicidad en los periódicos se mantiene constante

d) La utilidad = 83.2 + 2.29(3.5) + 1.30(1.8) = 93.56 o $93 560

6. *a*) La proporción ganada = 0.354 + 0.000888 HR
b) La proporción ganada = 0.865 − 0.0837 ERA
c) La proporción ganada = 0.709 + 0.00140 HR − 0.103 ERA

8. *a*) $\hat{y}$ = 31054 + 1328.7 de exactitud
b) $\hat{y}$ = 21313 + 136.69 de calificación − 1446.3 de exactitud
c) $26 643

10. *a*) PCT = −1.22 + 3.96 FG%
b) Un incremento de 1% en FG% aumentará 0.04 el PCT
c) PCT = −1.23 + 4.82 FG% − 2.59 Opp 3 Pt% + 0.0344 Opp TO
d) Aumenta FG%; disminuye Opp 3 Pt%; aumenta Opp TO
e) 0.638

12. *a*) $R^2 = \dfrac{\text{SCR}}{\text{STC}} = \dfrac{14\,052.2}{15\,182.9} = 0.926$

b) $R_a^2 = 1 - (1 - R^2)\dfrac{n-1}{n-p-1}$

$= 1 - (1 - 0.926)\dfrac{10-1}{10-2-1} = 0.905$

c) Sí; después de ajustar el número de las variables independientes en el modelo, se observa que 90.5% de la variabilidad en *y* ha sido tomada en cuenta

14. *a*) 0.75 b) 0.68

15. *a*) $R^2 = \dfrac{\text{SCR}}{\text{STC}} = \dfrac{23.435}{25.5} = 0.919$

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

$$= 1 - (1 - 0.919)\frac{8-1}{8-2-1} = 0.887$$

b) Se prefiere el análisis de regresión múltiple debido a que tanto R^2 como R_a^2 muestran un aumento en el porcentaje de variabilidad explicada de y cuando las dos variables independientes son utilizadas

16. *a*) No, $R^2 = 0.153$

b) Un mejor ajuste con regresión múltiple

18. *a*) $R^2 = 0.564$; $R_a^2 = 0.511$

b) El ajuste no es muy bueno

19. *a*) $\text{CMR} = \dfrac{\text{SCR}}{p} = \dfrac{6\,216.375}{2} = 3\,108.188$

$$\text{CME} = \frac{\text{SCE}}{n-p-1} = \frac{507.75}{10-2-1} = 72.536$$

b) $F = \dfrac{\text{CMR}}{\text{CME}} = \dfrac{3\,108.188}{72.536} = 42.85$

De la tabla F (2 grados de libertad en el numerador y 7 en el denominador), el valor-p es menor de 0.01

Utilizando Excel o Minitab, el valor-p correspondiente a $F = 42.85$ es 0.0001

Debido a que el valor-$p \le \alpha$, el modelo general es significativo

c) $t = \dfrac{b_1}{s_{b_1}} = \dfrac{0.5906}{0.0813} = 7.26$

El valor-$p = 0.0002$

Debido a que el valor-$p \le \alpha$, β_1 es significativo

d) $t = \dfrac{b_2}{s_{b_2}} = \dfrac{0.4980}{0.0567} = 8.78$

El valor-$p = 0.0001$

Debido a que el valor-$p \le \alpha$, β_2 es significativo

20. *a*) Significativo; el valor-$p = 0.000$

b) Significativo; el valor-$p = 0.000$

c) Significativo; el valor-$p = 0.002$

22. *a*) $\text{SCE} = 4\,000$; $s^2 = 571.43$
$\text{CMR} = 6\,000$

b) Significativo; el valor-$p = 0.008$

23. *a*) $F = 28.38$

El valor-$p = 0.002$

Debido a que el valor-$p \le \alpha$, existe una relación significativa

b) $t = 7.53$

El valor-$p = 0.001$

Debido a que el valor-$p \le \alpha$, β_1 es significativo y x_1 no debe ser eliminada del modelo

c) $t = 4.06$

El valor-$p = 0.010$

Debido a que el valor-$p \le \alpha$, β_2 es significativa y x_2 no debe ser eliminada del modelo

24. *a*) $\hat{y} = -0.682 + 0.0498$ ingresos $+ 0.0147$ % Victorias

b) Significativa; el valor-$p = 0.001$

c) El ingreso es significativo; el valor-$p = 0.001$
El % de victorias es significativo; el valor-$p = 0.025$

26. *a*) Significativa; el valor-$p = 0.000$

b) Todas son significativas: los valores-p son todos $< \alpha = 0.05$

28. *a*) Utilizando Minitab, 95% del intervalo de confianza es 132.16 a 154.16

b) Utilizando Minitab, 95% del intervalo de predicción es de 111.13 a 175.18

29. *a*) Observe los resultados de Minitab en la figura D15.5*b*)

$\hat{y} = 83.23 + 2.29(3.5) + 1.30(1.8) = 93.555$ o \$93 555

b) Los resultados de Minitab: 92.840 a 94.335, o \$92 840 a \$94 335

c) Los resultados de Minitab: 91.774 a 95.401, o \$91 774 a \$95 401

30. *a*) 46.758 a 50.646

b) 44.815 a 52.589

32. *a*) $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

donde $x_2 = \begin{cases} 0 \text{ si el nivel es 1} \\ 1 \text{ si el nivel es 2} \end{cases}$

b) $E(y) = \beta_0 + \beta_1 x_1 + \beta_2(0) = \beta_0 + \beta_1 x_1$

c) $E(y) = \beta_0 + \beta_1 x_1 + \beta_2(1) = \beta_0 + \beta_1 x_1 + \beta_2$

d) $\beta_2 = E(y \mid \text{nivel 2}) - E(y \mid \text{nivel 1})$
β_1 es la variación en $E(y)$ por una variación de 1 unidad en x_1 manteniendo x_2 constante

34. *a*) \$15 300

b) $\hat{y} = 10.1 - 4.2(2) + 6.8(8) + 15.3(0) = 56.1$
La predicción de ventas: \$56 100

c) $\hat{y} = 10.1 - 4.2(1) + 6.8(3) + 15.3(1) = 41.6$
La predicción de ventas: \$41 600

36. *a*) $\hat{y} = 1.86 + 0.291$ Meses $+ 1.10$ Tipo $- 0.609$ personas

b) Significativo; el valor-$p = 0.002$

c) La persona no es significativa; el valor-$p = 0.167$

38. *a*) $\hat{y} = -91.8 + 1.08$ Edad $+ 0.252$ Presión $+ 8.74$ Fumador

b) Significativo: el valor-$p = 0.01$

c) El 95% del intervalo de predicción es de 21.35 a 47.18, o una probabilidad de 0.2135 a 0.4718; dejar de fumar y comenzar algún tratamiento para reducir la presión arterial

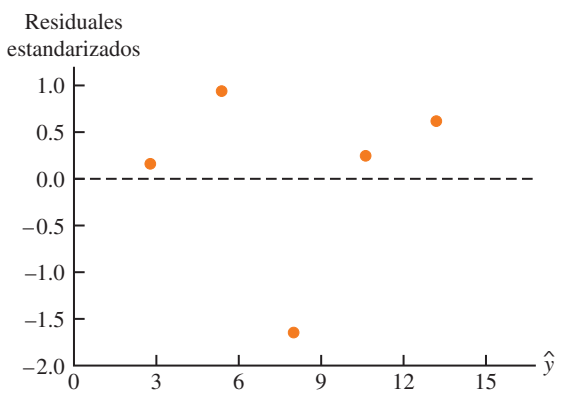
39. *a*) Los resultados de Minitab se muestran en la figura D15.39

b) Minitab proporciona los siguientes valores:

x_i	y_i	$\hat{y}_i$	**Residual estandarizado**
1	3	2.8	0.16
2	7	5.4	0.94
3	5	8.0	−1.65
4	11	10.6	0.24
5	14	13.2	0.62

FIGURA D15.39

```
The regression equation is
Y = 0.20 + 2.60 X

Predictor      Coef    SE Coef       T        p
Constant      0.200      2.132    0.09    0.931
X            2.6000     0.6429    4.04    0.027

S = 2.033     R-sq = 84.5%     R-sq(adj) = 79.3%

Analysis of Variance
SOURCE            DF        SS        MS        F        p
Regression         1    67.600    67.600    16.35    0.027
Residual Error     3    12.400     4.133
Total              4    80.000
```



El punto (3,5) no parece seguir la tendencia de los datos restantes; sin embargo, el valor del residual estandarizado para este punto, −1.65, no es lo suficientemente mayor para concluir que (3.5) es un valor atípico

c) Minitab proporciona los siguientes valores:

x_i	y_i	Eliminación de residuales estudentizados
1	3	0.13
2	7	0.91
3	5	−4.42
4	11	0.19
5	14	0.54

$t_{0.025} = 4.303$ ($n - p - 2 = 5 - 1 - 2 = 2$ grados de libertad)

Debido a que la eliminación de residuales estudentizados para (3,5) es $-4.42 < -4.303$, se concluye que la tercera observación es un valor atípico

40. *a*) $\hat{y} = -53.3 + 3.11x$

b) −1.94, −0.12, 1.79, 0.40, −1.90; no

c) 0.38, 0.28, 0.22, 0.20, 0.92; no

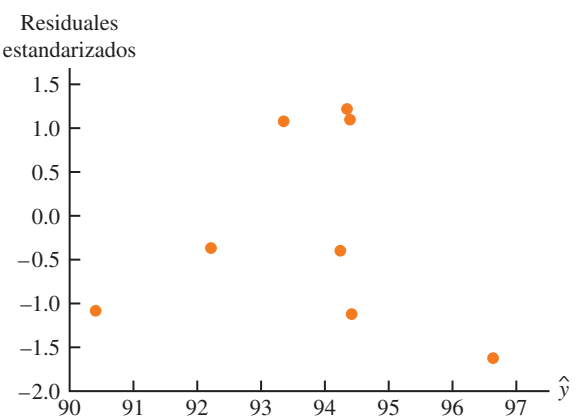
d) 0.60, 0.00, 0.26, 0.03, 11.09; sí, la quinta observación

41. *a*) El resultado de Minitab aparece en la figura D15.5b; la ecuación de regresión estimada es

Ingreso = 83.2 + 2.29 TVAdv + 1.30 Anuncios en periódico

b) Minitab proporciona los siguientes valores:

$\hat{y}_i$	Residual estandarizado	$\hat{y}_i$	Residual estandarizado
96.63	−1.62	94.39	1.10
90.41	−1.08	94.24	−0.40
94.34	1.22	94.42	−1.12
92.21	−0.37	93.35	1.08



En relación con algunas observaciones, es difícil determinar si algunos de los supuestos con respecto a ϵ han sido violados. Por ejemplo, un argumento podría ser que no parece haber ningún patrón en la gráfica; alternativamente, otro argumento indicaría que no hay un patrón curvilíneo en la gráfica

c) Los valores de los residuales estandarizados son mayores de −2 y menores que +2; por tanto, al utilizar esta prueba, no hay valores atípicos

Como un examen posterior para los valores atípicos, se utilizó Minitab para calcular la siguiente eliminación de residuales estudentizados:

Observación	Eliminación de residuales estudentizados	Observación	Eliminación de residuales estudentizados
1	−2.11	5	1.13
2	−1.10	6	−0.36
3	1.31	7	−1.16
4	−0.33	8	1.10

$t_{0.025} = 2.776$ ($n - p - 2 = 8 - 2 - 2 = 4$ grados de libertad)

Debido a que ninguno de los residuales estudentizados eliminados son menor que −2.776 o mayor que a 2.776, se concluye que no hay valores atípicos en los datos

d) Minitab proporciona los siguientes valores:

Observación	h_i	D_i
1	0.63	1.52
2	0.65	0.70
3	0.30	0.22
4	0.23	0.01
5	0.26	0.14
6	0.14	0.01
7	0.66	0.81
8	0.13	0.06

La ventaja del valor crítico es

$$\frac{3(p+1)}{n} = \frac{3(2+1)}{8} = 1.125$$

Debido a que ninguno de los valores excedieron 1.125, se concluye que no hay observaciones influyentes. Sin embargo, al utilizar la medida de distancia de Cook, se observa que $D_1 > 1$ (regla práctica del valor crítico); por tanto, se concluye que la primera es una observación influyente

Conclusión final: la observación 1 es influyente

42. *b*) Tendencia inusual
c) No hay valores atípicos
d) La observación 2 es de tipo influyente

44. *a*) $E(y) = \dfrac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$

b) Estimación de la probabilidad de que efectúe una compra un cliente que no tenga una tarjeta de crédito de Simmons

c) $\hat{g}(x) = -0.9445 + 1.0245x$

d) 0.28 de los clientes que no tienen una tarjeta de crédito de Simmons
0.52 para los clientes que tienen una tarjeta de crédito de Simmons

e) Cociente de las posibilidades estimado = 2.79

46. *a*) $E(y) = \dfrac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$

b) $E(y) = \dfrac{e^{-2.6355+0.22018x}}{1 + e^{-2.6355+0.22018x}}$

c) Significativo; el valor-$p = 0.0002$
d) 0.39
e) $1 200
f) Cociente de posibilidades estimado = 1.25

48. *a*) $E(y) = \dfrac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$

b) $\hat{g}(x) = -2.805 + 1.1492x$

c) 0.86

d) Cociente de posibilidades estimado = 3.16

50. *b*) 67.39

52. *a*) $\hat{y} = -1.41 + 0.0235x_1 + 0.00486x_2$

b) Significativo; el valor-$p = 0.0001$

c) Los dos son significativos

d) $R^2 = 0.937$; $R^2_a = 9.19$; un buen ajuste

54. *a*) Comprar nuevamente = −7.522 + 1.8151 Llanta

b) Sí

c) Comprar nuevamente = −5.388 + 0.6899 Llanta + 0.9113 Desgaste

d) Significativo: el valor-$p = 0.001$

56. *a*) $\hat{y} = 4.9090 + 10.4658$ FundDE + 21.6823 FundIE

b) $R^2 = 0.6144$; reasonably good fit

c) $\hat{y} = 1.1899 + 6.8969$ FundDE + 17.6800 FundIE + 0.0265 Net Asset Value ($) + 6.4564 Expense Ratio (%)
El valor del activo neto ($) no es significativo y puede ser eliminado

d) $\hat{y} = -4.6074 + 8.1713$ FundDE + 19.5194 FundIE + 5.5197 Expense Ratio (%) + 5.9237 3StarRank + 8.2367 4StarRank + 6.6241 5StarRank

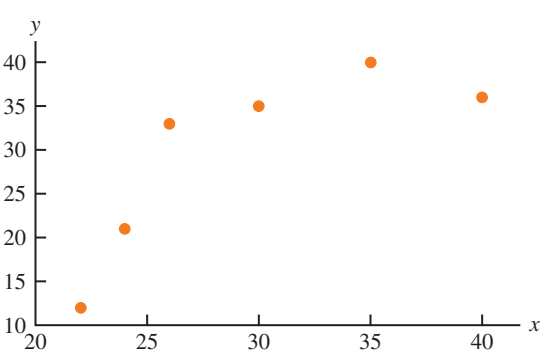
e) 15.28%

Capítulo 16

1. *a*) El resultado de Minitab se muestra en la figura D16.1*a*

b) Debido a que el valor-p correspondiente a $F = 6.85$ es $0.059 > \alpha = 0.05$, la relación no es significativa

c)



El diagrama de dispersión sugiere que puede ser apropiada una relación curvilínea

d) El resultado de Minitab se muestra en la figura D16.1*d*

e) Debido a que el valor-p correspondiente a $F = 25.68$ es $0.013 < \alpha = 0.05$, la relación es significativa

f) $\hat{y} = -168.88 + 12.187(25) - 0.17704(25)^2 = 25.145$

FIGURA D16.1*a*

```
The regression equation is
Y = - 6.8 + 1.23 X

Predictor      Coef    SE Coef        T        p
Constant      -6.77      14.17    -0.48    0.658
X            1.2296     0.4697     2.62    0.059

S = 7.269     R-sq = 63.1%     R-sq(adj) = 53.9%

Analysis of Variance

SOURCE            DF        SS        MS        F        p
Regression         1    362.13    362.13     6.85    0.059
Residual Error     4    211.37     52.84
Total              5    573.50
```

FIGURA D16.1*d*

```
The regression equation is
Y = - 169 + 12.2 X - 0.177 XSQ

Predictor        Coef    SE Coef        T        p
Constant      -168.88      39.79    -4.74    0.024
X              12.187      2.663     4.58    0.020
XSQ          -0.17704    0.04290    -4.13    0.026

S = 3.248     R-sq = 94.5%     R-sq(adj) = 90.8%

Analysis of Variance

SOURCE            DF        SS        MS        F        p
Regression         2    541.85    270.92    25.68    0.013
Residual Error     3     31.65     10.55
Total              5    573.50
```

2. *a*) $\hat{y} = 9.32 + 0.424x$; el valor- $p = 0.117$ indica una relación débil entre x y y
b) $\hat{y} = -8.10 + 2.41x - 0.0480x^2$
$R_a^2 = 0.932$; un buen ajuste
c) 20.965

4. *a*) $\hat{y} = 943 + 8.71x$
b) Significativo; el valor-$p = 0.005 < \alpha = 0.01$

5. *a*) El resultado de Minitab se muestra en la figura D16.5*a*
b) Debido a que el valor-p correspondiente a $F = 73.15$ es $0.003 < \alpha = 0.01$, la relación es significativa; H_0 es rechazada; $\beta_1 = \beta_2 = 0$
c) Vea la figura D16.5*c*

6. *b*) No, la relación parece ser curvilínea
c) Varios posibles modelos; por ejemplo
$\hat{y} = 2.90 - 0.185x + 0.00351x^2$

8. *a*) Parece que un modelo de regresión lineal simple no es apropiado
b) Price = 33 829 − 4 571 Rating + 154 RatingSq
c) logPrice = −10.2 + 10.4 logRating
d) Inciso *c*); se explica un mayor porcentaje de la variación

10. *a*) Significativo: valor-$p = 0.000$
b) Significativo: valor-$p = 0.000$

11. *a*) SCE = 1 805 − 1 760 = 45

$$F = \frac{\text{CMR}}{\text{CME}} = \left(\frac{1\,760/4}{45/25}\right) = 244.44$$

Debido a que el valor-$p = 0.000$, la relación es significativa
b) SCE(x_1, x_2, x_3, x_4) = 45
c) SCE(x_2, x_3) = 1805 − 1705 = 100
d) $F = \dfrac{(100 - 45)/2}{1.8} = 15.28$

Debido a que el valor-$p = 0.000$, x_1 y x_2 son significativos

FIGURA D16.5*a*

```
The regression equation is
Y = 433 + 37.4 X -0.383 XSQ

Predictor       Coef    SE Coef        T        p
Constant       432.6      141.2     3.06    0.055
X             37.429      7.807     4.79    0.017
XSQ          -0.3829     0.1036    -3.70    0.034

S = 15.83     R-sq = 98.0%     R-sq(adj) = 96.7%

Analysis of Variance

SOURCE            DF       SS       MS        F        p
Regression         2    36643    18322    73.15    0.003
Residual Error     3      751      250
Total              5    37395
```

FIGURA D16.5*c*

```
    Fit    Stdev.Fit              95% C.I.              95% P.I.
1302.01     9.93        (1270.41, 1333.61)    (1242.55, 1361.47)
```

12. *a*) El resultado de Minitab se muestra en la figura D16.12*a*
b) El resultado de Minitab se muestra en la figura D16.12*b*

c) $$F = \frac{[\text{SCE(reducido)} - \text{SCE(completo)}]/(\#\text{ términos extra})}{\text{MCE(completo)}}$$
$$= \frac{(7.2998 - 4.3240)/2}{0.1663} = 8.95$$

El valor-p asociado con $F = 8.95$ (2 grados de libertad en el numerador y 26 en el denominador) es 0.001; con un valor-$p < \alpha = 0.05$, la suma de las dos variables independientes es significativa

14. *a*) $\hat{y} = -111 + 1.32$ Edad $+ 0.296$ Presión
b) $\hat{y} = -123 + 1.51$ Edad $+ 0.448$ Presión $+ 8.87$ Fumador $- 0.00276$ EdadPresión
c) Significativo: el valor-$p = 0.000$

16. *a*) Semanas $= -8.9 + 1.51$ Edad
b) Semanas $= -0.07 + 1.73$ Edad $- 2.7$ Gerente $- 15.1$ Director $- 17.4$ Ventas
c) El mismo que el inciso *b*)
d) El mismo que el inciso *b*)
e) Semanas $= 13.1 + 1.64$ Edad $- 9.76$ Casado $- 19.4$ Director $- 29.0$ Gerente $- 19.0$ Ventas

18. *a*) RPG $= -4.05 + 27.6$ OBP
b) Una gran cantidad de modelos proporcionará un buen ajuste; el modelo de cinco variables identificadas utiliza el procedimiento de regresión de Minitab con alpha to enter $= 0.10$ y alpha to remove $= 0.10$ de la siguiente manera:

RPG $= -0.0909 + 32.2$ OBP $+ 0.109$ HR $- 21.5$ AVG $+ 0.244$ 3B $- 0.0223$ BB

20.

x_1	x_2	x_3	Tratamiento
0	0	0	A
1	0	0	B
0	1	0	C
0	0	1	D

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

22. Factor A: $x_1 = 0$ si es el nivel 1 y 1 si es el nivel 2
Factor B:

x_2	x_3	Nivel
0	0	1
1	0	2
0	1	3

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1 x_3$$

24. *a*) No es significativo con un nivel de significancia 0.05; el valor-$p = 0.093$
b) 139

26. Significativo en general; el valor-$p = 0.029$
Individualmente, ninguna de las variables es significativa en el nivel de significancia 0.05; una muestra de mayor tamaño sería útil

28. $d = 1.60$; la prueba no es concluyente

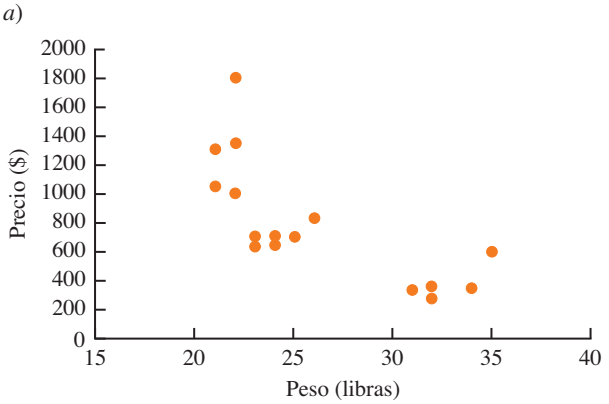
FIGURA D16.12*a*

```
The regression equation is
Scoring Avg) = 46.3 + 14.1 Putting Avg)

Predictor          Coef    SE Coef      T       p
Constant         46.277      6.026   7.68   0.000
Putting Avg)     14.103      3.356   4.20   0.000

S = 0.510596     R-Sq = 38.7%     R-Sq(adj) = 36.5%

Analysis of Variance

SOURCE            DF        SS        MS       F        p
Regression         1    4.6036    4.6036   17.66   0.0000
Residual Error    28    7.2998    0.2607
Total             29   11.9035
```

FIGURA D16.12*b*

```
The regression equation is
Scoring Avg) = 59.0 - 10.3 Greens in Reg)
  + 11.4 Putting Avg - 1.81 Sand Saves

Predictor            Coef    SE Coef         T        p
Constant           59.022      5.774     10.22    0.000
Greens in Reg)    -10.281      2.877     -3.57    0.001
Putting Avg)       11.413      2.760      4.14    0.000
Sand Saves         -1.8130     0.9210     -1.97    0.060

S = 0.407808     R-Sq = 63.7%     R-Sq(adj) = 59.5%

Analysis of Variance

Source             DF        SS        MS        F        p
Regression          3    7.5795    2.5265    15.19    0.000
Residual Error     26    4.3240    0.1663
Total              29   11.9035
```

30. *a*)



Parece existir una relación curvilínea entre el peso y el precio

b) Una parte del resultado de Minitab es la siguiente:

```
The regression equation is
Price = 11376 - 728 Weight + 12.0 WeightSq

Predictor    Coef   SE Coef       T       p
Constant    11376      2565    4.43   0.000
Weight     -728.3     193.7   -3.76   0.002
WeightSq   11.974     3.539    3.38   0.004

S = 242.804    R-Sq = 77.0%    R-Sq(adj) = 74.1%

Analysis of Variance

SOURCE           DF        SS        MS       F       p
Regression        2   3161747   1580874   26.82   0.000
Residual Error   16    943263     58954
Total            18   4105011
```

Los resultados obtenidos apoyan la conclusión de que existe una relación curvilínea entre el peso y el precio

c) Una parte del resultado de Minitab es la siguiente:

```
The regression equation is
Price = 1284 - 572 Type_Fitness - 907 Type_Comfort

Predictor         Coef   SE Coef       T       p
Constant       1283.75     95.22   13.48   0.000
Type_Fitness    -571.8     153.5   -3.72   0.002
Type_Comfort    -907.1     145.5   -6.24   0.000

S = 269.328    R-Sq = 71.7%    R-Sq(adj) = 68.2%

Analysis of Variance

SOURCE           DF        SS        MS       F       p
Regression        2   2944410   1472205   20.30   0.000
Residual Error   16   1160601     72538
Total            18   4105011
```

El tipo de bicicleta parece ser un factor significativo en la predicción del precio, pero la ecuación de regresión estimada obtenida en el inciso *b*) parece ofrecer un ajuste menor

d) Una parte del resultado de Minitab se muestra a continuación. En este resultado, WxF denota la interacción entre el peso de la bicicleta y la variable ficticia Type_Fitness, y WxX denota la interacción entre el peso de la bicicleta y la variable ficticia Type_Comfort

```
The regression equation is
Price = 5924 - 214 Weight - 6343 Type_Fitness - 7232
       Type_Comfort + 261 WxF + 266 WxC

Predictor         Coef   SE Coef       T       p
Constant          5924      1547    3.83   0.002
Weight         -214.56     71.42   -3.00   0.010
Type_Fitness     -6343      2596   -2.44   0.030
```

```
Type_Comfort     -7232      2518    -2.87    0.013
WxF              261.3     111.8     2.34    0.036
WxC              266.41     93.98    2.83    0.014

S = 224.438    R-Sq = 84.0%    R-Sq(adj) = 77.9%

Analysis of Variance

SOURCE           DF        SS       MS       F       p
Regression        5   3450170   690034   13.70   0.000
Residual Error   13    654841    50372
Total            18   4105011
```

Teniendo en cuenta el tipo de bicicleta, el peso y la interacción entre ambos factores, esta ecuación de regresión estimada proporciona un excelente ajuste

32. *a*) Delay = 63.0 + 11.1 Industry; autocorrelación positiva no significativa

34. Diferencias significativas entre los niveles de comodidad para los tres tipos de compradores; el valor-p = 0.034

Capítulo 17

1. *a*)

Artículo	Precio relativo
A	103 = (7.75/7.50)(100)
B	238 = (1500/630)(100)

b) $I_{2009} = \dfrac{7.75 + 1\,500.00}{7.50 + 630.00}(100) = \dfrac{1\,507.75}{637.50}(100) = 237$

c) $I_{2009} = \dfrac{7.75(1\,500) + 1\,500.00(2)}{7.50(1500) + 630.00(2)}(100)$

$= \dfrac{14\,625.00}{12\,510.00}(100) = 117$

d) $I_{2009} = \dfrac{7.75(1\,800) + 1\,500.00(1)}{7.50(1800) + 630.00(1)}(100)$

$= \dfrac{15\,450.00}{14\,130.00}(100) = 109$

2. *a*) 32%
b) \$8.14

3. *a*) Precios relativos de A = (6.00/5.45)100 = 110
B = (5.95/5.60)100 = 106
C = (6.20/5.50)100 = 113

b) $I_{2009} = \dfrac{6.00 + 5.95 + 6.20}{5.45 + 5.60 + 5.50}(100) = 110$

c) $I_{2009} = \dfrac{6.00(150) + 5.95(200) + 6.20(120)}{5.45(150) + 5.60(200) + 5.50(120)}(100)$
$= 109$
Incremento de 9% en el próximo periodo de dos años

4. $I_{2009} = 114$

6.

Artículo	Precio relativo	Periodo base: Precio	Periodo base: Uso	Peso	Precio relativo ponderado
A	150	22.00	20	440	66 000
B	90	5.00	50	250	22 500
C	120	14.00	40	560	67 200
			Totales	1250	155 700

$$I = \frac{155\,700}{1250} = 125$$

7. *a*) Los precios relativos de A = (3.95/2.50)100 = 158
B = (9.90/8.75)100 = 113
C = (0.95/.99)100 = 96

b)

Artículo	Precio relativo	Precio base	Cantidad	Peso $P_{i0}Q_i$	Precio relativo ponderado
A	158	2.50	25	62.5	9 875
B	113	8.75	15	131.3	14 837
C	96	0.99	60	59.4	5 702
			Totales	253.2	30 414

$$I = \frac{30\,414}{253.2} = 120$$

El costo de las materias primas aumentó 20% para la industria química

8. I = 105; el portafolio aumentó 5%

10. *a*) Salarios deflactados en 1996: $\dfrac{\$11.86}{154.9}(100) = \$7.66$

Salarios deflactados en 2009: $\dfrac{\$18.55}{212.2}(100) = \$8.74$

b) $\dfrac{18.55}{11.86}(100) = 156.4$; el aumento porcentual en los salarios reales es de 56.4%

c) $\dfrac{8.74}{7.66}(100) = 114.1$; el aumento porcentual en los salarios reales es de 14.1%

12. *a*) 2428, 2490, 2451
Los pedidos de la industria aumentaron ligeramente en términos de dólares constantes
b) 3043, 3132, 3050
c) PPI

14. $I = \dfrac{300(18.00) + 400(4.90) + 850(15.00)}{350(18.00) + 220(4.90) + 730(15.00)}(100)$
$= \dfrac{20110}{18328}(100) = 110$

15. $I = \dfrac{95(1200) + 75(1800) + 50(2000) + 70(1500)}{120(1200) + 86(1800) + 35(2000) + 60(1500)}(100)$
$= 99$
Las cantidades disminuyeron ligeramente

16. $I = 83$

18. *a*) 151, 197, 143, 178
b) $I = 170$

20. $I_{\text{Jan}} = 73.5$, $I_{\text{Mar}} = 70.1$

22. $I = 182.7$

24. \$36082; \$32528; \$27913; \$34387; \$40551; \$42651; \$46458; \$56324

26. $I = 143$; la cantidad aumentó 43%

Capítulo 18

1. La siguiente tabla muestra los cálculos de los incisos *a*), *b*) y *c*).

Semana	Valor de la serie de tiempo	Pronóstico	Error de pronóstico	Valor absoluto del error de pronóstico	Error de pronóstico cuadrado	Error porcentual	Valor absoluto del error porcentual
1	18						
2	13	18	−5	5	25	−38.46	38.46
3	16	13	3	3	9	18.75	18.75
4	11	16	−5	5	25	−45.45	45.45
5	17	11	6	6	36	35.29	35.29
6	14	17	−3	3	9	−21.43	21.43
			Totales	22	104	−51.30	159.38

a) EAM $= \dfrac{22}{5} = 4.4$

b) CME $= \dfrac{104}{5} = 20.8$

c) EPAM $= \dfrac{159.38}{5} = 31.88$

d) El pronóstico para la semana 7 es 14

2. La siguiente tabla muestra los cálculos para los incisos *a*), *b*) y *c*).

Semana	Valor de la serie de tiempo	Pronóstico	Error de pronóstico	Valor absoluto del error de pronóstico	Error de pronóstico cuadrado	Error porcentual	Valor absoluto del error porcentual
1	18						
2	13	18.00	−5.00	5.00	25.00	−38.46	38.46
3	16	15.50	0.50	0.50	0.25	3.13	3.13
4	11	15.67	−4.67	4.67	21.81	−42.45	42.45
5	17	14.50	2.50	2.50	6.25	14.71	14.71
6	14	15.00	−1.00	1.00	1.00	−7.14	7.14
			Totales	13.67	54.31	−70.21	105.86

a) $\text{EAM} = \dfrac{13.67}{5} = 2.73$

b) $\text{CME} = \dfrac{54.31}{5} = 10.86$

c) $\text{EPAM} = \dfrac{105.89}{5} = 21.18$

d) El pronóstico para la semana 7 es
$$\frac{18 + 13 + 16 + 11 + 17 + 14}{6} = 14.83$$

4. *a*) $\text{CME} = \dfrac{363}{6} = 60.5$

El pronóstico para el mes 8 es 15

b) $\text{CME} = \dfrac{216.72}{6} = 36.12$

El pronóstico para el mes 8 es 18

c) El promedio de todos los valores anteriores es mejor debido a que CME es menor

5. *a*) El dato parece seguir un patrón horizontal

b) Promedio de tres semanas móviles.

Semana	Valor de la serie de tiempo	Pronóstico	Error de pronóstico	Error de pronóstico cuadrado
1	18			
2	13			
3	16			
4	11	15.67	−4.67	21.78
5	17	13.33	3.67	13.44
6	14	14.67	−0.67	0.44
			Total	35.67

$$\text{CME} = \frac{35.67}{3} = 11.89$$

$$\text{El pronóstico para la semana 7} = \frac{(11 + 17 + 14)}{3} = 14$$

c) Una constante de suavizamiento = 0.2

Semana	Valor de la serie de tiempo	Pronóstico	Error de pronóstico	Error de pronóstico cuadrado
1	18			
2	13	18.00	−5.00	25.00
3	16	17.00	−1.00	1.00
4	11	16.80	−5.80	33.64
5	17	15.64	1.36	1.85
6	14	15.91	−1.91	3.66
			Total	65.15

$$\text{CME} = \frac{65.15}{5} = 13.03$$

El pronóstico para la semana 7 es $0.2(14) + (1 - 0.2)15.91 = 15.53$

d) El promedio cambiante de tres semanas proporciona un mejor pronóstico, ya que éste tiene un MCE menor

e) Una constante de suavizamiento = 0.4

Semana	Valor de la serie de tiempo	Pronóstico	Error de pronóstico	Error de pronóstico cuadrado
1	18			
2	13	18.00	−5.00	25.00
3	16	16.00	0.00	0.00
4	11	16.00	−5.00	25.00
5	17	14.00	3.00	9.00
6	14	15.20	−1.20	1.44
			Total	60.44

$$\text{CME} = \frac{60.44}{5} = 12.09$$

El pronóstico de suavizamiento exponencial que utiliza $\alpha = 0.4$ proporciona un mejor pronóstico que el que utiliza $\alpha = 0.2$, ya que éste tiene un CME menor

6. *a*) El dato parece seguir un patrón horizontal

b) $\text{CME} = \dfrac{110}{4} = 27.5$

El pronóstico para la semana 8 es 19

c) $\text{CME} = \dfrac{252.87}{6} = 42.15$

El pronóstico para la semana 7 es 19.12

d) El promedio móvil de tres semanas proporciona un mejor pronóstico, ya que éste tiene un CME inferior

e) CME = 39.79

El pronóstico de suavizamiento exponencial que utiliza $\alpha = 0.4$ proporciona un mejor pronóstico que el que utiliza $\alpha = 0.2$, ya que éste tiene un CME menor

8. *a*)

Semana	4	5	6	7	8	9	10	11	12
Pronóstico	19.33	21.33	19.83	17.83	18.33	18.33	20.33	20.33	17.83

b) CME = 11.49

Prefiere el promedio móvil no ponderado; éste tiene un CME menor

c) Siempre se puede determinar un promedio ponderado en movimiento al menos tan bueno como el no ponderado; en realidad el promedio no ponderado en movimiento es un caso especial de los ponderados donde los pesos son iguales

10. *b*) Los datos más recientes reciben el mayor peso o importancia al determinar el pronóstico; el método de promedios en movimiento pondera el último valor del dato *n* igualmente al determinar el pronóstico

12. *a*) Los datos parecen seguir un patrón horizontal

b) CME(3 meses) = 0.12
CME(4meses) = 0.14
Utilice los promedios móviles a 3 meses

c) 9.63

13. *a*) Los datos parecen seguir un patrón horizontal

b)

Mes	Valor de la serie de tiempo	Promedio móvil de pronóstico del mes	$(\text{Error})^2$	$\alpha = 0.2$	$(\text{Error})^2$
1	240				
2	350			240.00	12 100.00
3	230			262.00	1 024.00
4	260	273.33	177.69	255.60	19.36
5	280	280.00	0.00	256.48	553.19
6	320	256.67	4 010.69	261.18	3 459.79
7	220	286.67	4 444.89	272.95	2 803.70
8	310	273.33	1 344.69	262.36	2 269.57
9	240	283.33	1 877.49	271.89	1 016.97
10	310	256.67	2 844.09	265.51	1 979.36
11	240	286.67	2 178.09	274.41	1 184.05
12	230	263.33	1 110.89	267.53	1 408.50
		Totals	17 988.52		27 818.49

CME (3 meses) = 17 988.52/9 = 1 998.72

CME ($\alpha = 0.2$) = 27 818.49/11 = 2 528.95

Con base en los valores CME anteriores, los promedios móviles a 3 meses parecen mejores; sin embargo, el suavizamiento exponencial fue penalizado mediante la inclusión del mes 2, que fue difícil predecir por cualquier método. Utilizando sólo los errores de los meses 4 a 12, el CME para el exponente suavizado es

$$\text{MSE}(\alpha = 0.2) = 14\,694.49/9 = 1632.72$$

Por tanto, el suavizamiento exponencial fue mejor al considerar los meses 4 a 12

c) Al utilizar el suavizamiento exponencial,

$$F_{13} = \alpha Y_{12} + (1 - \alpha)F_{12}$$
$$= 0.20(230) + 0.80(267.53) = 260$$

14. *a*) Los datos parecen seguir un patrón horizontal

b) Los valores para los meses 2 a 12 son los siguientes.

105.00 114.00 115.80 112.56 105.79 110.05
120.54 126.38 118.46 106.92 104.85

$$\text{CME} = 510.29$$

c) Los valores para los meses 2 a 12 son los siguientes.

105.00 120.00 120.00 112.50 101.25 110.63
127.81 133.91 116.95 98.48 99.24

$$\text{CME} = 540.55$$

Conclusión: una constante de suavizamiento de 0.3 es mejor que de 0.5, ya que el CME es menor de 0.03

16. *a*) La gráfica de la serie de tiempo indica una posible tendencia lineal en los datos; esto puede deberse a la disminución del interés de los espectadores por ver el Master, pero una inspección más detallada de los datos indica que las dos posiciones más altas corresponden a los años 1997 y 2001, en los que Tiger Woods ganó el torneo. El patrón observado puede deberse simplemente al efecto que Tiger Woods tiene en la posición y no necesariamente a un decremento a largo plazo del interés del espectador

b) Los métodos estudiados en esta sección sólo son aplicables para una serie de tiempo que tiene un patrón horizontal, como si existiera realmente una tendencia lineal a largo plazo en los datos. Los métodos descritos en esta sección no son apropiados

c) La gráfica de las series de tiempo para los datos de los años 2002 a 2008 muestran un patrón horizontal; parece razonable concluir que los valores extremos observados en 1997 y 2001 son más atribuibles al interés de los espectadores en el desempeño de Tiger Woods. Basando el pronóstico en los años 2002-2008 parece razonable, pero debido a la lesión que Tiger Woods experimentó en 2008, la razón es si es capaz de jugar en el Master, entonces la posición para 2009 puede ser significativamente mayor que la sugerida por los datos para los años 2002 a 2008

17. *a*) La gráfica de la serie de tiempo muestra una tendencia lineal

b) $\bar{t} = \dfrac{\sum_{t=1}^{n} t}{n} = \dfrac{15}{5} = 3 \quad \overline{Y} = \dfrac{\sum_{t=1}^{n} Y_t}{n} = \dfrac{55}{5} = 11$

$$\Sigma(t - \bar{t})(Y_t - \overline{Y}) = 21 \quad \Sigma(t - \bar{t})^2 = 10$$

$$b_1 = \frac{\sum_{t=1}^{n}(t - \bar{t})(Y_t - \overline{Y})}{\sum_{t=1}^{n}(t - \bar{t})^2} = \frac{21}{10} = 2.1$$

$$b_0 = \overline{Y} - b_1\bar{t} = 11 - (2.1)(3) = 4.7$$

$$T_t = 4.7 + 2.1t$$

c) $T_6 = 4.7 + 2.1(6) = 17.3$

18. El pronóstico para la semana 6 es de 21.16

20. *a*) La gráfica de las series de tiempo exhibe una tendencia curvilínea

b) $T_t = 107.857 - 28.9881t + 2.65476t^2$

c) 45.86

21. *a*) La gráfica de la serie de tiempo muestra una tendencia lineal

b) $\bar{t} = \dfrac{\sum_{t=1}^{n} t}{n} = \dfrac{45}{9} = 5 \quad \overline{Y} = \dfrac{\sum_{t=1}^{n} Y_t}{n} = \dfrac{108}{9} = 12$

$$\Sigma(t - \bar{t})(Y_t - \overline{Y}) = 87.4 \quad \Sigma(t - \bar{t})^2 = 60$$

$$b_1 = \frac{\sum_{t=1}^{n}(t - \bar{t})(Y_t - \overline{Y})}{\sum_{t=1}^{n}(t - \bar{t})^2} = \frac{87.4}{60} = 1.4567$$

$$b_0 = \overline{Y} - b_1\bar{t} = 12 - (1.4567)(5) = 4.7165$$

$$T_t = 4.7165 + 1.4567t$$

c) $T_{10} = 4.7165 + 1.4567(10) = 19.28$

22. *a*) La gráfica de la serie de tiempo muestra una tendencia lineal descendente

b) $T_t = 13.8 - 0.7t$

c) 8.2

d) Si SCF puede continuar disminuyendo el porcentaje de los fondos destinados a gastos administrativos y mantiene una recaudación de 0.7% al año, el pronóstico de gastos para 2015 es de 4.70%

24. *a*) La gráfica de la serie de tiempo muestra una tendencia lineal
b) $T_t = 7.5623 - 0.07541t$
c) 6.7328
d) Dada la incertidumbre en las condiciones del mercado mundial, no es recomendable realizar un pronóstico de diciembre utilizando únicamente el tiempo

26. *a*) La tendencia lineal no es apropiada
b) $T_t = 5.702 + 2.889t - 1618t^2$
c) 17.90

28. *a*) La gráfica de series del tiempo muestra un patrón horizontal, pero existe un patrón estacional en los datos; por ejemplo, en cada año el valor menor se presenta en el trimestre 2 y el valor mayor en el trimestre 4
b) Una parte del resultado de la regresión de Minitab es la siguiente.

```
The regression equation is
Value = 77.0 - 10.0 Qtr1 - 30.0
        Qtr2 - 20.0 Qtr3
```

c) Los pronósticos trimestrales para los siguientes años son los siguientes.

Pronóstico del trimestre 1 = 77.0 − 10.0(1) − 30.0(0) − 20.0(0) = 67
Pronóstico del trimestre 2 = 77.0 − 10.0(0) − 30.0(1) − 20.0(0) = 47
Pronóstico del trimestre 3 = 77.0 − 10.0(0) − 30.0(0) − 20.0(1) = 57
Pronóstico del trimestre 4 = 77.0 − 10.0(0) − 30.0(0) − 20.0(0) = 77

30. *a*) Parece haber un patrón estacional en los datos y quizá una tendencia lineal ascendente moderada
b) La siguiente es una muestra de una parte del resultado de la regresión de Minitab.

```
The regression equation is
Value = 2492 - 712 Qtr1 - 1512
        Qtr2 + 327 Qtr3
```

c) Los pronósticos trimestrales para el próximo año son los siguientes.

El pronóstico del trimestre 1 es 1 780
El pronóstico del trimestre 2 es 980
El pronóstico del trimestre 3 es 2 819
El pronóstico del trimestre 4 es 2 492

d) La siguiente es una parte del resultado de la regresión de Minitab.

```
The regression equation is
Value = 2307 - 642 Qtr1 - 1465
        Qtr2 + 350 Qtr3 + 23.1 t
```

Los pronósticos trimestrales para el próximo año son los siguientes.

El pronóstico del trimestre 1 es 2 058
El pronóstico del trimestre 2 es 1 258
El pronóstico del trimestre 3 es 3 096
El pronóstico del trimestre 4 es 2 769

32. *a*) Las gráficas de series de tiempo muestran una tendencia lineal y los efectos estacionales
b) La siguiente es una parte del resultado de la regresión de Minitab.

```
The regression equation is
Revenue = 70.0 + 10.0 Qtr1 + 105
          Qtr2 + 245 Qtr3
```

El pronóstico del trimestre 1 es 80
El pronóstico del trimestre 2 es 175
El pronóstico del trimestre 3 es 315
El pronóstico del trimestre 4 es 70

c) Una parte del resultado de la regresión de Minitab es la siguiente.

```
The regression equation is
Revenue = -70.1 + 45.0 Qtr1 + 128
          Qtr2 + 257 Qtr3 + 11.7 Period
```

El pronóstico del trimestre 1 es 221
El pronóstico del trimestre 2 es 315
El pronóstico del trimestre 3 es 456
El pronóstico del trimestre 4 es 211

34. *a*) Las gráficas de series de tiempo muestran los efectos de una tendencia estacional y lineal
b) *Nota*: Jan = 1 si enero, 0 en caso contrario; Feb = 1 si es febrero, 0 en caso contrario, y así sucesivamente
Una parte del resultado de regresión de Minitab es el siguiente.

```
The regression equation is
Expense = 175 - 18.4 Jan - 3.72 Feb +
          12.7 Mar + 45.7 Apr + 57.1
          May + 135 Jun + 181 Jul + 105
          Aug + 47.6 Sep + 50.6 Oct +
          35.3 Nov + 1.96 Period
```

c) *Nota*. El siguiente periodo en la serie de tiempo es Periodo = 37 (Enero del año 4); el pronóstico para enero a diciembre es 229; 246; 264; 299; 312; 392; 440; 366; 311; 316; 302; 269

35. *a*) La gráfica de la serie de tiempo indica una tendencia lineal y un patrón estacional
b)

Año	Trimestre	Valor de la serie de tiempo	Promedio móvil del cuarto trimestre	Promedio móvil centrado
1	1	4		
	2	2		
			3.50	
	3	3		3.750
			4.00	
	4	5		4.125
			4.25	
2	1	6		4.500
			4.75	

Año	Trimestre	Valor de la serie de tiempo	Promedio móvil del cuarto trimestre	Promedio móvil centrado
	2	3		5.000
			5.25	
	3	5		5.375
			5.50	
	4	7		5.875
			6.25	
3	1	7		6.375
			6.50	
	2	6		6.625
			6.75	
	3	6		
	4	8		

c)

Año	Trimestre	Valor de la serie de tiempo	Promedio móvil centrado	Componente estacional irregular
1	1	4		
	2	2		
	3	3	3.750	0.800
	4	5	4.125	1.212
2	1	6	4.500	1.333
	2	3	5.000	0.600
	3	5	5.375	0.930
	4	7	5.875	1.191
3	1	7	6.375	1.098
	2	6	6.625	0.906
	3	6		
	4	8		

Trimestre	Valor estacional irregular		Índice estacional	Índice estacional ajustado
1	1.333	1.098	1.216	1.205
2	0.600	0.906	0.752	0.746
3	0.800	0.930	0.865	0.857
4	1.212	1.191	1.201	1.191
		Total	4.036	

$$\text{Ajuste para índice estacional} \quad \frac{4.000}{4.036} = 0.991$$

36. *a*)

Año	Trimestre	Valor desestacionalizado
1	1	3.320
	2	2.681
	3	3.501
	4	4.198
2	1	4.979
	2	4.021
	3	5.834
	4	5.877
3	1	5.809
	2	8.043
	3	7.001
	4	6.717

b) Sea Periodo = 1 el valor de las series de tiempo en el Año 1-Trimestre1; Periodo = 2 que denota el valor de las series de tiempo en el Año 1-Trimestre 2, y así sucesivamente. Una parte del resultado de regresión de Minitab trata el periodo como una variable independiente y los valores desestacionalizados como los valores de la variable dependiente de la siguiente forma.

```
The regression equation is
Deseasonalized Value = 2.42 + 0.422
                       Period
```

c) La tendencia trimestral desestacionalizada pronostica para el año 4 (periodos 13, 14, 15 y 16) lo siguiente:

El pronóstico para el trimestre 1 es 7.906
El pronóstico para el trimestre 2 es 8.328
El pronóstico para el trimestre 3 es 8.750
El pronóstico para el trimestre 4 es 9.172

d) El ajuste de la tendencia trimestral desestacionalizada proporciona las siguientes estimaciones trimestrales:

El pronóstico para el trimestre 1 es 9.527
El pronóstico para el trimestre 1 es 6.213
El pronóstico para el trimestre 1 es 7.499
El pronóstico para el trimestre 1 es 10.924

38. *a*) La gráfica de la serie de tiempo muestra una tendencia lineal y efectos estacionales

b) 0.71 0.78 0.83 0.97 1.02 1.30 1.50 1.23 0.98 0.99 0.93 0.79

c)

Mes	Gasto desestacionalizado
1	239.44
2	230.77
3	246.99
4	237.11
5	235.29
6	242.31
7	240.00
8	235.77
9	244.90
10	242.42
11	247.31

(*Continúa*)

Mes	Gasto desestacionalizado
12	246.84
13	253.52
14	262.82
15	259.04
16	252.58
17	259.80
18	253.85
19	266.67
20	272.36
21	265.31
22	272.73
23	274.19
24	278.48
25	274.65
26	269.23
27	277.11
28	288.66
29	284.31
30	300.00
31	280.00
32	268.29
33	295.92
34	297.98
35	301.08
36	316.46

d) Sea Periodo = 1 el valor de las series de tiempo en Enero-Año 1; Periodo = 2 el valor de las series de tiempo en Febrero-Año 2; y así sucesivamente. Una parte del resultado de regresión de Minitab trata al periodo como una variable independiente y los valores destacionalizados como los valores de la variable dependiente de la siguiente forma:

```
The regression equation is
Deseasonalized Expense = 228 + 1.96
                         Period
```

e)

Mes	Pronóstico mensual
Enero	213.37
Febrero	235.93
Marzo	252.69
Abril	297.21
Mayo	314.53
Junio	403.42
Julio	486.42
Agosto	386.52
Septiembre	309.88
Octubre	314.98
Noviembre	297.71
Diciembre	254.44

40. *a*) La gráfica de las series de tiempo indican un efecto estacional; el consumo de energía es menor en el periodo de 12 a 4 a.m.; se incrementa constantemente al valor más alto de 12 a 4 p.m., y después decrece nuevamente. Puede haber también alguna tendencia lineal en los datos

b)

Periodo de tiempo	Índice estacional ajustado
12–4 A.M.	0.3256
4–8 A.M.	0.4476
8–12 a medio día	1.3622
12–4 P.M.	1.6959
4–8 P.M.	1.4578
8–12 media noche	0.7109

c) El siguiente resultado de Minitab muestra una ecuación con tendencia lineal apropiada a la series de tiempo destacionalizadas:

```
The regression equation is
Deseasonalized Power = 63108 + 1854 t
```

Energía destacionalizada (t = 19) = 63 108 + 1854(19) = 98 334
Pronóstico para 12-4 p.m. = 1.6959(98,334) = 166 764.63 o aproximadamente 166 765 kWh
Energía destacionalizada (t = 20) = 63 108 + 1854(20) = 100 188
Pronóstico para 4-8 p.m. = 1.4578(100,188) = 146 054.07 o aproximadamente 146 054 kWh
Por tanto, el pronóstico de consumo de energía desde medio día hasta 8 p.m. es 166 765 + 146 054 = 312 819 kWh

42. *a*) La gráfica de series de tiempo indica un patrón horizontal

b) $\text{CME}(\alpha = 0.2) = 1.40$
$\text{CME}(\alpha = 0.3) = 1.27$
$\text{CME}(\alpha = 0.4) = 1.23$

Una constante suavizada de $\alpha = 0.4$ proporciona el mejor pronóstico debido a que este tiene un CME inferior

c) 31.00

44. *a*) Parece ser una tendencia de incremento en los datos

b) Una parte del resultado de regresión de Minitab es el siguiente (*Nota*: $t = 1$ corresponde a 2001, $t = 2$ corresponde a 2002, y así sucesivamente)

```
The regression equation is
Balance($) = 1984 + 146 t
```

El pronóstico para 2009 ($t = 9$) es Balance($) = 1984 + 146(9) = $3298

c) Una parte del resultado de regresión de Minitab es el siguiente (*Nota*: $t = 1$ corresponde a 2001, $t = 2$ corresponde a 2002, y así sucesivamente)

```
The regression equation is
Balance($) = 2924 - 419 t + 62.7 tsq
```

El pronóstico para 2009 ($t = 9$) es Balance ($) = $2924 - 419(9) + 62.7(9)^2 = \$4232$

d) La ecuación de tendencia cuadrática proporciona el mejor pronóstico exacto para los datos históricos
e) Ecuación de tendencia lineal

46. *a*) El pronóstico para julio es 236.97
El pronóstico para agosto, usando el pronóstico para julio como las ventas reales en julio, es 236.97
El suavizamiento exponencial proporciona el mismo pronóstico para cada periodo en lo futuro; éste es el porqué generalmente no se recomienda para la elaboración de pronósticos a largo plazo
b) Utilizando el procedimiento de regresión de Minitab se obtuvo una ecuación de tendencia lineal

$$T_t = 149.72 + 18.451t$$

El pronóstico para julio es 278.88
El pronóstico para agosto es 297.33
c) El acuerdo propuesto no es justo, ya que éste no toma en cuenta la tendencia a la alza en las ventas; con base en la proyección de tendencia, el acuerdo debe basarse en el pronóstico de la perdida de ventas de $278 880 en julio y $297 330 en agosto

48. *a*) La gráfica de las series de tiempo muestra una tendencia lineal
b) $T_t = -5 + 15t$
La pendiente de 15 indica que el promedio del alza en las ventas es de 15 pianos por año
c) 85, 100

50. *a*)

Trimestre	Índice estacional ajustado
1	1.2717
2	0.6120
3	0.4978
4	1.6185

Nota. El ajuste para el índice estacional $= \dfrac{4}{3.8985} = 1.0260$
b) El efecto mayor está en el trimestre 4; parece razonable, ya que el las ventas al menudeo son generalmente mayores durante octubre, noviembre y diciembre

52. *a*) Sí, un patrón de tendencia lineal parece estar presente
b) La siguiente es una parte del resultado de regresión de Minitab.

```
The regression equation is
Number Sold = 22.9 + 15.5 Year
```

c) El pronóstico en el año 8 es aproximadamente de 147 unidades

54. *b*) Los valores del promedio móvil centrado suavizan las series de tiempo mediante la eliminación de efectos estacionales y algunas de las variabilidades aleatorias; el promedio móvil centrado de las series de tiempo muestra la tendencia en los datos
c)

Trimestre	Índice estacional ajustado
1	0.899
2	1.362
3	1.118
4	0.621

d) Hudson Marine experimenta el mayor incremento estacional en el trimestre 2, ya que este trimestre se produce antes de la temporada alta de verano en navegación; este resultado parece razonable, pero el mayor efecto estacional es la disminución estacional en el trimestre 4, lo que también es razonable debido a la disminución de la navegación en el otoño y en el invierno

Capítulo 19

1. $n = 27$ con un valor distinto de 150
La aproximación normal $\mu = 0.5n = 0.5(27) = 13.5$
$\sigma = \sqrt{0.25\,n} = \sqrt{0.25(27)} = 2.5981$
Con el número de signos positivos = 22 en la cola superior, utilice el factor de continuidad de corrección como sigue

$$P(x \geq 21.5) = P\left(z \geq \frac{21.5 - 13.5}{2.5981}\right) = P(z \geq 3.08)$$

El valor-p = (1.0000 − 0.9990) = 0.0010
El valor-$p \leq 0.01$; H_0 es rechazada; concluye que la mediana de la población > 150

2. La eliminación de la no preferencia, las probabilidades binomiales para $n = 9$ y $p = 0.50$ son las siguientes

x	Probabilidad	x	Probabilidad
0	0.0020	5	0.2461
1	0.0176	6	0.1641
2	0.0703	7	0.0703
3	0.1641	8	0.0176
4	0.2461	9	0.0020

Número de signos positivos = 7

$$\begin{aligned} P(x \geq 7) &= P(7) + P(8) + P(9) \\ &= 0.0703 + 0.0176 + 0.0020 \\ &= 0.0899 \end{aligned}$$

Valor-p de dos colas = 2(0.0899) = 0.1798
El valor-$p > 0.05$, H_0 no es rechazada; se concluye que no hay ninguna indicación de que existan algunas diferencias

4. *a*) H_0: Mediana ≥ 15
H_a: Mediana < 15
b) $n = 9$; número de signos positivos = 1
El valor-p = 0.0196
H_0 es rechazada; los fondos de inversión en bonos tienen una mediana menor

6. $n = 48$; $z = 1.88$
El valor-p = 0.0301
H_0 es rechazada; la conclusión es que la mediana > $56.2 miles

8. *a*) $n = 15$
El valor-p = 0.0768
H_0 no es rechazada; no hay diferencia significativa para el ritmo de vida
b) 25%, 68.8%; se recomienda una muestra más grande

10. $n = 600$; $z = 2.41$
El valor-p = 0.0160
H_0 es rechazada; diferencia significativa; *American Idol* es preferida

12. H_0: la mediana para el aditivo 1 − la mediana para el aditivo 2 = 0
H_a: la mediana para el aditivo 1 − la mediana para el aditivo 2 ≠ 0

Diferencia	Diferencia absoluta	Rango	Rangos con signos Negativo	Positivo
2.07	2.07	9		9
1.79	1.79	7		7
−0.54	0.54	3	−3	
2.09	2.09	10		10
0.01	0.01	1		1
0.97	0.97	4		4
−1.04	1.04	5	−5	
3.57	3.57	12		12
1.84	1.84	8		8
3.08	3.08	11		11
0.43	0.43	2		2
1.32	1.32	6		6
			Suma del rango con signo positivo	$T^+ = 70$

$$\mu_{T^+} = \frac{n(n+1)}{4} = \frac{12(13)}{4} = 39$$

$$\sigma_{T^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{12(13)(25)}{24}} = 12.7475$$

$$P(T^+ \geq 70) = P\left(z \geq \frac{69.5 - 39}{12.7475}\right) = P(z \geq 2.39)$$

El valor-$p = 2(1.0000 - 0.9916) = 0.0168$

El valor-$p \leq 0.05$, H_0 es rechazada; se concluye una diferencia significativa entre los aditivos

13. H_0: la mediana del tiempo sin el relajante 1 − la mediana de tiempo con relajante ≤ 0

H_a: la mediana del tiempo sin el relajante 1 − la mediana del tiempo con el relajante > 0

Diferencia	Diferencia absoluta	Rango	Rangos con signo Negativo	Positivo
5	5	9		9
2	2	3		3
10	10	10		10
−3	3	6.5	−6.5	
1	1	1		1
2	2	3		3
−2	2	3	−3	
3	3	6.5		6.5
3	3	6.5		6.5
3	3	6.5		6.5
			Suma de intervalo con signo positivo	$T^+ = 45.5$

$$\mu_{T^+} = \frac{n(n+1)}{4} = \frac{10(11)}{4} = 27.5$$

$$\sigma_{T^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{10(11)(12)}{24}} = 9.8107$$

$$P(T^+ \geq 45.5) = P\left(z \geq \frac{45 - 27.5}{12.7475}\right) = P(z \geq 1.78)$$

El valor-$p = (1.0000 - 0.9925) = 0.0375$

El valor-$p \leq 0.05$; H_0 es rechazada; se concluye que sin el relajante tiene una mediana del tiempo mayor

14. $n = 11$; $T^+ = 61$; $z = 2.45$

El valor-$p = 0.0142$

H_0 es rechazada; se concluye que hay una diferencia significativa; a tiempo con mejor % que en 2006

16. $n = 10$; $T^+ = 12.5$; $z = -1.48$

El valor-$p = 0.1388$

H_0 no es rechazada; se concluye que no hay ninguna diferencia entre los promedios medianos

18. H_0: las dos poblaciones de los aditivos son idénticas

H_a: las dos poblaciones de los aditivos son idénticas

Aditivo 1	Rango	Aditivo 2	Rango
17.3	2	18.7	8.5
18.4	6	17.8	4
19.1	10	21.3	15
16.7	1	21.0	14
18.2	5	22.1	16
18.6	7	18.7	8.5
17.5	3	19.8	11
		20.7	13
		20.2	12
	$W =$	34	

$$\mu_W = \frac{1}{2}n_1(n_1 + n_2 + 1) = \frac{1}{2}7(7 + 9 + 1) = 59.5$$

$$\sigma_W = \sqrt{\frac{1}{12}n_1n_2(n_1 + n_2 + 1)} = \sqrt{\frac{1}{12}7(9)(7 + 9 + 1)}$$

$$= 9.4472$$

Con $W = 34$ en la cola inferior, utilice la corrección de continuidad

$$P(W \leq 34) = P\left(z \leq \frac{34.5 - 595}{9.4472}\right) = P(z \leq -2.65)$$

El valor-$p = 2(0.0040) = 0.0080$

El valor-$p < 0.05$; H_0 es rechazada; se concluye que los aditivos no son idénticos

El aditivo 2 tiende a proporcionar mayores millas por galón

19. *a*) H_0: las dos poblaciones de sueldos son idénticas

H_a: las dos poblaciones de sueldos no son idénticas

Contador público	Rango	Planeación financiera	Rango
50.2	5	49.0	2
58.8	19	49.2	3
56.3	16	53.1	10
58.2	18	55.9	15
54.2	13	51.9	8.8.5
55.0	14	53.6	11
50.9	6	49.7	4
59.5	20	53.9	12
57.0	17	51.8	7
51.9	8.8.5	48.9	1
	$W =$ 136.5		

$$\mu_W = \frac{1}{2}n_1(n_1 + n_2 + 1) = \frac{1}{2}10(10 + 10 + 1) = 105$$

$$\sigma_W = \sqrt{\frac{1}{12}n_1n_2(n_1 + n_2 + 1)} = \sqrt{\frac{1}{12}10(10)(10 + 10 + 1)}$$

$$= 13.2288$$

Con $W = 136.5$ en la cola superior, utilice la corrección de continuidad

$$P(W \geq 136.5) = P\left(z \geq \frac{136 - 105}{13.2288}\right) = P(z \geq 2.34)$$

El valor-$p = 2(1.0000 - 0.9904) = 0.0192$

El valor-$p \leq 0.05$; H_0 es rechazada; se concluye que las poblaciones no son idénticas

El contador público tiende a tener mayores sueldos

b) El contador público $\frac{(55.0 + 56.3)}{2} = \$55.65$ mil

Gerente de planeación financiera $\frac{(51.8 + 51.9)}{2} = \$51.85$ miles

20. *a*) \$54 900, \$40 400

b) $W = 69$; $z = 2.04$

El valor-$p = 0.0414$

H_0 es rechazada; se concluye que existe una diferencia entre las razones; el de los hombres es más alto

22. $W = 157$; $z = 2.74$

El valor-$p = 0.0062$

H_0 es rechazada; se concluye que existe una diferencia entre las razones: Japón tiende a ser el mayor

24. $W = 116$; $z = -.22$

El valor-$p = 0.8258$

H_0 no es rechazada; se concluye que no hay evidencia de que los precios difieran

26. H_0: todas las poblaciones de calificaciones de productos son idénticas

H_a: no todas las poblaciones de calificaciones de productos son idénticas

	A	B	C
	4	11	7
	8	14	2
	10	15	1
	3	12	6
	9	13	5
Suma de rangos	34	65	21

$$H = \left[\frac{12}{15(16)}\left(\frac{34^2}{5} + \frac{65^2}{5} + \frac{21^2}{5}\right)\right] - 3(16) = 10.22$$

La tabla de χ^2 con $gl = 2$, $\chi^2 = 10.22$; el valor-p está entre 0.005 y 0.01

El valor-$p \leq 0.01$; H_0 es rechazada; se concluye que las poblaciones de calificación no son idénticas

28. H_0: todas las poblaciones de calorías quemadas son idénticas

H_a: no todas las poblaciones de calorías quemadas son idénticas

	Natación	Tenis	Ciclismo
	8	9	5
	4	14	1
	11	13	3
	6	10	7
	12	15	2
Suma de rangos	41	61	18

$$H = \left[\frac{12}{15(16)}\left(\frac{41^2}{5} + \frac{61^2}{5} + \frac{18^2}{5}\right)\right] - 3(16) = 9.26$$

La tabla de χ^2 con $gl = 2$, $\chi^2 = 9.26$; el valor-p está entre 0.005 y 0.01

El valor-$p \leq 0.05$ H_0 es rechazada; se concluye que la población de calorías quemadas no son idénticas

30. $H = 8.03$; con $gl = 3$

El valor-p está entre 0.025 y 0.05

H_0 es rechazada; se concluye que hay una diferencia entre la calidad de los cursos

32. *a*) $\Sigma d_i^2 = 52$

$$r_s = 1 - \frac{6\Sigma d_i^2}{n(n^2 - 1)} = 1 - \frac{6(52)}{10(99)} = 0.685$$

b) $$\sigma_{r_s} = \sqrt{\frac{1}{n - 1}} = \sqrt{\frac{1}{9}} = 0.3333$$

$$z = \frac{r_s - 0}{\sigma_{r_s}} = \frac{0.685}{0.3333} = 2.05$$

El valor-$p = 2(1.0000 - 0.9798) = 0.0404$

El valor-$p \leq 0.05$ H_0 es rechazada; se concluye una correlación significativa de rango positivo

34. $\Sigma d_i^2 = 250$

$$r_s = 1 - \frac{6\Sigma d_i^2}{n(n^2 - 1)} = 1 - \frac{6(250)}{11(120)} = -0.136$$

$$\sigma_{r_s} = \sqrt{\frac{1}{n - 1}} = \sqrt{\frac{1}{10}} = 0.3162$$

$$z = \frac{r_s - 0}{\sigma_{r_s}} = \frac{-0.136}{0.3162} = -0.45$$

El valor-$p = 2(0.3336) = 0.6672$

El valor-$p > 0.05$ H_0 no es rechazada; no puede concluir que exista una relación significativa

36. $r_s = -0.709$, $z = -2.13$

El valor-$p = 0.0332$

H_0 es rechazada; se concluye que hay una correlación significativa de rango negativo

38. Número de signos de más = 905, $z = -3.15$

El valor-p menor a 0.0020

H_0 es rechazada; se concluye que hay una diferencia significativa entre las preferencias

40. $n = 12$; $T^+ = 6$; $z = -2.55$

El valor-$p = 0.0108$

H_0 es rechazada; se concluye que hay una diferencia significativa entre los precios

42. $W = 70$; $z = -2.93$

El valor-$p = 0.0034$

H_0 es rechazada; se concluye que las poblaciones de pesos no son idénticas

44. $H = 12.61$ con $gl = 2$

El valor-p es menor de 0.005

H_0 es rechazada; se concluye que las poblaciones de clasificaciones no son idénticas

46. $r_s = 0.757$, $z = 2.83$

El valor-$p = 0.0046$

H_0 es rechazada; se concluye que hay una correlación de rango positivo

Capítulo 20

2. *a*) 5.42
b) UCL = 6.09, LCL = 4.75

4. *Gráfica R:*
$\text{UCL} = \bar{R}D_4 = 1.6(1.864) = 2.98$
$\text{LCL} = \bar{R}D_3 = 1.6(0.136) = 0.22$
Carta $\bar{x}$:
$\text{UCL} = \bar{\bar{x}} + A_2\bar{R} = 28.5 + 0.373(1.6) = 29.10$
$\text{LCL} = \bar{\bar{x}} - A_2\bar{R} = 28.5 - 0.373(1.6) = 27.90$

6. 20.01, 0.082

8. *a*) 0.0470
b) UCL = 0.0989, LCL = −0.0049 (utilice LCL = 0)
c) $\bar{p} = 0.08$; en control
d) UCL = 14.826, LCL = −0.726 (utilice LCL = 0)
El proceso está afuera de control si hay más de 14 defectuosos
e) En control con 12 defectuosos
f) Gráfica *np*

10. $f(x) = \dfrac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$
Si $p = 0.02$, la probabilidad de aceptar el lote es
$$f(0) = \frac{25!}{0!(25-0)!}(0.02)^0(1-0.02)^{25} = 0.6035$$
Si $p = 0.06$, la probabilidad de aceptar el terreno es
$$f(0) = \frac{25!}{0!(25-0)!}(0.06)^0(1-0.06)^{25} = 0.2129$$

12. $p_0 = 0.02$; riesgo del productor = 0.0599
$p_0 = 0.06$; riesgo del productor = 0.3396
El riesgo del productor disminuye a medida que el número de aceptación de *c* se aumenta

14. $n = 20$, $c = 3$

16. *a*) 95.4
b) UCL = 96.07, LCL = 94.73
c) No

18.

	Gráfica *R*	Gráfica $\bar{x}$
UCL	4.23	6.57
LCL	0	4.27

La estimación de desviación estándar = 0.86

20.

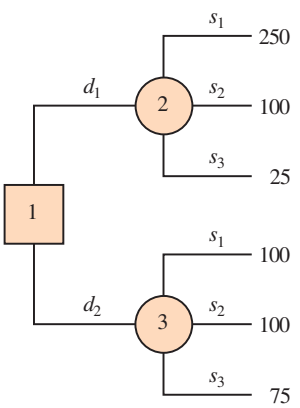
	Gráfica *R*	Gráfica $\bar{x}$
UCL	0.1121	3.112
LCL	0	3.051

22. *a*) UCL = 0.0817, LCL = −0.0017 (utilice LCL = 0)

24. *a*) 0.03
b) $\beta = 0.0802$

Capítulo 21

1. *a*)



b) $\text{VE}(d_1) = 0.65(250) + 0.15(100) + 0.20(25) = 182.5$
$\text{VE}(d_2) = 0.65(100) + 0.15(100) + 0.20(75) = 95$
La decisión óptima es d_1

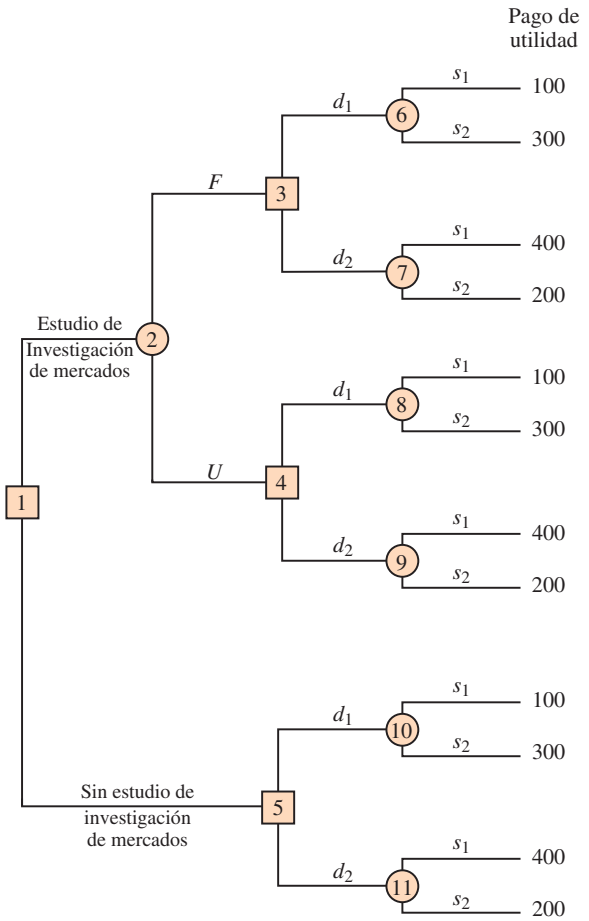
2. *a*) d_1; $\text{VE}(d_1) = 11.3$
b) d_4; $\text{VE}(d_4) = 9.5$

3. *a*) VE(personal propio) = 0.2(650) + 0.5(650) + 0.3(600) = 635
VE(proveedor externo) = 0.2(900) + 0.5(600) + 0.3(300) = 570
VE(combinación) = 0.2(800) + 0.5(650) + 0.3(500) = 635
Decisión óptima: contratar un proveedor externo con un costo esperado de $570 000
b) VEciP = 0.2(650) + 0.5(600) + 0.3(300) = 520
VEIP = $|520 - 570| = 50$, o $50 000

4. *b*) Precio bajo; VE = 565
c) Precio normal; VE = 670

6. *c*) Chardonnay únicamente; VE = 42.5
d) Las dos uvas; VE = 46.4
e) Las dos uvas; VE = 39.6

8. *a*)



b) VE (nodo 6) = 0.57(100) + 0.43(300) = 186
VE (nodo 7) = 0.57(400) + 0.43(200) = 314
VE (nodo 8) = 0.18(100) + 0.82(300) = 264
VE (nodo 9) = 0.18(400) + 0.82(200) = 236
VE (nodo 10) = 0.40(100) + 0.60(300) = 220
VE (nodo 11) = 0.40(400) + 0.60(200) = 280

VE (nodo 3) = Máx(186 314) = 314 d_2
VE (nodo 4) = Máx(264 236) = 264 d_1
VE (nodo 5) = Máx(220 280) = 280 d_2

VE (nodo 2) = 0.56(314) + 0.44(264) = 292
VE (nodo 1) = Máx(292,280) = 292

∴ Investigación de mercados
Si es favorable, la decisión d_2
Si es desfavorable, la decisión d_1

10. *a*) 5 000 − 200 − 2 000 − 150 = 2 650
3 000 − 200 − 2 000 − 150 = 650

b) Los valores esperados en nodos
8: 2350 5: 2350 9: 1100
6: 1 150 10: 2 000 7: 2 000
4: 1 870 3: 2 000 2: 1 560
1: 1 560

c) El costo tendría que disminuir por lo menos $130 000

12. *b*) d_1, 1 250
c) 1 700
d) Si *N*, d_1
Si *U*, d_2; 1 666

14.

Estado	$P(s_j)$	$P(I \mid s_j)$	$P(I \cap s_j)$	$P(s_j \mid I)$
s_1	0.2	0.10	0.020	0.1905
s_2	0.5	0.05	0.025	0.2381
s_3	0.3	0.20	0.060	0.5714
	1.0		$P(I) = 0.105$	1.0000

16. *a*) 0.695, 0.215, 0.090
0.98, 0.02
0.79, 0.21
0.00, 1.00
c) Si *C*, autopista
Si *O*, autopista
Si *R*, City Queen
26.6 minutos

18. *a*) El sector de tecnología proporciona el máximo rendimiento anual esperado de 16.97%. Al utilizar esta recomendación, rendimiento anual mínimo es −20.1% y el máximo es 93.1%
b) 15.20%; 1.77%
d) Debido a que el sector de tecnología del fondo de inversión muestra una mayor variación en la rentabilidad anual, se considera que tienen más riesgo
d) Esta es una recomendación de juicio y de opiniones que pueden variar, pero debido a que el inversionista se describe como un conservador, se recomienda financiar el menor riesgo de los fondos de inversión de menor capitalización

20. *a*) Estrategia óptima:

Inicio del proyecto R&D
Si es exitoso, construir las instalaciones

Valor esperado = $10M millones
b) En el nodo 3, el pago de derechos de venta tendría que ser de $25 millones o más, con el fin de recuperar el costo de $5 millones R&D, el precio de venta tendría que ser de $30 millones o más

Apéndice E Uso de las funciones de Excel

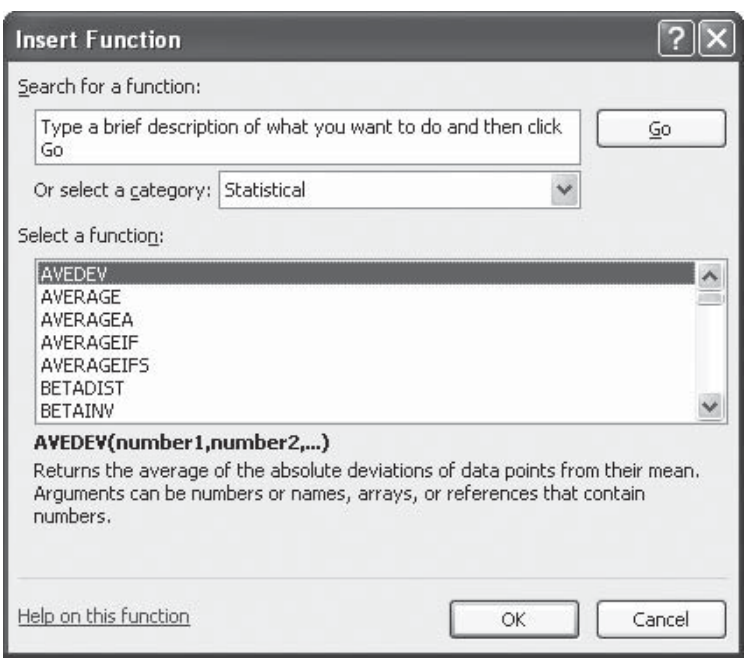
Excel proporciona una gran cantidad de funciones para el manejo de datos y el análisis estadístico. Si se sabe cuál es la función que se necesita y cómo utilizarla, simplemente se ingresa en la celda de la hoja de cálculo correspondiente. Sin embargo, si no se sabe cuáles funciones están disponibles para realizar una tarea o no se está seguro de cómo utilizar una función determinada, Excel proporciona asistencia.

Encontrar la función adecuada en Excel

Para identificar las funciones disponibles en Excel, seleccione la pestaña **Formulas** de la barra de herramientas y haga clic. En el grupo Function Library (biblioteca de funciones) haga clic en **Insert Function** (Insert Function). Otra alternativa es hacer clic en f_x en el botón de la barra de fórmulas. Desde cualquier acceso se proporciona el cuadro **Insert Function** que se muestra en la figura 1.

El cuadro **Search for a function** (buscar una función) en la parte superior del cuadro de diálogo Insert Function permite escribir una pequeña descripción de lo que se desea hacer. Después haga clic en **Go** (ir). Excel buscará y mostrará, en el cuadro **Select a function** (seleccionar una función), las funciones que pueden cumplir con la tarea. Sin embargo, en muchas ocasiones es posible que desee echar un vistazo a todas las categorías de funciones para ver con cuáles se puede contar. Para esta tarea es de utilidad el cuadro **Or select a category** (o seleccionar una categoría). Éste contiene una lista desplegable de varias categorías de funciones que ofrece Excel. La figura 1 muestra que se eligió **Statistical** (Estadística). Como resultado, las funciones

FIGURA 1 Cuadro de diálogo Insert Function



estadísticas de Excel aparecen en orden alfabético en el cuadro *Select a function*. Observe que la función AVEDEV figura en primer lugar de la lista, seguido por la función AVERAGE, y así sucesivamente.

La función AVEDEV está resaltada en la figura 1 indicando que ha sido seleccionada en ese momento. La sintaxis correcta para la función y una breve descripción de la misma aparecen debajo del cuadro *Select a function*. Uno puede desplazarse por la lista en el cuadro para desplegar la sintaxis y una breve descripción de cada una de las funciones estadísticas disponibles. Por ejemplo, al desplazarse hacia abajo se selecciona la función COUNTIF como se muestra en la figura 2. Observe que COUNTIF está ahora resaltada, y que inmediatamente debajo del cuadro *Select a function* se observa **COUNTIF(range,criteria)**, que indica que contiene dos argumentos: rango y criterios. Además, se observa que la descripción de COUNTIF es "Cuente el número de celdas con un rango que cumpla la condición dada."

Si la función seleccionada (resaltada) se desea utilizar, haga clic en **OK** y aparece el cuadro de diálogo **Function Arguments** (argumentos de la función). Los argumentos de la función en el cuadro de diálogo para COUNTIF se muestran en la figura 3. Este cuadro ayuda a crear los argumentos adecuados para la función seleccionada. Cuando termine de ingresar los argumentos, haga clic en **OK**; después Excel ingresa la función en una celda de la hoja de cálculo.

Inserción de una función dentro de una celda de la hoja de cálculo

Ahora se mostrará cómo utilizar los cuadros de diálogo *Insert Function* y *Functions Arguments* para elegir una función, proporcionar los argumentos e insertar la función en la celda de la hoja de cálculo.

En el apéndice 2.2 se utilizó la función COUNTIF de Excel para elaborar una función de distribución de frecuencias para la compra de bebidas refrescantes. La figura 4 muestra una

FIGURA 2 Descripción de la función COUNTIF en el cuadro de diálogo Insert Function

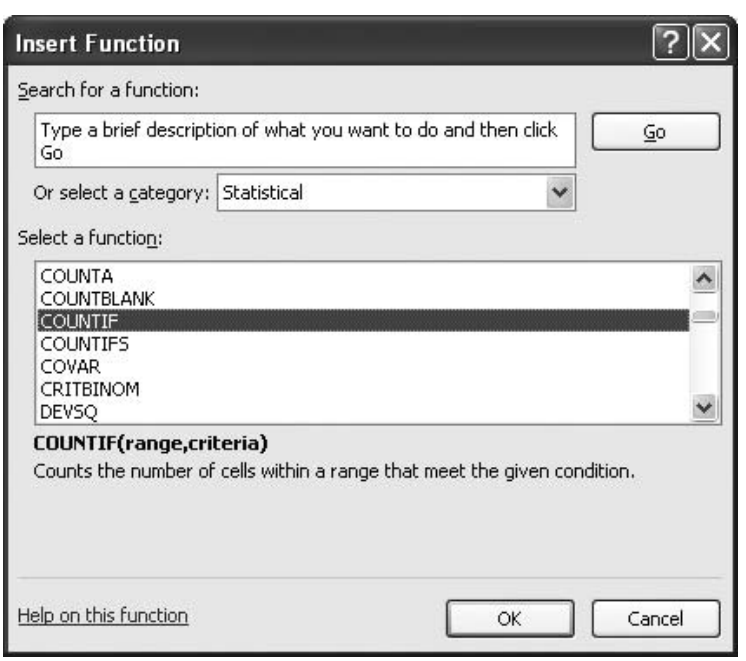


FIGURA 3 Cuadro de diálogo Function Arguments para la función COUNTIF

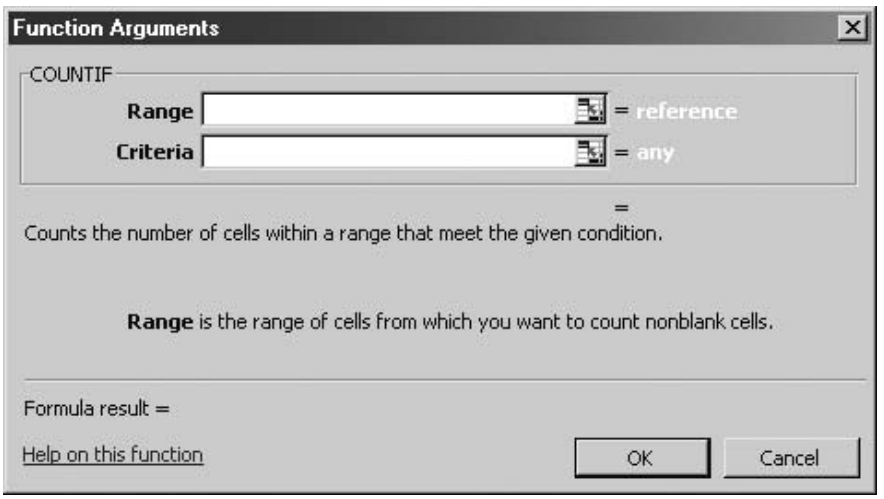


FIGURA 4 Hoja de cálculo de Excel con los datos de las bebidas refrescantes y las etiquetas para la distribución de frecuencias que se desea construir

WEB archivo
SoftDrink

	A	B	C	D	E
1	**Brand Purchased**		**Soft Drink**	**Frequency**	
2	Coke Classic		Coke Classic		
3	Diet Coke		Diet Coke		
4	Pepsi		Dr. Pepper		
5	Diet Coke		Pepsi		
6	Coke Classic		Sprite		
7	Coke Classic				
8	Dr. Pepper				
9	Diet Coke				
10	Pepsi				
45	Pepsi				
46	Pepsi				
47	Pepsi				
48	Coke Classic				
49	Dr. Pepper				
50	Pepsi				
51	Sprite				
52					

Nota. Las filas 11 a 44 están ocultas.

hoja de cálculo de Excel que contiene los datos sobre las bebidas refrescantes y las etiquetas para la distribución de frecuencias que se quiere construir. Observe que la frecuencia de compra de Coke Classic irá en la celda D2, la de la compra de Diet Coke en la celda D3, y así sucesivamente. Suponga que desea utilizar la función COUNTIF para calcular las frecuencias de estas celdas y se desea contar con un poco de ayuda de Excel.

Paso 1. Seleccione la celda D2.
Paso 2. Haga clic en f_x en la barra de fórmulas (o en la ficha **Formulas** en la barra de herramientas, y en **Insert Function**, en el grupo Function Library).
Paso 3. Cuando el cuadro de diálogo **Insert Function** aparezca:
Seleccione **Statistical** en la casilla **Or select a category box**.
Elija **COUNTIF** en el cuadro **Select a function box**.
Haga clic en **OK**.
Paso 4. Cuando el cuadro **Function Arguments** aparezca (figura 5):
Ingrese $A$2:$A$51 en el cuadro **Range**.
Ingrese C2 en el cuadro **Criteria** (en este momento el valor de la función aparecerá en la penúltima fila del cuadro de diálogo. Su valor es 19).
Haga clic en **OK**.
Paso 5. Copie la celda D2 en las celdas D3:D6.

La hoja de cálculo se verá como en la figura 6. La fórmula aparece en la hoja del fondo; la hoja de cálculo con los valores aparece al frente. En la hoja de la fórmula se observa que la función COUNTIF fue insertada en la celda D2. El contenido de esta celda se copia dentro de las celdas D3:D6. En la hoja de cálculo con los valores aparecen las frecuencias adecuadas según los cálculos.

Se ha ilustrado el uso de Excel para proporcionar apoyo al utilizar la función COUNTIF. El procedimiento es similar para todas las demás funciones. Esta capacidad es especialmente útil si no se sabe qué función utilizar o se olvida su nombre propio y/o sintaxis.

FIGURA 5 Cuadro de diálogo para proporcionar los argumentos de la función COUNTIF

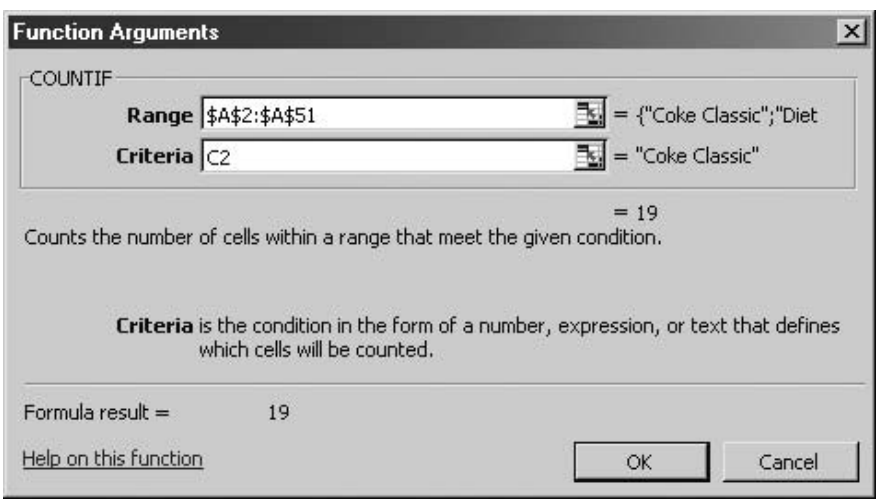


FIGURA 6 Hoja de cálculo de Excel en la que se muestra el uso de la función COUNTIF para elaborar una distribución de frecuencia

	A	B	C	D	E
1	**Brand Purchased**		**Soft Drink**	**Frequency**	
2	Coke Classic		Coke Classic	=COUNTIF($A$2:$A$51,C2)	
3	Diet Coke		Diet Coke	=COUNTIF($A$2:$A$51,C3)	
4	Pepsi		Dr. Pepper	=COUNTIF($A$2:$A$51,C4)	
5	Diet Coke		Pepsi	=COUNTIF($A$2:$A$51,C5)	
6	Coke Classic		Sprite	=COUNTIF($A$2:$A$51,C6)	
7	Coke Classic				
8	Dr. Pepper				
9	Diet Coke				
10	Pepsi				
45	Pepsi				
46	Pepsi				
47	Pepsi				
48	Coke Classic				
49	Dr. Pepper				
50	Pepsi				
51	Sprite				
52					

Nota. Las filas 11 a 44 están ocultas.

	A	B	C	D	E
1	**Brand Purchased**		**Soft Drink**	**Frequency**	
2	Coke Classic		Coke Classic	19	
3	Diet Coke		Diet Coke	8	
4	Pepsi		Dr. Pepper	5	
5	Diet Coke		Pepsi	13	
6	Coke Classic		Sprite	5	
7	Coke Classic				
8	Dr. Pepper				
9	Diet Coke				
10	Pepsi				
45	Pepsi				
46	Pepsi				
47	Pepsi				
48	Coke Classic				
49	Dr. Pepper				
50	Pepsi				
51	Sprite				
52					

Apéndice F Cálculo de los valores-*p* utilizando Minitab y Excel

Aquí se describe cómo se pueden utilizar Minitab y Excel para calcular los valores-*p* de los estadísticos z, t, χ^2 y F que se manejan en las pruebas de hipótesis. Como se analiza en el texto, los valores-*p* aproximados correspondientes a los estadísticos t, χ^2 y F sólo se pueden obtener mediante el uso de tablas. Este apéndice es útil para una persona que ha calculado manualmente estadísticos de prueba, o por otros medios, y desea utilizar software para calcular el valor-*p* exacto.

Usando Minitab

Minitab puede utilizarse para proporcionar la probabilidad acumulada relacionada con z, t, χ^2 y el estadístico de prueba F. Así que el valor-*p* en la cola inferior se obtiene directamente, mientras que en la cola superior se calcula restando 1 del valor-*p* de la cola inferior. El valor-*p* en las dos colas se obtiene duplicando el menor de los valores-*p* de las colas superior e inferior.

Estadístico de prueba *z* En la sección 9.3 se utiliza como ejemplo la cola menor de la hipótesis de prueba del Café Hilltop; el valor del estadístico de prueba es $z = -2.67$. Los pasos de Minitab utilizados para calcular la probabilidad acumulada correspondiente a $z = -2.67$ son los siguientes.

Paso 1. Seleccione el menú **Calc**.
Paso 2. Elija **Probability Distributions**.
Paso 3. Selecciona **Normal**.
Paso 4. Cuando el cuadro de diálogo Normal Distribution aparezca:
Seleccione **Cumulative probability**.
Ingrese 0 en el apartado **Mean**.
Ingrese 1 en el apartado **Standard deviation**.
Seleccione **Input Constant**.
Ingrese −2.67 en el apartado **Input Constant**.
Haga clic en **OK**.

Minitab proporcionará la probabilidad acumulada de 0.0038. Ésta es el valor-*p* para la cola inferior que se utilizó en la prueba de hipótesis de Hilltop Coffee.

Para una prueba de cola superior, el valor-*p* se calcula a partir de la probabilidad acumulada obtenida por Minitab como sigue.

$$\text{El valor-}p = 1 - \text{probabilidad acumulada}$$

Por ejemplo, el valor-*p* en la cola superior correspondiente a un estadístico de prueba de $z = -2.67$ es $1 - 0.0038 = 0.9962$, y para las dos colas correspondiente a un estadístico de prueba de $z = -2.67$ es el doble del mínimo de los valores-*p* de las colas inferior y superior; es decir, el valor-*p* para las dos colas correspondiente a $z = -2.67$ es $2(.0038) = 0.0076$.

Estadístico de prueba *t* El ejemplo del aeropuerto Heathrow de la sección 9.4 se utiliza para ilustrar esta prueba; el valor del estadístico de prueba es $t = 1.84$ con 59 grados de libertad. Los pasos de Minitab utilizados para calcular la probabilidad acumulada correspondiente a $t = 1.84$ son los siguientes.

Paso 1. Seleccione el menú **Calc**.
Paso 2. Elija **Probability Distributions**.

Paso 3. Elija **t**.
Paso 4. Cuando el cuadro de diálogo t Distribution aparezca:
Seleccione **Cumulative probability**.
Ingrese 59 en el cuadro **Degrees of freedom**.
Seleccione **Input Constant**.
Ingrese 1.84 en el cuadro **Input Constant**.
Haga clic en **OK**.

Minitab proporciona la probabilidad acumulada de 0.9646, y por tanto el valor-*p* en la cola inferior es = 0.9646. El ejemplo del aeropuerto Heathrow es una prueba de cola superior; el valor-*p* de la cola superior es 1 − 0.9646 = 0.0354. En el caso de una prueba de dos colas se utilizará el mínimo entre 0.9646 y 0.0354 para calcular el valor-*p* = 2(0.0354) = 0.0708.

Estadístico de prueba χ^2 Como ilustración se utiliza el ejemplo del metrobús de San Luis de la sección 11.1; el valor del estadístico de prueba es $\chi^2 = 28.18$, con 23 grados de libertad. Los pasos de Minitab utilizados para calcular la probabilidad acumulada correspondiente a $\chi^2 = 28.18$ son los siguientes.

Paso 1. Seleccione el menú **Calc**.
Paso 2. Elija **Probability Distributions**.
Paso 3. Elija **Chi-Square**.
Paso 4. Cuando el cuadro de diálogo Chi-Square Distribution aparezca:
Seleccione **Cumulative probability**.
Ingrese 23 en el cuadro **Degrees of freedom**.
Seleccione **Input Constant**.
Ingrese 28.18 en el cuadro **Input Constant**.
Haga clic en **OK**.

Minitab proporciona la probabilidad acumulada, 0.7909, el valor-*p* correspondiente a la cola inferior. El valor-*p* en la cola superior es = 1 − probabilidad acumulada, o 1 − 0.7909 = 0.2091. El valor-*p* para una prueba de dos colas es el mínimo del valor-*p* de las colas inferior y superior multiplicado por 2; por tanto, es 2(0.2091) = 0.4182. El ejemplo del metrobús de San Luis involucra la prueba de cola superior, así que el valor-*p* = 0.2091.

Estadístico de prueba *F* Como ilustración se utiliza el ejemplo de las escuelas de Dullus County de la sección 11.2; el estadístico de prueba es $F = 2.40$ con 25 grados de libertad en el numerador y 15 grados de libertad en el denominador. Los pasos de Minitab para calcular la probabilidad acumulada correspondiente a $F = 2.40$ son los siguientes.

Paso 1. Seleccione el menú **Calc**.
Paso 2. Elija **Probability Distributions**.
Paso 3. Elija **F**.
Paso 4. Cuando el cuadro de diálogo de F Distribution aparezca:
Seleccione **Cumulative probability**.
Ingrese 25 en el cuadro **Numerator degrees of freedom**.
Ingrese 15 en el cuadro **Denominator degrees of freedom**.
Seleccione **Input Constant**.
Ingrese 2.40 en el cuadro **Input Constant**.
Haga clic en **OK**.

Minitab proporciona la probabilidad acumulada y, por tanto, el valor-*p* en la cola inferior es 0.9594, y en la cola superior es 1 − 0.9594 = 0.0406. Debido a que el ejemplo de las escuelas de Dullus County es una prueba de dos colas, se utiliza el mínimo entre 0.9594 y 0.0406 para calcular el valor-*p* = 2(0.0406) = 0.0812.

Usando Excel

WEB archivo
p-Value

Las funciones y fórmulas de Excel pueden utilizarse para calcular los valores-*p* relacionados con los estadísticos de prueba *z*, *t*, χ^2 y *F*. Se proporciona una plantilla en los archivos de datos titulada valor-*p* para ser utilizada en el cálculo de estos valores. Con la plantilla, únicamente es necesario introducir el valor del estadístico de prueba y, si es necesario, los grados de libertad adecuados. Consulte la figura D.1 a medida que se lee la descripción de cómo utilizar la plantilla. Para los usuarios interesados en el uso de las funciones y fórmulas de Excel, sólo haga clic en la celda adecuada.

Estadístico de prueba *z* Como ilustración se utiliza la prueba de hipótesis de cola inferior de Hilltop Café en la sección 9.3; el valor del estadístico de prueba es $z = -2.67$. Para utilizar la plantilla del valor-*p* para esta hipótesis, simplemente introduzca −2.67 en la celda B6 (vea la figura F.1). Después aparecerán los valores-*p* para los tres tipos de pruebas de hipótesis. Para el Café Hilltop se utilizará el valor-*p* de cola inferior = 0.0038 en la celda B9. Para una prueba de cola superior se utilizará el valor-*p* en la celda 10, y para la prueba de dos colas en la celda B11.

Estadístico de prueba *t* Como ilustración se utiliza el ejemplo del aeropuerto Heathrow de la sección 9.4; el valor del estadístico de prueba es $t = 1.84$ con 59 grados de libertad. Para utilizar la plantilla del valor-*p* en esta prueba de hipótesis introduzca 1.84 en la celda E6 y 59 en la celda E7 (vea la figura F.1). Después aparecerán los valores-*p* de los tres tipos de prueba de hipótesis. El ejemplo del aeropuerto Heathrow se trata de una prueba de cola superior, así que

FIGURA F.1 Hoja de cálculo para calcular los valores-*p*

	A	B	C	D	E
1	**Computing *p*-Values**				
2					
3					
4	**Using the Test Statistic *z***			**Using the Test Statistic *t***	
5					
6	**Enter *z* -->**	−2.67		**Enter *t* -->**	1.84
7				***df* -->**	59
8					
9	***p*-value (Lower Tail)**	0.0038		***p*-value (Lower Tail)**	0.9646
10	***p*-value (Upper Tail)**	0.9962		***p*-value (Upper Tail)**	0.0354
11	***p*-value (Two Tail)**	0.0076		***p*-value (Two Tail)**	0.0708
12					
13					
14					
15					
16	**Using the Test Statistic Chi Square**			**Using the Test Statistic *F***	
17					
18	**Enter Chi Square -->**	28.18		**Enter *F* -->**	2.40
19	**df -->**	23		**Numerator df -->**	25
20				**Denominator df -->**	15
21					
22	***p*-value (Lower Tail)**	0.7909		***p*-value (Lower Tail)**	0.9594
23	***p*-value (Upper Tail)**	0.2091		***p*-value (Upper Tail)**	0.0406
24	***p*-value (Two Tail)**	0.4181		***p*-value (Two Tail)**	0.0812

se podrá utilizar el valor-*p* de la cola superior, que es = 0.0354, proporcionado en la celda E10 para la prueba de hipótesis.

Estadístico de prueba χ^2 Como ilustración se utiliza el ejemplo del metrobús de San Luis de la sección 11.1; el valor del estadístico de prueba es $\chi^2 = 28.18$, con 23 grados de libertad. Para utilizar la plantilla del valor-*p* en esta prueba de hipótesis, introduzca 28.18 en la celda B18 y 23 en la celda B19 (vea la figura F.1). Después aparecerán los valores-*p* de los tres tipos de pruebas de hipótesis. El ejemplo del metrobús de San Luis se trata de una prueba de cola superior, así que se utiliza el valor-*p* de cola superior = 0.2091 proporcionado en la celda B23 para la prueba de hipótesis.

Estadístico de prueba *F* Como ilustración se utiliza el ejemplo de las escuelas del condado de Dullus de la sección 11.2; el estadístico de prueba es $F = 2.40$, con 25 grados de libertad en el numerador y 15 grados de libertad en el denominador. Para utilizar la plantilla del valor-*p* en esta prueba de hipótesis, introduzca 2.40 en la celda E18, 25 en la celda E19 y 15 en la celda E20 (figura F.1). Después aparecerán los valores-*p* para los tres tipos de prueba de hipótesis. El ejemplo de las escuelas del condado de Dullus involucra una prueba de dos colas, así que se utiliza el valor-*p* para dos colas, que es 0.0812, proporcionado en la celda E24 para la prueba de hipótesis.

Índice analítico

Nota: Los números de página seguidos por una ***n*** indican una nota al margen o en la sección de notas y comentarios.

A

Agencias gubernamentales, 10-11
Ajuste de Bonferroni, 527-528
Ajustes estacionales, 836
Aleatorización, 508, 513*n*1
Alfa para eliminar, 739-740, 743*n*l
Alfa para ingresar, 739-740, 743*n*1
Alliance Data Systems, 561
Almacenamiento de datos, 17
American Military Standard Table (MIL-STD-105D), 929
American Society for Quality (ASQ), 904
American Statistical Association "Lineamientos éticos para la práctica estadística", 18-19
Análisis de datos exploratorios, 48-51, 109-114, 112*n*1
Análisis de decisiones
  árboles de decisión, 940-941
  con información muestral, 949-956
  estrategia de decisión, 951-954
  formulación del problema, 939-941
  tablas de pagos, 940
  toma de decisiones con probabilidades, 941-945
Análisis de regresión múltiple, 644, 692*n*2
Análisis de regresión, 562, 565*n*1, 618*n*1
  análisis de un problema mayor, 735-738
  autocorrelación y prueba de Durbin-Watson, 750-754
  determinación de cuándo agregar o eliminar variables, 729-732
  método de regresión múltiple para el diseño de experimentos, 745-748
  modelo lineal general, 714-725
  procedimientos de selección de variables, 739-743
  residual en el, 793
  solución por computadora, 600-601
Análisis de residuales, 605-614, 612*n*2
  del modelo de regresión múltiple, 676-677
  desviación estándar del residual *i*, 676
  detección de observaciones atípicas, 614-616, 678
  detección de observaciones influyentes, 616-618
  gráfica de probabilidad normal, 610-612
  gráfica de residuales contra $\hat{y}$, 607
  gráfica de residuales contra *x*, 606-607
  observaciones influyentes, 679
  residual de la observación *i*, 605
  residual estandarizado de la observación *i*, 676
  residuales estandarizados, 607-610
Análisis de varianza (ANOVA), 508-537, 513*n*3, 513*n*4
  diseño completamente aleatorizado y, 513-524
  diseño de bloques aleatorizado, 532-533
  diseño de experimentos y, 508-513
  para experimentos factoriales, 539-540
  resultados de computadora para el, 519-520
  supuestos para el, 510
Análisis estadístico, 17
Ancho de clase aproximado, fórmula para, 65
ANOVA, *Véase* análisis de varianza (ANOVA)
Aptitud para el uso, 905
Árboles de decisión, 940-941, 941*n*1, 942*n*2, 950-951
Área como medida de la probabilidad, 235-236
Asociación entre dos variables, medidas de, 115-124
Atributos de los planes de muestreo, 930*n*3
Autocorrelación, 750

B

Baldridge, Malcolm, 906
Baldridge National Quality Program (BNQP), 906
Banco de datos, 5
Barnett, Bob (Motorola), 906
Bell Labs, caso de 218
Bell Telephone Laboratories, 905
Bernoulli, Jakob, 208
Bloques, diseño de, 530, 531
Burke Marketing Services, Inc., 507
BusinessWeek, 2
Butler Trucking Company, un ejemplo, 646-648

C

Cálculo de índices de estacionalidad, 830-834, 837*n*1
Calidad total (TQ), 904
Casos a resolver
  Análisis de estadísticas de la PGA Tour, 758-759
  Aportaciones de exalumnos, 705
  Compensación para profesionales de ventas, 553-554
  Comportamiento ético de los estudiantes de negocios en la Universidad de Bayview, 397-398
  Consumer Research, Inc., 704-705
  Departamento del Transporte de Estados Unidos, 632-633
  Donaciones de exalumnos, 705, 633
  Escuelas de negocios de Asia-Pacífico, 139
  Estadísticas del PGA Tour, 633-635, 705-707
  Estrategia de defensa de una demanda, 969
  Gulf Real Estate Properties, 339-341
  Industria del cine, 72-73, 138-139
  Jueces del condado de Hamilton, 190-192

Medición del riesgo en el mercado bursátil, 631-632
Metropolitan Research, Inc., 341
Par, Inc., 441-442
Pelican Stores, 71-72, 137-138
Predicción del porcentaje de triunfos de la NFL, 708-709
Programa de entrenamiento de la fuerza aérea, 469
Pronóstico de pérdidas de ventas, 847-848
Pronóstico de ventas de alimentos y bebidas, 846-847
Quality Associates, Inc., 396-397
Rendimiento de combustible en los automóviles, 759-760
Revista Y*oung Profesional*, 338-339
Specialty Toys, 261-262
Transacciones del sitio web de Heavenly Chocolates, 139-141
Una agenda bipartidista para el cambio, 501-502
Wentworth Medical Center, 552-553
Causas comunes, 909
Causas imputables en el control estadístico de procesos, 909
Censo, 15
Cincinnatti Enquirer, 190
Citibank, 194
Clase de extremo abierto, 45*n*3
Clases, 39, 40
ancho de, 40
límites de, 40
número de, 39
punto medio de, 41
CME. *Véase* Cuadrado medio debido al error (CME)
CMR. *Véase* Cuadrado medio debido a la regresión (CMR);
CMTR. *Véase* Cuadrado medio debido a los tratamientos (CMTR)
Cociente de posibilidades de Odds, 688-691, 692*n*1
Coeficiente de confianza, 313
Coeficiente de correlación del producto-momento de Pearson, 119-120, 889*n*1
Coeficiente de correlación por rangos de Spearman, 887-889, 889*n*1
Software de hoja de cálculo y, 887-889, 889*n*1
Coeficiente de correlación, 119-120, 579-580
Coeficiente de determinación, 576-583, 579, 580*n*1, 692*n*2
Coeficiente de determinación múltiple, 654-655
Coeficiente de determinación múltiple ajustado, 655, 655*n*1
Coeficiente de variación, 99
Coeficientes, interpretación de los, 648-649
Colgate-Palmolive Company, 32
Combinaciones, 154
Complementos, 164, 165
Computadoras, 17
Confiabilidad del modelo, 18
Conglomerados o clusters, 298
Consecuencias, 939
Consistencia, 297
Constante de suavizamiento, 800, 801
Contabilidad, 3
Control de calidad, 905-908
Control estadístico de procesos, 908-920
causas imputables, 909
causas comunes, 909
gráfica $\bar{x}$, 909-915
gráfica *np*, 919-920
gráfica *p*, 917-919
gráfica *R*, 915-917
Correlación serial, 750
Covarianza, 115-119
Cravens, David W., 735
Criterio de aceptación, 924
Criterio de mínimos cuadrados, 567, 569*n*1, 645
Crosby, Philip B., 905
Cuadrado medio de la regresión (CMR), 588
Cuadrado medio debido a los tratamientos (CMTR), 514-515
Cuartiles, 91-92
Curva Bell, *Véase* también Curva normal, 238-240
Curva normal. *Véase* también Curva de Bell, 238-240
Curva característica de operación (OC), 925
Curvas de potencia, 385
Customer's Afternoon Letter, 772

D

Datos
aplicaciones de, 580*n*l
bimodal y multimodal, 89
fuentes de, 10-13
tipos de, 5-8
Datos agrupados, 125-127
media muestral para, 126
media poblacional para, 127
varianza muestral para, 126
varianza poblacional para, 127
Datos bimodales, 89
Datos categóricos o cualitativos, 7, 33-39
Datos cuantitativos, 7, 8*n*2, 33
límites de clase con, 45*n*2
resumen, 39-45
Datos de corte transversal, 7
Datos multimodales, 89
Deflactación de una serie, 773-775
DelGuzzi, Kristen, 190
Deming, W. Edwards, 905
Descomposición de series de tiempo, 829-837
ajustes estacionales, 836
cálculo de índices estacionales, 830-834
desestacionalización de una serie de tiempo, 834
modelo de descomposición aditiva, 829-830
modelo de descomposición multiplicativa, 830
modelos basados en datos mensuales, 837
patrón cíclico, 837
Desviación estándar estimada de b_1, 586
Desviación estándar poblacional (σ), 99, 310
Desviación estándar, 99, 204
de $\bar{p}$, 290
de $\bar{x}$, 280-281, 304-305
del residual *i*ésimo, 609
Desviación respecto de la media, 97
Diagrama de árbol, 152
Diagrama de Venn, 164
Diagramas de caja, 110-111, 112*n*1
Diagramas de dispersión, 57-59, 565

Diagramas de puntos, 41
Diagramas de tallo y hoja, 48-51
Diferencia de la suma de cuadrados de la regresión, 732*n*1
Diferencia mínima significativa de Fisher (LSD), 524-527
Diseño de bloques aleatorizado, 530-537, 535*n*1
Diseño de bloques completo, 534
Diseño de bloques incompleto, 534
Diseño de experimentos, 508-513
  Introducción al, y al análisis de varianza (ANOVA), 508-513
  método de regresión múltiple para el, 745-749
  recolección de datos, 509-510
Diseño de una muestra independiente, 426*n*2
Diseños de bloques aleatorizado, 508, 513-524
Distribución binomial
  para el muestreo de aceptación, 930*n*1
  valor esperado y varianza de la, 214-215
Distribución de frecuencia porcentual, 34, 41
Distribución de probabilidad, 197
Distribución de probabilidad binomial, 208
Distribución de probabilidad de Poisson, 218-220
  prueba de bondad de ajuste, 487-491
  relación entre la, y exponencial, 255
Distribución de probabilidad exponencial, 253-256, 256*n*1, 258
Distribución de probabilidad hipergeométrica, 221-223, 223*n*1
Distribución de probabilidad normal, 238-248
Distribución de probabilidad normal estándar, 240-245, 245-248
Distribución de probabilidad uniforme, 234-236
Distribución de probabilidad uniforme discreta, 199
Distribución *F*, 460, 464*n*1, 516
Distribución ji-cuadrada, 450-454
Distribución muestral binomial, 861*n*2
Distribución normal
  prueba de bondad de ajuste, 491-495
Distribución *t*, 316, 317
  clases, 39-41
  número de clases en una, 36*n*1
  suma de las, 36*n*2
Distribuciones de frecuencia, 33-34
Distribuciones de frecuencia acumulada, 43-44, 45*n*4
Distribuciones de frecuencia porcentual acumulada, 44
Distribuciones de frecuencia relativa acumulada, 44
Distribuciones de probabilidad discreta, 197-200
Distribuciones muestrales o de muestreo, 276-286
  de dos varianzas poblacionales, 460
  de b_1, 586
  de $(n-1)s^2/\sigma^2$, 450
  de $\overline{p}$, 289-293
  de $\overline{x}$, 278-279, 281-286
  muestreo con remplazo y, 270
  muestreo sin remplazo y, 269
Distribuciones sesgadas, 256*n*1
Doctrina de las probabilidades, La (Moivre), 238-240
Dow Chemical Company 904
Dow, Charles Henry, 772
dunnhumby, 643

E

Economía, 4
Ecuaciones de regresión, 563-564, 565*n*2
Ecuación de regresión estimada mediante el método de mínimos cuadrados, 580*n*1
Ecuación de regresión logística estimada, 685-687
Ecuación de regresión múltiple, 644
Ecuación de tendencia cuadrática, 814-816
Ecuación de tendencia exponencial, 816
Ecuaciones de regresión estimada, 563-565, 567, 594, 612*n*2
Ecuaciones de regresión múltiple estimada, 644-645, 665-666
Eficiencia relativa, 296
Elaboración de pronósticos usando el método de promedios móviles, 797-800, 804*n*2
Elementos, 5-6,
Eliminación hacia atrás, procedimiento de, 741
Encuesta por muestreo, 15,
Error de pronóstico, 792
Error estándar
  de dos muestras aleatorias independientes, 409
  de $\overline{p}_1 - \overline{p}_2$, 430
  de $\overline{p}_1 - \overline{p}_2$ cuando $p_1 = p_2 = p$, 432
Error estándar de estimación, 585
Error estándar de la proporción, 290
Error tipo I, 353-355, 355*n*1
  tasa de, por comparación, 527
  tasa de, por experimentación, 527
Error tipo II, 353-355, 355*n*1
  probabilidad de cometer un, 382-385
Errores de redondeo, 100*n*3
Errores en la adquisición de datos, 13
Errores en los datos, 681*n*1
Escala de intervalo, 6
Escala de razón, 6
Escala nominal, 6
Escala ordinal, 6
Escalas de medición, 6-7
Espacio muestral, 150
Estacionalidad y tendencia, 820-826
  estacionalidad sin tendencia, 820-823
  modelos basados en datos mensuales, 825-826
Estadística, 3
Estadística descriptiva, 13-15, 127*n*1
Estadístico de prueba, 357-358
  en las pruebas de hipótesis para la media poblacional: σ conocida, 358
  para la bondad de ajuste, 475
  para la igualdad de k medias poblacionales, 516
  para las pruebas ji-cuadrada, 483*n*1
  para pruebas de hipótesis acerca de $\mu_1 - \mu_2$: σ_1 y σ_2 conocidas, 411
  para pruebas de hipótesis acerca de $p_1 - p_2$, 432
  para pruebas de hipótesis acerca de dos varianzas poblacionales, 461
  para pruebas de hipótesis acerca de una varianza poblacional, 454
  para pruebas de hipótesis con muestras pareadas, 425
  para pruebas de hipótesis $\mu_1 - \mu_2$: σ_1 y σ_2 desconocidas, 417-419
Estadístico de prueba *F*, 461
Estadístico *F*, 732*n*1

Estadístico G, 692*n*l
Estadístico muestral, 87, 273-274
Estados de la naturaleza, 939
Estimación conjunta o dentro de los tratamientos de σ^2, 512
Estimación de la varianza poblacional dentro de los tratamientos, 515-516
Estimación de la varianza poblacional entre tratamientos, 514-515
Estimación de σ^2 dentro de los tratamientos, 512
Estimación de σ^2 entre tratamientos, 511-512, 521*n*2
Estimación por intervalo, 309, 310-314, 594
  de la diferencia entre dos medias poblacionales, 430
  de una varianza poblacional, 450-454
  procedimientos de, 322-323
Estimación por intervalo, 314*n*1, 409
  de la diferencia entre dos medias poblacionales: σ_1 y σ_2 conocidas, 410
  de la diferencia entre dos medias poblacionales: σ_1 y σ_2 desconocidas, 416
  de la media poblacional: σ conocida
  de la proporción poblacional, 329, 330
  de $\mu_1 - \mu_2$, 407-412, 415
  para pruebas de hipótesis y relación con la, 366-367
Estimación por intervalo de la media poblacional: σ, 313
Estimación por intervalo de la proporción poblacional, 329, 330
Estimación puntual, 273-275, 274, 594
Estimador combinado de p, 432
Estimador puntual, 87, 274
  de la diferencia entre dos medias poblacionales, 409
  de la diferencia entre dos proporciones poblacionales, 430
  e insesgadez, 295-296
  y consistencia, 297
  y eficiencia, 296-297
Estimadores insesgados, 295-296
Estrategia de decisión, 951-954
Estudio observacional, 12, 507
  pruebas para la igualdad de k medias poblacionales: un, 520-521
Estudios estadísticos, 11-13
Estudios experimentales u observacionales, 11-12, 507
Eventos aleatorios, 939
Eventos excluyentes, 175*n*1
Eventos independientes, 174, 175, 175*n*1
Eventos mutuamente excluyentes, 168, 175*n*1
Eventos, 160-162, 162*n*1, 164, 174
Exactitud del pronóstico, 792-797, 799, 800, 802
  cuadrado medio debido al error (CME), 793
  error absoluto medio (EAM), 793
  error porcentual absoluto medio (EPAM), 794
Excel
  Análisis de regresión con, 638-640
  análisis de varianza con, 555-557
  correlación de rango de Spearman, 900-901
  diagrama de dispersión, 81-84
  diferencia entre dos medias poblacionales: σ_1 y σ_2 conocidas, 444-445
  diferencia entre dos medias poblacionales: σ_1 y σ_2 desconocidas, 445
  diferencia entre la media de dos poblaciones con muestras pareadas, 445-446
  diseño completamente aleatorizado, 555
  diseño de bloques aleatorizado, 555
  distribución de frecuencia, 75-76, 77-79
  distribuciones de probabilidad continua con, 263-264
  distribuciones de probabilidad discretas con, 230-231
  elaboración de pronósticos con, 851-852
  estadística descriptiva usando, 143-146
  estimación por intervalo usando, 343-346
  experimento factorial, 556-557
  gráficas de barras, 76-77
  gráfico dinámico, 77-79
  herramienta Descriptive Statistics, 145-146
  histogramas, 77-79
  inferencias acerca de dos poblaciones usando, 444-446
  interpretación de los resultados de la ecuación de regresión estimada, 639-640
  interpretación de los resultados de los estadísticos de regresión, 640
  interpretación de los resultados del ANOVA, 640
  media poblacional: σ conocida, 343, 400-401
  media poblacional: σ desconocida, 344, 402-403
  métodos no paramétricos con, 899-900
  muestreo aleatorio con, 306-307
  PrecisionTree, 970-974
  promedios móviles, 851
  proporción poblacional, 345-346, 403-404
  proyección de tendencia, 852
  prueba de bondad de ajuste, 503, 504
  prueba de independencia, 503, 505
  prueba de signos, 899-900
  pruebas de hipótesis con, 400-404
  regresión múltiple con, 709-710
  suavizamiento exponencial, 851-852
  tabla dinámica, 77-79
  tabulación cruzada, 79-81
  uso de funciones de, 143-145
  varianzas poblacionales con, 470-471
  StatTools de, *Véase* StatTools, 17
Experimento binomial, 208-209
Experimento de un solo factor, 508
Experimento factorial, 537-544
  estadísticos F, 539-542
  procedimiento ANOVA, 539
Experimentos, 150, 158*n*1
Experimentos aleatorios, 158*n*1
Experimentos doble ciego, 513*n*2
Experimentos estadísticos, 158*n*1
Extremos de clases, 45*n*3

F

$F(x)$, 234
Factor de corrección de continuidad, 251
Factor de corrección de una población finita, 280
Factor de interés, 531
Factores, 508
Feigenbaum, A. V., 905
Finanzas, 4

Fisher, Ronald Alymer, Sir, 508
Food Lion, 309
Forma de la distribución, 102-103
Fórmulas de mínimos cuadrados, 635-636
Frecuencia relativa,
  distribuciones de, 34-36, 39-41
  fórmula para la, 65
Frecuencias, 13*t*1.4
Frecuencias de porcentaje, 13*f*1.4
Función de densidad de probabilidad, 234, 237*n*1
Función de densidad de probabilidad exponencial, 258
Función de densidad de probabilidad normal, 239, 258
Función de densidad de probabilidad uniforme, 234, 258
Función de la probabilidad binomial, 209, 212
Función de probabilidad, 197
Función de probabilidad de Poisson, 218, 488
Función de probabilidad hipergeométrica, 221-222
Función de probabilidad uniforme discreta, 199
Funciones de probabilidad discreta, 198

G

Galton, Francis, Sir, 562
Garantía de la calidad, 908
Gauss, Carl Friedrich, 567
Gosset, William Sealy, 316
Grados de creencia, 156
Grados de libertad, 316, 317, 319, 416, 535*n*1
Grados de libertad del error, 535*n*1
Gráfica $\bar{x}$, 909, 920*n*1
  proceso de la media y la desviación estándar conocida, 910-912
  proceso de la media y la desviación estándar desconocida, 912-915
Gráfica circular o de pastel, 35-36
Gráfica de barras, 14*f*1.5, 34-36, 45*n*1
Gráfica de residuales, 606, 612*n*1
  contra *x*, 606-607
  contra *y*, 607
Gráfica *np*, 910, 919-920, 920*n*2
Gráfica *p*, 910
Gráfica *R*, 910, 915-917, 920*n*1
Gráficas de control, 909-910
  gráfica $\bar{x}$, 910-915
  gráfica *np*, 919-920
  gráfica *p*, 917-919
  gráfica *R*, 915-917
  interpretación de las, 920
Gráficas de probabilidad normal, 610-612, 612*n*1
Gráficas de series de tiempo, 786-792

H

Hipótesis alternativa, 349
  como hipótesis de investigación, 350-351
Hipótesis nula, 349-353
Histograma, 14*f*1.6, 41-43, 45*n*1

I

Índice Baldridge, 906
Índice de Laspeyres, 767
Índice de Paasche, 767
Índices de precios
  deflactación de una serie por, 773-775
  índice de precios al consumidor (IPC), 771
  índice de precios al productor (IPP), 771
  promedios Dow Jones, 772
  selección de artículos, 777
  selección de un periodo base, 777
  variaciones en la calidad, 777-778
Índice de precios agregado ponderados, 766
Índice de precios al consumidor (IPC), 764, 771
Índice de producción industrial, 779
Indice industrial Down Jones (DJIA), 772
Índices de cantidad, 778-779
Índices de precios agregados, 765-767
  cálculo de precios relativos, 769, 770
Inferencia estadística, 15-16
Influencia de la observación *i*, 617, 676
Información muestral, 949
  valor esperado de la (VEIM), 954-956
Ingeniería de la calidad, 908
Instituto de Normas y Tecnología (NIST) del Departamento de Comercio de Estados Unidos, 906
Interacciones, 538-539, 718-720
Intersección de dos eventos, 166
Intervalo de confianza, 313, 594
  estimaciones por, 323*n*2
  para β_1, 587-588
  para el valor medio de *y*, 595-596
Intervalo de predicción para un solo valor de y, 596-598
Intervalo de predicción, 594
Intervalos de distancia o longitud, 220
Intervalos de tiempo
  distribución de probabilidad de Poisson y, 218-220
Inversión en acciones y fondos de acciones, 100*n*2
Investigación sobre la probabilidad de los veredictos en materia penal y civil (Poisson), 218
Ishikawa, Karou, 905
ISO 9000, 906

J

John Morrell & Company, 349
Juran Joseph, 905

K

k medias poblacionales, 513*n*3

L

Ley de la adición, 165-166
Ley de la multiplicación, 174-175
Límite de calidad promedio de salida (AOQL), 930*n*2
Límite de control superior (UCL), 910
Límites de clase, 45*n*2
Límites de control inferior (LCL), 910
Línea de tendencia, 57-59
Lineamientos éticos, 18-19
Lineamientos éticos para la práctica estadística (ASA), 18-19
Logit, 691
Logit estimado, 691
Lote, 922, 924

M

Malcolm Baldrige National Quality Award, 906
Marco, 267
Margen de error, 309, 310-314, 323*n*1, 331*n*1
Marketing, 4
MeadWestvaco Corporation, 266
Media, 14-15, 87-88, 124-125, 219
Media muestral, 126, 267, 297*n*1, 521*n*1
Media muestral general, 511
Media poblacional
  estimación del tamaño de la muestra cuando σ conocida, 310-314
  inferencia acerca de la diferencia entre muestras pareadas, 423-425
  inferencias acerca de la diferencia entre σ_1 y σ_2 conocidas, 407-412
  inferencias acerca de la diferencia entre σ_1 y σ_2 desconocidas, 415-419
  para datos agrupados, 127
  σ conocida, 310-314
  σ desconocida, 316-323
Media poblacional: σ conocida
  estimación por intervalo, 310-314
  margen de error, 310-314
  prueba de una cola, 356-361
Media poblacional: σ desconocida
  estimación por intervalo, 317-320
  margen de error, 317-320
  prueba de dos colas, 372-373
  pruebas de hipótesis y, 370-374
Media ponderada, 124-125
Media recortada, 92*n*1
Mediana, 88-89
Medida de la distancia de Cook, 679-681, 681*n*2
Medidas de asociación entre dos variables, 115-124
Medidas de localización central, 297*n*1
Medidas de posición o localización, 87-92
Medidas de variabilidad, 95-102
Mejora continua, 909
Método de elaboración de pronósticos usando promedios móviles ponderados, 800
Método de los mínimos cuadrados, 565-575, 569*n*1, 645-649
Método de pronóstico de regresión de tendencia lineal, 807-812, 817*n*1
Método de suavizamiento exponencial, 800-804, 804*n*2
Método del valor crítico, 360-361
Método del valor esperado, 941-943
Método del valor-*p*, 358-360
Métodos de distribución libre, 857
Métodos de elaboración de pronósticos
  estacionalidad y tendencia, 820-829
  promedios móviles, 797-800
  promedios móviles ponderados, 800
  proyección de tendencia, 807-820
  suavizamiento exponencial, 800-804
Métodos no paramétricos, 857
Prueba de Kruskal-Wallis, 882-884
  coeficiente de correlación por rangos de Spearman, 887-889
  prueba de Mann-Whitney-Wilcoxon (MWW), 871-882
  prueba de rangos con signos de Wilcoxon, 865-871
  prueba de signos, 857-865, 861*n*1
Métodos paramétricos, 856
Minería de datos, 17-18
Minitab, 17
  alpha to enter, 739-740
  análisis de regresión con, 637-638
  análisis de varianza, 554-555
  correlación por rangos de Spearman, 899
  covarianza y correlación, 143
  descomposición de una serie de tiempo, 850-851
  diagrama de caja, 143
  diagrama de puntos, 73
  diagrama de tallo y hoja, 74
  diagramas de dispersión, 74
  diferencia entre dos medias poblacionales con muestras pareadas, 443
  diferencia entre dos medias poblacionales: σ_1 y σ_2 desconocidas, 442-443
  diferencia entre dos proporciones poblacionales, 443-444
  diseño completamente aleatorizado, 554
  diseño de bloques aleatorizado, 554
  distribuciones de probabilidad continua con, 262-263
  distribuciones de probabilidad discreta con, 230
  elaboración de pronósticos con, 848-851
  estadística descriptiva usando, 142-143
  estimación por intervalo con, 341-343
  experimento factorial, 554-555
  gráficas de control con, 935
  histograma, 73-74
  inferencias acerca de dos poblaciones usando, 442-444
  media poblacional: σ conocida, 341-342, 398-399
  media poblacional: σ desconocida, 342, 399
  métodos no paramétricos con, 896-899
  muestreo aleatorio con, 306
  procedimiento de eliminación hacia atrás utilizando, 761
  procedimiento de los mejores subconjuntos usando, 761
  procedimiento de selección hacia adelante utilizando, 761
  procedimiento por pasos usando, 760
  procedimientos de selección de variables con, 760-761
  promedios móviles, 848-849
  proporción poblacional, 342-343, 399-400
  proyección de tendencia, 849-850
  prueba de bondad de ajuste, 502
  pruebas de hipótesis con, 398-400
  prueba de independencia, 503
  prueba de Kruskal-Wallis, 898-899
  prueba de Mann-Whitney-Wilcoxon, 898
  prueba de rangos con signo de Wilcoxon con muestras pareadas, 897-899
  prueba de signos para una prueba de hipótesis acerca de una mediana poblacional, 896-897
  prueba de signos para una prueba de hipótesis con muestras pareadas, 897
  regresión logística con, 710
  regresión múltiple con, 708-709
  suavizamiento exponencial, 849
  suavizamiento exponencial lineal de Holt, 850

tabulación cruzada, 74-75
uso de, para presentaciones tabulares y gráficas, 73-75
varianzas poblacionales con, 470
Moda, 89
Modelo de descomposición multiplicativa, 830
Modelo de regresión, 562, 743*n*3
Modelo de regresión múltiple, 644, 657
Modelo de segundo orden con una variable predictora, 715
Modelo lineal general, 714-729
interacción, 718-720
modelado de relaciones curvilíneas, 714-717
modelo de segundo orden con una variable predictora, 715
modelo simple de primer orden con una variable predictora, 714
modelos no lineales que son intrínsecamente lineales, 724-725
transformaciones que involucran la variable dependiente, 720-724
Modelo simple de primer orden con una variable predictora, 714
Modelos de descomposición aditiva, 829-830
Modelos estadísticos, 18
Moivre, Abraham de, 238
Monsanto Company, 713
Motorola, Inc., 906
Muestra aleatoria, 158*n*2, 270, 271*n*1
Muestra, 15, 271*n*1
Muestras aleatorias independientes simples, 407
Muestras aleatorias simples, 271*n*1, 271*n*2, 271-272*n*3
población finita, 268-270
Muestras pareadas, 423, 426*n*1, 426*n*2
prueba de rangos con signo de Wilcoxon, 865-871
Muestras probabilísticas, 271*n*2, 513*n*1
Muestreo aleatorio estratificado, 297-298, 300*n*1
Muestreo de aceptación, 922-931
cálculo de la probabilidad de aceptación de un lote, 924-927
función de probabilidad binomial, 925
KALI, Inc., ejemplo de, 924
planes de muestreo múltiple, 930
selección de un plan de, 928-929
Muestreo de conveniencia, 299, 300*n*1
Muestreo de criterio, 209, 300*n*1
Muestreo por conglomerados, 300*n*1
Muestreo probabilístico, 299, 300*n*1
Muestreo sistemático, 298-299, 300*n*1
Multicolinealidad, 662, 663*n*1

N

Nivel de calidad aceptable (AQL), 930*n*2
Nivel de calidad de indiferencia (IQL), 930*n*2
Nivel de calidad rechazable (RQL), 930*n*2
Nivel de confianza, 313
Nivel de significancia, 354
Nodos, 940
Nodos aleatorios, 940
Nodos de decisión, 940
Norris, Electronics, 15-16, 19
Números índice
cálculo de un índice de precios agregado a partir de precios relativos, 769-779
deflactación de una serie mediante índices de precios, 773-775
índice de precios agregado, 765-767
índices de cantidad, 778-779
índices de precios, 771-773
precios relativos, 765

O

Observación 6, 8*n*1
Observaciones atípicas, 106, 614-616, 678-679, 681*n*1
Observaciones influyentes, 616-618, 679, 681*n*1
uso de la medida de la distancia de Cook para identificar, 679-681
Occupational Health Clinic de Nevada, 785
Oceanwide Seafood, 149
Oficina de Estadísticas Laborales, Departamento del Trabajo de Estados Unidos, 764
Ohio Edison Company, 938
Ojiva, 44-45
Organización Internacional de Estandarización (ISO), 906

P

Pagos, 940
Paradoja de Simpson, 56-57
Parámetros, 268
Parámetros poblacionales, 87
Partición, 518
Patrón cíclico, 789-791
Patrón de tendencia, 788
Patrón estacional, 788-789
Patrón horizontal, 786-788
Patrones de tendencia y estacional, 789
deflactación mediante índices de precios, 773-775
gráficas de, 9*f*1.2
Patrones de una serie de tiempo, 786-792
cíclico, 789-791
patrón de tendencia, 788
patrón estacional, 788-789
patrón horizontal, 786-788
patrones de tendencia y estacional, 789
selección de un método de elaboración de pronósticos, 791-792
Pearson, Karl, 562
Percentiles, 90-91
Permutaciones, 154-155
Plan de muestreo sencillo, 930
Planes de muestreo de variables, 930*n*3
Planes de muestreo múltiple, 930
Población, 15
Población muestreada, 267
Población multinomial, 474
Población objetivo, 275
Poblaciones sesgadas, 323*n*2
Poisson, Simeón, 218
Posibilidades a favor de que ocurra un evento, 688
Potencia, 385
Precios relativos, 765, 769-770

PrecisionTree (complemento de Excel), 970-974
Probabilidad condicional, 171-175, 960
Probabilidad previa, 178, 949
Probabilidad, 150
  condicional, 171-175
  conjunta, 172
  de éxito, 215*n*1, 215*n*2
  marginal, 172
  método clásico de asignación de, 155-156, 162*n*1
  método de frecuencia relativa, 156
  método subjetivo (de asignación), 156-155
  posterior, 178
  previa, 178
Probabilidades binomiales
  aproximación normal, 250-252
  tablas de, uso de, 213-214, 215*n*1, 215*n*2
Probabilidades conjuntas, 172, 962
Probabilidades marginales, 172
Probabilidades posteriores, 178, 949
Problema de Grear Tire Company, 246-248
Problema de Martin Clothing Store, 209-213
Problema EAI, 283
Procedimiento de regresión por pasos, 739-740, 743*n*1
Procedimiento de Turkey, 528
Procedimientos de comparación múltiple
  diferencia mínima significativa de Fisher (LSD), 524-527
  tasas de error tipo I, 527-528
Procedimientos de selección de variables
  alpha to remove, 739-740
  eliminación hacia atrás, 741
  regresión de los mejores subconjuntos, 741-742
  regresión por pasos, 739-740
  selección hacia adelante, 740-741
Proceso de Bernoulli, 208
Procter & Gamble, 233
Producción, 4
Promedio o media, 14-15
Promedios Dow Jones, 772
Pronósticos, 785
Proporción poblacional, 328-331, 331*n*1
  aproximación normal a la distribución de muestreo de una, 328
  estadístico de prueba en las pruebas de hipótesis para la, 378
  estimación por intervalo de una, 329
  estimación por intervalo para $p_1 - p_2$, 429-431
  inferencias acerca de la diferencia entre dos, 429-433
  pruebas de hipótesis y, 376-379
  tamaño de la muestra para una estimación del intervalo de la, 330
Proyección de tendencia
  regresión de tendencia lineal, 807-812
  regresión de tendencia no lineal, 814-816
  suavizamiento exponencial lineal de Holt, 812-814
Prueba de bondad de ajuste, 476-477
  distribución de Poisson, 487-491
  distribución multinomial, 476-477
  distribución normal, 491-495
  estadístico de prueba para, 475
Prueba de bondad de ajuste para una distribución poblacional multinomial, 476-477
Prueba de cola superior, 356, 361, 461
Prueba de dos colas, 362-367
  cálculo del valor-*p* en una, 364
  media poblacional: caso σ conocida, 362-365
  media poblacional: caso σ desconocida, 372-373
  método del valor crítico, 364
  método del valor-*p*, 363-364
Prueba de hipótesis, 861*n*1
  acerca de $\mu_1 - \mu_2$, 410-412, 417-419
  acerca de $p_1 - p_2$, 431-433
  acerca de una media poblacional, 857-861
  con muestras pareadas, 862-863
  errores tipo I y tipo II, 353-356
  hipótesis nula y alternativa, 349-353
  media poblacional: σ desconocida, 370-376
  método del intervalo de confianza, 366
  para la varianza poblacional, 454-457
  pasos en la, 365
  proporción poblacional, 376-381
  prueba de una cola, 356-361, 371-372
  pruebas de dos colas, 362-365
  relación entre estimación por intervalo y, 366-367
  y toma de decisiones, 381-382
Prueba de hipótesis para las varianzas de dos poblaciones, 461
  para pruebas de hipótesis con muestras pareadas, 425
Prueba de independencia, 479-483
Prueba de cola inferior, 356, 361
Prueba de la igualdad de *k* medias poblacionales, 517, 520-521
Prueba de Mann-Whitney-Wilcoxon (MWW), 871-882, 878*n*1
Prueba de rangos con signo de Wilcoxon, 865-868, 868*n*1, 868*n*2
Prueba de rangos múltiples de Duncan, 528
Prueba de significancia, 585-591
  usando correlación, 636-637
Prueba de significancia, 585-591, 591*n*1, 591*n*3, 636-637, 658-663, 687
Prueba de signos, 857-861, 861*n*2
  prueba de hipótesis acerca de una mediana poblacional, 857-863
  prueba de hipótesis con muestras pareadas, 862-863
Prueba de una cola, 371-372, 475
Prueba de Durbin-Watson, 751
Prueba *F*, 516, 588-590
  para modelos de regresión múltiple, 658-661
Prueba de Kruskal-Walis, 882-884, 884*n*1
Prueba *t*, 586, 661-662
  modelos de regresión múltiple para la significancia individual, 661-662
  significancia para la regresión lineal simple, 587
Prueba *t*, 586-587
Prueba *z*, 692*n*1
Pruebas de estrés para controladores de tráfico aéreo, 531-532
Pruebas de hipótesis acerca de μ_1 y μ_2, 417-419
Puntos de gran influencia, 617
Punto de la muestra, 150
Puntos normales, 610-612

R

Rango, 96
Rango intercuartílico (RIC), 96-97
Registros internos de una empresa, 10
Regla empírica, 105-106
Regla para el rechazo en una prueba de cola inferior: método del valor crítico, 361
Regla para el rechazo usando el valor-*p*, 360
Reglas para el cálculo
  para combinaciones, 154
  para experimentos de pasos múltiples, 151
  para permutaciones, 154-155
Regresión de corte transversal, 786
Regresión de los mejores subconjuntos, 741-742
Regresión de series de tiempo, 786
Regresión de tendencia no lineal, 814-816
Regresión lineal simple, 562, 565*n*2
  prueba *F* de significancia en el caso de la, 589
Regresión logística, 683-691, 692*n*2
Regresión simple, 692*n*2
Replicación, 509, 538
Residual, 793
Residual estandarizado de la observación *i*, 610
Residual *i*th, 576
Residuales eliminados estudentizados, 678-679
Residuales estandarizados, 607-610
Resumen de cinco números, 109-110
Reynolds, Inc., 714-717
Riesgo del consumidor, 923
Riesgo del productor, 923

S

Selección de una muestra, 268-271
  muestra aleatoria, 270
  muestreo con remplazo, 270
  muestreo de una población infinita, 270-271
  muestreo sin remplazo, 269
Selección hacia adelante, 740-741
Serie de tiempo, 786-792
Serie de tiempo desestacionalizada, 834-835, 837*n*2
Series de tiempo estacionarias, 787, 804*n*2
Series de tiempo no estacionarias, 804*n*2
Sesgadez, 102-103, 256*n*1, 323*n*2
Shewhart, Walter A., 905
Significancia estadística frente a significancia práctica, 591*n*2
Six Sigma, 906-908
  límites y defectos por millón de oportunidades (dpmo), 907-908
Small Fry Design, 86
Software de hoja de cálculo, 804*n*1
Software, 17, 18
Sofware estadístico, 100*n*1, 272*n*4
StatTools
  análisis de regresión con, 640-641
  análisis de regresión múltiple con, 711
  análisis de diseño completamente aleatorizado usando, 557-559
  cómo empezar a usar, 28-30
  covarianza y correlación, 147
  desviación estándar poblacional simple, 471
  determinación del tamaño de la muestra, 346-347
  diagramas de caja, 147
  diagramas de dispersión, 84
  elaboración de pronósticos con, 852-854
  estadística descriptiva, 146-147
  estimación por intervalo con, 346-347
  estimación por intervalo de la media poblacional: caso de σ desconocida, 346
  estimación por intervalos de μ_1 y μ_2, 446
  gráficas de control al utilizar, 935-936
  histograma, 84
  inferencias acerca de dos poblaciones usando, 446-447
  inferencias acerca de la diferencia entre dos medias poblaciones: muestras pareadas, 447
  media poblacional: σ desconocida, 404-405
  métodos no paramétricos con, 901-902
  muestreo aleatorio con, 307
  procedimiento de selección de variables con, 761-762
  promedios móviles, 852-853
  prueba de Mann-Whitney-Wilcoxon, 901-902
  prueba de rangos con signo de Wilcoxon para muestras pareadas, 901
  pruebas de hipótesis acerca de μ_1 y μ_2, 446-447
  pruebas de hipotésis con, 404-405
  suavizamiento exponencial, 853
  suavizamiento exponencial lineal de Holt, 853-854
  uso de, para presentaciones tabulares y gráficas, 84
Suavizamiento exponencial lineal de Holt, 812-814, 817*n*1
Suma de cuadrados debido a la regresión (SCR), 577
Suma de cuadrados debido a los tratamientos (SCTR), 515
Suma de cuadrados debido al error (SCE), 515-516, 576
Suma de los cuadrados de las desviaciones (diferencias), 566
Suma total de cuadrados (STC), 577
Supuesto de estacionariedad, 209
Supuestos del modelo
  acerca del término del error ϵ en el modelo de regresión, 583, 584
  intervalo de confianza para β_1, 587-588
  para el modelo de regresión, 584, 585
  prueba *F*, 588-590
  prueba *t*, 586

T

Tabla de ANOVA, 518-519, 589-590
Tabla de contingencia, 480
Tablas de pagos, 940
Tabulaciones cruzadas, 53-55
Taguchi, Genichi, 905
Tamaño de la muestra
  determinación del, 325-327
  en una prueba de hipótesis para la media poblacional, 387-390
  muestra pequeña, 320-322
  observaciones atípicas y, 320
  para una estimación por intervalo de la media poblacional, 326

para una estimación por intervalo de la proporción poblacional, 330
relación entre el, y la distribución de muestreo de $\bar{x}$ 285-286
sesgo y, 320
Tasa de error tipo I por experimentación, 527-528
Tasa de error tipo I por comprobación, 527
Teorema de Bayes, 157*n*, 178-182, 183*n*1, 183*n*2
cálculo de probabilidades mediante el, 960-963
caso de dos eventos, 181
método tabular, 182
Teorema de Chebyshev, 104-105, 106-107*n*1
Teorema del límite central, 281-283, 286*n*2
Thearling, Kurt, 17
Tolerancia de porcentaje de defectuos en el lote (LPTD), 930*n*2
Toma de decisiones, 381-382, 941-945
Transformación logit, 691
Tratamiento, 508

U

U.S. Food and Drug Administration (FDA), 407
U.S. Government Accountability Office (GAO), 449
UEPS (último en entrar, primero en salir) método de valuación de inventarios, 309
Unidades experimentales, 508
Unión de dos eventos, 165
United Way, 473

V

Validez de los datos, 107*n*2
Valor crítico, 360
Valor esperado (VE), 942, 943-945
Valor esperado, 202-203
de $\bar{x}$, 279-280, 304
de $\bar{p}$, 289-290
de la información muestral (VEIM), 954-956
de la información muestral, 954-956
varianza de la distribución binomial, 214-215
Valor planeado, 326
Valor-*p*, 358, 367*n*1
Valor *z*, 103-104, 106
Variable aleatoria normal estándar, 245, 258
Variable cuantitativa, 7
Variable de respuesta, 508
Variable dependiente, 562, 720-724
Variable indicadora, 669
Variables, 5-6
aleatorias 194-196
determinación de cuándo agregar o eliminar, 729-732
uso de valores-*p* y, 732
Variables aleatorias, 194-196, 196*n*1
Variables aleatorias continuas, 196
Variables aleatorias discretas, 195
Variables categóricas, 7
Variables ficticias (*dummy*), 669
Variables independientes cualitativas o categóricas, 668-673
Variables independientes, 508, 562, 662, 663*n*1, 668-673, 743*n*2
Varianza, 97-99, 203-204
de la distribución binomial, 214-215
distribución de probabilidad de Poisson y, 219
Varianza muestral, 97, 100*n*4, 126
Varianza muestral combinada, 419*n*1
Varianza poblacional, 97
estimación de la, dentro de los tratamientos, 515-516
estimación de la, entre tratamientos, 514-515
inferencias acerca de una, 450-457
para datos agrupados, 127
pruebas de hipótesis y, 454-457
Varianzas de dos poblaciones
distribución de muestreo de, 460
inferencias acerca de las, 460-464
pruebas de hipótesis de una cola, 461

W

West Shell Realtors, 856
Western Electric Company, 905

El propósito de este libro es proporcionar al lector una introducción conceptual al campo de la estadística. Su orientación se dirige a las aplicaciones y fue escrito tomando en cuenta las necesidades de los lectores que no cuentan con conocimientos profundos de matemáticas.

Las aplicaciones del análisis de datos y la metodología estadística son parte integral de la organización y presentación del material. El análisis y desarrollo de cada técnica se presentan en el escenario de una aplicación que permite comprender los resultados estadísticos.

Cambios en la 11a. ed.

- **Actualización del capítulo 18 *Análisis de series de tiempo y elaboración de pronósticos*** Este capítulo fue reescrito por completo considerando el uso de patrones en una gráfica de serie de tiempo para seleccionar un método de elaboración de pronósticos apropiado.
- **Actualización del capítulo 19 *Métodos no paramétricos*** Se contrasta cada método no paramétrico con su contraparte paramétrica y se explica que se requieren algunos supuestos para el procedimiento no paramétrico.
- **Complemento StatTools® para Excel** StatTools® es un complemento (*add-in*) comercial de Excel 2007, desarrollado por Palisades Corporation, que amplía la variedad de opciones estadísticas para los usuarios de Excel. En un apéndice del capítulo 1 se muestra cómo descargarlo, y la mayoría de los capítulos incluye un apéndice que describe los pasos requeridos para realizar un procedimiento estadístico usando esta herramienta.
- **Apéndice de Excel actualizado para la estadística descriptiva de tablas y gráficas** El apéndice de Excel del capítulo 2 muestra cómo se usan las herramientas Chart Tools, PivotTable Report y PivotChart Report para mejorar las capacidades de mostrar la estadística descriptiva en tablas y gráficas.
- **Análisis comparativo con diagramas de caja** El tratamiento de diagramas de caja del capítulo 2 se ha ampliado para incluir comparaciones relativamente rápidas y fáciles de dos o más bases de datos.
- **Nuevo Software PrecisionTree® para el análisis de decisiones** PrecisionTree® es otro complemento de Excel desarrollado por Palisades Corporation, muy útil en el análisis de decisiones. El capítulo 21 contiene un apéndice nuevo que muestra cómo usarlo.
- **Ejemplos y ejercicios nuevos basados en datos reales** Al usar datos reales, los lectores se interesarán más en el material y podrán aprender tanto sobre la metodología estadística como sobre sus aplicaciones. Esta edición contiene más de 350 ejemplos y ejercicios basados en información real.
- **Archivos de datos que acompañan el libro** Más de 200 archivos de datos están disponibles en el sitio web **http://latinoamerica.cengage.com/anderson**. Las bases de datos se encuentran tanto en formato de Minitab como de Excel.

Estadística para negocios y economía



ISBN-13: 978-607481750-8
ISBN-10: 607481750-2



CENGAGE Learning™

http://latinoamerica.cengage.com