A Biostatistics Toolbox for Data Analysis

STEVE SELVIN

A Biostatistics Toolbox for Data Analysis

This sophisticated package of statistical methods is for advanced master's degree and Ph.D. students in public health and epidemiology who are involved in the analysis of data. It makes the link from statistical theory to data analysis, focusing on the methods and data types most common in public health and related fields. Like most toolboxes, the statistical tools in this book are organized into sections with similar objectives. Unlike most toolboxes, however, these tools are accompanied by complete instructions, explanations, detailed examples, and advice on relevant issues and potential pitfalls – conveying skills, intuition, and experience.

The only prerequisite is a first-year statistics course and familiarity with a computing package such as R, Stata, SPSS, or SAS. Though the book is not tied to a particular computing language, its figures and analyses were all created using R. Relevant R code, data sets, and links to public data sets are available from www.cambridge.org/9781107113084.

STEVE SELVIN is a professor of biostatistics in the School of Public Health at the University of California, Berkeley, and was the head of the division from 1977 to 2004. He has published more than 250 papers and authored several textbooks in the fields of biostatistics and epidemiology. His book *Survival Analysis for Epidemiologic and Medical Research* was published by Cambridge University Press in 2008.

A Biostatistics Toolbox for Data Analysis

Steve Selvin
University of California, Berkeley

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

32 Avenue of the Americas, New York, NY 10013-2473, USA

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107113084

© Steve Selvin 2015

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2015

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication Data
Selvin, S.
A biostatistics toolbox for data analysis / Steve Selvin, University of California, Berkeley.
pages cm
Includes bibliographical references and index.
ISBN 978-1-107-11308-4 (hardback)
1. Medical statistics. 2. Biometry. I. Title.
RA409.S327 2015
610.2′1 – dc23 2015012679

ISBN 978-1-107-11308-4 Hardback

Additional resources for this publication at www.cambridge.org/9781107113084

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party Internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

For Nancy, David, Liz, Ben, and Eli

Contents

Preface xv

Basics

1 Statistical Distributions 3
The Normal Probability Distribution 3
The *t*-Distribution 7
The Chi-Square Probability Distribution 9
Illustration of the Pearson Chi-Square Statistic – Simplest Case 12
The *f*-Distribution 13
The Uniform Probability Distribution 13
The *p*-value 15
A Few Relationships among Probability Distributions 16

2 Confidence Intervals 19
Four Properties of an Estimated Confidence Interval 20
Confidence Intervals for a Function of an Estimate 21
Confidence Intervals Based on Estimates near Zero 21
Exact Confidence Interval (Computer-Estimated) 22
A Confidence Interval Based on an Estimated Median Value 23
A Confidence Interval and a Confidence Band 25
Details: A Confidence Interval and a Statistical Test 27

3 A Weighted Average 29
A Basic Application 29
Ratios and Weighted Averages 30
Estimates Weighted by Reciprocal Variances 32
A Puzzle 37
Age-Adjusted Rates Using Weighted Averages 39
Smoothing – A Weighted Average Approach 40
Example: Weighted Average Smoothing of Hodgkin's Disease Mortality Data 44

4 Two Discrete Probability Distributions 47
Binomial Probability Distribution 49

Two Applications of the Binomial Distribution 53
The Geometric Probability Distribution 57
A Poisson Probability Distribution 58
Two Applications of the Poisson Probability Distribution 62
A Note on Rare Events 67

5 Correlation 69
Spearman's Rank Correlation Coefficient 73
Point Biserial Correlation Coefficient 74
Nonparametric Measure of Association: The γ-Coefficient 77
A Special Case: The $2 \times k$ Table 80
Chi-Square–Based Measures of Association 81
Proportional Reduction in Error Criterion 83

Applications

6 The 2 × 2 Table 87
The Analysis of a 2×2 Table 89
Measures of Association in a 2×2 Table 92
Odds Ratio and Relative Risk Ratio 94
A Correction to Improve the Normal Distribution Approximation 96
The Hypergeometric Probability Distribution 98
Fisher's Exact Test 101
Correlation in a 2×2 Table 102
A 2×2 Table with a Structural Zero 106
Assessing the Accuracy of a Diagnostic Test 107

7 Linear Bivariate Regression Model 111
The Bivariate Linear Regression Model 111
Additivity 114
Coefficients 114
Multiple Correlation Coefficient 116
Adjustment 118
Interaction 119
Confounder Bias 121
Collinearity 123

8 The 2 × *k* Table 126
Wilcoxon (Mann-Whitney) Rank Sum Test 126
Nonparametric Analysis of a $2 \times k$ Table 132
A Chi-Square Analysis of a $2 \times k$ Table 134
Another Example: Childhood Cancer and X-Ray Exposure 137

9 The Loglinear Poisson Regression Model 141
A Simple Poisson Regression Model 141

Poisson Regression Model: Analysis of Vital Statistics Data 144
Rate Ratios Adjusted for Several Variables 146
A Test for Linear Trend 148
A Graphical Display of a Table 150
Implementation of Poisson Models Applied to Tabular Data 151
Analysis of Tables with Incomplete Data 153
Another Illustration of the Analysis of an Incomplete Table 154
Quasi-independence: Analysis of an Incomplete Table 156
A Case Study – Adjusted Mortality Rates: Black-White Infant Mortality 158
First Approach: Weight-Specific Comparisons 161
Second Approach: A Model-Free Summary 163
Third Approach: Poisson Regression Model 165

10 Two-Way and Three-Way Tables and Their Analysis 170
Analysis of Tables – Continuous Data 177
Matched Pairs – The Categorical Variable Case 182
Analysis of Tables – Count Data 185
Three-Way Tables 189
The Analysis of the Three-Way Table 190
Complete Independence 190
Joint Independence 191
Conditional Independence 191
No Pairwise Independence 192
Log-Linear Models for Four $2 \times 2 \times 2$ Three-Way Tables 193
Example of Completely Independent Variables 193
Example of Jointly Independent Variables 194
Example of Conditional Independent Variables 195
Example of Additive Relationships 196
A Case Study – Joint Independence 197
A Case Study – Conditional Independence in a $2 \times 2 \times 4$ Table 201

11 Bootstrap Analysis 204
Example: Analysis of Paired Data 209
Example: Evaluation of a Difference between Two Harmonic Mean Values 211
An Example of Bootstrap Estimation from Categorical Data 214
Kappa Statistic – A Bootstrap-Estimated Confidence Interval 216
A Graphic Application of Bootstrap Estimation 218
A Property of Bootstrap Estimation 220
Randomization and Bootstrap Analysis Applied to Two-Sample Data 220
Randomization Test 222
Bootstrap Analysis 223

12	**Graphical Analysis**	227
	A Confidence Interval and a Boxplot	227
	Multiple Comparisons – A Visual Approach	229
	The Cumulative Probability Distribution Function	229
	Inverse Functions for Statistical Analysis	232
	Kolmogorov One-Sample Test	241
	Kolmogorov-Smirnov Two-Sample Test	243
	Simulation of Random "Data" with a Specific Probability Distribution	245
	Simulation of "Data" with a Continuous Probability Distribution	248
	Simulation of "Data" with a Discrete Probability Distribution	249
	A Case Study – Poisson Approximation	249
	A Graphical Approach to Regression Analysis Diagnostics	250
	A Graphical Goodness-of-Fit for the Logistic Regression Model	257
13	**The Variance**	260
	A Brief Note on Estimation of the Variance	260
	A Confidence Interval	261
	Test of Variance	262
	Homogeneity/Heterogeneity	264
	Analysis of Variance – Mean Values	265
	Another Partitioning of Variability	268
	Two-Sample Test of Variance	269
	F-Ratio Test of Variance	270
	Bartlett's Test of Variance	271
	Levene's Test of Variance	273
	Siegel-Tukey Two-Sample Test of Variance	274
	Comparison of Several Variances	275
14	**The Log-Normal Distribution**	278
	Example: Lognormal Distributed Data	281
	A Left-Censored Log-Normal Distribution	283
	An Applied Example	284
15	**Nonparametric Analysis**	287
	The Sign Test	289
	The Wilcoxon Signed Rank Test	291
	Kruskal-Wallis Nonparametric Comparison of *k* Sample Mean Values	294
	Three-Group Regression Analysis	297
	Tukey's Quick Test	300
	Friedman Rank Test	302
	Survival	
16	**Rates**	309
	An Average Mortality Rate	309

An Approximate Average Rate 314
A Rate Estimated from Continuous Survival Times 318

17 Nonparametric Survival Analysis 324
Cumulative Hazard Function 333
Description of the Median Survival Time 334
Comparison of Two Survival Probability Distributions – The Log-Rank Test 335
Proportional Hazards Rates 339
An Example of Survival Analysis 342
The Cox Analysis of Proportional Hazards Models 345
Proportional Hazards Model – A Case Study 346

18 The Weibull Survival Function 352
Two-Sample Comparison – The Weibull Model 358
The Shape Parameter 360
Multivariable Weibull Survival Time Model – A Case Study 363

Epidemiology

19 Prediction, a Natural Measure of Performance 371
Net Reclassification Index (Binary Case) 373
Net Reclassification Index (Extended Case) 375
Integrated Discrimination Improvement 379
Summary Lines as Measures of Improvement in Model Prediction 380
Covariance and Correlation 382
Epilogue 383
Least Squares Estimation – Details 383

20 The Attributable Risk Summary 387
Incidence Rates 391
Two Risk Factors 393
Adjustment by Stratification 394
Adjustment by Model 395
Issues of Interpretation 396
A Few Less Specific Issues of Interpretation 397

21 Time-Space Analysis 399
Knox's Time-Space Method 399
The Statistical Question Becomes: Is $m = 0$? 400
Test of Variance Applied to Spatial Data 403
Mantel's Time-Space Regression Method 404
Performance and Time-Space Methods 405

22 ROC Curve and Analysis 407
An ROC Curve 407
An ROC Analysis Example 414
Nonparametric Estimation of an ROC Curve 417
An Example – Nonparametric ROC Analysis 418
How to Draw a Nonparametric ROC Curve 419
Construction of the ROC Curve 419
Implementation 425

Genetics

23 Selection: A Statistical Description 433
Stable Equilibrium 434
A Description of Recombination 435
Two Examples of Statistical Analyses 439
Selection – Recessive Lethal 442
A More General Selection Pattern 444
Selection – A Balanced Polymorphism 445
Fitness 448

24 Mendelian Segregation Analysis 450
Ascertainment 450
Truncation 451
Estimation of a Segregation Probability: Complete Ascertainment 451
Estimation of a Segregation Probability: Single Ascertainment 455

25 Admixed Populations 458
A Model Describing Admixture 459
Estimation of the Admixture Rate 461
An Example: Estimation of the Extent of Admixture 462

26 Nonrandom Mating 465
Genotype Frequencies – Correlation and Variance 465
Genetic Variance 469
Wahlund's Model 470
Random Genetic Drift 473
A Description: Consecutive Random Sampling of a Genetic Population 473
Random Genetic Drift – A Description 474
The Statistics of Random Genetic Drift 474
Mutation/Selection Equilibrium 477
The Story of Mr. A and Mr. B 477
One-Way Mutation 479
Mutation/Selection Balance 480

Assortative Mating 481
Assortative Mating Model 482

Theory

27 Statistical Estimation 489
The Sample 489
Covariance 490
The Population 492
Infinite Series with Statistical Applications 496
Example of Infinite Power Series 497
Binomial Theorem 498
Functions of Estimates 499
Approximate Values for the Expectation and Variance of a Function 500
Partitioning the Total Sum of Squares 503
Expected Value and Variance of Wilcoxon Signed Rank Test 504
Maximum Likelihood Estimation 505
Properties of a Maximum Likelihood Estimate 510
Example Maximum Likelihood Estimates 512
Method of Moments Estimation 515

Appendix: R Code 519
Index 557

Preface

Books on statistical methods largely fall into two categories: elementary books that describe statistical techniques and mathematically advanced texts that describe the theory underling these techniques. This text is about the art and science of applying statistical methods to the analysis of collected data and provides an answer to the question, after a first course in statistics, What next? Like most toolboxes, the statistical tools in the text are loosely organized into sections of similar objectives. Unlike most toolboxes, these tools are accompanied by complete instructions, explanations, detailed examples, and advice on relevant issues and potential pitfalls. Thus, the text is a sophisticated introduction to statistical analysis and a necessary text for master's and Ph.D. students in epidemiology programs as well as others whose research requires the analysis of data.

The text employs a pattern of describing sampled data and providing examples followed by a discussion of the appropriate analytic strategies. This approach introduces the reader to data and analytic issues and then explores the logic and statistical details. A large number of examples and illustrations are included to appeal to a diverse audience. It is often the case that examples produce substantial insight into the problem in hand. Most statistical texts reverse this approach and describe the statistical method first followed by examples.

The level of this text is beyond introductory but far short of advanced.

Fundamental topics, such as the median values, simple linear regression methods, correlation coefficients, and confidence intervals found in most introductory books are not repeated but frequently reinforced. The explanations, illustrations, and examples require little or no mathematics to completely understand the presented material. Nevertheless, simple mathematical notation and arguments are included because they are unambiguous, and frequently their clarity leads to better understanding.

The complexity of the mathematics that supports the discussion of the statistical tools is no more than high school algebra. Mostly the symbols and notation from mathematics are used to describe the various approaches to data analysis. An exception is the last chapter, which is intended to enrich the applied nature of the text with a bit of statistical theory.

The choices of the statistical tools included in the text are largely based on two criteria: the usefulness in the analysis of data from human populations as well as a variety of other methods because they simply demonstrate general statistical data analysis strategies. The level of presentation of the techniques discussed evolved from a second-year course in biostatistics for epidemiology graduate students taught over the last decade at the University of California, Berkeley. Also, much of the material has been presented in summer courses at the Graduate Summer Session in Epidemiology at the University of Michigan School of Public Health and more recently at The Johns Hopkins Bloomberg School of Public Health as

part of the Summer Institute of Epidemiology and Biostatistics. In other words, the material has been thoroughly "classroom tested."

The text is organized in six sections.

Section I: BASICS is a review of fundamental statistical distributions and methods: for example, confidence intervals and correlation including important details not typically included in introductory texts.

Section II: APPLICATIONS builds on these basic statistical tools to create more advanced strategies used to analyze specific kinds of issues and data. Linear models, contingency table analysis, graphical methods, and bootstrap estimation are examples of methods that are frequently useful in unraveling the sometimes complicated issues that data are collected to explore. Also included are parallel and often neglected nonparametric methods.

The next three sections continue to develop new methods as well as include further extensions of basic statistical tools but applied to specific subject matter areas: survival, epidemiologic, and genetic data. These sections are a "proving grounds" for statistical techniques applied to a variety of challenging kinds of data.

Section III: SURVIVAL contains discussions of three important statistical techniques designed for the analysis of survival time data. Specifically, they consist of an extensive exploration of rates and their properties, followed by descriptions of nonparametric methods and regression models specifically designed to analyze survival data.

Section IV: EPIDEMIOLOGY similarly explores statistical methods particularly designed for analysis of epidemiologic data such as attributable risk analysis, cluster analysis, ROC curves, and reclassification methods.

Section V: GENETICS presents statistical tools applied to several fundamental topics to give a sense of the role of statistics and data analysis in a genetics context. Topics include a statistical description of selection/mutation dynamics, sibship analysis, and application to several ways that statistical tools identify the consequences of nonrandom mating.

Section VI: THEORY is focused on the underlying principles of a few statistical tools frequently used in data analysis settings. This last section, however, is not about data analysis but is a "user friendly" introduction to how important data analysis tools work. Many "tools" of modern society are effectively used without even a hint of how and why they work. Statistical methods are not an exception. This readily accessible "beginner's guide to statistical theory" is intended to enrich the focus on applications. Primarily, these theoretical details demystify the sometimes baffling expressions that appear when the focus is entirely on description and application. With surprisingly little effort, such mysterious expressions and the sometimes obscure logic of statistical techniques disappear with use of high school algebra and a bit of first semester calculus.

Key Features

The text examples mostly consist of small subsets of real data so that the analytic results described are easily verified or explored with alternative analyses. These examples also range over a number of methods and kinds of data, from estimating missing values to special

techniques potentially useful in the analysis of genetic data. Several statistical methods are illustrated with the analysis of complete and again real data in terms of comprehensive case studies.

The introduction of many of the statistical methods discussed begins with an example made up of only a few artificial observations. These miniature examples allow simple "hands-on" illustrations of the more complicated computational and technical details. In addition, more than 150 figures provide visual displays of analytic concepts, and issues adding yet another important dimension to the use of statistical tools to analyze data. The distinguished statistician John W. Tukey stated, "The greatest value of a picture is it forces us to notice what we never see."

It is not possible to characterize statistical methods in a definite linear or logical sequence. The discussed techniques are indexed to create a "road map" so that the interconnections among the discussed material can be traced back and forth throughout the text.

Nonparametric methods are typically treated as a separate topic. These methods are integrated into the text where they are natural partners to parallel parametric methods. These important techniques enrich the practice and understanding of most statistical approaches and are particularly important when small sample sizes are encountered.

The text does not contain references to parallel published material. The ability to use online searches to locate references certainly replaces the tradition of annotating a text. Using Google, for example, easily produces not only specific material referenced in the text but a wide range of other perhaps worthwhile sources on the same topic.

An appendix contains the computer code that produced the worked-out examples in the text. The statistical software used is entitled simply "R." Free and easily downloaded, it provides an extensive statistical analysis system. The explanations and details in the appendix are not comprehensive and are left to manuals and books created exactly for this purpose. The R system, however, is completely self-documenting with direct access to descriptions of the R language, computer code, and examples. In addition, this popular data analysis system is well documented in numerous books, and a huge number of online computer sites exist that provide specific instructions and illustrations. The R analyses in the appendix can be used in variety of ways in conjunction with the text. First, using the presented code to verify the text material provides an extremely detailed description of the application of specific methods and statistical approaches. Furthermore, with minor modification, additional data sets can be analyzed to furnish alternative examples or assess results from larger sample sizes or gauge the influences of extreme values. The text does not refer to or depend on the R code, and the appendix can be completely ignored. At the other extreme, this appendix presents the opportunity to be part of learning a useful statistical analysis software language. R is available from www.r-project.org.

Remember, the study of statistics makes it possible to acquire the ability to state with great certainty the degree of uncertainty.

Basics

1

Statistical Distributions

The Normal Probability Distribution

The normal probability distribution has a long and rich history. Three names are always mentioned: Abraham de Moivre (b. 1667), Carl Friedrich Gauss (b. 1777), and Pierre-Simon Laplace (b. 1799). De Moivre is usually credited with the discovery of the normal distribution. Gauss introduced a number of important mathematical and statistical concepts derived from a normal distribution (1809). Adolphe Quetelet (b. 1796) suggested that the normal distribution was useful for describing social and biologic phenomena. In his study of the "average man" Quetelet characterized heights of army recruits with a normal distribution (1835). Twentieth-century statisticians Karl Pearson (b. 1857) and R. A. Fisher (b. 1890) added a few details, producing the modern normal distribution. Today's normal probability distribution has other less used names: Gaussian distribution, "bell-shaped curve," and Gauss-Laplacian distribution.

The algebraic expression of the normal probability distribution for a value denoted x is

$$f(x) = \frac{1}{\sigma_X\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma_X}\right]^2}.$$

The expression shows that the value of the function $f(x)$ is defined by two parameters: a mean value represented by μ and a variance represented by σ_X^2 (Chapter 27). The mean value μ determines location, and variance σ_X^2 determines spread or shape of the normal distribution (Figure 1.1). The usual estimates of these two parameters are the sample mean value (denoted $\bar{x}$) and the sample variance (denoted S_X^2). For a sample of n values $\{x_1, x_2, \ldots, x_n\}$,

$$\text{sample estimated mean value} = \bar{x} = \frac{1}{n}\sum x_i$$

and

$$\text{sample estimated variance} = S_X^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 \quad i = 1, 2, \ldots, n \ \ [13].$$

From the mathematical expression, the height $f(x - \mu)$ equals the height $f(\mu - x)$, making the normal distribution symmetric relative to the mean value μ. In addition, again seen from the expression $f(x)$, the normal distribution is always positive because e^{-z^2} is always positive for any value of z. The expression $f(x)$ dictates that a single maximum value occurs at the value $x = \mu$ and is $f(x_{\max}) = f(\mu) = 1/(\sigma_X\sqrt{2\pi})$ because e^{-z^2} is always less than 1 except

Note: the chapter number in parentheses indicates the chapter where the discussed statistical tool is further described and applied in a different context.

Table 1.1 *Description: A Few Selected Critical Values and Their Cumulative Probabilities[a] from a Standard Normal Distribution ($\mu = 0$ and $\sigma = 1$)*

	Standard normal distribution								
Critical values (z)	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0
Probabilities ($1-\alpha$)	0.023	0.067	0.159	0.309	0.500	0.691	0.841	0.933	0.977
Probabilities (α)	0.977	0.933	0.841	0.691	0.500	0.309	0.159	0.067	0.023

[a] $P(Z \leq z) = 1 - \alpha$ making the value z a critical value (percentile/quantile) for a specific cumulative probability denoted $1 - \alpha$.

when $z = x - \mu = 0$. Therefore, the symmetric normal distribution has mean value, median value, and mode at the center of the distribution and "tails" that extend indefinitely for both positive and negative values of x (Figure 1.1). The total area enclosed by a normal probability distribution is 1.

A principal role of the normal distribution in statistics is the determination of probabilities associated with specific analytic results. In this context, the term *critical value* is frequently used to identify values from a normal distribution that are otherwise called quantiles or percentiles. These values and their associated probabilities typically arise as part of a statistical evaluation. For example, value $z = 1.645$ is the 95th percentile of a normal distribution with mean value $= \mu = 0$ and variance $= \sigma^2 = 1$ but is usually referred to as the critical value at the 95% significance level when applied to test statistics, confidence intervals, or other statistical summaries.

An essential property of the normal probability distribution is that a standard distribution exists; that is, a single normal distribution can be used to calculate probabilities for values from any normal distribution using the parameters μ and σ_X^2. This standard normal distribution has mean value $\mu = 0$ and variance $\sigma^2 = 1$. Table 1.1 gives a sense of the relationship between the critical values (quantiles/percentiles) and their associated probabilities from this standard normal distribution. Of course, more extensive tables exist, and typically these probabilities are computer calculated as part of a statistical analysis.

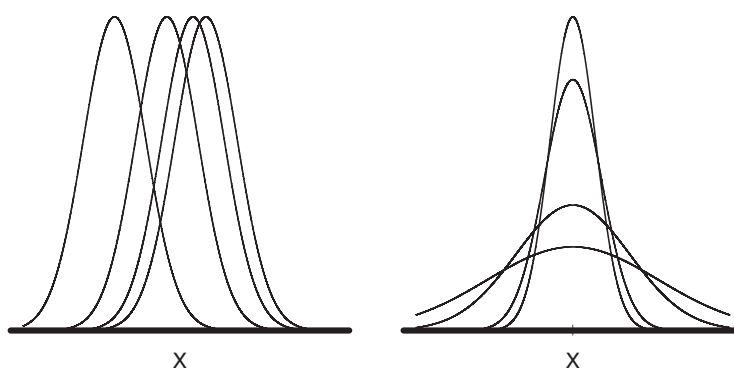


Figure 1.1 Four Normal Probability Distributions with Different Mean Values and the Same Variance (Left) and Four Normal Distributions with the Same Mean Value and Different Variances (Right)

For example, the probability that a random value from the standard normal distribution is less than a critical value 1.5 is 0.933, called a *cumulative normal probability*. In symbols, for a random value represented by Z from a standard normal distribution, then

$$\text{cumulative normal probability} = P(Z \leq z) = P(Z \leq 1.5) = 0.933.$$

Geometrically, the cumulative probability $P(Z \leq z)$ is the area enclosed by the normal distribution to the left of the value z.

A miraculous property of the normal distribution allows this standard normal distribution to be used to calculate probabilities for any normal distribution. The process of obtaining a probability for a specific value is simply a change in measurement units. When a normally distributed value is measured in terms of standard deviations (square root of the variance, denoted σ_X) relative to above or below the mean value, the associated cumulative probabilities from all normal distributions are the same. For example, the probability that a value is more than two standard deviations below a mean value μ is 0.023 for all normal distributions. Thus, from any normal distribution $P(X \leq \mu - 2\sigma_X) = 0.023$. Therefore, to find a probability for a specific value, the units of the original value are converted to units of standard deviations. For example, an observation of $x = 25$ feet sampled from a normal probability distribution with mean value $\mu = 10$ and variance $\sigma_X^2 = 100$ is $Z = (x - \mu)/\sigma_X$ standard deviations above the mean, and Z has a normal probability distribution with mean $\mu = 0$ and standard deviation $\sigma_X = 1$ (Table 1.1). Therefore, the probabilities associated with Z are the same as the probabilities associated with X when the value x is converted to standard deviations. Specifically, the value $x = 25$ feet is $z = (25 - 10)/10 = 1.5$ standard deviations above the mean, making $P(X \leq 25.0) = P(Z \leq 1.5) = 0.933$ because $P(X \leq x$ natural units$) = P(Z \leq z$ standard deviation units$)$ from a standard normal distribution (Table 1.1).

To determine a value in natural units associated with a specific probability, the process is reversed. The value Z from the standard normal distribution is converted to original units. The critical value $z_{1-\alpha}$ associated with the probability $1 - \alpha$ is used to calculate $X = \mu + z_{1-\alpha}\sigma_X$ where $z_{1-\alpha}$ represents a value from the standard normal distribution. The symbol $1 - \alpha$ traditionally represents the cumulative probability $P(X \leq x) = 1 - \alpha$ making x the $(1 - \alpha)$ – level quantile or the $(1 - \alpha) \times 100$ percentile of the normal probability distribution of a variable X. For the example, when $1 - \alpha = 0.933$, then $z_{0.933} = 1.5$ (Table 1.1) and $X = 10 + 1.5(10) = 25$ feet for a normal probability distribution with mean value $\mu = 10$ and variance ${\sigma^2}_X = 100$. As required, the associated probability is again $P(Z \leq 1.5) = P(X \leq 25.0) = 0.933$. Figure 1.2 schematically displays the relationship between cumulative probabilities, observed values, and standard deviations for all normal distributions.

From the symmetry of the normal probability distribution $f(x)$, it follows that the probabilities are symmetric. For a normally distributed value X, then $P(X \geq c) = P(X \leq -c)$. For example, for a standard normal distribution $P(Z \geq 1.0) = P(Z \leq -1.0) = 0.159$ because $P(Z \leq -1.0) = 0.159$ (Table 1.1 and Figure 1.2). In symbols, for example,

$$P(X \leq \mu - 2\sigma) = P(X \geq \mu + 2\sigma).$$

The ubiquitous role of the normal distribution in statistical analysis stems from the *central limit theorem*. Probability theory is not simple, and neither is the central limit theorem. Leaving out considerable detail, a statement of this indispensable theorem is as follows:

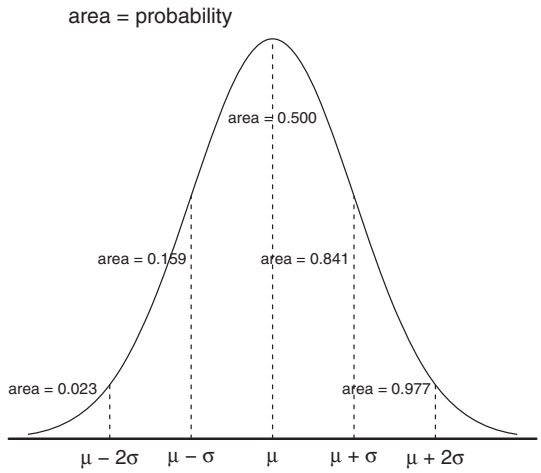


Figure 1.2 Normal Distribution Probabilities (Area to the Left = Cumulative Probability = $1 - \alpha$) and Their Associated Critical Values (Percentiles/Quantities)

When a sample mean value represented by $\bar{X}$ is estimated from n independent random observations $\{x_1, x_2, \ldots, x_n\}$ sampled from the same distribution with mean value μ and variance σ_X^2, the distribution of

$$Z = \sqrt{n}\left[\frac{\bar{X} - \mu}{\sigma_X}\right]$$

converges to a normal distribution with mean = 0 and variance = 1 as the sample size n increases.

The most important feature of the central limit thereon is what is missing. No mention is made of the properties of the sampled population. The extraordinary usefulness of this theorem comes from the fact that it applies to many kinds of mean values from many situations. Thus, for a sample size as few as 20 or 30 observations, a normal distribution is frequently an accurate description of the distribution of a large variety of summary statistics. Therefore, different kinds of statistical summaries can be evaluated with approximate normal distribution probabilities for moderate sample sizes. In fact, data sampled from symmetric distributions likely produce mean values with close to symmetric distributions and, therefore, are often accurately approximated by a normal distribution for a sample sizes as small as 10. Ratio summaries are similarly evaluated because the logarithm of a ratio frequently produces values with an approximate symmetric distribution. Of note, when values are sampled from a normal distribution, sums, and means of these values have exactly normal distributions for any sample size.

To illustrate the central limit theorem, consider a sample of $n = 10$ independent and randomly sampled values distributed between 0 and 1. One such a sample is

{0.593, 0.726, 0.370, 0.515, 0.378, 0.418, 0.011, 0.532, 0.432 and 0.094}

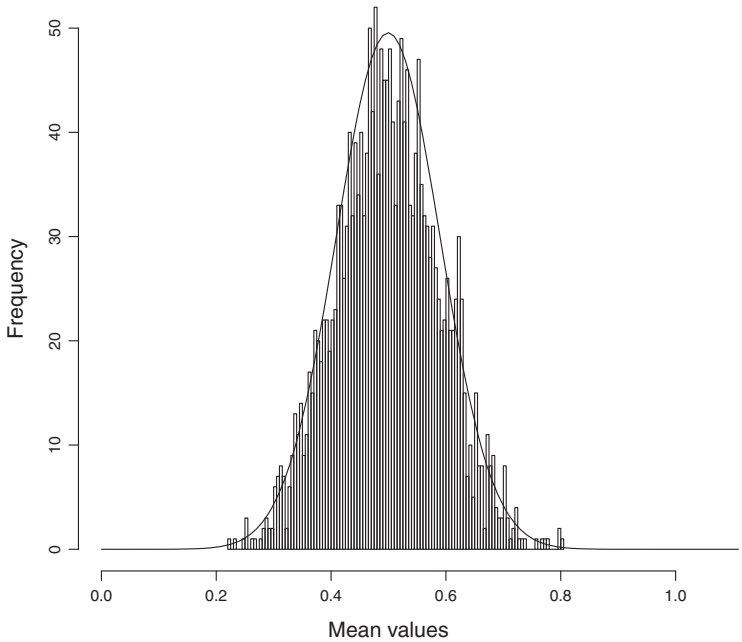


Figure 1.3 An Illustration of the Central Limit Theorem Consisting of 1000 Mean Values Each Estimated from 10 Randomly Sampled Values between 0 and 1 ($n = 10$)

with mean value $\bar{x} = 0.406$. Thus, a distribution of 1000 such mean values each calculated from 10 observations randomly sampled from the same distribution is accurately approximated by a normal distribution with a mean value $\mu = 0.5$ and variance $\sigma_{\bar{x}}^2 = \sigma_x^2/n = (1/12)/n = 0.0083$ (Figure 1.3, solid line).

The convergence described by the central limit theorem becomes slower and less accurate as the population sampled becomes less symmetric. Transformations such as the square root or logarithm or other specialized functions of observations produce values with an approximate normal distribution (Chapter 6). These transformations typically make large reductions in extreme values and relatively smaller reductions in small values tending to create a more symmetric distribution. When the distribution of data is extremely asymmetric, the mean value is not a useful summary, and alternatives such as the median value or other statistical measures are more meaningful. Thus, when a mean value is a worthwhile summary, even for modest sample sizes of 20 or 30 observations, it is likely to have at least an approximate normal distribution.

The *t*-Distribution

A probability from a *t*-distribution describes the properties of a test statistic designed to evaluate the sample mean value $\bar{x}$ consisting of n independently sampled observations from a normal distribution with mean μ. Thus, the expression

$$t \text{ statistic} = T = \frac{\bar{x} - \mu}{S_{\bar{x}}} = \sqrt{n}\left[\frac{\bar{x} - \mu}{S_X}\right]$$

Table 1.2 *Description: Standard Normal and t-Distribution Critical Values for Five Selected Probabilities (Degrees of Freedom = df = {2, 10, 20, 40, and 60})*

Normal distribution		*t*-Distributions				
Probabilities $(1-\alpha)$	$z_{1-\alpha}$	$df=2$	$df=10$	$df=20$	$df=40$	$df=60$
0.990	2.326	6.965	2.764	2.528	2.423	2.390
0.975	1.960	4.303	2.228	2.086	2.021	2.000
0.950	1.645	2.920	1.812	1.725	1.684	1.671
0.900	1.282	1.886	1.372	1.325	1.303	1.296
0.800	0.842	1.061	0.879	0.860	0.851	0.848

has a *t*-distribution (Table 1.2) with a mean value of 0.0. The evaluation of the sample mean value $\bar{x}$ is much like the *Z*-value based on the central limit theorem. The difference is that, unlike the normal distribution test statistic Z where the variance is a known value, the variance used in the calculation of the t statistic is estimated from the sampled data that generated the mean value (in symbols, $S^2_{\bar{X}} = S^2_X/n$) (Chapter 27).

The *t*-distribution is defined by a single parameter called the degrees of freedom (denoted *df*) determined by the number of observations used to estimate the mean value and its variance. Thus, a different *t*-distribution exists for every sample size. Figure 1.4 displays the

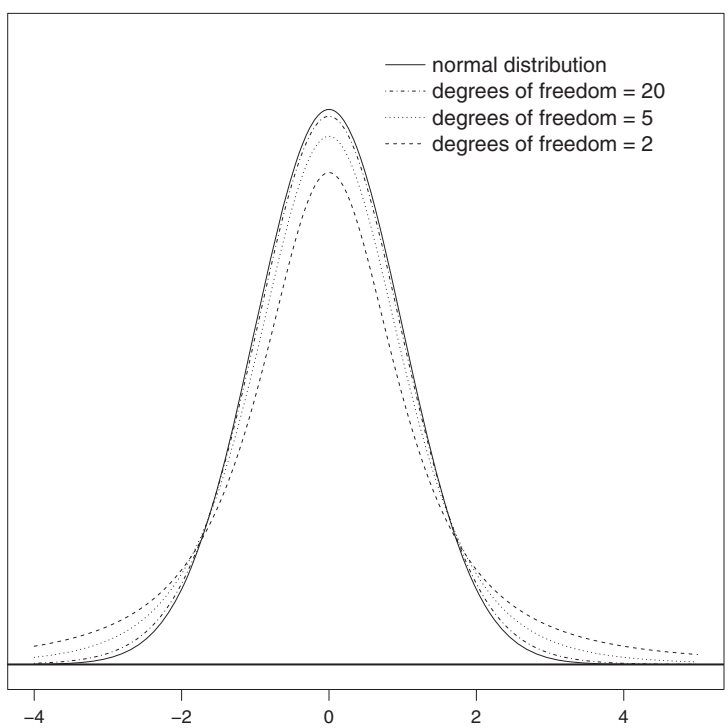


Figure 1.4 Three Representative *t*-Distributions with Degrees of Freedom of 2, 5, and 20 and a Standard Normal Distribution ($\mu = 0$ and $\sigma^2 = 1$)

standard normal distribution and three *t*-distributions. Because this statistical distribution accounts for the added variability due to an estimated variance, the cumulative probabilities from the *t*-distribution are larger than the corresponding cumulative probabilities from a standard normal distribution. For example, this increased variability, for degrees of freedom = 30, causes the modest difference $P(T \geq 2.0) = 0.027$ and $P(Z \geq 2.0) = 0.023$.

The critical values from a standard normal distribution $Z_{1-\alpha}$ are approximately equal to critical values from a *t*-distribution (denoted $t_{1-\alpha,df}$) for degrees of freedom (*df*) greater than 30 or so. As the sample size increases, the difference decreases (Figure 1.4). In symbols,

$$Z_{1-\alpha} \approx t_{1-\alpha,df} \quad \text{for } df = \text{degrees of freedom} > 30.$$

Table 1.2 illustrates a few selected *t*-distribution cumulative probabilities. The mean of the symmetric *t*-distribution is 0, and the variance is $df/(df-2)$ for $df > 2$. Thus, as the degrees for freedom (sample size) increases, these two parameters more closely correspond to the mean value of 0.0 and variance of 1.0 of the standard normal distribution.

Table 1.2 further indicates that differences between a *t*-distribution and a normal distribution are small and generally negligible for sample sizes greater than 30 or 40 observations. For a sample size where the degrees of freedom are 60, then $t_{0.95,60} = 1.671$ and $z_{0.95} = 1.645$. Thus, for large sample sizes, ignoring the difference between a *t*-distribution and normal distribution has negligible effects, as noted in Figure 1.4. From a practical point of view, this similarity indicates that the difference between the estimated variance (S_X^2), and the variance estimated (σ_X^2) becomes unimportant unless the sample is small ($n < 30$).

For small sample sizes, the *t*-distribution provides an analysis of a mean value when the data are sampled from a normal distribution. The central issue, however, for a small sample size is bias. An error in measuring a single observation or a loss of a single observation, for example, can have considerable influence when the sample size is small. If 10 observations are collected, a single biased or missing value represents 10% of the data. When a student described an experiment based on six observations to statistician R. A. Fisher, he is said to have replied, "You do not have an experiment, you have an experience." Thus, for small sample sizes, the accuracy of the usually unsupported assumption that the data consist of independent and randomly sampled unbiased observations from a normal distribution becomes critically important. Furthermore, exact statistical analyses exist for small samples sizes that do not depend on the properties of the population sampled (Chapters 8 and 15).

The Chi-Square Probability Distribution

Karl Pearson (circa 1900) introduced the chi-square probability distribution as a way to evaluate a test statistic that combines estimates of variability from different sources into a single statistical summary. His chi-square distribution is defined by the following theorem:

> If $Z_1, Z_2, \ldots, Z_m$ are *m* independent and normally distributed random variables each with mean value = 0 and variance = 1, then the sum of squared *z*-values,
>
> $$X^2 = Z_1^2 + Z_2^2 + \cdots + Z_m^2,$$
>
> has a chi-square distribution with *m* degrees of freedom.

Table 1.3 *Description: A Few Selected Critical Values and Their Cumulative Probabilities from Chi-Square Probability Distributions (df = {1, 2, 10, 30, 50, and 100})*

	Degrees of freedom (*df*)					
Probabilities $(1-\alpha)$	1	2	10	30	50	100
0.990	6.635	9.210	23.209	50.892	76.154	135.807
0.975	5.024	7.378	20.483	46.979	71.420	129.561
0.950	3.841	5.991	18.307	43.773	67.505	124.342
0.900	2.706	4.605	15.987	40.256	63.167	118.498
0.800	1.642	3.219	13.442	36.250	58.164	111.667

Each chi-square distribution is a member of a family of probability distributions identified by an associated degree of freedom (again denoted *df*). The degrees of freedom of a chi-square distribution completely define its location and shape and, therefore, its properties. The degrees of freedom associated with a specific chi-square distribution depend on the number of independent *z*-values in the sum that makes up the chi-square statistic denoted X^2. For most chi-square statistics the values $\{Z_1^2, Z_2^2, \ldots, Z_m^2\}$ are not independent ($df < m$). This lack of independence is dealt with by adjusting the degrees of freedom. Thus, the degrees of freedom are occasionally difficult to determine but are usually part of the description of a specific statistical application or are computer generated with statistical software. Table 1.3 contains a few chi-square critical values (denoted $X^2_{1-\alpha,df}$) and their corresponding cumulative probabilities giving a sense of this probability distribution. For example, the chi-square distribution 90th percentile ($1-\alpha = 0.90$) value $X^2_{0.90,2}$ is 4.605 when the degrees of freedom are two ($df = 2$).

Thus, for the chi-square variable represented by X^2, the associated cumulative probability is $P(X^2 \leq 4.605) = 0.90$. Figure 1.5 displays four representative chi-square distributions.

The mean value of a chi-square distribution equals the degrees of freedom (*df*), and the variance is 2*df*. To evaluate a chi-square value directly, it is handy to know that a chi-square distributed value less than its mean value (*df*) has a cumulative probability always greater than 0.3 for all chi-square distributions or always $P(X^2 \leq df) > 0.3$.

The essence of a chi-square statistic is that it combines a number of summary values each with a standard normal distribution into a single measure of variability. For example, consider four independent sample mean values $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$, and $\bar{x}_4$ each estimated from four samples consisting of n_j normally distributed observations. In addition, the mean values and the variances of each sampled source are the same, represented by μ and $\sigma^2_{\bar{X}}$. A chi-square comparison of these sample mean values addresses the statistical question: Are the differences among the four sample mean values likely to have occurred by chance alone? If the answer is yes, the test statistic has a chi-square distribution with four degrees of freedom. If the answers is no, it is likely that a larger and less probable test statistic X^2 occurs (Chapter 13). Typically, a significance probability (*p*-value; to be discussed) is useful in choosing between these two alternatives. The probability calculated from a test statistic X^2 with a chi-square probability distribution indicates the likelihood that the observed variability among the mean values occurred by chance alone. In other words, the chi-square distribution summarizes the observed variation relative to a known and fixed population variance (Chapter 13). Like many test statistics, it is a comparison of data to theoretical

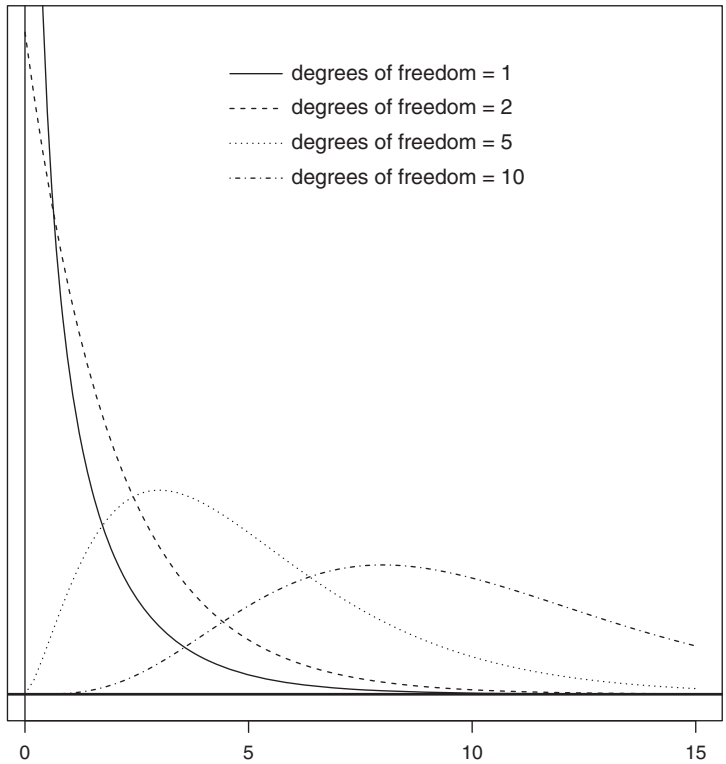


Figure 1.5 Four Representative Chi-Square Distributions for Degrees of Freedom of 1, 2, 5, and 10

values:

$$\begin{aligned} X^2 &= Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 \\ &= \left[\frac{\bar{x}_1 - \mu}{\sqrt{\sigma_X^2/n_1}}\right]^2 + \left[\frac{\bar{x}_2 - \mu}{\sqrt{\sigma_X^2/n_2}}\right]^2 + \left[\frac{\bar{x}_3 - \mu}{\sqrt{\sigma_X^2/n_3}}\right]^2 + \left[\frac{\bar{x}_4 - \mu}{\sqrt{\sigma_X^2/n_4}}\right]^2 \\ &= \sum \frac{n_j(\bar{x}_j - \mu)^2}{\sigma_X^2} \quad j = 1, 2, 3, \text{and } 4. \end{aligned}$$

A version of the chi-square test statistic used to compare observed to theoretical counts is often called the *(Karl) Pearson goodness-of-fit test.* This test statistic and the accompanying chi-square distribution apply in numerous situations. In fact, *Science* magazine published the top 10 scientific contributions in the twentieth century, and the Pearson goodness-of-fit statistic placed seventh. The expression for this simple, effective, and extensively used test statistic is

$$X^2 = \sum \frac{(o_i - e_i)^2}{e_i} \quad i = 1, 2, \ldots, m = \text{number of comparisons}$$

where o_i represents an observed count and e_i represents a theoretical count of the same observation producing a summary evaluation of the comparisons o_i versus e_i. The test

statistic X^2 has an approximate chi-square distribution when data and theoretical counts differ by chance alone. The relationship between this expression and the definition of a chi-square distributed value is not simple. Specifically, the explanation as to why e_i appears in the denominator of each term of the goodness-of-fit chi-square test statistic when the theorem calls for a variance requires a sophisticated mathematical argument. Nevertheless, the simplest case illustrates this relationship and requires no more than a small amount of algebra.

Illustration of the Pearson Chi-Square Statistic – Simplest Case

Consider the following observed counts (denoted o_1 and o_2) and theoretical counts (denoted e_1 and e_2) of a binary variable (denoted $X = 0$ and $X = 1$):

	$X = 0$	$X = 1$	Total
Data			
Observed counts	o_1	o_2	n
Estimates	$\hat{p} = o_1/n$	$1 - \hat{p} = o_2/n$	1.0
Theory			
Expected counts	e_1	e_2	N
Probabilities	$p_0 = e_1/n$	$1 - p_0 = e_2/n$	1.0

The value of p_0 is a fixed and theoretical probability. The comparison of a proportion estimated from data (denoted $\hat{p}$) to this theoretical probability (denoted p_0) is typically accomplished with the test statistic based on n observations where

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

has an approximate standard normal distribution (Chapter 6). Therefore, because the value Z^2 has an approximate chi-square distribution ($df = 1$), then

$$\begin{aligned} Z^2 = X^2 &= \left[\frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}\right]^2 \\ &= [n(\hat{p} - p_0)]^2 \left[\frac{1}{np_0} + \frac{1}{n(1 - p_0)}\right] = \frac{[n(\hat{p} - p_0)]^2}{np_0} + \frac{[n(\hat{p} - p_0)]^2}{n(1 - p_0)} \\ &= \frac{[n(\hat{p} - p_0)]^2}{np_0} + \frac{[n(1 - \hat{p}) - n(1 - p_0)]^2}{n(1 - p_0)} = \frac{(o_1 - e_1)^2}{e_1} = \frac{(o_2 - e_2)^2}{e_2}. \end{aligned}$$

In general, to repeat,

$$X^2 = \sum \frac{(o_i - e_i)^2}{e_i} \quad i = 1, 2, \ldots, m$$

has a chi-square distribution when counts o_i and e_i differ only due to random variation. Thus, this chi-square statistic is an assessment of the variation among estimated counts relative to corresponding theoretical counts (Chapter 10). For the simplest case $m = 2$, there is only

Table 1.4 *Description: A Few Selected Critical Values and Their Cumulative Probabilities from f-Probability Distributions*

	f-Distributions															
df_1	1	1	1	1	2	2	2	2	10	10	10	10	30	30	30	30
df_2	10	20	30	60	10	20	30	60	10	20	30	60	10	20	30	60
0.990	10.04	8.10	7.56	7.08	7.56	5.85	5.39	4.98	4.85	3.37	2.98	2.63	4.25	2.78	2.39	2.03
0.975	6.94	5.87	5.57	5.29	5.46	4.46	4.18	3.93	3.72	2.77	2.51	2.27	3.31	2.35	2.07	1.82
0.950	4.96	4.35	4.17	4.00	4.10	3.49	3.32	3.15	2.98	2.35	2.16	1.99	2.70	2.04	1.84	1.65
0.900	3.29	2.97	2.88	2.79	2.92	2.59	2.49	2.39	2.32	1.94	1.82	1.71	2.16	1.74	1.61	1.48
0.800	1.88	1.76	1.72	1.68	1.90	1.75	1.70	1.65	1.73	1.53	1.47	1.41	1.66	1.44	1.36	1.29

one independent value ($df = 1$), namely, $\hat{p}$. As noted, the degrees of freedom depend on the structure of data that produce the observed values.

The *f*-Distribution

A theorem defining the *f*-distribution is simple, but the algebraic description of this probability distribution and its properties is not. The definition of an *f*-distribution is

> when a variable denoted X_1^2 has a chi-square distribution with degrees of freedom denoted df_1 and a second independent variable denoted X_2^2 has a chi-square distribution with degrees of freedom denoted df_2, then the ratio, denoted F, has an *f*-distribution with degrees of freedom df_1 and df_2 when

$$F = \frac{X_1^2/df_1}{X_2^2/df_2}.$$

This probability distribution describes probabilities associated with a ratio statistic calculated to compare two chi-square distributed measures of variation and is defined by two parameters, namely, the degrees of freedom df_1 and df_2 from the compared chi-square statistics. A small table again gives a sense of the relationship between critical values and their associated probabilities for a few selected pairs of degrees of freedom (Table 1.4). A plot of four *f*-distributions characterizes the properties of this probability distribution (Figure 1.6). The mean value and variance of an *f*-distribution, determined by the parameters df_1 and df_2, are

$$\text{Mean value} = \frac{df_2}{df_2 - 2} \text{ for } df_2 > 2 \text{ and variance} = \frac{2df_2^2(df_1 + df_2 - 2)}{df_1(df_2 - 2)^2(df_2 - 4)} \text{ for } df_2 > 4.$$

The Uniform Probability Distribution

A uniform probability distribution provides a simple and useful description of observations with the same probability of occurrence. Unlike the previous probability distributions, the uniform probability distribution is not generally used to find critical values or evaluate estimated summary statistics. A uniform distribution accurately describes of the properties of sampled populations in several important situations. For example, the probabilities of

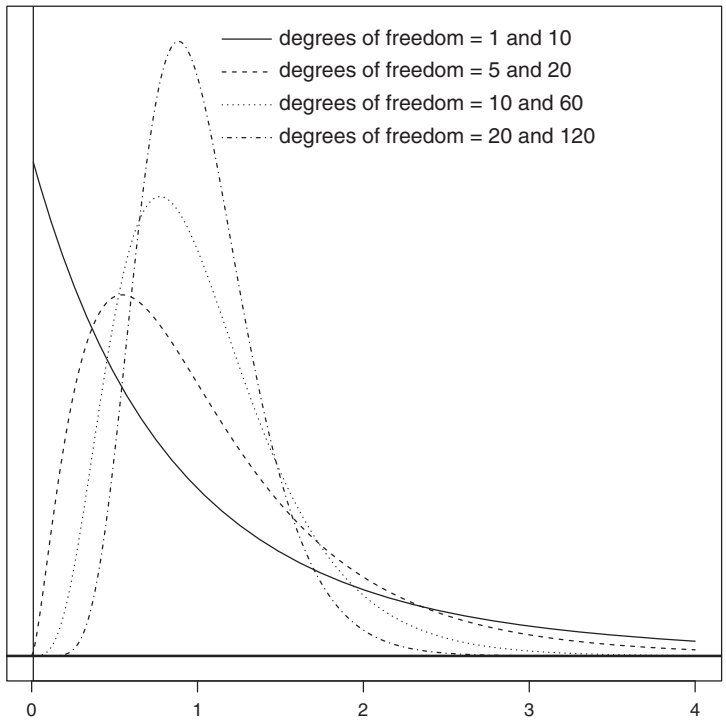


Figure 1.6 Four Representative *f*-Distributions for Pairs of Degrees of Freedom – (df_1, df_2) = {(1, 10), (5, 20), (10, 60), and (20, 120)}

human death or disease occurring within a short time or age interval frequently are at least approximately equal. Thus, uniform distributions play a central role in the estimation of mortality and disease risk (Chapters 16 and 17). In another entirely different context, random uniform probabilities are used to create simulated random "data" with known statistical properties (Chapter 12).

A uniform probability distribution applies to an interval with a minimum value represented by a and a maximum value represented by b (Figure 1.7). The expression for a uniform

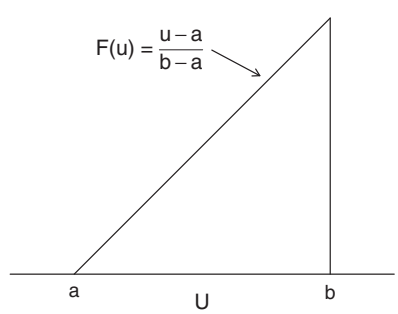


Figure 1.7 The Uniform Cumulative Distribution Function of a Variable U with Parameters a and b

cumulative probability distribution defined by the parameters a and b for a variable denoted u is

$$F(u) = P(U \leq u) = \frac{u-a}{b-a}.$$

Thus, uniform probabilities produce a linear cumulative probability distribution function:

$$\left(\text{intercept} = \frac{a}{a-b} \text{ and slope} = \frac{1}{b-a}\right).$$

(See Figure 1.7.)

The mean value of a uniform probability distribution is

$$\text{mean value} = EU = \frac{b-a}{2}$$

and the variance is

$$\textit{variance}\,(u) = \frac{(b-a)^2}{12}.$$

The time of birth of an infant during a 24-hour period is said to have a uniform distribution. The probability of a birth between 2:00 and 4:00 A.M. is then

$$P(2 \text{ to } 4 \text{ a.m.}) = F(4) - F(2) = \frac{4-0}{24-0} - \frac{2-0}{24-0} = \frac{2}{24} = 0.0833$$

where $a = 0$ and $b = 24$ hours.

An important uniform probability distribution occurs when parameters $a = 0$ and $b = 1$ describe random values u, denoted p. This special case yields the description of uniformly distributed probabilities. The mean value is $\textit{mean}\,(p) = 1/2 = 0.5$ with $\textit{variance}\,(p) = 1/12 = 0.083$ for this frequently used special case of a uniform probability distribution. A direct consequence of this cumulative distribution function is

$$\text{cumulative uniform probability distribution} = F(p) = P(U \leq p) = p.$$

That is, for values sampled from this uniform probability distribution, the cumulative probability that a value is less than p is the probability p itself. For example, the probability of a random uniform probability p less than 0.75 is $F(0.75) = P(U \leq 0.75) = 0.75$.

The *p*-value

A discussion of statistical probability distributions is not complete without at least a brief mention of a specific probability called *a significance probability* or more often referred to as a p-value. In the classic sense, the theory and purpose of a statistical test are to produce the simple inference that the data analysis presents evidence of systematic influences or that it does not. There are only two choices. An improvement on this not very informative inference is the p-value. A p-value is an estimate of the probability that a test statistic calculated from the observed data reflects only random variation. More technically, the p-value is the probability that a test statistic value more extreme than the one observed would occur by chance alone.

The historical origins of the *p*-value are not clear. Undoubtedly, the "*p*" in *p*-value stands for probability. Perhaps a more meaningful interpretation of "*p*" would be that it stands for plausibility. When a *p*-value is large, say, greater than probability 0.5, then little or no evidence exists that the observed variation reflects more than random variation. Random variation is a plausible explanation. As the *p*-value decreases, say, less than 0.05, a traditionally chosen value, the plausibility of random variation as the only explanation becomes questionable. For yet smaller *p*-values, random variation as the only explanation becomes increasingly less plausible, and the likelihood that the observed variation is increased by systematic (nonrandom) influences becomes a more and more tenable explanation.

A jury trial begins with the assumption that the defendant is innocent. The prosecution then presents evidence to convince the jury that this assumption is unlikely. A verdict of guilty is reached when the plausibility of innocence is reduced to "beyond a reasonable doubt." A *p*-value plays a parallel role in statistical inference. At some point, plausibility, measured by a *p*-value, becomes so unlikely that random variation as the only explanation of the observed value is rejected in favor of more reasonable alternatives. Unlike the courtroom, the *p*-value is a rigorously derived measure of the extent of circumstantial evidence and is usually one of a number contributors to the often difficult decision whether data do or do not indicate systematic influences.

Additionally, it is important to note, a courtroom verdict of not guilty does not prove the defendant innocent. Similarly a moderate or large *p*-value does not prove that random variation is the only explanation.

A Few Relationships among Probability Distributions

An *f*-distribution describes a ratio of two chi-square distributed variables, and a chi-square distribution describes a sum of normally distributed variables. Naturally, special cases of these three distributions are related. These relationships are useful in several situations.

The squared value Z from a standard normal distribution with a $(1 - \alpha/2)$-significance level equals a chi-square value X^2 with one degree of freedom with a $(1 - \alpha)$ significance level. In symbols,

$$Z^2_{1-\alpha/2} = X^2_{1-\alpha,1}.$$

For example, if the value of a normally distributed test statistic is $Z_{0.975} = 1.960$ for $\alpha = 0.025$, then $X^2_{0.95,1} = (1.960)^2 = 3.841$ for $\alpha = 0.05$ from a chi-square distribution.

A convention frequently applied to a normally distributed test statistic is to use the chi-square version to determine a significance probability using the test statistic $Z^2 = X^2$ (degrees of freedom = 1). The squared value does not account for the direction of the test statistic. Both positive and negative values of Z have the same chi-square probability. Situations certainly arise where an improved assessment is created by taking the sign into consideration. For example, in a study of spatial distances among a sample of diseased individuals, it is likely that only small distances are of interest (Chapter 21). Such a test statistic is called a *one-tail test*. A test statistic that does not account for the sign is called a *two-tail test*. Typically, statistical computer software present results in terms of a two-tail test. The chi-square version of a normally distributed test statistic is always a two-tail test derived from one tail of the

distribution. Mechanically, the two-sided significance probability is simply twice the one-sided value.

The critical value from an *f*-distribution $F_{1-\alpha,1,df}$ for the $(1-\alpha)$ probability with one degree of freedom is identical to the squared critical value from a *t*-distribution with the $(1-\alpha/2)$ probability with the same degrees of freedom (*df*). In symbols,

$$F_{1-\alpha,1,df} = t^2_{1-\alpha/2,df}.$$

For example, for the value $F_{0.95,1,10} = 4.964$ and $t_{0.975,10} = 2.228$, then $(2.228)^2 = 4.964$, or $F_{0.90,1,60} = 2.792$ and the value $t_{0.95,60} = 1.671$, then $(1.671)^2 = 2.792$. This relationship plays a role in identifying the differences between a confidence interval and a confidence band (Chapter 2). A novel property of the *F*-statistic is

$$F_{\alpha,df_1,df_2} = \frac{1}{F_{1-\alpha,df_2,df_1}}.$$

The critical value from a chi-square distribution has an approximate normal distribution with a mean value *df* and a variance *2df*. In symbols,

$$X^2_{1-\alpha,df} \approx df + Z_{1-\alpha}\sqrt{2df} \quad df > 50 \text{ or so.}$$

For example, for degrees of freedom $= df = 100$, the exact chi-square 95% significance level $X^2_{0.95,100} = 124.34$, and the approximate value is $X^2_{0.95,100} \approx 100 + 1.645\sqrt{2(100)} = 123.26$ where $z_{0.95} = 1.645$.

Using a normal distribution with mean value $= df$ and variance $= 2df$ to approximate a critical value for a chi-square test statistic is no longer much of an issue because the exact values are readily available from computer statistical analysis programs or websites. In the past, approximations to easily calculate critical values and probabilities were developed for most all statistical distributions. The fact that a normal distribution can be used to calculate a reasonably accurate corresponding chi-square distributed value is a reminder that mean values and sums regardless of origins often have approximate normal distributions for large sample sizes.

The critical value of a chi-square distribution with degrees of freedom df_1 is approximately equal to the critical value of an *f*-distribution with degrees of freedom df_1 and df_2 multiplied by the degrees of freedom df_1 when degrees of freedom df_2 are greater than 40 or so. In symbols,

$$X^2_{1-\alpha,df_1} \approx F_{1-\alpha,df_1,df_2} \times df_1 \quad \text{for degrees of freedom } df_2 > 40.$$

For example, the exact chi-square value is $X^2_{0.95,10} = 18.31$, and the approximate value is

$$F_{0.95,10,60} = 1.993 \times 10 = 19.93 \quad \text{for degrees of freedom } df_1 = 10 \text{ and } df_2 = 60.$$

This approximate relationship is a direct result of the definition of the *f*-distribution. The *F*-statistic is a ratio of two chi-square distributed values (X^2_1 and X^2_2). When the number of observations used to estimate the chi-square statistic in the denominator is large, the influence of sampling variation is reduced. That is, the chi-square distributed value in the denominator of the *F*-ratio becomes approximately equal to its mean value $df_2/(df_2 - 2) \approx 1$

leaving only $F \approx X_1^2/df_1$ in the numerator. The F-statistic multiplied by df_1 then becomes approximately a chi-square distributed value, namely, X_1^2 (degrees of freedom $= df_1$). This seemly obscure relationship becomes useful in certain nonparametric analyses (Chapter 15).

2

Confidence Intervals

At a Paris meeting of statisticians in 1938, statistician Jerzy Neyman introduced a new analytic technique he called a *confidence interval*. Others at the meeting were not enthusiastic about his novel idea. One statistician called this new approach to enriching and describing properties of estimated values a "confidence game." Nevertheless, this simple and effective technique is today a cornerstone of data analysis. In fact, a number of prominent scientific and medical journals require authors to include a confidence interval along with all estimated values.

A general form a confidence interval, with estimated bounds denoted $\hat{A}$ and $\hat{B}$, based on a data-estimated value with at least an approximate normal distribution is

$$\hat{A} = \text{lower bound} = \text{estimated value} + z_{\alpha/2} \times \sqrt{variance(\text{estimated value})}$$

and

$$\hat{B} = \text{upper bound} = \text{estimated value} + z_{1-\alpha/2} \times \sqrt{variance(\text{estimated value})}.$$

The symbol $z_{1-\alpha}$ represents the cumulative probability $P(Z \leq z_{1-\alpha}) = 1 - \alpha$ when Z has a standard normal distribution (Chapter 1). Nevertheless, almost all confidence intervals are constructed using values $z_{1-\alpha/2} = z_{0.975} = 1.960$ and $z_{\alpha/2} = z_{0.025} = -1.960$ making the probability $P(|Z| \leq 1.960) = 0.95$. The lower and upper bounds of a confidence interval then identify a region with a 0.95 probability of containing an underlying and unknown parameter based on its estimated value. In addition, the length of the confidence interval provides a sense of precision of the estimated value.

Although a 95% confidence interval is conceptually simple and practically effortless to construct, its interpretation causes confusion. The source of confusion is a failure to identify which values are estimated from sampled data and subject to random variation and which values are population values and not subject to random variation. A parameter value is a fixed quantity, and the estimated confidence interval bounds, like all estimates, vary from sample to sample. Because the estimated bounds vary, an estimated confidence interval can be viewed as a single occurrence among many possible confidence intervals, each with a probability of 0.95 of containing the unknown and fixed underlying parameter value. Even in statistics textbooks, a confidence interval is occasionally incorrectly described as "an interval with a 95% chance whether the true parameter will be contained between the lower and upper bounds of a confidence interval." Probabilities do not apply to fixed quantities. The probability of 0.95 applies to the variation associated with the estimated interval bounds $\hat{A}$ and $\hat{B}$.

A confidence interval enriches the presentation of an estimate by identifying a range of likely locations of the value estimated (accuracy) and at the same time the length of the interval naturally indicates the extent of variation associated with the estimated value (precision). The basic purpose of a confidence interval is to identify boundaries between unlikely and likely parameter values. The traditional but arbitrary probability of less than 0.05 is usually selected to define "unlikely." Thus, based on sampled data, an estimated confidence interval simply defines a range of the likely locations of an unknown parameter value and simultaneously communicates the influence from variation associated with the estimated value. Amazingly, a confidence interval reflects both accuracy and precision of the estimation of an unknown value, a seemly impossible task.

A rifleman shoots at a bull's-eye on a target five times. His friend asks, "How did you do?" The shooter replies, "I need to look at the target." Neyman's confidence interval provides the answer to the same question when a target is not available.

A confidence interval and a statistical test are related (technical details are at the end of the chapter). In many situations a statistical test can be constructed from a confidence interval. Both techniques are frequently based on the same estimated value and standard error. Parenthetically, the term standard error of an estimate is used instead of the exactly equivalent phrase "the standard deviation of the probability distribution of the value estimated." Because a confidence interval identifies likely and unlikely values of a theoretical parameter, a parameter value included within the confidence interval bounds can be viewed as likely consistent with the data and a value outside the confidence interval bounds as likely inconsistent with the sampled data. Therefore, a confidence interval and a test of hypothesis typically produce essentially the same inference.

Because a confidence interval and a statistical test are frequently based on the same estimates, the differences that emerge are in the description of the influence of sampling variation on the estimated value. General agreement exists that a confidence interval is a more descriptive quantity than a p-value generated by a statistical test. A statistical test alone gives no indication of the influence from random variation and frequently leads to a simplistic "yes/no" inference. In addition, a statistical test can produce a small p-value with no biological or physical importance simply because the sample size is large. That is, the result of a statistical test does not directly indicate the relevant influence of sample size. Gertrude Stein famously said, "A difference, to be a difference, must make a difference."

Four Properties of an Estimated Confidence Interval

Entire books are devoted to confidence intervals and their extensive variety of applications. Four properties of a confidence interval important to data analysis that are not always emphasized are the following:

1. A confidence interval based on a function of an estimate
2. A confidence interval for estimates equal or close to zero
3. A confidence interval for the medium value and
4. A series of intervals versus a confidence band.

Table 2.1 *Examples: Six* 95% *Confidence Intervals for Functions Based on an Estimate* $\hat{p} = 0.40$ *and Its Standard Error* $= 0.077$

Functions	Lower bounds	Upper bounds
$\hat{p} = 0.4$	$\hat{A} = 0.25$	$\hat{B} = 0.55$
$100 \times \hat{p} = 40\%$	$100 \times \hat{A} = 25\%$	$100 \times \hat{B} = 55\%$
$1 - \hat{p} = 0.60$	$1 - \hat{B} = 0.45$	$1 - \hat{A} = 0.75$
$1/\hat{p} = 2.50$	$1/\hat{B} = 1.81$	$1/\hat{A} = 4.03$
$log(\hat{p}) = -0.92$	$log(\hat{A}) = -1.39$	$log(\hat{B}) = -0.59$
$log[\hat{p}/(1-\hat{p})] = -0.41$	$log[\hat{A}/(1-\hat{A})] = -1.11$	$log[\hat{B}/(1-\hat{B})] = 0.21$
$log(1-\hat{p}) = -0.51$	$log(1-\hat{B}) = -0.80$	$log(1-\hat{A}) = -0.29$

Confidence Intervals for a Function of an Estimate

A function of the confidence interval bounds is usually the bounds of the same function of the estimated value. In symbols, for an estimate represented by $\hat{g}$ with confidence interval bounds $(\hat{A}, \hat{B})$, the lower bound $f(\hat{A})$ and upper bound $f(\hat{B})$ are the bounds of the confidence interval for the same function of the estimated value, namely, $f(\hat{g})$. Table 2.1 displays six examples of confidence intervals of a function of a probability p based on the estimated value $\hat{p} = 0.40$ $(n = 40)$ and its 95% confidence interval $(\hat{A}, \hat{B})$ where

$$\hat{p} \pm 1.960\ \sqrt{variance(\hat{p})} = 0.40 \pm 1.960(0.077) \rightarrow (\hat{A}, \hat{B}) = (0.25, 0.55).$$

In symbols, Table 2.1 presents an application of the often used general pattern:

when P(estimated lower bound $\leq$ parameter $\leq$ estimated upper bound) $= 1 - \alpha$

then $P(f$[estimated lower bound] $\leq f$[parameter] $\leq f$[estimated upper bound]) $= 1 - \alpha$.

Confidence Intervals Based on Estimates near Zero

Confidence intervals are typically constructed from estimates that have at least approximate normal distributions. For an estimated value such as a probability or a proportion or a rate close to zero, a symmetric normal distribution no longer produces accurate confidence interval bounds in this asymmetric situation, particularly when the sample size is small. Four commonly used methods to improve the accuracy of a confidence interval calculated from the estimated probability $\hat{p}$ for values $\hat{p} = 0.10$, 0.20 and 0.50 illustrate (sample size of $n = 30$):

1. Unadjusted: $\hat{p} \pm 1.960\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
2. Adjusted [6]: $\hat{p} \pm 1.960\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \pm \frac{1}{2n}$
3. Log odds: $\hat{l} \pm 1.960\sqrt{\frac{(n+1)(n+2)}{n(x+1)(n-x+1)}}$ where $\hat{l} = \log[\frac{\hat{p}}{1-\hat{p}}]$,

then the confidence interval bounds for the parameter p are

$$\frac{1}{1 + e^{\pm(log-odds\ bounds)}} \quad \text{and}$$

4. Exact confidence interval (computer estimated).

Table 2.2 *Example: Comparisons of Four Different Methods to Calculate* 95% *Confidence Interval Bounds* ($n = 30$) *from Estimated Probability* $\hat{p}$

	$\hat{p} = 0.10$		$\hat{p} = 0.20$		$\hat{p} = 0.50$	
Unadjusted	−0.007	0.207	0.057	0.343	0.321	0.679
Adjusted	0.009	0.191	0.074	0.326	0.338	0.662
Logistic	0.037	0.244	0.096	0.370	0.331	0.669
Exact	0.021	0.265	0.077	0.386	0.313	0.687

Exact Confidence Interval (Computer-Estimated)

The examples indicate two issues (Table 2.2). First, for $\hat{p} = 0.20$ or greater and a sample size of $n = 30$ observations, only moderate and slightly inconsistent differences appear among the estimated confidence interval bounds. For $p = 0.10$, however, a negative lower bound indicates the inaccuracy of a normal distribution based approximation for small values of p. For estimates in the neighborhood of zero, availability of modern computer capabilities and statistical software makes the tedious and sometimes extensive computation to produce exact confidence interval bounds unnecessary, and the accuracy of approximate bounds is no longer an issue (Table 2.2, last row).

It is instructive, nevertheless, to explore the construction of a confidence interval for the special case where the sampled value observed is zero ($x = 0$) among n observations. Viewing a confidence interval as a technique to establish likely and unlikely values of an underlying parameter, an exact confidence interval for the parameter p estimated by $\hat{p} = x/n = 0$ requires only the solution to a simple equation.

Values of a parameter p close to zero are likely to have produced an observed value of zero. As possible values postulated for p increase, they become increasing less plausible values for the underlying parameter that produced estimated value $\hat{p} = 0$ (Figure 2.1). Simply, when p_1 is less than p_2, parameter value p_1 is more likely than parameter p_2 to have produced the estimate $\hat{p} = 0$. Thus, a 95% confidence interval consists of a bound (denoted P_{bound}) that divides likely from unlikely parameter values. Such a bound is the solution to the equation

$$(1 - P_{bound})^n = 0.05,$$

making the 95% confidence interval bound $P_{bound} = 1 - 0.05^{1/n}$. The 95% confidence interval then becomes $(0, P_{bound})$. As usual, "unlikely" is defined as a parameter that has less than 0.05 probability of producing the observed estimate $\hat{p} = 0$ (Figure 2.1). Any small confidence level could be used. An accurate but approximate solution to the equation that produces the 95% bound is $P_{bound} = 3/n$ (occasionally called the rule of three) based on the two approximations $log(0.05) \approx -3$ and $log(1 - p) \approx -p$ that are accurate for small values of p such as 0.05.

For example, from a community program of cancer prevention in west Oakland, California, during a single year, no cases of colon cancer were observed among $n = 20{,}183$ African American women ($\hat{p} = 0$). A 95% confidence interval based on this estimate is $(0, P_{bound}) = (0, 3/n) = (0, 3/20{,}183) = (0, 14.9)$ per 100,000 persons (Figure 2.1). Therefore, the

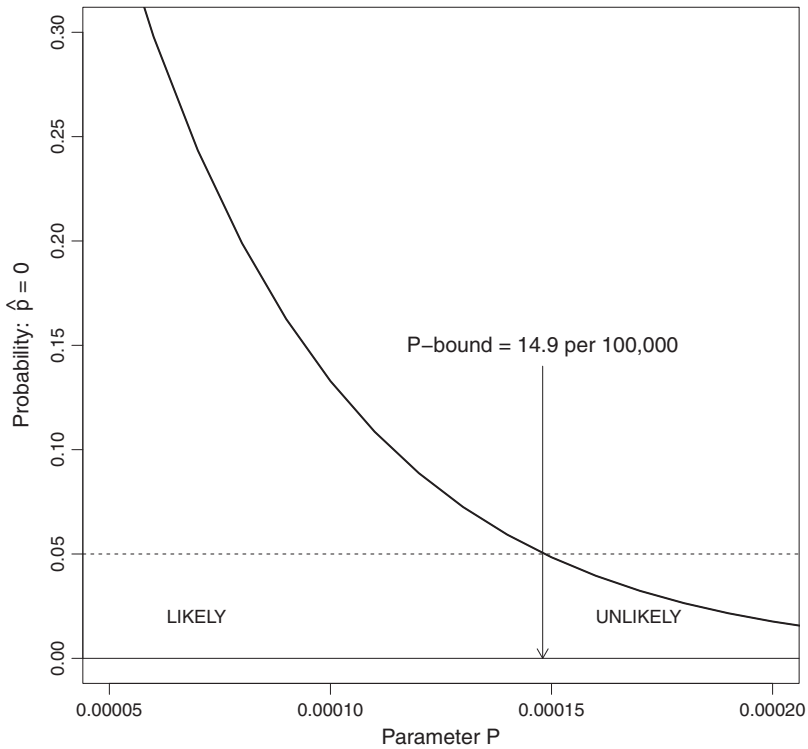


Figure 2.1 A 95% Confidence Interval Calculated from the Estimate of $\hat{p} = 0$ Cases of Colon Cancer among $n = 20{,}183$ African American Women

confidence interval (0, 14.9/100,000) has a 0.95 probability of containing the underlying parameter p that produced the estimated value zero (Figure 2.1).

A Confidence Interval Based on an Estimated Median Value

A mean value characterizes a "typical" value from a sample of observations. When the observations sampled have a normal distribution, the mean value is simply interpreted, has optimum properties, and is easily assessed in terms of a statistical test or a confidence interval. For samples from asymmetric distributions or data likely to contain extreme or outlier values, the median value is frequently a more useful summary value. A practical difference between these two estimates is that techniques to evaluate and interpret a median value are not as simple and generalizable as those of a mean value. In fact, discussions of the statistical properties of an estimated median value rarely appear in introductory statistical textbooks. Unlike normal distribution–based confidence intervals, a confidence interval based on estimated median value is simply created without knowledge or assumptions about the properties of the population distribution sampled.

The characteristics of an estimated mean value and many other estimates are accurately estimated and described based on a normal probability distribution. The parallel properties of an estimated median value are similarly estimated and described but based on the more complicated and less flexible binomial probability distribution (Chapter 4). Nevertheless,

Table 2.3 *Example: Calculation of Approximate* 95% *Confidence Interval Bounds for the Median Value* ($n = 15$ *Observations)*

Probabilities[a]	0.000	0.000	0.004	0.018	0.059	0.151	0.304	0.500	0.696	0.849	0.941	0.982	0.996	1.00	1.00
Data (ordered)	3.04	3.70	3.80	3.81	3.84	3.89	3.98	4.03	4.10	4.25	4.26	4.32	4.57	4.84	5.18
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

[a]Cumulative binomial probabilities with parameters $n = 15$ and $p = 0.5$.

construction of a confidence interval from an estimated median value is an important addition to its description.

The binomial probability distribution relevant to construct a confidence interval based on median value has parameters $n = \{$the number of sampled observations$\}$ and probability $p = 0.5$. Like an estimated median value itself, the estimated confidence interval bounds do not depend on the properties of the population sampled. The first step in constructing a 95% confidence interval is to select from this binomial distribution the value with the largest cumulative probability less than 0.025 and the value with the smallest cumulative probability greater than 0.975. These two selected integer values then locate the bounds of an interval that contains close to but greater than 95% of the binomial distributed values. The second step is to rank the sampled values from smallest to largest. The approximate 95% confidence bounds then become the two observations with ranks corresponding to the selected binomial distribution determined limits. Like all confidence intervals, the interval then contains the true underlying median value with a known probability, namely, approximately 0.95. In the case of the median value, this probability is the difference between the two selected cumulative binomial probabilities that are used to form the confidence interval bounds and is slightly greater than 0.95 creating *a conservative confidence interval.*

In symbols, the confidence interval bounds are the observed values with the ranks corresponding to the integer values from a binomial distribution with parameters $= n$ and $p = 0.5$ determined by

$$\hat{A} = \text{lower bound} = \text{largest binomial probability where } P(X \leq 0.025) \text{ and}$$

$$\hat{B} = \text{upper bound} = \text{smallest binomial probability where } P(X \geq 0.975)$$

producing an approximate confidence interval with a slightly larger confidence level than 95%.

The $n = 15$ observations in Table 2.3 provide an example.

The lower and upper bound rank values determined from the binomial probability distribution with parameters $n = 15$ and $p = 0.5$ are 4 (largest cumulative probability $= 0.018 < 0.025$) and 12 (smallest cumulative probability $= 0.982 > 0.975$). The two corresponding data values are the lower bound $= 3.81$ (rank $= 4$) and the upper bound $= 4.32$ (rank $= 12$) and, therefore, estimate a conservative 95% confidence interval (Table 2.3, vertical lines) associated with the estimated median value of 4.03 (rank $= (n + 1)/2 = 8$). Thus, the estimated median value 4.03 has an associated approximate 95% confidence interval (3.81, 4.32).

An approximate 95% confidence interval created for a median value estimated from a moderately large number of observations ($n > 30$) is again based on the binomial distribution with parameters of sample size $= n$ and $p = 0.5$ but the binomial probabilities are approximated

Table 2.4 *Data: Age and Cholesterol Levels for n = 28 Men at Extremely High Risk for Coronary Disease (Subset from a Large Study)*

Age (x)	50	54	51	44	53	46	43	49	59	45	56	39	47	42
Cholesterol (y)	238	274	349	226	218	276	254	240	241	274	232	247	209	270
Age (x)	46	48	48	44	45	39	48	45	43	44	56	51	41	54
Cholesterol (y)	311	261	223	222	213	289	271	319	282	253	220	227	296	212

by the simpler normal distribution (Chapter 6). The approximate 95% confidence interval bounds become

$$\text{lower bound rank} = \frac{n+1}{2} - 1.960\frac{\sqrt{n}}{2}$$

and

$$\text{upper bound rank} = \frac{n+1}{2} + 1.960\frac{\sqrt{n}}{2}$$

where the observed value with rank $(n + 1)/2$ is the median value. The calculated bounds are then rounded to integer values and again determine the ranks of the two observed values that become the approximate 95% confidence interval bounds for a median value. A novel property of this approximate confidence interval is that it depends entirely on the sample size n.

An example application of this approximate approach for $n = 30$ normally distributed ($\mu = 10$ and $\sigma = 4$) sample values yields the same result whether based on the binomial or approximate normal distribution derived bounds (data in the R Appendix). In both cases, the lower and upper bounds produce ranks of 10 and 21 or, from the approximation,

$$\text{bound ranks} = 15.5 \pm 1.960(2.739) \rightarrow (10.132, 20.868) \rightarrow (10, 21)$$

making the approximate 95% confidence interval 10th and the 21th ranked observed data values or (8.48, 12.04) associated with the estimated median value of 10.515 (mean of the observed values ranked 15 and 16).

One last note: The construction of a confidence interval for any estimated quantile value follows the same pattern as the estimated median value again based on the binomial distribution with again parameter $n = \{\text{sample size}\}$ but with probability parameter q corresponding to the quantile level estimated.

A Confidence Interval and a Confidence Band

A frequently ignored distinction exists between a series of confidence intervals and a confidence band. This often important distinction arises in variety of situations, but to illustrate, the comparison of these two statistical summaries is discussed in detail for the estimation of the simple linear regression line, $y_x = a + bx$. The data from $n = 28$ men at extremely high risk for coronary heart disease and the relationship between their age and cholesterol levels provides a concrete example (Table 2.4).

Table 2.5 *Results: Linear Regression Analysis Summarizing the Relationship between Age* (x) *and Cholesterol Levels* (y) *in Men at High Risk for a Coronary Event* ($y_x = a + bx$)

	Regression analysis		
	Estimates	*s.e.*	*p*-value
Intercept ($\hat{a}$)	351.816	–	–
Age ($\hat{b}$)	−2.033	1.279	0.124

The estimated regression line describing the age-cholesterol association is summarized in Table 2.5 and displayed in Figure 2.2. The specific estimated summary line is $\hat{y}_x = 351.816 - 2.033x$ (solid line).

The expression for a value estimated from a regression line for a selected value x_0 (denoted $\hat{y}_0$) is

$$\hat{y} = \hat{a} + \hat{b}x_0 = \bar{y} + \hat{b}(x_0 - \bar{x})$$

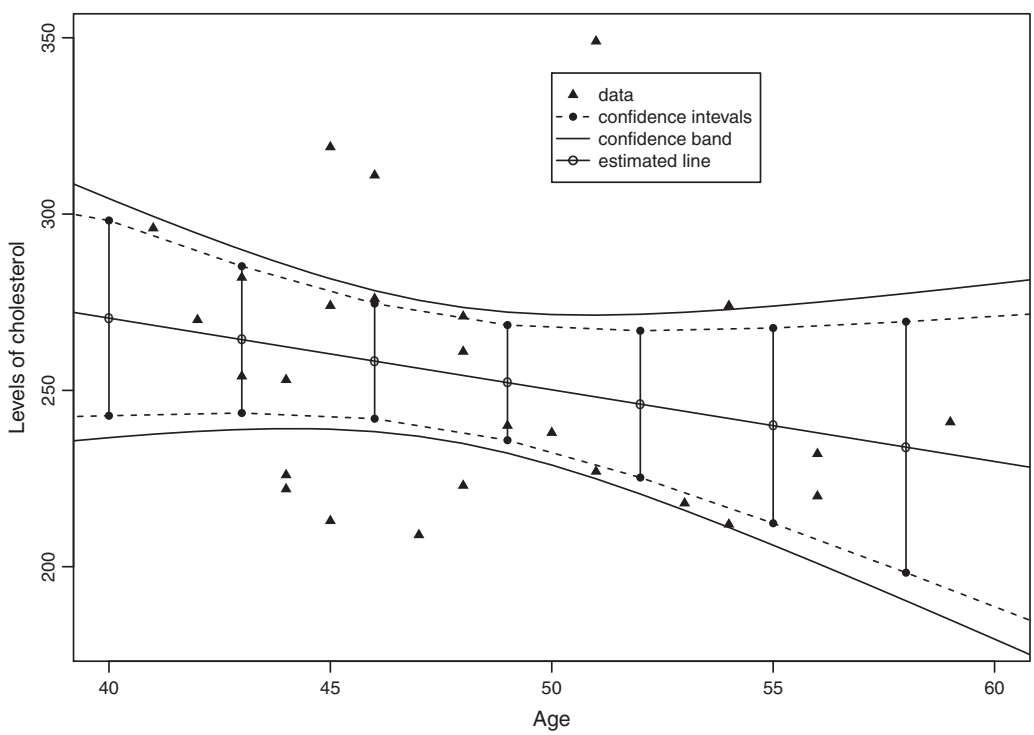


Figure 2.2 The Regression Line Estimated from Age and Cholesterol Data (Table 2.4): A Series of Connected Pointwise 95% Confidence Intervals (Dashed Line) and a 95% Confidence Band (Solid Line)

with estimated variance

$$\hat{V} = variance\,(\hat{y}_0) = variance\,[\bar{y} + \hat{b}(x_0 - \bar{x})] = variance\,(\bar{y}) + (x_0 - \bar{x})^2\ variance(\hat{b})$$
$$= S^2_{Y|x}\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]$$

where $S^2_{Y|x} = \sum (y_i - \hat{y}_i)^2/(n-2)$. Then a 95% pointwise confidence interval based on the estimate $\hat{y}_0$ for a specific selected value x_0 is

$$\hat{y}_0 \pm t_{0.975,\,26} \times \sqrt{\hat{V}} \qquad \text{where } \hat{y}_0 = \hat{a} + \hat{b}x_0.$$

The value $t_{0.975,\,26}$ is the 97.5th percentile from a *t*-distribution with $n - 2 = 26$ degrees of freedom. An identical 95% confidence interval is

$$\hat{y}_0 \pm \sqrt{F_{0.95,1,26}} \times \sqrt{\hat{V}}$$

where $F_{0.95,1,26}$ represents the 95th percentile from a *f*-distribution with 1 and 26 degrees of freedom because $t_{1-\alpha/2,n-2} = \sqrt{F_{1-\alpha,n-2}}$ for any sample size n (Chapter 1).

The estimated boundaries of a 95% confidence band for the estimated line y_x are calculated from the expression

$$\hat{y}_x \pm \sqrt{2F_{0.95,2,26}} \times \sqrt{\hat{V}}$$

where $F_{0.95,2,26}$ is the 95th percentile from a *f*-distribution with 2 and 26 degrees of freedom (Figure 2.1, solid line). This confidence band has a 0.95 probability that the entire line estimated by $\hat{y}_x$ is contained between these two bounds for all values of x (Figure 2.1). The justification of this confidence band requires a sophisticate mathematical argument.

Connecting the bounds of a series of pointwise confidence intervals produces a "band" but no accurate probabilistic interpretation exists (Figure 2.2, dashed line). The probability that a series of pointwise 95% confidence intervals contains the line estimated by $\hat{y}_x$ for a series of values x is not 0.95 but something less. Furthermore, the significance level of such a "band" would depend on the number of estimated pointwise confidence intervals and, in addition, no guarantee exits that the entire estimated line lies between the "bounds." To repeat, a confidence band dictates a 0.95 probability that the line estimated by $\hat{y}_x$ is complete contained between two confidence band boundaries for all values x.

The bounds of a confidence band are greater than the bounds constructed by connecting a series of pointwise confidence intervals. The difference is determined by the difference between $\sqrt{F_{1-\alpha/2,1,n-2}}$ and $\sqrt{2F_{1-\alpha,2,n-2}}$. The cost of simultaneity is the increased bounds of the confidence band. The proportional difference between the widths of a 95% confidence band and a series of connected 95% pointwise confidence intervals at any specific value x is close the 23% when the sample size is greater than $n = 30$.

Details: A Confidence Interval and a Statistical Test

The general form of a statistical test of the mean value $\bar{x}$ estimated from normally distributed data requires determination of two bounds based on a specific probability. The test statistic

bounds generated by a hypothesis that the sample mean value is an estimate of the population mean value denoted μ_0 are

$$\textit{lower bound} = a_0 = \mu_0 - 2\sigma_{\bar{x}} \quad \text{and} \quad \textit{upper bound} = b_0 = \mu_0 + 2\sigma_{\bar{x}}.$$

The hypothesis that $\bar{x} - \mu_0$ differs from zero by chance alone is then incorrectly rejected if $\bar{x} < a_0$ or $\bar{x} > b_0$ with probability close to 0.05 (0.046) when in fact the underlying mean value is μ_0.

The general form of a confidence interval based on the same estimated mean value $\bar{x}$ requires the determination of two bounds again based on a specific probability. The bounds are

$$\textit{lower bound} = \hat{A} = \bar{x} - 2\sigma_{\bar{x}} \quad \text{and} \quad \textit{upper bound} = \hat{B} = \bar{x} + 2\sigma_{\bar{x}}.$$

The underlying parameter μ, estimated by $\bar{x}$, will not be contained in the confidence interval $(\hat{A}, \hat{B})$ with probability also close to 0.05 (0.046).

Therefore, when $\bar{x}$ exceeds the upper bound of the test statistic or $\bar{x} > \mu_0 + 2\sigma_{\bar{x}}$, then $\bar{x} - 2\sigma_{\bar{x}} > \mu_0$ and the parameter μ_0 is not contained in the confidence interval $(\mu_0 < \hat{A})$. Similarly, when $\bar{x}$ is less that the lower bound of the test statistic or $\bar{x} < \mu_0 - 2\sigma_{\bar{x}}$, then $\bar{x} + 2\sigma_{\bar{x}} < \mu_0$ and again the parameter μ_0 is not contained in the confidence interval $(\mu_0 > \hat{B})$. Observing whether the theoretical mean value μ_0 is or is not contained in the confidence interval, therefore, produces the same statistical inference as a statistical test of the mean value $\bar{x}$ as a plausible estimate of μ_0. This description is exact when data are sampled from a normal distribution with a known variance. Otherwise, the relationships described are approximate but rarely misleading, particularly for mean values based on at least moderate sample sizes.

3

A Weighted Average

A weighted average is the statistical "jack-of-all-trades" or, as they say in baseball, a utility infielder. That is, this statistic plays important roles in a number of different situations because the selection of weights creates an exceptionally flexible analytic tool. Among a variety of possibilities, four useful applications demonstrate this versatility.

A weighted average is used:

1. To combine observations creating a single summary value
2. To estimate optimal summary values by combining strata-specific statistical estimates
3. To provide summary comparisons among disease or mortality rates classified into strata, for example, age strata and
4. To construct smooth parsimonious summary curves to describe relationships within data.

The definition of a weighted average is

$$\text{weighted average} = \frac{\sum w_i \hat{g}_i}{\sum w_i} \quad i = 1, 2, \ldots, n = \text{number of values } \hat{g}_i$$

where $\{\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_n\}$ represents a sample of n observations or ratios or rates or estimates or most any summary value. The choice of weights (denoted w_i) dictates the properties of a weighted average.

A Basic Application

For k mean values denoted $\bar{x}_j$ estimated from n_j observations, the selection of weights as the k sample sizes ($w_j = n_j$) produces an unbiased estimate of the overall mean value. In symbols, the weighted average is

$$\bar{x} = \frac{\sum n_j \bar{x}_j}{\sum n_j} = \frac{1}{N} \sum \sum x_{ij} \quad j = 1, 2, \ldots, k \quad \text{and} \quad i = 1, 2, \ldots, n_j$$

where the total number of observations is represented by $N = \sum n_j$. When k groups are the same size ($n_j = n$), then $N = kn$; for example, for $N = 12$ values $x_j = j$ sorted into $k = 3$ groups of $n_j = n = 4$ observations, then

Group 1	Group 2	Group 3
1 2 3 4	5 6 7 8	9 10 11 12
$\bar{x}_1 = 2.5$	$\bar{x}_2 = 6.5$	$\bar{x}_3 = 10.5$

Table 3.1 *Ratios: Three Mean Ratios That Are Weighted Averages*

	Weights	Ratio estimates	Properties
1	$w_i = 1$	$\bar{r}_1 = \frac{1}{n}\sum y_i/x_i$	Mean of ratio values
2	$w_i = x_i$	$\bar{r}_2 = \sum y_i / \sum x_i = \bar{y}/\bar{x}$	Ratio of mean values
3	$w_i = x_i^2$	$\bar{r}_3 = \sum x_i y_i / \sum x_i^2$	Change in y per change in x

The weighted average is

$$\bar{x} = \frac{4(2.5) + 4(6.5) + 4(10.5)}{4+4+4}$$

$$= 0.33\,(2.5) + 0.33\,(6.5) + 0.33\,(10.5) = 6.5 \quad \text{using weights} = w_i = \frac{n_i}{N} = \frac{4}{12} = 0.33$$

or directly

$$= \frac{10+26+42}{12} = \frac{78}{12} = 6.5$$

or

$$\bar{x} = \frac{1+2+3+4+5+6+7+8+9+10+11+12}{12} = \frac{78}{12} = 6.5.$$

Similarly, a weighted average of k probabilities $\hat{p}_j$ each estimated from n_j observations from each of k groups follows the same pattern, and

$$\hat{p} = \frac{\sum w_j \hat{p}_j}{\sum w_j} = \frac{\sum n_j \hat{p}_j}{\sum n_j} = \sum \left[\frac{n_j}{N}\right] \hat{p}_j \quad \text{with again weights } w_j = \frac{n_j}{N}$$

is the summary estimated probability produced by combining k estimated probabilities $\hat{p}_j$.

Ratios and Weighted Averages

A fundamental summary characterizing the joint influence from two variables is a ratio. In symbols, such a ratio is $r_i = y_i/x_i$ when x_i and y_i represent members of a pair of n observations (x_i, y_i). Different mean ratios emerge when a weighted average is used to combine a series of ratio values into a single summary depending on the choice of weights (again denoted w_i). In general, a weighted average summary ratio is

$$\textit{mean ratio} = \bar{r} = \frac{\sum w_i (y_i/x_i)}{\sum w_i} = \frac{\sum w_i r_i}{\sum w_i} \quad i = 1, 2, \ldots, n = \text{number of pairs.}$$

Three different summary mean ratios $\bar{r}$ are created by the choice of weights $w_i = 1$ or $w_i = x_i$ or $w_i = x_i^2$ (Table 3.1).

An ratio important in a study of pregnancy outcomes is the mother's body mass index (denoted bmi_i). These ratios estimated from $n = 779$ mothers measured for height (ht_i) and weight at delivery (wt_i) illustrate the three weighted average ratios. Estimated ratios for each women are

$$bmi_i = wt_i / ht_i^2 = 1, 2, \ldots, 779$$

Table 3.2 *Results: Three Mean Ratios Estimated from n = 779 Mothers Calculated from Their Heights (Meters) and Delivery Weights (Kilograms)*

	BMI		
Ratios	Weights	Means	*s. e.*
$\bar{r}_1$	$w_i = 1.0$	24.2	0.44
$\bar{r}_2$	$w_i = x_i$	24.1	0.44
$\bar{r}_3$	$w_i = x_i^2$	24.1	0.40

where maternal weight is measured in kilograms and height in meters. Three choices for weights produce three summary mean ratios and their standard errors (*s.e.*) that have essentially the same values (Table 3.2).

The similarity observed among the three estimates is not always the case. Data occur where important differences arise. The optimum choice of a summary mean value is then not obvious. Subject matter considerations might dictate a useful choice. Ratio estimates traditionally used for specific situations are another possibility. For example, mortality rates are almost always estimated as a ratio of mean values ($\bar{r}_2$) (Chapter 16). The ratio $\bar{r}_2$ is often chosen because it is the more mathematically tractable summary ratio. The ratio $\bar{r}_3$ produces the slope of a line describing the x/y-relationship. Thus, like many decisions required in a statistical analysis, no specific guidelines exist that lead to an unequivocal choice.

The estimated mean ratio created by weights that are the squared denominator values ($\bar{r}_3$ where $w_i = x_i^2$) creates an expression for the estimated coefficient of the simplest linear regression model. For the linear model $y_i = bx_i$ or ratio $y_i/x_i = b$, the estimate of the slope b or mean ratio b is the weighted average $\hat{b} = \sum x_i y_i / \sum x_i^2 = \bar{r}_3$ (Table 3.2).

In fact, the ratio

$$r_i = \frac{y_i - \bar{y}}{x_i - \bar{x}} = \hat{b}_i$$

using weights $w_i = (x_i - \bar{x})^2$ produces the weighted average

$$\bar{r}_3 = \frac{\sum w_i r_i}{\sum w_i} = \frac{\sum w_i \hat{b}_i}{\sum w_i} = \frac{\sum (x_i - \bar{x})^2 \left[\frac{y_i - \bar{y}}{x_i - \bar{x}}\right]}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \hat{b}.$$

The weighted average $\bar{r}_3$ is the ordinary least squares estimate of the regression coefficient b from the simple linear regression model $y_i = a + bx_i$ (Chapters 7 and 9). In addition, the estimate $\bar{r}_3$ is the maximum likelihood estimate of the slope b (Chapters 19 and 27).

Eight pairs of artificial values (x_i, y_i) illustrate:

$$(x_i, y_i) = \{(1, 3), (9, 8), (8, 9), (6, 7), (3, 2), (1, 1), (3, 6), \text{ and } (9, 4)\}.$$

A plot (Figure 3.1) displays the estimated slope $\hat{b} = \bar{r}_3 = 0.610$ as a weighted average of eight slopes/ratios ($\hat{b}_i$ or r_i).

Probably the most commonly used mean ratio is $\bar{r}_2$ where weights $w_i = x_i$ cause the weighted average to become the mean of the values y_i (numerator is $\bar{y}$) divided by the mean

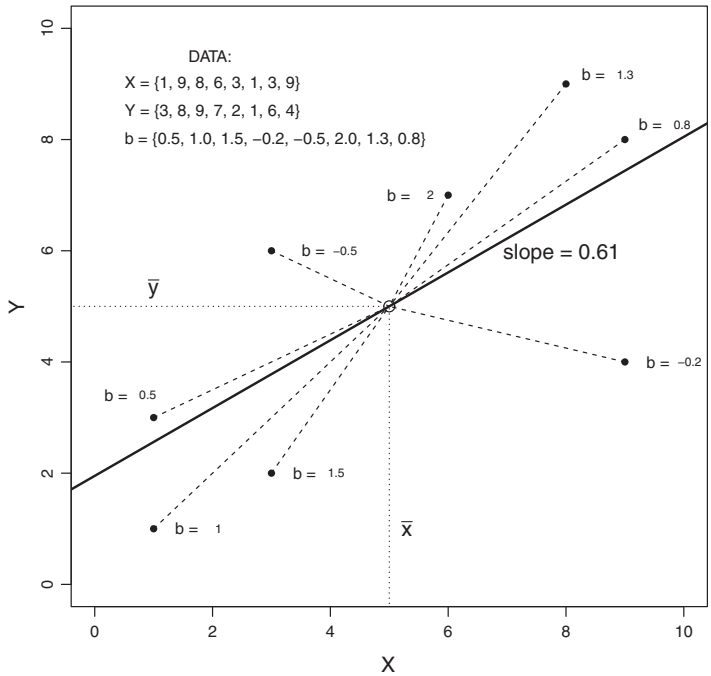


Figure 3.1 Eight Slopes $\hat{b}_i$ Estimated from Each of $n = 8$ Pairs of Observations (x_i, y_i) to Demonstrate That a Weighted Average of Individual Slopes $\hat{b}_i$ (Dashed Lines) Estimates the Slope $\hat{b}$ (Solid Line)

of the values x_i (denominator is $\bar{x}$). For the mean ratio value estimated by $\bar{r}_2$, a natural relationship between the estimate $\bar{r}_2$ and the value estimated r is

$$\bar{r}_2 - r = \frac{\bar{y}}{\bar{x}} - r = \frac{1}{\bar{x}}(\bar{y} - r\bar{x})$$

leading to an expression for the estimation of the variance of the distribution of the weighted average $\bar{r}_2$ as

$$variance(\bar{r}_2) = \frac{1}{n\bar{x}^2} \sum (y_i - \bar{r}_2 x_i)^2.$$

Again n represents the number of pairs (x_i, y_i) that make up the weighted average $\bar{r}_2$.

Estimates Weighted by Reciprocal Variances

The choice of weights to create a weighted average is essentially unrestricted. When a weighted average is created to summarize a series of independent estimated values, one choice, among many possibilities, is to weight each value included in the average by the reciprocal of the estimated variance of that value. Thus, the influence of precise estimates is large, and the influence of less precise estimates is relatively smaller because the reciprocal of the variance has exactly the same property. In fact, such a weighted average is optimum in the sense it has minimum variance among all possible unbiased weighted averages. Thus, for

k independent summary statistics, again represented by $\hat{g}_i$, an expression for this weighted average is

$$\text{mean weighed average} = \bar{g} = \frac{\sum w_i \hat{g}_i}{\sum w_i} i = 1, 2, \ldots, k = \text{number of estimates}$$

where the k weights are $w_i = 1/variance(\hat{g}_i)$. Furthermore, the variance of the weighted average $\bar{g}$ is estimated by a simple function of weights given by the expression

$$\hat{v} = variance(\bar{g}) = \frac{1}{\sum w_i}.$$

Formally, the estimate of the variance of $\bar{g}$ is

$$\begin{aligned}
\hat{v} &= variance(\bar{g}) = variance\left[\frac{\sum w_i \hat{g}_i}{\sum w_i}\right] \\
&= \left[\frac{1}{\sum w_i}\right]^2 variance\left(\sum w_i \hat{g}_i\right) \\
&= \left[\frac{1}{\sum w_i}\right]^2 \sum w_i^2 \; variance(\hat{g}_i) = \left[\frac{1}{\sum w_i}\right]^2 \sum w_i^2 \frac{1}{w_i} \\
&= \left[\frac{1}{\sum w_i}\right]^2 \sum w_i = \frac{1}{\sum w_i} \quad i = 1, 2, \ldots, k = \text{number of estimates } \hat{g}_i.
\end{aligned}$$

The reciprocal of the sum of the weights reflects the variance of the weighted average from two points of view. First, the more values (larger k) included in a summary mean value, the more precise the estimate. This property is a property of all mean values. Second, the weights reflect the relative importance of the contribution of each component. Thus, the sum of the weights accounts for the influence from the number of values, and each contribution is proportional to its effectiveness in the sum. Together, they determine the variance of the weighted average. Other choices of weights create a weighted average but with a larger variance.

A simple example of a weighted average using weights that are the reciprocal of the variance is the sample mean value. In this case, each sampled value has the same variance when the estimated mean value consists of n independent observations sampled from a single population. For such a sample $\{x_1, x_2, \ldots, x_n\}$ from a distribution with variance σ_X^2, then

$$\text{mean value} = \frac{\sum w_i x_i}{\sum w_i} = \frac{\sum x_i}{n} = \bar{x} \quad \text{where weights} = w_i = w = \frac{1}{\sigma_X^2}$$

and the variance of the mean value is

$$variance(\bar{x}) = \frac{1}{\sum w_i} = \frac{\sigma_X^2}{n}.$$

The same principles apply to a summary mortality rate calculated from a series of age-specific strata (Table 3.3). Like many ratio estimates, the statistical properties of a rate are simpler and more accurately evaluate when the logarithm of a rate is used. Therefore, strata-specific

Table 3.3 *Data: Age-Specific Childhood Leukemia Mortality Rates per Million at Risk (age < 20 Years – California Vital Records, 1999 to 2007)*

Ages	At risk (p_i)	Deaths (d_i)	Rates (r_i)[a]
<1	4,731,047	33	6.98
1–4	18,368,181	181	9.85
5–9	23,326,046	231	9.90
10–14	23,882,120	260	10.89
15–19	23,156,490	245	10.58

[a]Rates per million population $= r_i = (d_i/p_i) \times 1{,}000{,}000$ persons at risk.

log-rates, denoted $log(r_i)$, create the weighted average

$$\text{mean } log\text{-rate} = \overline{log(r)} = \frac{\sum w_i \, log(r_i)}{\sum w_i} \quad \text{where } rate = r_i = \frac{d_i}{P_i}$$

when d_i deaths occur among P_i persons at risk within the ith strata. The reciprocal variance weights are $w_i = 1/variance(log[r_i]) = d_i$ because the estimated $variance(log[r_i]) = 1/d_i$ (Chapter 27). The weighted average becomes

$$\overline{log(r)} = \frac{\sum d_i \, log(r_i)}{\sum d_i} = \frac{1}{D} \sum d_i \, log(r_i)$$

where $D = \Sigma d_i$ represents the total number of deaths.

For childhood leukemia data, the weighted average mortality *log*-rate is $\overline{log(r)} = -11.493$ or $r = e^{-11.493} \times 1{,}000{,}000 = 10.203$ deaths per million (Table 3.3, last column). The estimated variance is $\hat{v} = 1/\Sigma w_i = 1/\Sigma d_i = 1/D = 1/950 = 0.00105$. An approximate 95% confidence interval based on $\overline{log(r)}$ is then

$$\hat{A} = \overline{log(r)} - 1.960\sqrt{\hat{v}} = -11.493 - 1.960(0.0324) = -11.556$$

and

$$\hat{B} = \overline{log(r)} + 1.960\sqrt{\hat{v}} = -11.493 + 1.960(0.0324) = -11.429.$$

For the California data, an approximate 95% confidence interval based on the estimated rate 10.203 leukemia deaths per 1,000,000 children at risk becomes $(e^{-11.556}, e^{-11.430}) \times 1{,}000{,}000 = (9.575, 10.873)$. Note that the precision of the estimated summary rate is entirely determined by the total number of deaths ($D = \Sigma d_i = 950$) despite the fact the data consist of more than 93 million at-risk children (Chapter 27).

A summary rate ratio is a statistic frequently used to compare mortality and disease rates between two groups or populations. Table 3.4 contains age-specific childhood leukemia mortality rates r_i again from California vital records (repeated) as well as national rates R_i from U.S. vital records.

Table 3.4 *Data: Age-Specific Counts of Childhood Leukemia Deaths and Mortality Rates per* 1,000,000 *Persons at Risk from California and United States Populations (Age* < 20 *Years – 1999–2007)*

	California			United States		
Age	Deaths (d_i)	At risk (p_i)	Rates (r_i)[a]	Deaths (D_i)	At risk (P_i)	Rates (R_i)[a]
<1	33	4,731,047	6.98	238	36,243,470	6.57
1–4	181	18,368,181	9.85	1118	142,332,392	7.85
5–9	231	23,326,046	9.90	1253	179,720,009	6.97
10–14	260	23,882,120	10.89	1505	186,897,294	8.05
15–19	245	23,156,490	10.58	1950	185,997,993	10.48
Total	950	93,463,884	10.16	6064	731,191,158	8.29

[a]Per million children at risk.

An effective summary comparison is a weighted average of age-specific rate ratios (denoted $\hat{rr}_i$). A weighted average estimate yields a summary rate ratio denoted $\overline{rr}$ and a 95% confidence interval, following the pattern of the previous weighted average. The mean *log*-rate ratio is

$$\text{mean } log\text{-rate ratio} = \overline{log(rr)} = \frac{\sum w_i \, log(\hat{rr}_i)}{\sum w_i} \quad \text{where } \hat{rr}_i = \frac{r_i}{R_i} = \frac{d_i/p_i}{D_i/P_i} \text{[27].}$$

The reciprocal variance weights are

$$\text{weights} = w_i = \frac{1}{variance(log[\hat{rr}_i])} \quad \text{where } variance(log[\hat{rr}_i]) = \frac{1}{d_i} + \frac{1}{D_i}.$$

The age-specific rate ratios and weights from the childhood leukemia data for each of the five age strata are

$$\text{rate ratios estimated by} = \hat{rr}_i = r_i/R_i : 1.062, 1.255, 1.420, 1.352, \text{ and } 1.001$$

and

$$\text{weights } w_i = 1/variance(log[\hat{rr}_i]) : 28.983, 155.780, 195.040, 221.700, \text{ and } 217.654.$$

The estimated mean *log*-rate ratio $\overline{log(rr)} = 0.213$ yields the summary ratio $\overline{rr} = e^{0.213} = 1.237$. The estimated variance, again based on the reciprocal of the sum of the weights, is $\hat{v} = variance\left[\overline{log(rr)}\right] = 1/\sum w_i = 1/819.15 = 0.00122$. An approximate 95% confidence interval becomes

$$\overline{log(rr)} \pm 1.960\sqrt{variance[\overline{log(rr)}]} = 0.213 \pm 1.960(0.035) \rightarrow (0.144, 0.281)$$

making the confidence interval for the rate ratio $(e^{0.144}, e^{0.281}) = (1.155, 1.325)$, based on the estimate $\overline{rr} = 1.237$. Again, note, the precision of the summary estimate $\overline{rr}$ is again entirely determined by the numbers of deaths.

An unreasonably popular measure applied to summarize association between two binary variables classified into a 2 × 2 tables is the odds ratio. Like many estimates, a series of odds ratios can be summarized by a weighted average. Data that were collected to study

Table 3.5 *Data: Mother/Infant Pairs Classified by Smoking Exposure, Birth Weight, and Ethnicity (Four* 2×2 *Tables – University of California, San Francisco Perinatal Database, 1980 to 1990)*

White	<2500 g	≥2500 g	Total
smokers	98	832	930
nonsmokers	169	3520	3689
Total	267	4352	4619
African American	<2500 g	≥2500 g	Total
Smokers	54	227	281
Nonsmokers	55	686	741
Total	109	913	1022
Hispanic	<2500 g	≥2500 g	Total
Smokers	11	85	96
Nonsmokers	61	926	987
Total	72	1011	1083
Asian	<2500 g	≥2500 g	Total
Smokers	7	102	109
Nonsmokers	90	1936	2026
Total	97	2038	2135

the influence of smoking on the likelihood of a low-birth-weight infant among four ethnic groups illustrate (Table 3.5). The notation for an ethnic-specific 2×2 table is presented in Table 3.6 (Chapter 6). These 2×2 tables produce four estimates of ethnic-specific odds ratios (Table 3.7).

To create a summary odds ratio from these ethnic-specific odds ratio estimates, statisticians N. Mantel and W. Haenszel suggest a weighted average using weights $w_i = b_i c_i / n_i$. From the strata-specific estimated odds ratios, represented by $\hat{or}_i$, the summary odds ratio becomes

$$\bar{or}_{mh} = \frac{\sum w_i \hat{or}_i}{\sum w_i} = \frac{\sum \frac{b_i c_i}{n_i}\left[\frac{a_i d_i}{b_i c_i}\right]}{\sum \frac{b_i c_i}{n_i}} = \frac{\sum a_i d_i / n_i}{\sum b_i c_i / n_i} \quad \text{where } \hat{or}_i = \frac{a_i d_i}{b_i c_i}.$$

The Mantel-Haenszel estimated summary odds ratio from the low-birth-weight data is $\bar{or}_{mh} = 2.448$ (Table 3.7).

Alternatively, following again the previous pattern, logarithms of the odds ratios can be used to estimate a weighted average summary and its variance. Specifically, the estimated

Table 3.6 *Notation: A* 2×2 *Table for the Smoking and Birth Weight Data for the* i*th Strata*

Strata i	<2500 g	≥2500 g
Smokers	a_i	c_i
Nonsmokers	b_i	d_i

Table 3.7 *Data/Estimates: Odds Ratios and Their Variances from Infants Tabulated by Maternal Smoking Exposure (Smokers and Nonsmokers), Low Birth Weight (<2500 g and ≥2500 g), and Ethnicity (Tables 3.5 and 3.6)*

	Data				Estimates			
Ethnicity	a_i	b_i	c_i	d_i	$\hat{or}_i$	$log(\hat{or}_i)$	$\hat{v}_i$	w_i
White	98	832	169	3520	2.453	0.897	0.018	56.795
African American	54	227	55	686	2.967	1.088	0.043	23.494
Hispanic	11	85	61	926	1.965	0.675	0.120	8.323
Asian	7	102	90	1936	1.476	0.390	0.164	6.087

log-mean summary value is

$$\overline{log(or)} = \frac{\sum w_i \, log(\hat{or}_i)}{\sum w_i} \quad i = 1, 2, \ldots, k = \text{number of odds ratios}$$

with weights chosen as the reciprocal of the variances of $log(\hat{or}_i)$, estimated (Chapter 27) by

$$variance(log[\hat{or}_i]) = \hat{v}_i = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}.$$

From the birth weight data (Table 3.7), the mean *log*-odds ratio is $\overline{log(or)} = 0.892$, and the associated estimated variance is $\hat{v} = 1/\Sigma w_i = 1/94.699 = 0.0106$ where $w_i = 1/\hat{v}_i$ (Table 3.7). The estimated summary odds ratio follows as $\overline{or} = e^{0.892} = 2.441$. The test statistic with an approximate normal distribution

$$z = \frac{\overline{log(or)} - log(1)}{\sqrt{\hat{v}}} = \frac{0.892 - 0}{0.103} = 8.685$$

yields a p-value < 0.001 for the birth weight data. The parallel approximate 95% confidence interval is 95% confidence interval $= \overline{log(or)} \pm 1.960\sqrt{\hat{v}} = 0.892 \pm 1.960(0.103) \rightarrow (0.691, 1.094)$.

Based on the estimate $\overline{or} = e^{0.892} = 2.441$, the estimated confidence interval becomes $(e^{0.691}, e^{1.094}) = (1.996, 2.986)$.

A Puzzle

A number of years ago the registrar at the University of California, Berkeley, received a letter critical of the fact that the university admitted more male than female first-year students. The registrar looked into the situation and found indeed that the percentage of males admitted was greater than females. Upon further investigation, it was discovered that a vast majority of academic departments admitted more female than male applicants. This apparent contradiction didn't appear to be possible.

The following artificial data present the issues that created this apparent contradiction and are used to describe a simple weighted average that usefully reflects admission policy (Table 3.8).

The percentage of males admitted is $(1000/2700) \times 100 = 37.0\%$, and for females the percentage is $(600/1750) \times 100 = 34.3\%$. Furthermore, similar to the actual situation, all

Table 3.8 *Data: Artificial Data to Explore Admission Data That Appear Contradictory*

	Males			Females			
	Applied	Admitted	%	Applied	Admitted	%	Total
Sciences	800	480	60.0	100	70	70.0	900
Professional	600	300	50.0	50	30	60.0	650
Social Sciences	300	120	40.0	600	300	50.0	900
Liberal Arts	1000	100	10.0	1000	200	20.0	2000
Total	2700	1000	37.0	1750	600	34.3	4450

four disciplines admitted a higher percentage of female than male applicants. Specifically, in the four disciplines 10% more female applicants were admitted than males. Therefore, at first glance, it seems that the percentage of males admitted cannot possibly exceed the percentage of females admitted.

A key issue is the influence of considerably different male and female application distributions. For example, the percentage of male applicants to the sciences is $(800/2700) \times 100 = 30\%$, and the same value for females is $(100/1750) \times 100 = 6\%$. The total male and female percentages admitted depends on both the distribution of applicants and admissions. Without accounting for differing influences from the distribution of applicants, a summary male-female comparison will not accurately reflect the pattern of admissions.

A simple method to account for the differing application distributions is to create a single "standard application rate" to calculate both male and female admission rates with a weighted average. Choosing as a standard the total application distribution (Table 3.8, last column) as weights produces two adjusted admission rates that reflect only the difference in admissions. The created "application" distribution is identical for both groups. The adjusted admission percentages are then

$$\text{male-adjusted} = \frac{1}{4450}\,[900(0.6) + 650(0.5) + 900(0.4) + 2000(0.1)] = 0.320$$

and

$$\text{female-adjusted} = \frac{1}{4450}\,[900(0.7) + 650(0.6) + 900(0.5) + 2000(0.2)] = 0.420.$$

This comparison clearly shows that the female admissions rates are, as expected, higher than male rates (difference = 10%) when the influence of the differing application patterns is removed. To repeat, it is made identical for both males and females. The process of comparing male and female admissions rates based on a single standard distribution is an example of a technique traditionally called *direct adjustment.*

The comparison of estimates with differing distributions of observations among strata is occasionally a source of misinterpretation. For example, the infant mortality rate of African American infants divided by the infant mortality rate of white infants yields a value of 1.63, leading to the widely held belief that African American infants have higher infant mortality risk. The same ratio of adjusted rates is 0.75. This close to complete reversal is caused by not accounting for the substantially different black-white distributions of infant birth weights that are a central element in the risk of infant mortality (Chapter 9).

Age-Adjusted Rates Using Weighted Averages

Parametric comparisons of age-adjusted mortality and disease rates are efficiently made using Poisson regression models (Chapter 9). Simple and model-free age-adjusted comparisons, however, are created with weighted averages. The notation for age-specific numbers of deaths and mortality rates is the following:

Population 1	p_i = number of at-risk individuals
	d_i = number of deaths
	$r_i = d_i/p_i$ = estimated rate
Population 2	P_i = number of at-risk individuals
	D_i = number of deaths
	$R_i = D_i/P_i$ = estimated rate

where $i = 1, 2, \ldots,$ and $k =$ number of strata.

Relevant estimated mortality rates from these two population are the following:

$$\text{Rate from population 1: } r = \frac{\sum w_i r_i}{\sum w_i} \text{ with weights } w_i = p_i,$$

$$\text{Rate from population 2: } R = \frac{\sum w_i R_i}{\sum w_i} \text{ with weights } w_i = P_i.$$

A common summary comparison of age-specific rates between two populations is a statistic called a *standard mortality ratio* (*smr*). The estimate $\hat{smr}$ is expressed as

$$\text{standard mortality ratio} = \hat{smr} = \frac{d}{E}$$

where d represents the observed total number of deaths in one population and E represents the estimated total number of deaths calculated as if a second population has the identical age distribution as the first population ($p_i = P_i$).

For a comparison of California and U.S. childhood leukemia mortality rates, an expected number of deaths (denoted E) is the weighted average $E = \Sigma p_i R_i$ calculated as if the United States has the identical age-specific population distribution as California (Table 3.4). For the California and U.S. childhood leukemia mortality data, the expected number of deaths is then $E = 773.06$ based on the U.S. age-specific rates (R_i) with California age-specific populations (p_i). The observed number of California deaths $\Sigma p_i r_i = \Sigma d_i = d = 950$. The estimated leukemia standard mortality ratio is $\hat{smr} = 1.229$ ($\hat{smr} = d/E = 950/773.06 = 1.229$). The variance of the $\hat{smr}$ is estimated by $\hat{v}_{smr} = \hat{smr}/E = \hat{smr}^2/d$[27]. Specifically, the estimated variance is

$$\hat{v}_{smr} = variance(\hat{smr}) = variance\left(\frac{d}{E}\right) = \frac{1}{E^2} variance(d)$$

$$= \frac{d}{E^2} = \frac{\hat{smr}}{E} = \frac{d/E}{E} \times \left[\frac{d}{E} \times \frac{E}{d}\right] = \frac{\hat{smr}^2}{d}.$$

A confidence interval constructed from the logarithm of the estimated *smr* again identifies the accuracy and improves the interpretation of the estimated bounds. Thus, an approximate

95% confidence interval is

$$95\%\text{ confidence intervals bounds} = log(\hat{smr}) \pm 1.960\sqrt{\hat{v}_{log(\hat{smr})}}.$$

The *variance*($log[\hat{smr}]$) is estimated by $\hat{v}_{log(smr)} = 1/d + 1/D = 1/950 + 1/6064 = 0.00122$ (d = total California deaths and D = total U.S. deaths; see Table 3.4) (Chapter 27). Based on the estimate $log(\hat{smr}) = log(1.229) = 0.206$, the confidence interval is

$$0.206 \pm 1.960\,(0.035) \rightarrow (0.138, 0.275)$$

and the confidence interval based on the estimate $\hat{smr} = e^{0.206} = 1.229$ becomes ($e^{0.138}$, $e^{0.275}$) = (1.148, 1.316). The compared mortality rates summarized by a standard mortality ratio indicates that differences in the risk of childhood leukemia likely exist ($smr \neq 1$) between California and the rest of the nation that are not attributable to differences in age distributions.

The same comparison estimated as an adjusted rate ratio is

$$\hat{smr} = \frac{r}{r^*} = \frac{\sum p_i r_i / \sum p_i}{\sum p_i R_i / \sum p_i}.$$

Again the U.S. "rate" is calculated as if the U.S. age distribution is exactly the same as the California age distribution ($P_i = p_i$), yielding the California/U.S. ratio = $r/r^* = 1.229$; that is, the estimated *smr* is a special case of direct adjustment.

Smoothing – A Weighted Average Approach

Smoothing techniques produce often effective and assumption-free summaries of relationships within sampled data that are primarily descriptive. They paint a picture, and, like most pictures, the results depend a great deal on the artist. Without associated probability distributions, assumptions, or statistical hypotheses, smoothing techniques attempt to distill an underlying relationship from obscuring influences such as random or haphazard variation, bias, and misclassification. Smoothing techniques have the same goals as estimating a specific curve postulated to describe a relationship within sampled data. Estimates based on a specified curve, however, require sometimes tenuous and frequently unsupported assumptions. When a natural or justifiable theoretical curve does not exist, a smoothing process remains an assumption-free alternative estimate to represent the relationship under investigation, simply producing a frequently useful summary description. Payment for simplicity and utility is a lack of rigor, and certainly no p-values or other probabilistic assessments are possible. The sometimes strength of a smoothing approach is also its sometime weakness. Because of its unrestricted nature, a smoothing approach potentially incorporates artificial properties or extreme random variation into the resulting estimated curve that are difficult to separate from truly systematic influences. Unlike fitting a model-based curve, a comparison between the observed and theoretical values is not available, making relationships created from smoothed data vulnerable to misinterpretation. From the perspective that statistical analysis is about learning from collected data, smoothed curves can reveal features of the sampled observations that might otherwise go unnoticed.

A principle that underlies most smoothing techniques is a process to increase the similarity of observations in proximity. The phrase "to make values locally smooth" is frequently used. A large variety of techniques exist to locally smooth observations, and computer software

Table 3.9 *Example: An Illustration of Weighted Average Smoothing a Single Observation to Become Similar to Its Neighboring Values (Locally Smoothed)*

	Weights					
	w_1	w_2	w_3	w_4	w_5	$\bar{x}_3{}^a$
0	0.0	0.0	1.0	0.0	0.0	10.0
1	0.0	0.1	0.8	0.1	0.0	8.8
2	0.1	0.1	0.6	0.1	0.1	7.0
3	0.1	0.2	0.4	0.2	0.1	5.8
4	0.2	0.2	0.2	0.2	0.2	4.0

[a]Weighted average $= \bar{x} = \sum w_i x_i / \sum w_i =$ smoothed value.

allows some of these techniques to be complex. On the other hand, weighted averages are usually useful, simple, and directly applied. Often little is gained by more complicated and sophisticated methods.

A small example illustrates in the simplest terms, perhaps simplistic terms: the process of a weighted average used to locally smooth a specific value. Consider the sequence of values $x = \{1, 4, 10, 4, 1\}$ and a weighted average used to make the original value $x_3 = 10$ locally similar to the other four values.

Table 3.9 displays four choices for weights so that a weighted average makes the value $x_3 = 10$ increasingly more similar to its neighbors. The first set of weights, given by weights $= w_i = \{0.0, 0.0, 1.0, 0.0, 0.0\}$, produce a weighted average that does not change the value x_3. The second set of weights, $w_i = \{0.0, 0.1, 0.8, 0.1, 0.0\}$, starts the process of making the value $x_3 = 10$ similar to its neighbors. The value 10 is reduced to the weighted average of $\bar{x}_3 = 8.8$. As the weights become more similar among themselves, the smoothed value becomes correspondingly more locally similar. When the weights are equal ($w_i = 0.2$), the original value $x_3 = 10$ becomes the mean of all five values ($\bar{x}_3 = 4.0$), the mean of the original values. In a statistical sense, the mean value is the "closest" value possible. Technically, the mean value produces the smallest sum of squared deviations for the five values. Applying such a weighted average sequentially to each observation in a sample of data, the selected weights determine the degree of local smoothness by reducing local random variation as well as minimizing other potentially obscuring influences in the neighborhood of each value, typically producing a parsimonious summary curve.

A somewhat more realistic example continues to illustrate weighted average smoothing. A polynomial model given by

$$y_i = x_i^3 - 9x_i^2 + x_i + 150 + e_i$$

creates an artificial sample of observations. The symbol e_i represents a haphazard, perhaps random, influence on each x/y observation. The polynomial curve is displayed in Figure 3.2 (upper left: $e_i = 0$). A sample of $n = 10$ x_i-values produces the model generated y_i-values (upper right: $e_i \neq 0$). Specifically, these 10 selected example values are the following:

i	1	2	3	4	5	6	7	8	9	10
x_i	0.31	1.27	2.23	3.21	4.17	5.13	6.11	7.07	8.03	9.01
y_i	174.07	145.70	160.61	82.36	50.93	35.94	44.26	−21.78	86.27	242.46

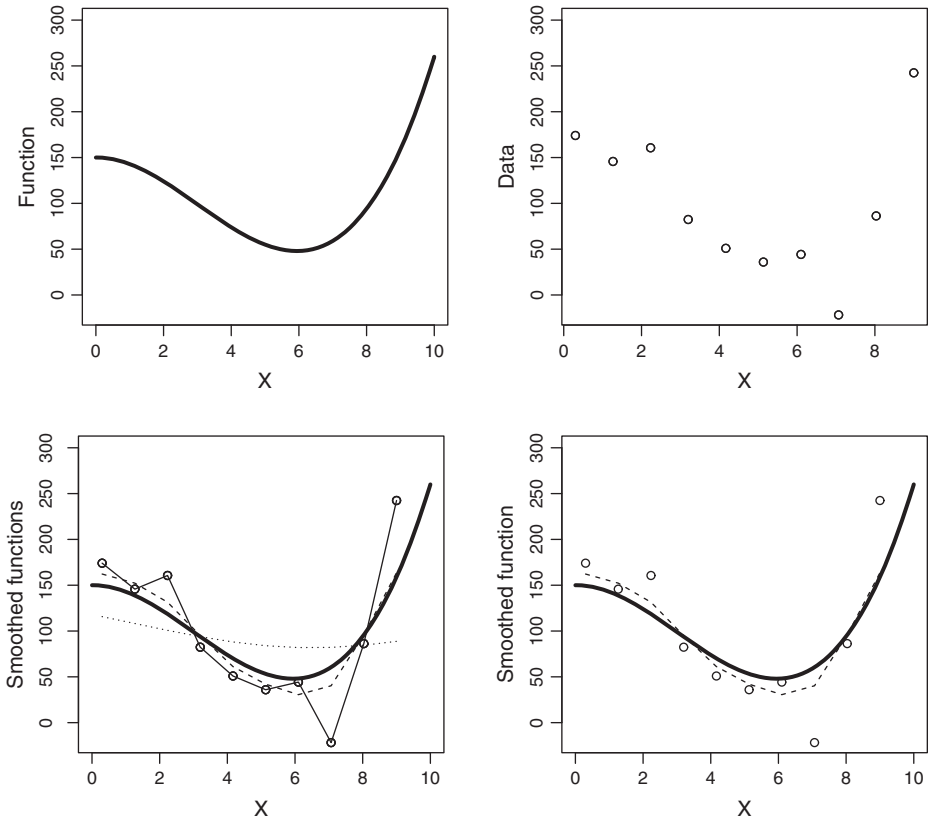


Figure 3.2 Polynomial-Generated "Data" and the Locally Smoothed Curves Based on Three Bandwidths: $h = \sigma = 0.1$, 1.0, and 3.0

A large number of choices for weights exists. A common and simple choice is the heights of a normal distribution with mean value equal to the observation to be locally smoothed. For the example, the fifth data value $x_5 = 4.17$ generates the 10 weights

$$weights = w_i = \{0.00, 0.01, 0.06, 0.25, 0.40, 0.25, 0.06, 0.01, 0.00, 0.00\}$$

from a normal distribution with mean value $\mu = x_5 = 4.17$ and standard deviation $= \sigma = 1.0$ (Table 3.10, row 5). The weighted average smoothed value associated with the observed value $y_5 = 50.93$ becomes $\bar{y}_5 = \sum w_i y_i / \sum w_i = 61.13$. Figure 3.3 displays the 10 weights corresponding to each sample value (dots). Specifically these weights are the heights for the normal distribution, mean value $x_5 = 4.17$, with standard deviation $= \sigma = 1.0$ (circles). Table 3.10 contains sets of 10 normal distribution-generated weights similarly applied sequentially to each of the 10 y-values. The resulting smoothed values are denoted $\bar{y}_i$ (Table 3.10, last column).

These weights are simply adjusted to create the degree of smoothness of the estimated curve. The choice of standard deviation of the normal distribution produces different degrees of smoothness. In the context of statistical smoothing, the choice of the distribution of the weights, such as the standard deviation of the normal distribution, is called the *bandwidth*

Table 3.10 *Results: Weights* (wt)[a] *Used to Smooth 10 Points* (y) *That Produce Smoothed Estimates* ($\bar{y}$) *at Each Point* (x)

Obs.	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	x_i	y_i	$\bar{y}_i$
1	0.40	0.25	0.06	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.31	174.07	162.17
2	0.25	0.40	0.25	0.06	0.01	0.00	0.00	0.00	0.00	0.00	1.27	145.70	152.36
3	0.06	0.25	0.40	0.25	0.06	0.01	0.00	0.00	0.00	0.00	2.23	160.61	131.76
4	0.01	0.06	0.25	0.40	0.25	0.06	0.01	0.00	0.00	0.00	3.20	82.36	94.57
5	0.00	0.01	0.06	0.25	0.40	0.25	0.06	0.01	0.00	0.00	4.17	50.93	61.13
6	0.00	0.00	0.01	0.06	0.25	0.40	0.25	0.06	0.01	0.00	5.13	35.94	42.09
7	0.00	0.00	0.00	0.01	0.06	0.25	0.40	0.25	0.06	0.01	6.11	44.26	30.54
8	0.00	0.00	0.00	0.00	0.01	0.06	0.25	0.40	0.25	0.06	7.07	−21.78	40.19
9	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.25	0.40	0.25	8.03	86.27	94.96
10	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.25	0.40	9.01	242.46	164.12

[a]Heights of a normal distribution with mean $\mu = x_i$ and standard deviation $\sigma = 1.0$.

(denoted h). Figure 3.2 (lower left) displays three choices of bandwidths where $h = \sigma = \{0.1, 1.0, \text{and } 3.0\}$. The choice of bandwidth (the standard deviation of the normal distribution) dictates the degree of smoothness of the summary curve. A small bandwidth ($h = 0.1$) produces a curve that differs little from the data (solid line).

A large bandwidth ($h = 3.0$) produces an almost horizontal line (dotted line). An intermediate choice ($h = 1.0$) accurately reflects underlying curve (dashed line). The smoothed curve generated by the choice of bandwidth $h = 1.0$ and the underlying model curve generated

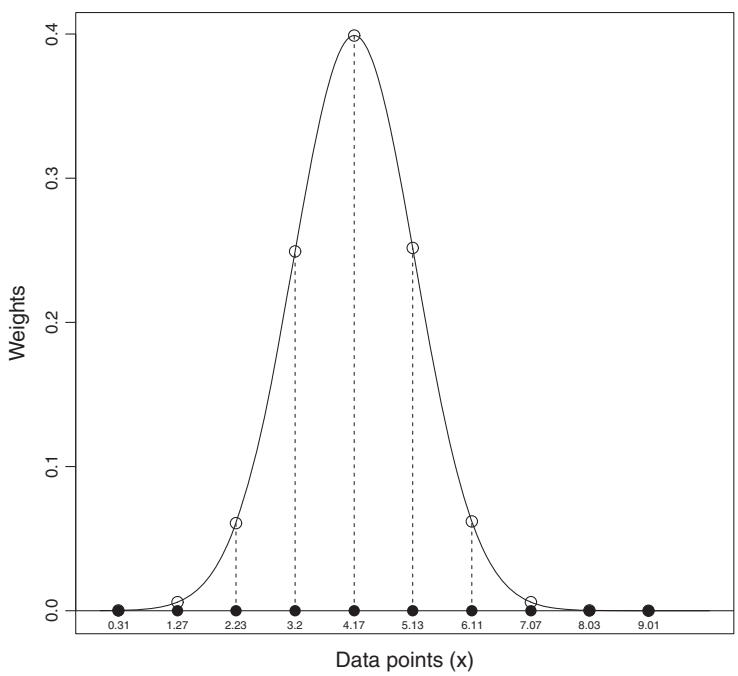


Figure 3.3 Weights (Circles) Used to Smooth the Single Value $y_5 = 50.93$ at $x_5 = 4.17$ to the Value $\bar{y}_5 = 61.13$

Table 3.11 *Data: Hodgkin's Disease Mortality Rates per* 1,000,000 *Person-Years among U.S. African Americans Males and Females (1973–1990) and Weighted Average Smoothed Estimates*

	Mortality rates/1,000,000								
Years	1973	1974	1975	1976	1977	1978	1979	1980	1981
Males									
Rates	14.0	12.0	13.0	12.0	10.0	10.0	9.0	10.0	8.0
Estimates	12.7	12.4	12.0	11.4	10.8	10.1	9.5	9.0	8.6
Females									
Rates	5.0	6.0	6.0	5.0	4.0	4.0	4.0	6.0	5.0
Estimates	5.4	5.4	5.2	5.0	4.7	4.6	4.6	4.7	4.6
Years	1982	1983	1984	1985	1986	1987	1988	1989	1990
Males									
Rates	7.0	8.0	10.0	8.0	7.0	8.0	6.0	8.0	7.0
Estimates	8.4	8.3	8.2	8.0	7.7	7.5	7.3	7.3	7.2
Females									
Rates	4.0	4.0	5.0	4.0	4.0	4.0	4.0	3.0	2.0
Estimates	4.5	4.4	4.3	4.2	4.0	4.0	4.0	4.0	4.0

by data are compared in Figure 3.2 (lower right, dashed line). Other choices of a bandwidth produce different degrees of smoothness. As noted, unlike the example, the degree of success from a weighted average smoothing strategy applied to actual data cannot be statistically evaluated. Simply, a "picture" frequently emerges that potentially suggests important properties of the sampled data.

Many choices of weights exist to smooth data into a summary curve. For example, a popular alternative set of weights is generated by the tricubic relationship

$$w_i = (1 - |x_i|^3)^3$$

for $-1 \leq \mathrm{x_i} \leq 1$. An example of these weights (w_i) is the following:

x_i	−1.0	−1.8	−0.6	−0.4	−0.2	0	0.2	0.4	0.6	0.8	1.0
w_i	0.000	0.116	0.482	0.820	0.976	1.000	0.976	0.820	0.482	0.116	0.000

Many other choices exist.

Example: Weighted Average Smoothing of Hodgkin's Disease Mortality Data

Hodgkin's disease mortality rates per 1,000,000 person-years among U.S. African Americans for the years 1973 to 1990 (Table 3.11) vary considerably, primarily because of the small numbers of deaths each year. In addition, even a small bias, misclassification, or lost data have strong influences when the sample size is small. Using weights derived from a normal distribution with standard deviation $= \sigma = 2.0$ (a bandwidth of $h = 2.0$) produces weighted average smoothed curves describing the temporal trend of male and female Hodgkin's disease mortality rates and *log*-rates from 1973 to 1990. In both plots, the weighted average smoothed

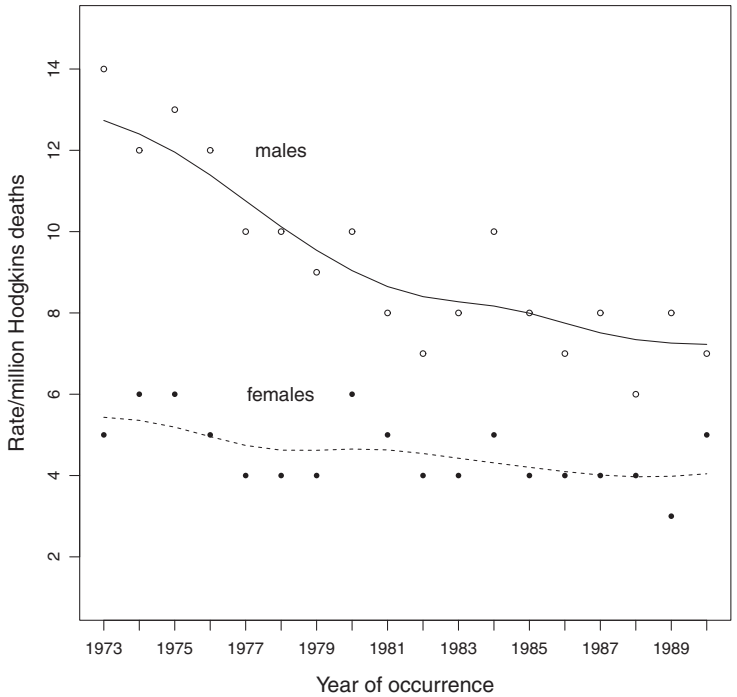


Figure 3.4 Smoothed Hodgkin's Disease Mortality Rates per 100,000 Person-Years among African Americans for Males and Females (1973–1990)

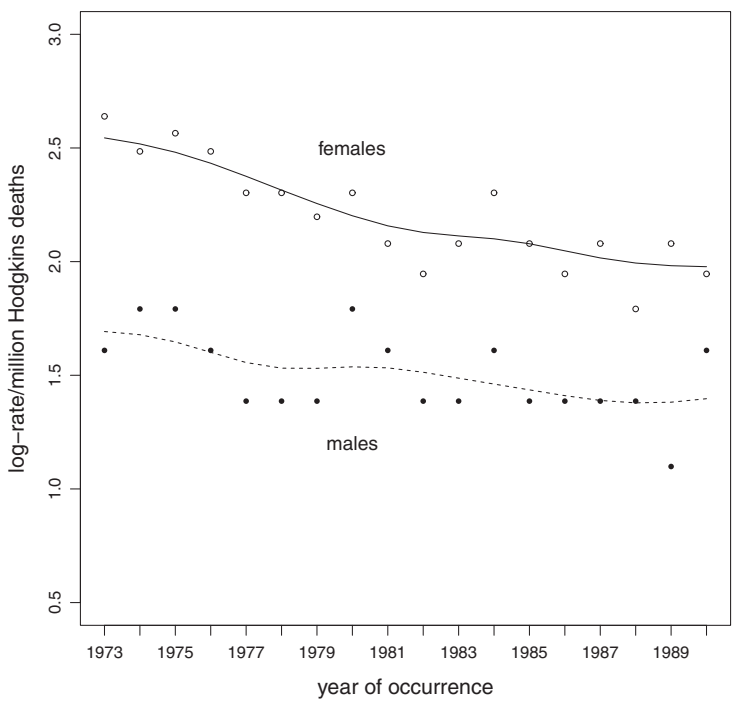


Figure 3.5 Smoothed Hodgkin's Disease Mortality Log-Rates per 100,000 Person-Years among African Americans for Males and Females (1973–1990)

curves clearly display males with higher mortality risk and trends for both males and females that continuously decrease over the observed period (Figures 3.4 and 3.5). In addition, the downward trends of the *log*-rate curves appear to be approximately parallel, indicating a constant male-to-female ratio of rates. The mean distance between the two smoothed curves is close to 0.7, making the constant male-to-female mortality rate ratio approximately 2.0 ($e^{0.7} \approx 2.0$) for the 17-year period. Note that the close to constant rate ratio is not obvious from the data or from the direct plot of the smoothed rates.

4

Two Discrete Probability Distributions

The normal probability distribution is essential for description and assessment of summary statistics calculated from continuous data (Chapter 1). Two probability distributions play a parallel role in the assessment of summary statistics calculated from discrete data, the *binomial* and *Poisson probability distributions.* Before exploring these two related distributions, an example illustrates several basic properties of discrete probability distributions.

Consider a discrete probability distribution consisting of a variable that takes on four values 1, 2, 3, and 4 (denoted x) with probabilities 0.1, 0.2, 0.3, and 0.4 (denoted p_x). Specifically, this four-value discrete probability distribution is

x	1	2	3	4	total
p_x	0.1	0.2	0.3	0.4	1.0

Like all discrete probability distributions, this distribution consists of a series of mutually exclusive and exhaustive outcomes. The probabilities p_x, therefore, add to 1.

Consider a random sample from this distribution of $n = 20$ observations and the accompanying frequencies (Table 4.1, row 2). The four estimated probabilities $\hat{p}_x$ based on the 20 random sampled observations vary considerably from the underlying values that generated the sample (last row). Naturally, as the sample size increases, these estimated probabilities increasingly resemble from the underlying probabilities of the distribution sampled. Thus, the sample estimated probabilities $\hat{p}_x$ ultimately become indistinguishable from the fixed underlying probabilities p_x as the sample size n increases or, in symbols, $\hat{p}_x \to p_x$ (Table 4.1, columns).

Fundamental elements of a probability distribution are the mean value and variance. Consider again the sample of $n = 20$ observations sampled from the example four-value probability distribution and the calculation of the mean value. The 20 observations are (Table 4.1, row 2):

$$\text{data: } \{3, 3, 3, 3, 2, 4, 1, 2, 4, 3, 4, 3, 3, 4, 4, 4, 2, 3, 4, 4\}.$$

The mean value can be calculated in a number of ways, for example:

$$\text{Mean value: } \bar{x} = \frac{1}{20}[3 + 3 + 3 + 3 + 2 + 4 + 1 + 2 + 4 + 3 + 4 + 3 + 3 + 4 + 4 + 4 + 2 + 3 + 4 + 4] = 3.150,$$

$$\text{Mean value: } \bar{x} = \frac{1}{20}[1 + 2 + 2 + 2 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4] = 3.150,$$

Table 4.1 *Example: Distributions, Probabilities, Means, and Variances from Samples of n values from Probability Distribution* $p_x = \{0.1, 0.2, 0.3, \text{and } 0.4\}$

	Sample estimates					
X	1	2	3	4	Means	Variances
$n = 10$	1	3	4	2	2.700	0.810
$\hat{p}_x$	0.100	0.300	0.400	0.200		
$n = 20$	1	3	8	8	3.150	0.727
$\hat{p}_x$	0.05	0.15	0.40	0.40		
$n = 50$	7	12	19	12	2.720	0.962
$\hat{p}_x$	0.140	0.240	0.380	0.240		
$n = 200$	14	45	55	86	3.065	0.931
$\hat{p}_x$	0.070	0.225	0.275	0.430		
$n = 500$	49	94	145	212	3.040	1.002
$\hat{p}_x$	0.098	0.188	0.290	0.424		
$n = 1000$	102	215	297	386	2.967	1.008
$\hat{p}_x$	0.102	0.215	0.297	0.386		
$n = 10{,}000$	1023	1979	3095	3983	2.996	0.998
$\hat{p}_x$	0.102	0.198	0.310	0.398		
p_x[a]	0.1	0.2	0.3	0.4	3.0	1.0

[a]Probability distribution sampled.

$$\text{Mean value: } \bar{x} = \frac{1}{20}[1(1) + 3(2) + 8(3) + 8(4)] = 3.150,$$

$$\text{Mean value: } \bar{x} = \frac{1}{20}(1) + \frac{3}{20}(2) + \frac{8}{20}(3) + \frac{8}{20}(4) = 3.150, \text{ and}$$

$$\text{Mean value: } \bar{x} = 0.050(1) + 0.150(2) + 0.400(3) + 0.400(4) = 3.150.$$

Therefore, for a discrete probability distribution, calculation of a sample mean value can be viewed as the sum of each observed value multiplied by the frequency of its occurrence in the sample, a weighted average (weights $= \hat{p}_x$) (Chapter 3). For the 20 example values, the mean value is $\sum \hat{p}_x x = 3.150$.

As the sample size increases, the estimated probabilities more precisely reflect the underlying probabilities. Thus, the sample mean value becomes increasingly a more precise reflection of the mean value of the probability distribution sampled, given the special name *the expected value* (denoted *EX*). This convergence is illustrated in Table 4.1 (columns 6 and 7).

In general terms, the expression for an expected value of a discrete theoretical probability distribution (mean value of the distribution sampled, denoted *EX*) is

$$\text{expected value} = EX = \sum p_x x.$$

Parallel to the estimation of the sample mean value, the expected value *EX* is also the sum of each possible value (x) but multiplied by its probability of occurrence (p_x). For the example four-value probability distribution, the expected value is the weighted average (weights $=$ px):

$$\text{expected mean value} = EX = \sum p_x x = 0.1(1) + 0.2(2) + 0.3(3) + 0.4(4) = 3.0.$$

The same logic applies to the variance of a discrete probability distribution. The general expression for the variance of a discrete probability distribution is

$$\text{probability distribution variance} = \textit{variance}(x) = \sum p_x(x - EX)^2.$$

The sum consists of each possible squared deviation from the expected value ($[x - EX]^2$) multiplied by the corresponding probability of its occurrence (p_x). The value $(x - EX)^2$ measures variability. The variance is also a weighted average. As the sample size increases, the sample variance converges to the variance of the distribution sampled because again the estimated probabilities become increasingly a more precise reflection of the underlying probabilities that generated the sample (Table 4.1, last row). The variance of the four-value probability distribution is

$$\begin{aligned}\text{expected variance} = \textit{variance}(x) &= \sum p_x(x - EX)^2 \\ &= p_1(1-3)^2 + p_2(2-3)^2 + p_3(3-3)^2 + p_4(4-3)^2 \\ &= 0.1(1-3)^2 + 0.2(2-3)^2 + 0.3(3-3)^2 + 0.4(4-3)^2 = 1.0,\end{aligned}$$

where $EX = 3.0$. The expected value and variance are values generated by specific fixed and known values of x and their respective probabilities (p_x) and, unlike the sample mean and variance, are entirely properties of the distribution sampled.

Binomial Probability Distribution

Binomial probabilities are described by the binomial theorem that first appeared in 10th-century China. For a probability represented as p ($q = 1 - p$), the theorem states (Chapter 27):

$$\begin{aligned}(p+q)^n = &\binom{n}{0} q^n + \binom{n}{1} pq^{n-1} + \binom{n}{2} p^2 q^{n-2} + \cdots + \binom{n}{n-2} p^{n-2} q^2 \\ &+ \binom{n}{n-1} p^{n-1} q + \binom{n}{n} p^n.\end{aligned}$$

The general expression for the probability of each of the $n + 1$ binomial outcomes (denoted x) is

$$p_x = \binom{n}{x} p^x q^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n.$$

Thus, two parameter values n and p completely define this discrete probability distribution and allow the calculation of the probability of any specific outcome x where $x = 0, 1, 2, \ldots, n$. As with all discrete probability distributions, the binomial distribution probabilities add to 1 ($\sum p_x = 1.0$). The $n + 1$ probabilities add to 1 because $p + q = 1$.

The value of the binomial coefficient $\binom{n}{x}$ is the number of different arrangements of x outcomes of one kind and $n - x$ outcomes of another kind. For example, three +'s and three o's {+, +, +, o, o and o} from $\binom{n}{x} = \binom{6}{3} = 20$ different arrangements. For $n = 6$, the probability of the occurrence of a specific outcome $x = 3$ and $n - x = 3$ is $p^3\, q^3$. The product of these two terms produces the binomial probability of occurrence of three events of one kind and three events of another kind or $p_x = P(X = 3) = 20p^3q^3$ regardless of the order

Table 4.2 *Example: Binomial Probability Distribution for Parameters $n = 6$ and $p = 0.4$ ($x = 0, 1, 2, \ldots, 6$)*

	Outcomes							
x	0	1	2	3	4	5	6	Total
$\binom{n}{x}$	1	6	15	20	15	6	1	64
$p^x q^{n-x}$	0.047	0.031	0.021	0.014	0.009	0.006	0.004	–
px	p^6	$6\,p^5 q$	$15\,p^4 q^2$	$20\,p^3 q^3$	$15\,p^2 q^4$	$6\,pq^5$	q^6	1.0
p_x	0.047	0.187	0.311	0.276	0.138	0.037	0.004	1.0

of occurrence. An example of the complete binomial probability distribution is displayed in Table 4.2 for parameters $n = 6$ and $p = 0.6$ ($q = 1 - p = 0.4$).

The value of a binomial coefficient $\binom{n}{x}$ directly calculated is $\frac{n!}{x!(n-x)!}$. For example, when $n = 6$ and $x = 3$, the binomial coefficient $\binom{6}{3} = \frac{6!}{3!(3!)} = \frac{720}{36} = 20$. Alternatively, a historical method used to find values of the binomial coefficients employs Pascal's triangle, named after the French mathematician Blaise Pascal (b. 1623). This pattern of binomial coefficients should be called the Precious Mirror of the Four Elements from a chart created by a Chinese scholar three centuries earlier (1303).

The rows in Pascal's triangle are the binomial coefficients $\binom{n}{x}$ for outcome x among $n + 1$ possibilities. Each row adds to 2^n. In addition, each binomial coefficient is the sum of the two values directly to the left and to the right in the above row or, in symbols,

Pascal's Triangle Generation of the Binomial Coefficients

n																					2^n
0										1											1
1									1		1										2
2								1		2		1									4
3							1		3		3		1								8
4						1		4		6		4		1							16
5					1		5		10		10		5		1						32
6				1		6		15		20		15		6		1					64
7			1		7		21		35		35		21		7		1				128
8		1		8		28		56		70		56		28		8		1			256
9	1		9		36		84		126		126		84		36		9		1		512

$$\binom{n}{x} = \binom{n-1}{x-1} + \binom{n-1}{x}.$$

Using this expression sequentially, the values in Pascal's triangle are easily calculated; that is, each row creates the row below by simple addition.

For example, for $n = 5$, the binomial distribution probabilities are

$$\text{binomial probability} = p_x = \binom{5}{x} p^x (1-p)^{5-x} \quad x = 0, 1, \ldots, 5.$$

Specifically, using the six coefficients $\binom{5}{x} = \{1, 5, 10, 10, 5, \text{ and } 1\}$ (row 5), then

$$(p+q)^5 = q^5 + 5pq^4 + 10p^2q^3 + 10p^3q^2 + 5p^4q + p^5 = 1.0 \,(\text{row } n = 5)$$

generates the binomial probabilities for the six possible outcomes of x.

The mean value (expected value) of this binomial probability distribution with parameters $n = 5$ and probability p is

$$\begin{aligned}\text{expected value} = EX &= \sum x\,p_x \\ &= (0)q^5 + (1)5pq^4 + (2)10p^2q^3 + (3)10p^3q^2 + (4)5p^4q + (5)p^5 \\ &= 5p[q^4 + 4pq^3 + 6p^2q^2 + 4p^3q + p^4] = 5p[(p+q)^4] = 5p.\end{aligned}$$

The sum in the square brackets is itself a binomial probability distribution ($n = 4$) and equals 1.0. In general, the expected value of a binomial probability distribution is

$$EX = \sum x\,p_x = \sum x \binom{n}{x} p^x q^{n-x} = np \quad x = 0, 1, 2, \ldots, n.$$

For the example binomial distribution, when $n = 6$ and $p = 0.4$, the mean value of the probability distribution is $EX = np = 6(0.4) = 2.4$ (Table 4.2).

The justification of the variance of a binomial probability distribution is algebraically a bit more complex but follows a similar pattern as the expected value and is

$$\text{binomial variance} = \textit{variance}(x) = \sum p_x(x - EX)^2 = np(1-p) \quad x = 0, 1, 2, \ldots, n.$$

For the binomial distribution with parameters $n = 6$ and $p = 0.4$, the variance of the probability distribution is $\textit{variance}(x) = np(1-p) = 6(0.4)(0.6) = 1.44$.

The simplest binomial distribution describes a single observation ($n = 1$) with two outcomes (coded 0 or 1) that occur with probabilities $1 - p$ and p. That is, a binary variable X takes on value 0 with probability $1 - p$ or value 1 with probability p. This special case is sometimes called a *Bernoulli variable.* The importance of the Bernoulli variable is that a sum of these 0 or 1 values is a statistical description for the everyday process of counting.

The expected value of this two-value probability distribution is

$$\text{expected value} = EX = \sum x\,p_x = (0)(1-p) + (1)p = p,$$

and the variance of a Bernoulli variable is

$$\begin{aligned}\textit{variance}(x) &= \sum p_x(x - EX)^2 = \sum p_x(x-p)^2 \\ &= (1-p)(0-p)^2 + p(1-p)^2 = p(1-p).\end{aligned}$$

Or, from the description of the general binomial distribution, $EX = np = p$ and $\textit{variance}(x) = np(1-p) = p(1-p)$ because $n = 1$.

It is important to note that a sum of n independent Bernoulli values ($X = \Sigma x_i$) has an expected value

$$EX = E\left(\sum x_i\right) = \sum E(x_i) = p + p + p + \cdots + p = np$$

and

$$\begin{aligned} variance(X) &= variance\left(\sum x_i\right) = \sum variance(x_i) \\ &= p(1-p) + p(1-p) + p(1-p) + \cdots + p(1-p) = np(1-p). \end{aligned}$$

These two expressions indicate that a binomial distribution can be viewed as the distribution of a sum of n independent 0 or 1 Bernoulli variables (Chapter 27). In symbols, when

$$X = \sum x_i, \quad i = 1, 2, \ldots, n$$

from n independent binary observations $x_i = \{0 \text{ or } 1\}$ with probabilities $1 - p$ or p, then the sum (denoted X) has a binomial distribution with parameters n and p; that is, the sum of Bernoulli values is identical to a single binomial outcome. In more specific terms, tossing a coin 10 times and counting the number of heads is the same as tossing 10 coins and counting the number of heads. Thus, the binomial probability distribution describes the properties of counts. Albert Einstein pointed out a problem with counts: "Not everything that counts can be counted and not everything that can be counted counts."

The most important application of a binomial probability distribution produces expressions for the expected value and variance of a proportion estimated from a sample of discrete observations. Such a sample of n observations consists of two kinds of outcomes where x represents the observed count of a binary outcome with a specific property and $n - x$ represents the count of the same binary outcome without the property. For example, a family of five ($n = 5$) can consist of $x = 3$ girls and $n - x = 2$ boys. The assumption or knowledge that the population sampled produces a count x with at least an approximate binomial distribution (constant value p) is the basis of numerous statistical analyses.

The expected value of a variable X multiplied by a constant value a is

$$\text{expected value of } aX = E(aX) = \sum ax\,p_x = a\sum x\,p_x = aEX.$$

Therefore, for a proportion or probability estimated by $\hat{p} = x/n$, the expected value becomes

$$E(\hat{p}) = E\left(\frac{x}{n}\right) = \frac{1}{n}EX = \frac{1}{n}np = p$$

where $a = 1/n$, when the observed count x has a binomial distribution with parameters n and p (Chapter 27). Similarly, the variance of a variable X multiplied by a constant value a is

$$\begin{aligned} variance\ of\ aX = variance(aX) &= \sum p_x(ax - aEX)^2 \\ &= a^2\sum p_x(x - EX)^2 = a^2 variance(X). \end{aligned}$$

Then the variance of the estimate of a proportion $\hat{p} = x/n$ from a random sample of n independent values with a binomial distribution becomes

$$variance(\hat{p}) = variance\left(\frac{x}{n}\right) = \frac{1}{n^2}variance(x) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

where again $a = 1/n$ (Chapter 27). These two expressions for the expected value and variance of the distribution of an estimated proportion or probability $\hat{p}$ are essential to evaluate and summarize data consisting of counts.

As noted, a probability of a specific outcome x estimated from a sample of n binary observations can be exactly calculated from a binomial distribution. Even for small samples ($n \approx 10$), the process is somewhat tedious and often unnecessary. A normal distribution with mean value μ set to the value of np(binomial mean value) and variance σ^2 set to the value $np(1-p)$ (binomial variance) allows a simple and easily applied method to calculate approximate binomial probabilities and is typically used to identify the extent of influence of random variation associated with the estimated probability $\hat{p}$, a proportion.

For example, directly from a binomial probability distribution with parameters $n = 10$ and $p = 0.5$, the exact probability $P(X \geq x) = P(X \geq 6) = 0.172$. For $EX = np = 10(0.5) = 5$ and $\sigma_X^2 = np(1-p) = 10(0.5)(0.5) = 2.5$, the approximation based on the normal distribution is

$$z = \frac{x - EX + 0.5}{\sqrt{\sigma_X^2}} = \frac{6 - 5 + 0.5}{\sqrt{2.5}} = 0.947$$

making $P(X \geq 6) \approx P(Z \geq 0.947) = 0.171$. A complete explanation of the origin and properties of this important normally distributed approximation of a binomial probability will be discussed in detail (Chapter 6).

Direct estimation of an exact binomial distribution–derived confidence interval generally requires a computer calculation. Therefore, a normal distribution is typically used as a simpler but approximate alternative to the cumbersome direct calculation; that is, the commonly estimated 95% confidence interval

$$\hat{p} \pm 1.960\sqrt{\hat{p}(1-\hat{p})/n}$$

is usually an accurate approximation based on a normal distribution with the mean value $\mu = p$ and variance $\sigma^2 = p(1-p)/n$ (Chapter 2). This approximate confidence interval is most accurate when binomial parameter p is in the neighborhood of 0.5 and works well for other values of p as long as the sample size is large. When the sample size is small, this approximation can be improved or exact computer calculations are used (Chapter 2). The idea of using the simpler normal distribution to approximate the more complicated binomial distribution originated in the 18th century and is sometimes referred to as the *de Moivre–Laplace theorem* (Chapters 2 and 6). Four binomial distributions ($p = 0.3$ and $n =$ 4, 10, 20, and 60) and their corresponding approximate normal distributions are illustrated in Figure 4.1 (dots = exact binomial probabilities).

Two Applications of the Binomial Distribution

An Example: Estimation of Viral Prevalence

The prevalence of a specific virus carried by mosquitos is of interest because it is frequently related to the likelihood of human disease. A natural estimate of a viral prevalence is to collect a sample from female mosquitos and count the number infected with virus. This approach is often not practical in terms of time, effort, and accuracy. Estimation of a usually low prevalence rate among a large number of mosquitos is improved by randomly pooling the collected mosquitos into separate groups and collectively determining whether virus is

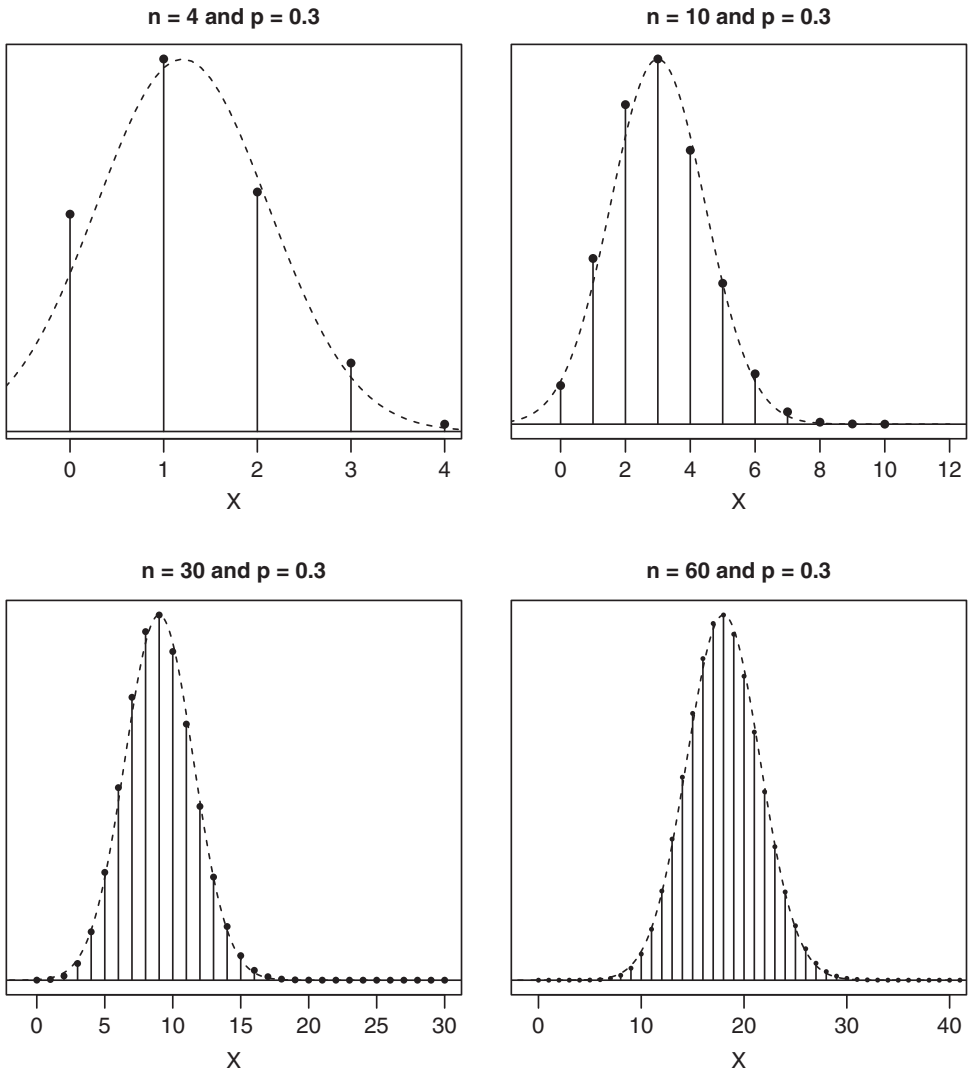


Figure 4.1 Four Binomial Distributions ($p = 0.3$) and Their Corresponding Approximate Normal Distributions ($\mu = np = \{1.2, 3.0, 9.0, \text{and } 18.0\}$ and $\sigma^2 = np(1-p) = \{0.84, 2.10, 6.30, \text{and } 12.60\}$, Dashed Lines)

present or absent in each group. Then an accurate prevalence rate is efficiently estimated from a laboratory assay based on a binomial distribution.

Specifically, mosquitos are pooled into k groups each containing n mosquitos. A single laboratory analysis then determines if one or more mosquitos are infected within each of the k groups. The number of determinations becomes k (one for each group) and not nk (one for each mosquito). The collected data used to estimate the proportion of infected mosquitos (denoted p) are now binary. Each of the k groups is either positive or negative for the presence of virus, a Bernoulli/binomial variable. The probability a pool contains no infected mosquitos is directly estimated. A negative pool consists entirely of n virus-free mosquitos. Therefore,

the probability of a negative pool (denoted Q) is $Q = q^n$ where $q = 1 - p$ represents the probability of a noninfected mosquito.

The estimated probability of a negative pool is $\hat{Q} = x/k$ where x represents the observed number of pools without infected mosquitos among the k laboratory determinations. The observed value x has a binomial distribution, because the k independent pools each containing n mosquitos are either a positive pool ($X = 1$) or a negative pool ($X = 0$). For example, when $k = 60$ pools of $n = 20$ mosquitos are formed and $x = 18$ pools are determined to be negative for the virus, then the estimated probability of a negative pool is $\hat{Q} = x/k = 18/60 = 0.300$ (Chapter 27). The estimate of the prevalence rate (denoted $\hat{p}$) becomes

$$\text{estimated prevalence rate} = \hat{p} = 1 - \hat{q} = 1 - \hat{Q}^{1/n}$$

because $q^n = Q$, then $q = Q^{1/n}$. For the example, the estimated probability of a negative pool is $\hat{Q} = 0.300$; therefore, the estimated virus prevalence rate becomes $\hat{p} = 1 - 0.300^{1/20} = 0.058$.

Because k independent binary determinations yield the estimate $\hat{Q}$, its estimated binomial distribution variance is $\hat{Q}\,(1 - \hat{Q})/k$. Based on a binomial distribution, an expression for the variance of the estimated prevalence $\hat{p}$ becomes

$$\textit{variance}(\hat{p}) = \left[\frac{\hat{Q}^{1/n}}{n\hat{Q}}\right]^2 \frac{\hat{Q}(1-\hat{Q})}{k}.$$

For the example mosquito data with $\hat{Q} = 0.300$, the estimated standard error is

$$\sqrt{\textit{variance}}(\hat{p}) = 0.0093[27].$$

To measure the influence from sampling variation on the prevalence rate $\hat{p}$, two essentially equivalent approximate 95% confidence intervals can be estimated.

Based on the estimate $\hat{Q}$, an approximate 95% confidence interval is

$$\text{lower bound} = \hat{A} = \hat{Q} - 1.960\sqrt{\frac{\hat{Q}(1-\hat{Q})}{k}} = 0.300 - 1.960(0.0592) = 0.184$$

and

$$\text{upper bound} = \hat{B} = \hat{Q} + 1.960\sqrt{\frac{\hat{Q}(1-\hat{Q})}{k}} = 0.300 + 1.960(0.0592) = 0.416$$

using the normal distribution approximation of a binomial distributed variable. The 95% confidence interval for the prevalence rate p becomes $(1 - \hat{B}^{1/n}, 1 - \hat{A}^{1/n}) = (1 - 0.416^{1/20}, 1 - 0.184^{1/20}) = (0.043, 0.081)$ based on the estimate $\hat{p} = 1 - \hat{Q}^{1/n} = 1 - 0.300^{1/20} = 0.058$ (Chapter 2).

Alternatively, with the confidence interval bounds calculated directly from the estimated variance of $\hat{p}$ and again applying a normal distribution approximation, then

$$(\hat{p}) \pm 1.960\sqrt{\textit{variance}}(\hat{p}) = 0.058 \pm 1.960(0.0093) \rightarrow (0.040, 0.077),$$

again based on the estimated prevalence rate $\hat{p} = 0.058$.

Table 4.3 *Model: Four Probabilities from a Randomized Response to an Interview Question*

	Subject response		
Device	Yes	No	Total
"Yes"	$P\pi$	$(1-P)\pi$	π
"No"	$P(1-\pi)$	$(1-P)(1-\pi)$	$1-\pi$
	P	$1-P$	1.0

An Example: Randomized Response Estimates from Survey Data

Survey questions occasionally call for responses that may not be answered accurately because of a reluctance on the part of the person interviewed. A natural reluctance frequently exists to answering personal questions such as Do you use recreational drugs? or Have you been a victim of sexual abuse? or Do you earn more than $350,000 a year? A method designed to improve the cooperation of a subject being interviewed is called a *randomization response survey technique*.

This technique employs a device that produces a random "yes" or "no" response with a known probability (denoted π). When a sensitive question is asked, the subject interviewed uses the device to produce a random "yes" or "no" response and replies to the interviewer only that the response given by the device is correct or not correct. Because the interviewer cannot see the device and subject's response does not reveal the answer to the question, the subject might be less evasive and, perhaps, more truthful. The process guarantees confidentiality because the actual answer to the question is not stated or recorded. The collected data are then a series of binary responses indicating only whether the subject's true answer agrees or does not agree with the random "answer" shown on the device. Each response is again a Bernoulli variable.

Random response data then consist of counts of four possible outcomes described by four probabilities (Table 4.3). Thus, the probability of agreement is $p = P\pi + (1 - P)(1 - \pi)$ and disagreement is $1 - p = P(1 - \pi) + (1 - P)\pi$, where P represents the probability of a true response and π is a known probability of a device-generated random response "yes." Using the predetermined fixed value $\pi = P(\textit{random}$ "*yes response*"), the probability of a true response P is estimated based on n interviews and the binomial distribution.

The number interviews n and the probability of agreement denoted p are the parameters defining a binomial distribution. An estimate of the proportion of true yes answers is achieved by equating the observed binomial proportion of agreement $\hat{p} = x/n$ to the model probability of agreement where x represents the observed number of the n subjects interviewed who stated that their answer agreed with the randomly produced yes/no response (Table 4.3). Specifically, equating the data-estimated value $\hat{p}$ to the corresponding model value of the probability of agreement,

$$\hat{p} = \frac{x}{n} = P\pi + (1 - P)(1 - \pi)$$

and, solving for the value P, then

$$\hat{p} = \frac{\hat{p} - (1 - \pi)}{2\pi - 1}$$

provides an estimate of the probability of a true yes response, namely, P (Chapter 27).

Adding a constant value to a variable does not change its variance, and the variance of the distribution of the variable aX is a^2 *variance*(X) (Chapter 27). Therefore, the expression for an estimate of the variance of the estimate of $\hat{P}$ based on the binomial probability distribution is

$$\begin{aligned} variance(\hat{p}) &= variance\left[\frac{\hat{p} - (\pi - 1)}{2\pi - 1}\right] = variance\left[\frac{\hat{p}}{2\pi - 1} - \frac{\pi - 1}{2\pi - 1}\right] \\ &= \frac{1}{(2\pi - 1)} 2 variance(\hat{p}) = \frac{1}{(2\pi - 1)} 2 \frac{\hat{p}(1 - \hat{p})}{n} \end{aligned}$$

where π represents a known probability and the estimated binomial distribution variance of the estimated proportion $\hat{p}$ is $\hat{p}(1 - \hat{p})/n$. Curiously, the value $\pi = 0.5$ cannot be used.

For example, when $n = 200$ subjects are interviewed and $x = 127$ replied that their answer agreed with the "answer" given by the device, the estimated probability of agreement is $\hat{p} = 127/200 = 0.635$. Therefore, because π was set at 0.3,

$$\hat{p} = \frac{0.635 - (1 - 0.3)}{2(0.3) - 1} = 0.163$$

is an estimate of the probability of P, the probability of a true yes response to the question. The estimated variance of the distribution of $\hat{p}$ is $variance(\hat{p}) = 0.0072$. Therefore, an approximate 95% confidence interval becomes

$$\hat{P} \pm 1.960\sqrt{variance(\hat{P})} = 0.163 \pm 1.960(0.085) \rightarrow (-0.004, 0.329)$$

based on the estimate $\hat{P} = 0.163$ and applying once again the normal distribution approximation. Like all such confidence intervals, the probability is approximately 0.95 that the true yes answer, estimated by $\hat{P}$, is contained in the interval.

The Geometric Probability Distribution

An occasionally useful and simple probability distribution, called the *geometric probability distribution*, provides an introduction to discrete probability distributions that consist of an infinite number of possible outcomes. Geometric probabilities describe the likelihood a specific event occurs after x occurrences of the event fail to occur. For example, the probability of tossing a coin x times before heads occurs is described by a geometric probability.

In specific statistical terms, consider a binary outcome denoted X where $P(X = 1) = p$ and $P(X = 0) = 1 - p = q$, then the probability that x independent events fail to occur $(X = 0)$ before the other event occurs $(X = 1)$ is

$$geometric\ probability = P(X = x) = p_x = q^x p$$

Table 4.4 *Geometric Probability Distribution: Occurrence of an Event after x Failures of the Event to Occur* ($q = 0.6$)

x	0	1	2	3	4	5	6	7	8	$\cdots$
Probability	p	qp	q^2p	q^3p	q^4p	q^5p	q^6p	q^7p	q^8p	$\cdots$
Example	0.400	0.240	0.144	0.086	0.052	0.031	0.019	0.011	0.007	$\cdots$

producing a geometric probability distribution for the outcomes $x = 0, 1, 2, 3, \ldots$. In theory, any number of events can occur. Table 4.4 displays the geometric probabilities for $q = 0.6$ $(1 - q = p = 0.4)$ for $x = 0, 1, 2, \ldots, 8, \ldots$.

As always, the sum of the probabilities of all possible outcomes is $\Sigma\, p_x = 1.0$, despite the fact the sum is infinite. For the expression $S = \Sigma\, q^x$, then

$$S = 1 + q + q^2 + q^3 + q^4 + \cdots \quad \text{and} \quad qS = q + q^2 + q^3 + q^4 + q^5 + \cdots.$$

Thus, the expression $S - qS = 1$, making $S = 1/(1 - q) = 1/p$. Therefore, the sum the of the geometric probabilities is $\sum q^x p = p \sum q^x = pS = p\frac{1}{p} = 1.0$. The expected value of a geometric probability distribution is

$$\text{expected mean value} = EX = \sum x\, p_x = \sum x q^x p = qp[1 + 2q + 3q^2 + 4q^3 + 5q^4 + \cdots]$$
$$= qp\left[\frac{1}{p^2}\right] = \frac{q}{p} x = 0, 1, 2, 3, \ldots.$$

The fact that the sum $\Sigma(k + 1)q^k = 1/p^2$ is a result from calculus (Chapter 27).

The justification of the variance of a geometric probability distribution follows a similar but more complex pattern:

$$\text{geometric distribution variance} = \textit{variance}(x) = \frac{q}{p^2} \quad \text{(Chapter 27)}.$$

This one-parameter probability distribution is displayed for four parameter values p in Figure 4.2. Parenthetically, the geometric probability distribution is the simplest case of a rather complicated discrete probability distribution called the *negative binomial probability distribution* (not discussed).

Consider a mythical country where all families are required to have children until a male is born, and then these families are not allowed to have additional children. The geometric probability distribution answers the question: What are the consequences of this severe family planning policy? When the probability of a male or female child is $p = q = 0.5$, Table 4.5 shows the details based on the geometric probability distribution. The geometric probability distribution dictates that the expected number of female births per family in the mythical country is $EX = q/p = 0.5/0.5 = 1.0$ or ratio of male to female births of 1.0. In addition, the expected family size is 2.0 (family size $= \Sigma(x + 1)q^x p = EX + 1 = 2.0$).

A Poisson Probability Distribution

The Poisson probability distribution, named after the French scholar Simenon Denis Poisson (b. 1781), can be derived from a variety of perspectives. Perhaps the most statistically relevant

Table 4.5 *Geometric Probability Distribution: Family Compositions from a Mythical Country* ($M = Male$, $F = Female$ *and* $p = 0.5$)

Family compositions	Females (x)	Probability (pq^x)
M	0	$\left(\frac{1}{2}\right)^1$
FM	1	$\left(\frac{1}{2}\right)^2$
FFM	2	$\left(\frac{1}{2}\right)^3$
FFFM	3	$\left(\frac{1}{2}\right)^4$
FFFFM	4	$\left(\frac{1}{2}\right)^5$
FFFFFM	5	$\left(\frac{1}{2}\right)^6$
FFFFFFM	6	$\left(\frac{1}{2}\right)^7$
FFFFFFFM	7	$\left(\frac{1}{2}\right)^8$
FFFFFFFFM	8	$\left(\frac{1}{2}\right)^9$
– –	–	–
– –	–	–
– –	–	–

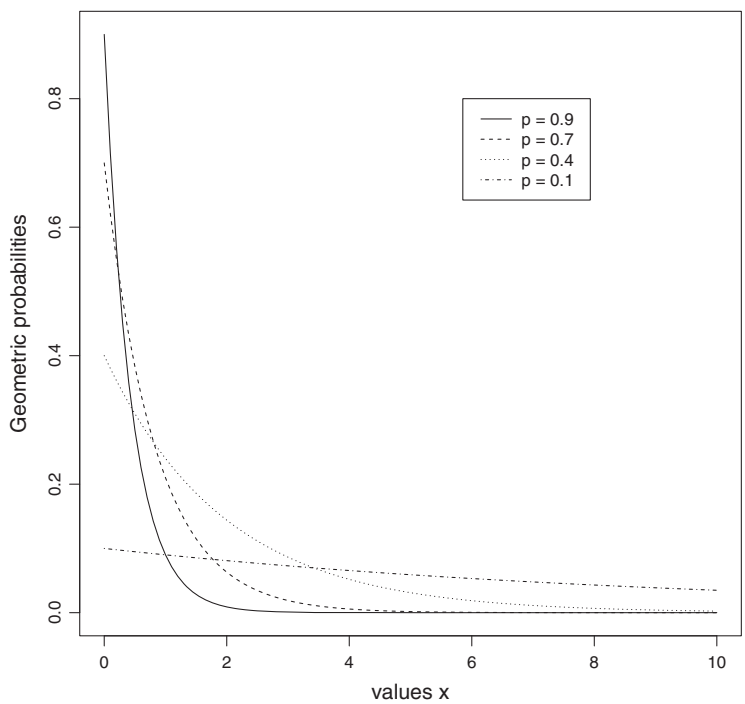


Figure 4.2 Four Geometric Probability Distributions ($p = 0.9$, 0.7, 0.4, and 0.1)

relates to the binomial distribution. As the binomial parameter p becomes smaller and the sample size n becomes larger, so that the product converges to a constant value (denoted $\lambda = np$), the expression for the resulting discrete *Poisson probability distribution* becomes

$$\textit{Poisson probability} = p_x = \frac{\lambda^x e^{-\lambda}}{x!}$$

defined by the single parameter $\lambda = np$.

The evolution of a Poisson distribution from a binomial distribution is

$$\begin{aligned}\text{binomial probability} = p_x &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \frac{n(n-1)(n-2)\cdots(n-[x+1])}{x!} p^x (1-p)^{n-x} \\ &= \frac{1(1-1/n)(1-2/n)\cdots(1-[x+1]/n)}{x!} (np)^x \left(1 - \frac{np}{n}\right)^{n-x}\end{aligned}$$

and when n becomes large (*approaches infinite*), p becomes small (*approaches zero*) so that λ represents the constant value np, then the corresponding discrete Poisson probability is

$$p_x = \frac{(np)^x e^{-np}}{x!} = \frac{\lambda^x e^{-\lambda}}{x!}$$

for a count x. Note that as the sample size n increases, the value $(1 - \frac{np}{n})^{n-x}$ converges to the value $e^{-np} = e^{-\lambda}$ (Chapter 27).

The case of $p = 0.1$ and $n = 50$ gives a sense of the binomial distribution converging to a Poisson distribution with parameter $\lambda = np = 50(0.1) = 5.0$; that is, specifically, the following:

	Binomial and poisson probabilities								
x	0	1	2	3	4	5	6	7	8
Binomial	0.005	0.029	0.078	0.139	0.181	0.185	0.154	0.108	0.064
Poisson	0.007	0.034	0.084	0.140	0.175	0.175	0.146	0.104	0.065
Difference	−0.002	−0.005	−0.006	−0.002	0.005	0.009	0.008	0.003	−0.001

Note that both distributions have expected value $EX = np = \lambda = 5.0$. The variance of the binomial distributed variable is $np(1-p) = 50(0.1)(0.9) = 4.5$. The mean and the variance of the Poisson distribution is $\lambda = 5.0$.

A less formal description of a Poisson distributed value X is

> For a large number (theoretically infinite) of independent binary events X, when the probability of occurrence of each of these events is $P(X = 0) = 1 - p$ and $P(X = 1) = p$ and p is small and constant, then
>
> $$p_x = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad x = 0, 1, 2, \ldots$$
>
> is the probability of x occurrences of the binary variable $X = 1$ ($np \rightarrow \lambda$).

Figure 4.3 displays four Poisson probability distributions and their corresponding normal distribution approximations ($\lambda = 1$, 2, 10, and 20).

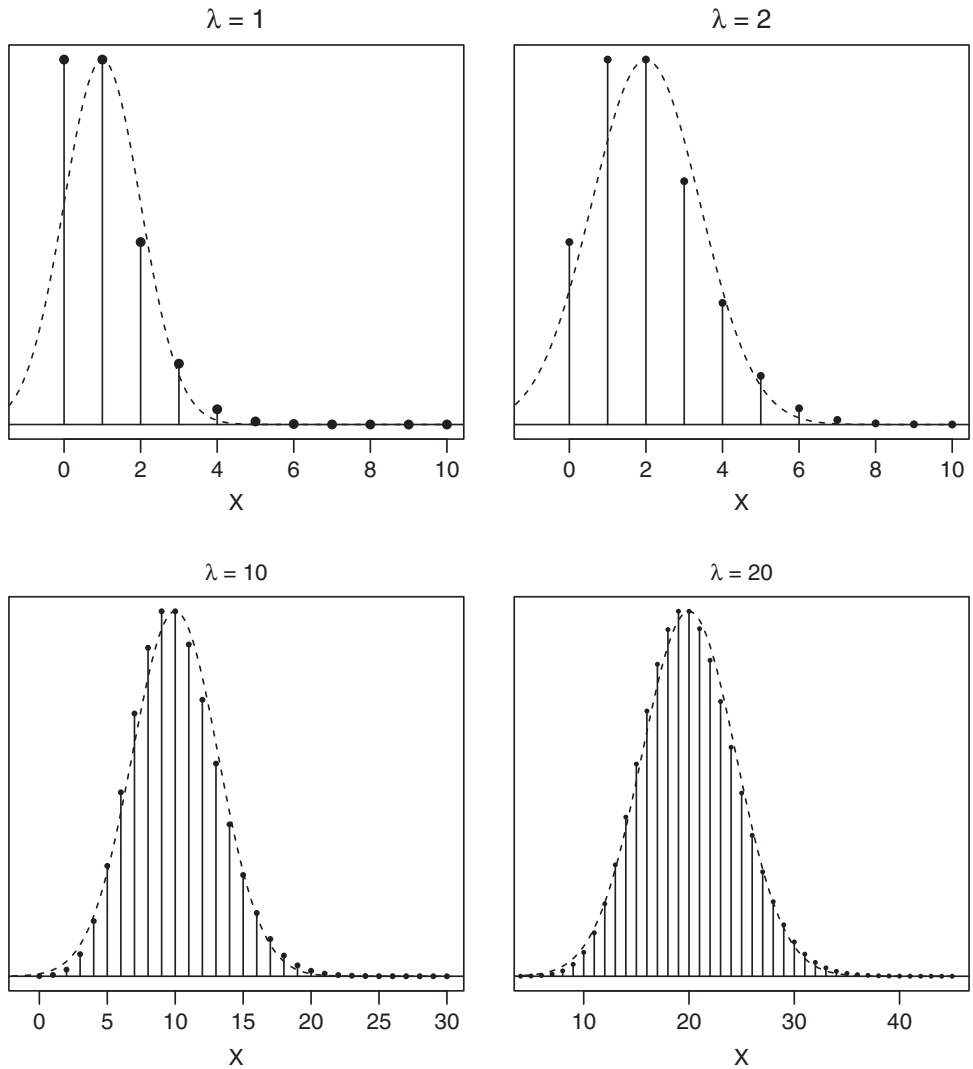


Figure 4.3 Four Poisson Distributions and Their Approximate Relationships to a Normal Distribution ($\lambda = 1$, 2, 10, and 20)

A Poisson probability distribution, like the geometric probability distribution, has theoretically an infinite number of discrete outcomes ($\lambda = 2$; Table 4.6).

Like all discrete probability distributions, the probabilities p_x sum to 1.0 ($\Sigma\, p_x = 1.0$). To illustrate, for the Poisson probability distribution ($\lambda = 2$),

$$\begin{aligned} Poisson(\lambda = 2) = \sum p_x &= \frac{2^0 e^{-2}}{0!} + \frac{2^1 e^{-2}}{1!} + \frac{2^2 e^{-2}}{2!} + \frac{2^3 e^{-2}}{3!} + \frac{2^4 e^{-2}}{4!} + \frac{2^5 e^{-2}}{5!} + \cdots \\ &= e^{-2}\left[\frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!} + \frac{2^3}{3!} + \frac{2^4}{4!} + \frac{2^5}{5!} + \cdots\right] = e^{-2}[e^2] = 1.0. \end{aligned}$$

Table 4.6 *Example: Poisson Probability Distribution p_x Defined by Parameter $\lambda = 2$*

x	0	1	2	3	4	5	6	7	8	9	$\cdots$
p_x	0.135	0.271	0.271	0.180	0.090	0.036	0.012	0.003	0.001	0.000	$\cdots$

Note: $p_x = \dfrac{2^x e^{-2}}{x!}$

The terms within the square brackets are an infinite series of values that sums to e^2 (Chapter 27). The expected value of a Poisson probability distribution ($\lambda = 2$) is

$$\text{expected value} = \sum x p_x = 0\frac{e^{-2}}{0!} + 1\frac{2e^{-2}}{1!} + 2\frac{2^2e^{-2}}{2!} + 3\frac{2^3e^{-2}}{3!} + 4\frac{2^4e^{-2}}{4!} + 5\frac{2^5e^{-2}}{5!} + \cdots$$
$$= 2\left[\frac{e^{-2}}{0!} + \frac{2e^{-2}}{1!} + \frac{2^2e^{-2}}{2!} + \frac{2^3e^{-2}}{3!} + \frac{2^4e^{-2}}{4!} + \frac{2^5e^{-2}}{5!} + \cdots\right] = 2.0.$$

The infinite series within the square brackets is a Poisson probability distribution and the sum is necessarily 1.0. In general, for a Poisson probability distribution, the mean value is

$$\text{expected value} = EX = \sum x p_x = \sum x\frac{\lambda^x e^{-\lambda}}{x!} = \lambda \quad x = 0, 1, 2, \ldots \text{(Chapter 27)}.$$

The estimate of the parameter λ is simply the sample mean value $\bar{x}$ calculated from the n observed counts x_i (Chapter 27). The expression for the variance of a Poisson probability distribution follows a similar pattern where

$$variance(x) = \sum p_x(x - EX)^2 = \sum \frac{\lambda^x e^{-\lambda}}{x!}(x - EX)^2 = \lambda \quad x = 0, 1, 2, \ldots.$$

The variance of a Poisson probability distribution is also λ and, not surprisingly, is also estimated by the sample mean value, $\bar{x}$. This result is expected. The expected value of the binomial distribution is np or λ and the variance is $np(1 - p) \approx np = \lambda$ for small values of p and large values of n. Both estimates are maximum likelihood estimates (Chapter 27).

Two Applications of the Poisson Probability Distribution

An Example: Poisson Analysis of Random Spatial Patterns

Figure 4.4 (upper left) displays $n = 200$ points randomly distributed over a square divided into $N = 100$ subsquares of equal area to illustrate a classic application of the Poisson probability distribution as an accurate statistical description of random spatial patterns. The probability that a specific subsquare contains a random point is $p = 1/N = 1/100$. Thus, each of n random points is distributed into one of a large number of subsquares (N) with a constant and small probability (p). A Poisson distribution then accurately describes the distribution of the counts of the subsquares (denoted $\hat{n}_k$) containing specific numbers of random points (denoted k). Tables 4.8 and 4.9 summarize $n = 200$ points randomly distributed into the $N = 100$ subsquares where $p = 1/100$ (Figure 4.4, upper left).

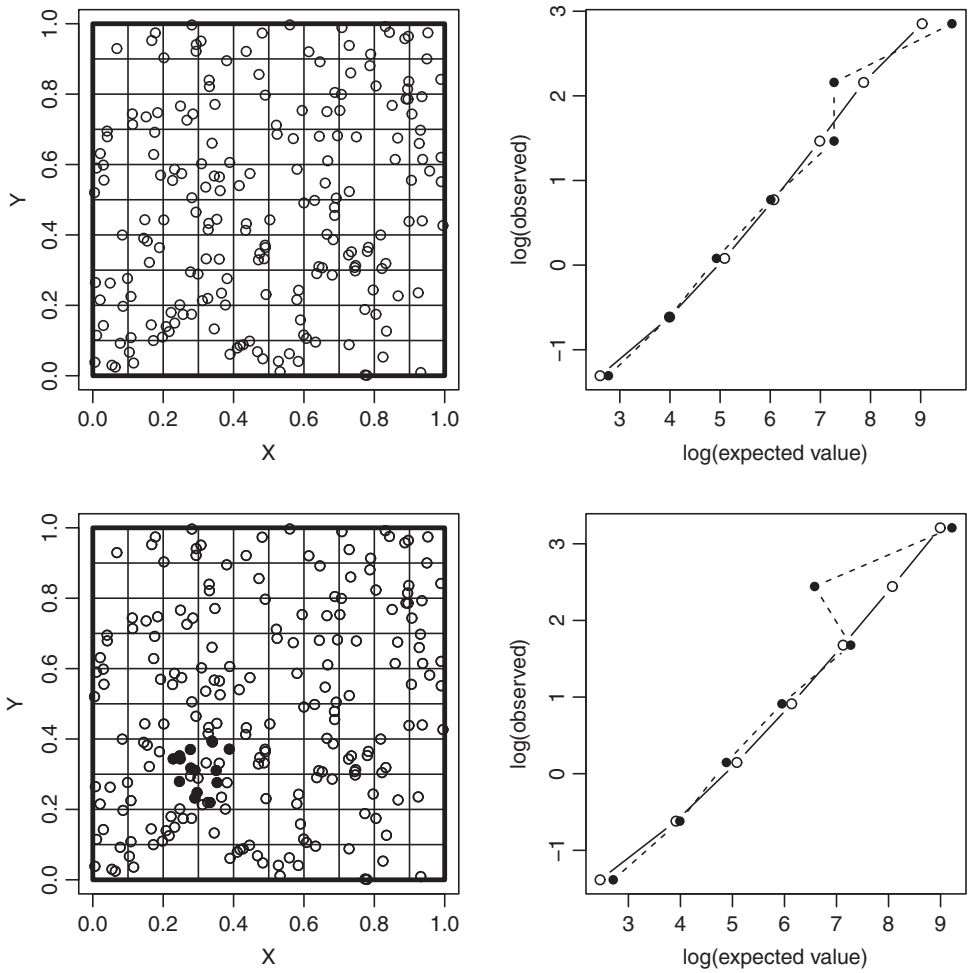


Figure 4.4 Two Spatial Distributions of Points Distributed over a Unit Square and Their Associated "Goodness-of-Fit" Plots: $n = 200$ Points (Random – Top) and 215 Points (Not Random – Bottom)

For example, from Figure 4.4 and Table 4.7, there are $\hat{n}_2 = 23$ subsquares that contain $k = 2$ points.

The Poisson parameter that generated these example data is the mean number of random points per subsquare $\lambda = np = 200/100 = 2$ (200 points/100 squares). Therefore, the Poisson probability that a subsquare contains k random points is

$$P(X = k) = p_k = \frac{\lambda^k e^{-\lambda}}{k!} = \frac{2^k e^{-2}}{k!} k = 0, 1, 2, \ldots.$$

For example, the probability that a subsquare contains two random points is $P(X = 2) = 2^2 e^{-2}/2! = 0.271$. This theoretical Poisson distribution ($\lambda = 2$) and the distribution of the expected number of subsquares with k random points (denoted $np_k = n_k$) are given in Table 4.8. Also included is the distribution of observed counts $\hat{n}_k$ (Figure 4.4).

Table 4.7 *Data: Number of Points in Each of n = 100 Subsquares (Figure 4.4)*

	Y									
X	0.0–0.1	0.1–0.2	0.2–0.3	0.3–0.4	0.4–0.5	0.5–0.6	0.6–0.7	0.7–0.8	0.8–0.9	0.9–1.0
0.0–0.1	1	2	4	1	2	1	1	3	4	2
0.1–0.2	0	0	0	3	1	0	2	2	3	1
0.2–0.3	0	4	3	1	1	2	1	2	3	2
0.3–0.4	3	2	0	3	0	2	3	1	2	4
0.4–0.5	4	1	4	4	2	1	2	1	0	3
0.5–0.6	0	1	2	3	2	2	4	0	1	2
0.6–0.7	1	4	0	2	5	0	3	6	3	0
0.7–0.8	4	1	3	5	1	2	2	2	1	1
0.8–0.9	3	3	6	1	0	2	1	1	2	0
0.9–1.0	4	3	0	1	6	4	1	3	1	1

An application of the Pearson chi-square test statistic produces a comparison of observed counts $\hat{n}_k$ to Poisson distribution generated theoretical counts $np_k = n_k$ and has an approximate chi-square distribution with six degrees of freedom ($k = 6$) yielding a p-value of 0.689. Thus, the obviously close correspondence between observed and theoretical Poisson counts produces no persuasive evidence of a nonrandom spatial pattern (Table 4.8).

$$X^2 = \sum(\hat{n}_k - n_k)^2/n_k = 3.911 k = 0, 1, 2, \ldots, 6.$$

A visual display of the correspondence between observed and Poisson model generate counts is usually worthwhile. Following the fundamental rule of visual comparisons that states "whenever possible create straight lines," a logarithmic transformation provides an approximate straight line for observations with a Poisson distribution. When n_k represents the expected count of the number of subsquares with k points, the expression

$$log(n_k) = -\lambda + klog(\lambda) - log(k!)$$

produces an approximate straight line. Plotting $x = log(\hat{n}_k) + log(k!)$ against $y = -\hat{\lambda} + k\, log(\hat{\lambda})$ displays the observed *log*-counts $\hat{n}_k$ (Figure 4.4, upper right). The same expression using the theoretical values from a Poisson distribution (n_k and $\lambda = 2$) produces an almost straight line and is also plotted (Figure 4.4). Thus, the difference in plotted values from the same log transformation applied to the observed counts $\hat{n}_k$ (dots) and theoretical values n_k (circles) yields a visual goodness-of-fit comparison (Table 4.8).

Table 4.8 *Data/Distribution: Observed Distribution of Subsquares with k Random Points (Table 4.7)*

k	0	1	2	3	4	5	≥6	Total
Probability (p_k)	0.135	0.271	0.271	0.180	0.090	0.036	0.017	1.0
Theoretical (n_k)	13.53	27.07	27.07	18.04	9.02	3.61	1.66	100
Observed ($\hat{n}_k$)	16	27	23	17	12	2	3	100

Table 4.9 *Data/distribution: Poisson Distribution* ($\lambda = 2.15$) *and the Observed Distribution of Subsquares with* k *Points from 200 Random Points and* 15 *Additional Nonrandom Points*

k	0	1	2	3	4	5	6	≥7	Total
Probability (p_k)	0.116	0.250	0.269	0.193	0.104	0.045	0.016	0.007	1.0
Theoretical (n_k)	11.65	25.04	26.92	19.29	10.37	4.46	1.60	0.66	100
Observed ($\hat{n}_k$)	15	27	22	16	12	1	6	1	100

When the distribution of the points is not random (p is not constant), deviations from the Poisson distribution–generated straight line are frequently an effective way to identify regions of nonrandom patterns. Figure 4.4 (lower right) includes an additional cluster of 15 points ($n = 215$). The spatial analysis again based on a Poisson distribution with the slightly increased value of $\lambda = 215/100 = 2.15$ produces the expected counts $n_k = np_k$ (Table 4.9).

An informal comparison of the Poisson estimated and observed counts (Table 4.9) identifies likely nonrandom clusters $\hat{n}_5 = 1$ and $\hat{n}_6 = 6$ where the corresponding theoretical Poisson values are $n_5 = 4.46$ and $n_6 = 1.60$. The accompanying goodness-of-fit plot also clearly identifies this nonrandom pattern. Although it is hardly necessary, the chi-square test statistic $X^2 = 17.818$ yields a p-value of 0.016 (degrees of freedom $= 8 - 1 = 7$).

An Example: A Truncated Poisson Distribution

The California Department of Fish and Game required captains of selected "party boats" to report the number of salmon caught on each outing. The reports for a specific month yield the data in Table 4.10. The mean number of salmon caught per boat is $307/116 = 2.647$ based on the 116 reports. This estimated mean value overstates the actual value because only boats that caught salmon were required to report. To provide a more accurate estimate of the mean value, the assumption is made that the number of salmon caught per boat has a Poisson distribution. That is, it is assumed that a large number of salmon exist and each one has a small and equal probability of being caught.

A Poisson distribution that does not included a value zero is called a *truncated.* For this truncated case, the Poisson distribution describing the number salmon caught

Table 4.10 *Data: Reported Number of Salmon Caught per Boat over a Specific Month (n =* 116 *Reports)*

Data:

2, 2, 2, 2, 1, 1, 2, 2, 2, 3, 2, 4, 1, 2, 3, 2, 5, 3, 1, 1, 2, 2, 1, 2, 2, 2, 1, 2, 2, 4, 3, 2, 2,
6, 1, 5, 2, 4, 1, 4, 2, 6, 2, 2, 2, 1, 5, 4, 3, 4, 1, 4, 3, 3, 2, 3, 1, 3, 1, 2, 3, 4, 3, 2, 3, 3,
1, 2, 5, 1, 3, 1, 2, 3, 3, 3, 3, 4, 3, 5, 1, 3, 3, 3, 3, 1, 4, 1, 6, 1, 2, 1, 3, 2, 3, 2, 4, 2, 4,
2, 2, 2, 1, 3, 3, 3, 6, 4, 3, 6, 4, 2, 3, 3, 3, 4

Or

	1	2	3	4	5	6	Total
Frequency	22	37	32	15	5	5	116

(denoted x) is

$$p_x = \left[\frac{1}{1 - e^{-\lambda}}\right] \frac{\lambda^x e^{-\lambda}}{x!} \quad x = 1, 2, 3, \ldots$$

where the probability of catching at least one salmon is $P(X \neq 0) = 1 - e^{-\lambda}$. As always, the sum of these discrete probabilities is $\Sigma\, p_x = 1.0$.

An estimate of the parameter λ from a truncated Poisson distribution that accounts for the "missing" reports (no salmon caught) is

$$\text{estimated mean value} = \hat{\lambda} = \frac{1}{n}\left[\sum x_i - S\right]$$

where $\Sigma\, x_i$ represents the total number of salmon caught, S represents the number of reports of one salmon caught, and n represents the total number of reports. Specifically, these summary values from the salmon boat data are $\Sigma\, x_i = 307$, $S = 22$, and $n = 116$ (Table 4.10). An unbiased estimate of the mean number of salmon caught is then

$$\text{estimated mean value} = \hat{\lambda} = \frac{1}{116}[307 - 22] = 2.457 \text{ per boat.}$$

This "singles removed" estimated mean value applied to truncated data is an unbiased estimate of the underlying mean value λ because the sample sum ($\Sigma\, x_i$) estimates

$$\sum x_i \rightarrow \frac{n\lambda}{1 - e^{-\lambda}}$$

adjusted for truncation and, similarly, the number of observed values equal to one (S) estimates

$$S \rightarrow \frac{n\lambda e^{-\lambda}}{1 - e^{-\lambda}}$$

also adjusted for truncation. The difference between these two expected summary values divided by n becomes

$$\frac{1}{n}\left[\frac{n\lambda}{1 - e^{-\lambda}} - \frac{n\lambda e^{-\lambda}}{1 - e^{-\lambda}}\right] = \lambda.$$

Therefore, the same relationship among corresponding observed values ($n = 116$, $\Sigma x_i = 307$ and $S = 22$) estimates the Poisson parameter λ accounting for the truncated nature of the collected data.

The Poisson distribution expected value, again adjusted for truncation, is

$$\text{expected value} = EX = \frac{\lambda}{1 - e^{-\lambda}}.$$

Using the observed mean value $\bar{x}$ in place of the unknown truncated Poisson distribution expected mean value EX and solving for the parameter λ produces an alternative estimate of the number of salmon caught (Chapter 27). That is, a numerical solution of the equation

$$\bar{x} = \frac{\hat{\lambda}}{1 - e^{-\hat{\lambda}}}$$

yields an estimate of the parameter λ. For the salmon data, this estimate of the mean number of salmon caught per boat is $\hat{\lambda} = 2.409$ and is the maximum likelihood estimate of the Poisson parameter λ. It is bit ironic that the Poisson distribution is central to the estimation of the number of fish caught.

A Note on Rare Events

Events such as identifying the fact that five of the last 12 U.S. presidents were left-handed are often interpreted as so unlikely by chance alone that a better explanation must exist. Since 1932, Presidents Truman, Ford, Bush, Clinton, and Obama have been left-handed.

The situation is more complicated than it appears on the surface. The last 12 U.S. presidents have been subject to numerous and extensive investigations of all aspects of their professional and personal lives. It is likely that some sort of "highly unlikely" similarity will be noticed among the huge number of comparisons made among these 12 presidents, and the "rare event" can be just about anything. If 12 people are selected at random and extensively searched for "remarkable" events, it is almost certain that a "rare event" would be found. For example, five of the 12 individuals could have a home address that begins with the number 7, or five of these people could have an uncle whose first name starts with the letter "B." The number of possibilities is huge. But they are not presidents of the United States, so no one makes extensive comparisons. Thus, there are two kinds of "rare" events, those that are noticed and those that are not. Furthermore, as the number of comparisons increases, it becomes inevitable that some remarkable "rare event" will be found. Of more importance, it is impossible to assess the likelihood of such "rare events" because the vast majority of rare events go undefined, unnoticed, and, therefore, unreported. Only a few selected special events become a focus of attention.

Nevertheless, it sometimes thought useful to estimate the probability of these "rare" events. The calculations are necessarily made under specific conditions that are almost always unrealistic. If the handedness among the last 12 presidents is analogous to tossing a biased coin 12 times with the probability of tails equal to 0.10 (probability of being left-handed), the binomial probability distribution yields the probability that five tails occurs among 12 tosses. Applying such a binomial distribution gives a "probability" that five presidents in the last 12 would be left-handed by chance alone as $P(X \geq 5) = 0.0005$. It should be kept in mind that there are also a large number of unremarkable and unnoticed "rare" events. For example, four of the last 12 presidents were previous governors of a state, and six of the last 12 have an "o" in their last name. Therefore, an accurate probability of such events as left-handedness of the last 12 presidents eludes a rigorous estimation.

The probability of 0.0005 is a reasonable value for 12 random coin tosses ($p = 0.10$) but completely fails to describe the likelihood of five left-handed president among the last 12. In addition, a philosophical question arises: Do probabilities apply to a single event that has already occurred? A probability is formally defined as the proportion of times a specific event occurs among a large number of possible events. Therefore, it is not clear that a probability such as 0.0005 sensibly applies to a one-time occurrence of a single event that has already occurred.

A lottery drawing presents another example of a "rare" event. In 2009 the Bulgarian lottery selected exactly the same six numbers on two consecutive weeks. The probability

of this event occurring by chance was said to be about 1 in 50 million. It is certain that a large number of unremarkable rare events with a probability 1 in 50 million also occurred during the same two weeks and were not noticed. In fact, the six lottery numbers selected the following week had the same probability of 1 in 50 million, but the results were unnoticed because they were unremarkable to everyone but the winner. Someone once said that an extremely rare event would be the occurrence of no "rare" events.

5

Correlation

Statistical techniques are designed to identify and describe differences. Equally, statistical techniques are designed to identify and describe similarities. The preeminent among these techniques is the product-moment correlation coefficient. This remarkable summary statistic was introduced by Sir Francis Galton (a pioneer in the use of statistical methods; b. 1822) and perfected by early statistician Karl Pearson (b. 1857). The statistical theory that justifies the properties of a correlation coefficient is sophisticated mathematically, but application to data follows a typical pattern leading to the utility and popularity of this measure of the strength of linear association between two variables.

A correlation coefficient applied to the relationship between body weight (denoted X) and diastolic blood pressure (denoted Y) measured for 20 men at high risk for coronary disease (*chd*) illustrates the properties of this fundamental statistical summary (Table 5.1).

Summary statistics relevant to the calculation of a correlation coefficient (denoted r_{XY}) for the *chd* data are the following:

means		standard deviations		covariance
$\bar{x}$	$\bar{y}$	S_X	S_Y	S_{XY}
184.3	82.3	36.219	6.242	87.168

An estimate of *the product-moment correlation coefficient* created to measure the association between two variables can be viewed as a standardizing of their estimated covariance (Chapter 27). The expression for its estimation, based on n pairs of observations represented as $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, is

$$r_{XY} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{S_{XY}}{S_X S_Y}$$

for $i = 1, 2, \ldots, n$ = number of independent sampled pairs. The correlation coefficient estimated from the example data is

$$r_{XY} = \frac{87.168}{36.219(6.242)} = 0.386.$$

Perhaps the most important property of a product-moment correlation coefficient is that the estimated covariance S_{XY}, a difficult measure of association to directly interpret, is scaled to create an easily interpreted summary value. Specifically, the strength of the linear association between two variables is characterized by a single estimated value between -1.0 and 1.0 (Chapter 27).

Table 5.1 *Data: Body Weight* (X) *and Diastolic Blood Pressure* (Y) *Measurements from* $n = 20$ *Men (Small Part of Large Study of Coronary Heart Disease Risk)*

Body weight (X)	144	154	142	152	250	204	164	259	190	175
Blood pressure (Y)	77	79	78	74	86	89	84	87	76	90
Body weight (X)	167	186	153	166	180	251	168	225	196	160
Blood pressure (Y)	91	92	73	85	80	88	83	81	71	82

An algebra trick demonstrates that the correlation coefficient r_{XY} is always greater or equal to -1 and always less than or equal to $+1$:

$$\begin{aligned} 0 &\le \frac{1}{n-1}\sum\left[\frac{(x_i-\bar{x})}{S_X} \pm \frac{(y_i-\bar{y})}{S_Y}\right]^2 \\ &\le \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{S_X}\right)^2 \pm \frac{2}{n-1}\sum\frac{(x_i-\bar{x})(y_i-\bar{y})}{S_X S_Y} + \frac{1}{n-1}\sum\left(\frac{y_i-\bar{y}}{S_Y}\right)^2 \\ &\le 1 \pm 2r_{XY} + 1, \end{aligned}$$

then $-2 \le \pm 2r_{XY}$ or $r_{\mathrm{XY}} \ge -1$ and $r_{XY} \le +1$.

Another property that makes a correlation coefficient an intuitive summary value is its relationship to a straight line. A straight line is described by the expression $y = a + bx$. Estimated values of the intercept a and slope b produce a linear description of the x/y relationship between body weight (x) and blood pressure (y) (Table 5.2).

For the *chd* data, the estimated line is $\hat{y} = 70.053 + 0.066x$. The estimated slope $\hat{b} = S_{XY}/S_X^2 = 0.066$ is yet another standardization of the estimated covariance and creates a direct and easily interpreted measure of association (Chapters 3 and 27).

Correlation between two variables x and y is identical to the correlation between the two variables $(Ax + B)$ and $(Cy + D)$, where A, B, C, and D represent constant values (Chapter 27). Less technically, a correlation coefficient is a unitless quantity that is not affected by the measurement units of either of the x/y variables; that is, in symbols,

$$r_{Ax+B,Cy+D} = r_{X,Y} \text{ (Chapter 27).}$$

Specific to an estimated linear relationship, then

$$r_{\hat{Y},Y} = r_{\hat{a}+\hat{b}X,Y} = r_{XY}$$

where $\hat{Y}$ represents values calculated from an estimated line. Therefore, a correlation coefficient between values X and Y (r_{XY}) and the correlation between linear estimates $\hat{Y}$ and observations $Y(r_{\hat{Y},Y})$ identically indicate the degree that a straight line summarizes a sample

Table 5.2 *Results: Estimate of the Linear Relationship between Body Weight and Blood Pressure (*$n = 20$ *White Males – Table 5.1)*

		Estimates	*s. e.*	*p*-value
Intercept	$\hat{a}$	70.053	–	
Slope	$\hat{b}$	0.066	0.038	0.093

of n pairs of observations (x_i, y_i). For example, when a line exactly represents the x/y relationship, then $r_{XY} = r_{Y\hat{Y}} = 1.0$ or $r_{XY} = r_{Y\hat{Y}} = -1.0$. If a straight line is useless as a summary measure of the relationship between variables x and y, then $\hat{b} = 0$ and the estimated correlation coefficient is also $r_{XY} = 0$. A bit of manipulation of the expression for a correlation coefficient shows

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{S_{XY}}{S_X S_Y} \times \frac{S_X}{S_X} = \frac{S_{XY} S_X}{S_X^2 S_Y} = \hat{b}\frac{S_X}{S_Y}.$$

For the example data, again $r_{XY} = 0.066(36.219)/6.242 = 0.386$. Unequivocally, when $r_{XY} = 0$, then $\hat{b} = 0$ and vice versa. Thus, a correlation coefficient succinctly addresses the statistical question: How well does a straight line represent the x/y data? The two estimates (r_{XY} and $\hat{b}$) are different statistical descriptions of the same relationship.

Furthermore, as might be expected, the statistical assessments of these two estimates are identical. Using the example *chd* data (Table 5.2), the comparison of estimate $\hat{b}$ to $b = 0$ (no linear association) produces the test statistic

$$z = \frac{\hat{b} - 0}{\sqrt{variance\,(\hat{b})}} = \frac{0.066}{0.038} = 1.773.$$

Again, for the case of no linear association or $\rho_{XY} = 0$, a test statistic based on the correlation coefficient also produces

$$z = \frac{r_{XY} - 0}{\sqrt{variance\,(r_{XY})}} = \frac{r_{XY} - 0}{\sqrt{\left(1 - r_{XY}^2\right)/(n-2)}} = \frac{0.386 - 0}{\sqrt{[1 - (0.386)^2]/18}} = 1.773$$

and a p-value of 0.077. The symbol ρ_{XY} represents the underlying population parameter estimated by the correlation coefficient r_{XY}. The estimated variance of the distribution of the estimate r_{XY} is

$$variance\,(r_{XY}) = \frac{1 - r_{XY}^2}{n - 2}$$

only when x and y are unrelated or, in symbols, only when $\rho_{XY} = 0$. Otherwise, the expression for an estimated variance is complicated, and the distribution of the estimate r_{XY} is not symmetric ($\rho_{XY} \neq 0$), particularly in the neighborhood of -1.0 or 1.0 (Figure 5.1). Lack of symmetry makes a direct evaluation with a normal distribution–based test statistic or confidence interval accurate only for the special case of no association.

There are two kinds of statistical transformations. One is based on empirical experience, and the other on mathematics. For example, a logarithmic transformation is justified by the fact that a variety of transformed summary statistics or data values with asymmetric distributions are observed to become more symmetric, improving the accuracy of normal distribution–based test statistics and confidence intervals. The transformation usually applied to the statistical description of a correlation coefficient is the second kind. Statistician R. A. Fisher derived a mathematical process to create transformations that cause values with asymmetric distributions to have approximately symmetric distributions. The usual normal distribution–based techniques then become more accurate statistical assessments. The transformation of the estimated correlation coefficient r_{XY}, resulting from Fisher's mathematics, dictates the use of the inverse hyperbolic tangent function.

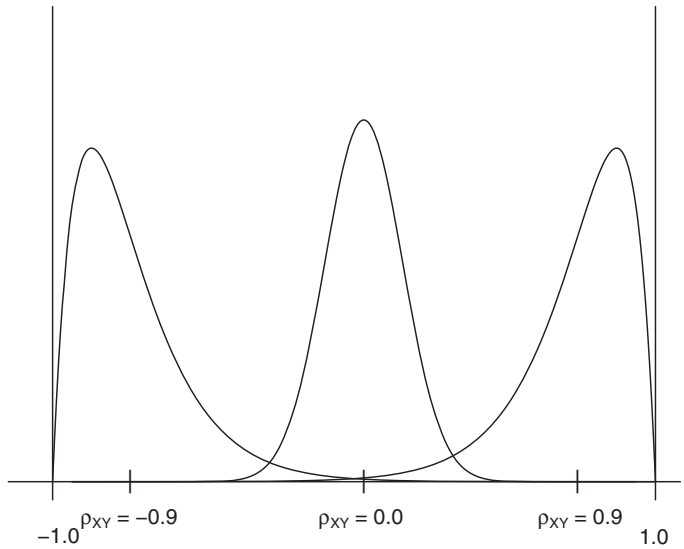


Figure 5.1 Relationship between Symmetry and Distribution of Estimated Correlation Coefficient r_{XY} for Three Values of $\rho_{XY} = \{-0.9, 0.0, \text{and } 0.9\}$

The transformation is

$$\hat{s} = arctanh\,(r_{XY}) = \frac{1}{2} log\left[\frac{1 + r_{XY}}{1 - r_{XY}}\right] \quad |r_{XY}| < 1.$$

Amazingly, Fisher's method also produces the variance of the transformed variable.

In the case of a correlation coefficient, the variance of the distribution of $\hat{s}$ is $variance\,(\hat{s}) = 1/(n-3)$, and, as before, n represents the number of sampled pairs. Thus, the distribution of $\hat{s}$ has an approximate normal distribution with a known variance. Notice that the estimated value and the variance of the $\hat{s}$ statistic are not related, which is a property of symmetric distributions.

An accurate but still approximate 95% confidence interval emerges because the distribution $\hat{s}$ has an approximate normal distribution for most values of r_{XY}. For the *chd* data (Table 5.1), the approximate 95% confidence interval $(\hat{A}, \hat{B})$ based on the transformed value $\hat{s}$ is

$$\hat{s} \pm 1.960\sqrt{variance\,(\hat{s})} = 0.407 \pm 1.960\sqrt{1/17} \rightarrow (\hat{A}, \hat{B}) = (-0.068, 0.882),$$

where $r_{XY} = 0.386$ and, as required,

$$\hat{s} = arctanh\,(r_{XY}) = arctanh\,(0.386) = \frac{1}{2} log\left[\frac{1 + 0.386}{1 - 0.386}\right] = 0.407.$$

The inverse function of the $arctanh(\hat{s})$ is

$$f^{-1}(\hat{s}) = \frac{e^{2\hat{s}} - 1}{e^{2\hat{s}} + 1} = r_{XY}.$$

Therefore, the bounds for an approximate 95% confidence interval $(\hat{A}, \hat{B})$ become

$$\text{lower bound} = f^{-1}(\hat{A}) = \frac{e^{2\hat{A}} - 1}{e^{2\hat{A}} + 1} = \frac{e^{2(-0.069)} - 1}{e^{2(-0.069)} + 1} = -0.069$$

Table 5.3 *Data: Ranked Values and Differences for Weight and Diastolic Blood Measurements from* $n = 20$ *Men at High Risk of a Coronary Event (Table 5.1)*

Body weight – ranks (X)	2	5	1	3	18	16	7	20	14	11
Blood pressure – ranks (Y)	5	7	6	3	14	17	12	15	4	18
Difference ($X - Y$)	−3	−2	−5	0	4	−1	−5	5	10	−7
Body weight – ranks (X)	9	13	4	8	12	19	10	17	15	6
Blood pressure – ranks (Y)	19	20	2	13	8	16	11	9	1	10
Difference ($X - Y$)	−10	−7	2	−5	4	3	−1	8	14	−4

and

$$\text{upper bound} = f^{-1}(\hat{B}) = \frac{e^{2\hat{B}} - 1}{e^{2\hat{B}} + 1} = \frac{e^{2(0.882)} - 1}{e^{2(0.882)} + 1} = 0.707$$

based on the estimated correlation coefficient $r_{XY} = 0.386$. The approximate 95% confidence interval is then (−0.069, 0.707). As required, the inverse function yields

$$f^{-1}(\hat{s}) = f^{-1}(0.407) = \frac{e^{2(0.407)} - 1}{e^{2(0.407)} + 1} = r_{XY} = 0.386.$$

Inverse functions and their role in data analysis are described in detail (Chapter 12).

Spearman's Rank Correlation Coefficient

As a general rule, parametric statistics have nonparametric versions (Chapters 8 and 15). Spearman's rank correlation coefficient (denoted r_s) is a nonparametric version of the parametric product-moment correlation coefficient. Like many nonparametric techniques, estimation begins with replacing the observed values with their ranks. The observed values x_i become $X_i =$ rank of x_i, and similarly, the observed values y_i become $Y_i =$ rank of y_i. For the example coronary heart disease data, the observed weight and blood pressure values are replaced with their ranked value, and the properties of the originally sampled distribution are no longer relevant to estimation and evaluation (Table 5.3).

Spearman's rank correlation coefficient is simply the previous product-moment expression for a correlation coefficient applied to pairs of ranked values, namely, pairs (X_i, Y_i).

Summary statistics relevant to the calculation of a rank correlation coefficient are the following:

means		standard deviations		covariance
$\bar{X}$	$\bar{Y}$	S_X	S_Y	S_{XY}
10.5	10.5	5.916	5.916	15.684

For the example *chd* data, Spearman's rank correlation coefficient is $r_s = S_{XY}/S_X S_Y = 15.684/[5.916(5.916)] = 0.448$. A usually presented expression that gives little indication of the origins of Spearman's correlation coefficient but provides a more easily calculated value is

$$r_S = \frac{S_{XY}}{S_X S_Y} = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6(734)}{7980} = 0.448 \qquad i = 1, 2, \ldots, n$$

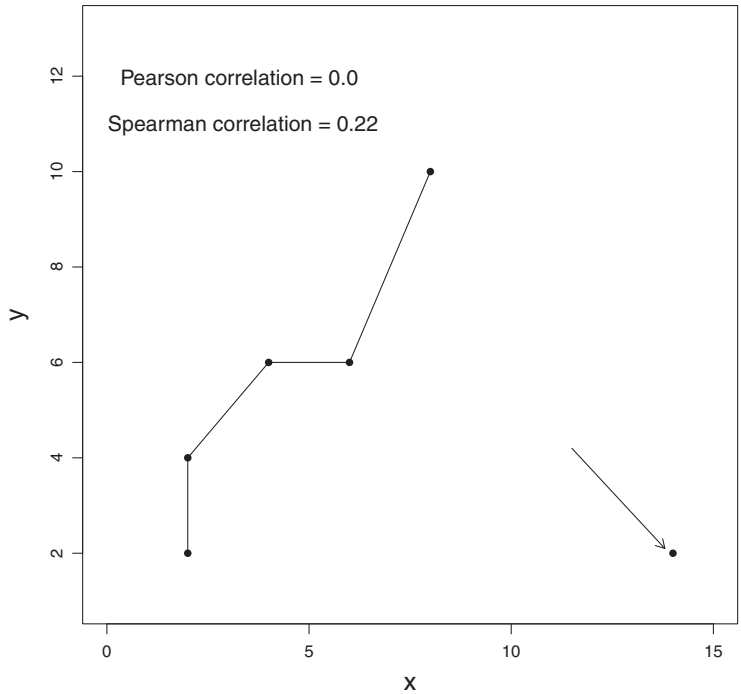


Figure 5.2 Contrast between Spearman's Rank Correlation and a Product-Moment Correlation Coefficient for an Exaggerated Situation ($r_s = 0.22$ and $r_{XY} = 0.0$)

where $d_i = rank(x_i) - rank(y_i) = X_i - Y_i$ (Table 5.3). Typically, a product-moment and Spearman's correlation coefficients produce similar values. Like all rank procedures, however, the Spearman estimate is insensitive to "out-and-out" outliers and other extreme patterns of sampled data. Two examples illustrate this property using artificial and exaggerated "data" (Figures 5.2 and 5.3).

Point Biserial Correlation Coefficient

The point biserial correlation coefficient (denoted r_{pb}) is another application of the expression for the product-moment correlation coefficient. In this case, values x_i represent measured observations, but values y_i represent binary values that take on the values 0 or 1.

For example, consider a representative sample of $n = 30$ pairs of observations from $n = 3154$ men participating in a coronary heart disease study. The value x_i represents a measure of the subject's cholesterol level and value y_i represents the subject's behavior type as a binary variable, type A behavior ($y_i = 1$) and type B behavior ($y_i = 0$) (Table 5.4). Summary values relevant to the calculation of a point biserial correlation coefficient (denoted r_{bp}) are the following:

means		standard deviations		covariance
$\bar{x}$	$\bar{y}$	S_X	S_Y	S_{XY}
206.4	0.533	44.533	0.507	7.952

Table 5.4 *Data: Cholesterol Levels* (x) *and Behavior Types* (y) $n = 30$ *Men (Again a Small Subset from a Large Coronary Heart Disease Study)*

Cholesterol	225	177	181	132	255	182	155	140	149	325	223	271	238	189	140
Behavior type[a]	1	1	0	0	0	0	0	1	0	1	1	1	1	0	0
Cholesterol	247	220	176	185	202	227	192	179	237	177	239	210	220	274	225
Behavior type[a]	0	0	1	0	1	1	1	0	1	1	1	1	0	1	0

[a]Type A = 1 and type B = 0.

Again direct application of the product-moment expression for r_{XY} using these summary values yields the estimate $r_{XY} = S_{XY}/S_X S_Y = 7.592/[44.533(0.507)] = 0.352$. Like the Spearman rank correlation coefficient, this version of a correlation coefficient is again the expression for the product-moment correlation coefficient applied to a specific kind of data.

An expression for the same estimate (r_{XY}) takes advantage of the fact that the binary y values simplify the expressions $\sum (y_i - \bar{y})^2$ and $\sum (x_i - \bar{x})(y_i - \bar{y})$. Specifically, components of the product-moment correlation coefficient that create this alternate expression called a *point biserial correlation coefficient* are the following:

$$SS_{XX} = \sum (x_i - \bar{x})^2 \qquad i = 1, 2, \ldots, n = n_0 + n_1 = \text{number of pairs},$$

$$SS_{YY} = \sum (y_i - \bar{y})^2 = n_0 \left(0 - \frac{n_1}{n}\right)^2 + n_1 \left(1 - \frac{n_1}{n}\right)^2 = \frac{1}{n^2}\left(n_0 n_1^2 + n_0^2 n_1\right) = \frac{n_1 n_0}{n}$$

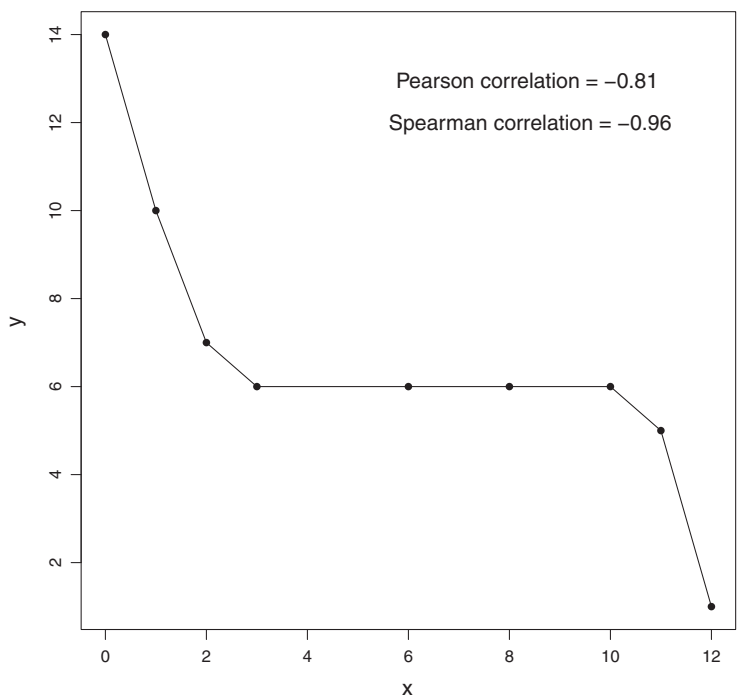


Figure 5.3 Another Contrast between Spearman's Rank Correlation and a Product-Moment Correlation Coefficient for an Exaggerated Situation ($r_s = -0.96$ and $r_{XY} = -0.81$)

and

$$SS_{XY} = \sum (x - \bar{x})(y - \bar{y}) = \sum (x_i - \bar{x}) \left(0 - \frac{n_1}{n}\right) + \sum (x_i - \bar{x}) \left(1 - \frac{n_1}{n}\right)$$
$$= \frac{1}{n}[-n_1 n_0 \bar{x}_0 + n_1 n_0 \bar{x} + n_1 n_0 \bar{x}_1 - n_1 n_0 \bar{x}] = \frac{n_1 n_0}{n}(\bar{x}_1 - \bar{x}_0)$$

for n_0 values of $y_i = 0$ and n_1 values of $y_i = 1$ (Chapter 8). Therefore, an expression for the point biserial correlation coefficient becomes

$$r_{pb} = \frac{S_{XY}}{\sqrt{S_X S_Y}} = \frac{SS_{XY}}{\sqrt{SS_X SS_Y}} = \frac{\sqrt{\frac{n_1 n_0}{n}}(\bar{x}_1 - \bar{x}_0)}{\sqrt{SS_X}}.$$

From the *chd* data, summary values relevant to the calculation of a point biserial correlation coefficient r_{bp} are the following:

mean values		sum of products		
$\bar{x}_1$	$\bar{x}_0$	SS_X	SS_Y	SS_{XY}
220.812	189.929	57,513.2	7.467	230.6

then again

$$r_{pb} = \frac{SS_{XY}}{\sqrt{SS_X SS_Y}} = \frac{\sqrt{\frac{14(16)}{30}}(220.812 - 189.929)}{\sqrt{57{,}513.2}} = 0.352$$

for $n_0 = 14$ and $n_1 = 16$ pairs of observations (Table 5.4).

The expression for r_{pb} indicates that the point biserial correlation coefficient is a scaling of the difference between two estimated mean values creating a summary value between -1 and $+1$. It should, therefore, not be surprising that the correlation coefficient r_{pb} is related to the two-sample *t*-test. In fact, the test statistic for the point biserial correlation coefficient

$$T = \frac{r_{pb} - 0}{\sqrt{(1 - r_{pb}^2)/(n - 2)}}$$

is identical to the *t*-test comparison of two estimated mean values. Student's *t*-test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{V_p \left[\frac{1}{n_1} + \frac{1}{n_0}\right]}}$$

where the expression for the pooled variance is

$$V_p = \frac{(\mathrm{n}_1 - 1)\ variance\,(x_1) + (n_0 - 1)\ variance\,(x_0)}{n_1 + n_0 - 2}.$$

For test statistics T and t, then $T = t$. Specifically, the test statistic is $T = t = 1.989$ (degrees of freedom = 28) with associated p-value = 0.057 for both versions of assessing the behavior-*chd* relationship (Table 5.4).

Nonparametric Measure of Association: The γ-Coefficient

The expression for the covariance estimated from variables x and y is

$$S_{XY} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}).$$

When $x_i - \bar{x}$ and $y_i - \bar{y}$ are both positive or both negative, their product contributes a positive value to the covariance. When $x_i - \bar{x}$ and $y_i - \bar{y}$ are positive and negative or negative and positive, their product contributes a negative value to the covariance. The sum of these products produces a summary covariance based on n observed pairs (x_i, y_i) that reflects the x/y association (Chapter 27). This estimated value is a parametric measure of the strength of association within pairs of observations. The contributions are proportional to the magnitude of the x/y product. Thus, the properties of the sampled distribution influence the measured association (correlation coefficient).

An alternative strategy produces a nonparametric measure of the association. Like the product-moment correlation coefficient, a nonparametric summary value between -1 and $+1$ measures the strength of association within pairs of observations. Measures of association, however, are created so that the properties of the sampled distribution do not influence the analytic results.

To measure the association within the x/y pairs, when x_i is greater than x_j and y_i is greater than y_j, the x/y pair is said to be concordant. When x_i is less than x_j and y_i is less than y_j, the x/y pair is also said to be concordant. Similarly, when x_i is greater than x_j and y_i is less than y_j, the x/y pair is said to be discordant. Again, when x_i is less than x_j and y_i is greater than y_j, the x/y pair is also said to be discordant. The total number of concordant pairs (denoted C) and the total number of discordant pairs (denoted D) produce a nonparametric measure of association. A measure of association created by comparing the proportion of concordant pairs to the proportion of discordant pairs is usually called the *gamma coefficient* or sometimes *Goodman and Kendall's γ coefficient* and, in symbols, is

$$\gamma = \frac{C}{C+D} - \frac{D}{C+D} = \frac{C-D}{C+D}.$$

The sum $C + D$ is $n(n-1)$ where n represents the total number of pairs compared.

A γ-coefficient of $+1$ occurs when all pairs are concordant ($D = 0$), and a γ coefficient of -1 occurs when all pairs are discordant ($C = 0$). As expected, the γ coefficient equals zero when the number concordant equals the discordant pairs ($C = D$). A value of the γ coefficient of zero indicates no association but does not mean that the product-moment coefficient also r_{XY} equals zero.

A table constructed from values of all possible differences of x-values $(x - x')$ and y-values $(y - y')$ produce the values C and D where positive differences are assigned a value $+1$ and negative differences are assigned a value -1 (Table 5.5). Consider four artificial x/y pairs of observations to illustrate one approach to calculating values C and D to estimate a γ-coefficient:

$$(x_i, y_i) = \{(1, 2), (6, 5), (5, 3), \text{ and } (3, 6)\}.$$

The products of differences $(x - x')$ and $(y - y')$ produce values $(x - x')(y - y')$ with either positive or negative signs. Total counts of positive (C) and negative (D) values are the

Table 5.5 *Example: Computation of Number of Concordant Pairs* (C) *and Number of Discordant Pairs* (D)

	$(x - x')$ − Differences					$(y - y')$ − Differences							
x/x'	1	6	5	3	y/y'	2	5	3	6		$(x - x')(y - y')^a$		
1	0	−1	−1	−1	2	0	−1	−1	−1	0	1	1	1
6	1	0	1	1	5	1	0	1	−1	1	0	1	−1
5	1	−1	0	1	3	1	−1	0	−1	1	1	0	−1
3	1	−1	−1	0	6	1	1	1	0	1	−1	−1	0

[a]The product +1 identifies concordant pairs, and the product −1 identifies discordant pairs.

components of the γ coefficient. For the example, there are $C = 8$ concordant pairs (positive) and $D = 4$ discordant pairs (negative) among $n(n - 1) = 4(3) = 12$ pairs compared. The estimate of the γ coefficient is then

$$\gamma = \frac{C}{C + D} - \frac{D}{C + D} = \frac{C - D}{C + D} = \frac{8 - 4}{8 + 4} = \frac{4}{12} = 0.333.$$

For large samples sizes, the calculation of C and D are best accomplished with computer statistical software. For example, the previous body weight (x) and blood pressure data (y) from $n = 20$ study participants yield the values $C = 250$ and $D = 130$ among the $n(n - 1) = 20(19) = 380$ comparison of pairs of observations (Table 5.1). The estimate γ-coefficient becomes $\gamma = (250 - 130)/380 = 0.316$. The variance of the γ-statistic is tedious to calculate but a bootstrap or computer estimation are routinely applied (Chapter 11).

The γ-coefficient nonparametric strategy also applies to measuring association between two categorical variables. Data on vitamin use and socioeconomic status (categorical variables) from a study of birth defects illustrate (Table 5.6).

A general notation for the calculation of counts of concordant pairs C and discordant pairs D from a two-way table is cumbersome and extensive. But the pattern becomes clear from a detailed description of the calculation from the 3 × 4 table of vitamin use and socioeconomic status. Each cell count in the table multiplied by the total counts in the blocks of cells below and to the right produce the count C. The value D is similarly calculated. Each cell count in the table multiplied by the total counts in the blocks of cells below and to the left produces

Table 5.6 *Data: Vitamin Use during Pregnancy (Rows = 3 Levels) and Socioeconomic status (Columns = 4 Levels) for n* = 826 *White Study Participants*

	Socioeconomic status				
	ses_1	ses_2	ses_3	ses_4	Total
Always used	78	84	120	101	383
Used during	50	90	89	122	351
Never used	13	31	21	27	92
Total	141	205	230	250	826

Table 5.7 *Computation: Pattern of Counts for Calculation of C and D from* 3×4 *Table (Blocks Are Indicated by Parentheses)*

$$
\begin{aligned}
C = \; & d_{11}(d_{22} + d_{23} + d_{24} + d_{32} + d_{33} + d_{34}) + \\
& d_{12}(d_{23} + d_{24} + d_{33} + d_{34}) + \\
& d_{13}(d_{24} + d_{34}) + \\
& d_{21}(d_{32} + d_{33} + d_{34}) + \\
& d_{22}(d_{33} + d_{34}) + \\
& d_{23}(d_{34}) \\
D = \; & d_{14}(d_{21} + d_{22} + d_{23} + d_{31} + d_{32} + d_{33}) + \\
& d_{13}(d_{21} + d_{22} + d_{31} + d_{32}) + \\
& d_{12}(d_{21} + d_{31}) + \\
& d_{24}(d_{31} + d_{32} + d_{33}) + \\
& d_{23}(d_{31} + d_{32}) + \\
& d_{22}(d_{31})
\end{aligned}
$$

the count D; that is, the blocks below and to the right contain the concordant pairs, and the blocks below and to the left contain the discordant pairs. The total counts C and D produce the estimated γ correlation coefficient.

A complete description of the calculation of C and D from the birth defects data is contained in an array denoted d (Tables 5.7).

Then, for concordant pairs

$$C = 78(380) + 84(259) + 120(149) + 50(79) + 90(48) + 89(27) = 79{,}949$$

and for discordant pairs

$$D = 101(294) + 120(184) + 84(63) + 122(65) + 89(44) + 90(13) = 70{,}082.$$

For the birth defects data, the γ-coefficient is

$$\gamma = \frac{C - D}{C + D} = \frac{79{,}949 - 70{,}082}{150{,}031} = 0.0658.$$

An additional consideration is the number of "tied" values in the table; that is, values in the same column or in the same row have the same value. The total number of these "tied" pairs of observations in the same rows is denoted t_x, and the total number of "tied" pairs of observations in the same columns is denoted t_y. Again, rather than describe the general notation, the number of row and column "tied" values calculated from the example birth defects data are contained in the array denoted d (Table 5.8).

Then, for the vitamin use birth defects data,

$$
\begin{aligned}
t_x &= 78(305) + 84(221) + 120(101) + 50(301) + 90(211) + 89(122) \\
&\quad + 13(79) + 31(48) + 21(27) \\
&= 102{,}454
\end{aligned}
$$

and

$$
\begin{aligned}
t_y &= 78(63) + 50(13) + 84(121) + 90(31) + 120(110) + 89(21) + 101(149) + 122(27) \\
&= 51{,}930.
\end{aligned}
$$

Table 5.8 *Computation: Pattern of Counts for Calculation of "Tied" Values t_X and t_Y for a* 3×4 *Table*

$$
\begin{aligned}
t_x = {} & d_{11}(d_{12} + d_{13} + d_{14}) + d_{12}(d_{13} + d_{14}) + d_{13}(d_{14}) + \\
& d_{21}(d_{22} + d_{23} + d_{24}) + d_{22}(d_{23} + d_{24}) + d_{23}(d_{24}) + \\
& d_{31}(d_{32} + d_{33} + d_{34}) + d_{32}(d_{33} + d_{34}) + d_{33}(d_{34}) \\
t_y = {} & d_{11}(d_{21} + d_{31}) + d_{21}(d_{31}) + \\
& d_{12}(d_{22} + d_{32}) + d_{22}(d_{32}) + \\
& d_{13}(d_{23} + d_{33}) + d_{23}(d_{33}) + \\
& d_{14}(d_{24} + d_{34}) + d_{24}(d_{34})
\end{aligned}
$$

A measure of association, called Kendall's τ coefficient, accounts for "tied" values by modifying the γ-coefficient and is given by the expression

$$\tau = \frac{C - D}{\sqrt{(C + D + t_x)(C + D + t_y)}}.$$

For the example data, then

$$\tau = \frac{79{,}949 - 70{,}082}{\sqrt{(150{,}031 + 102{,}454)(150{,}031 + 51{,}930)}} = 0.0437.$$

Kendall's τ is not restricted to characterizing association between categorical variables. When data consist of x/y pairs of measured observations ($t_x = t_y = 0$), then the nonparametric coefficient $\tau = (C - D)/(C + D) = \gamma$ for n data pairs.

A Special Case: The $2 \times k$ Table

A special case of the γ-coefficient measuring the strength of association between a binary variable and a numeric or an ordinal variable contained in a $2 \times k$ table is called *Somer's coefficient* (denoted γ_0). The data collected from a health attitudes survey provides an opportunity to measure the strength of association between views on longevity (a binary variable) and reported exercise (an ordinal variable; Table 5.9).

Table 5.9 *Data (Chapter 8): Responses* ($n = 165$) *to a Question*[a] *on Life Expectancy and Reported Physical Activity (an Ordinal Variable)*

	Reported exercise levels				
	None	Occasional	Moderate	Strenuous	Total
Yes	7	7	8	20	42
No	25	34	32	32	123
Total	32	41	40	52	165

[a]Answer to the question: Do you think exercise increases the length of life (yes/no)?

The calculation of values C, D, and t_x are again a count of concordant, discordant, and "tied" pairs. For a 2×4 table of survey data, contained in the array labeled m, then

$$\begin{aligned} C &= m_{11}(m_{22} + m_{23} + m_{24}) + m_{12}(m_{22} + m_{23}) + m_{13}(m_{23}) \\ &= 7(98) + 7(64) + 8(32) = 1309, \\ D &= m_{14}(m_{21} + m_{22} + m_{23}) + m_{13}(m_{21} + m_{22}) + m_{12}(m_{12}) \\ &= 20(91) + 8(59) + 7(25) = 2467 \end{aligned}$$

and

$$\begin{aligned} t_x &= m_{11}(m_{21}) + m_{12}(m_{22}) + m_{13}(m_{23}) + m_{14}(m_{24}) \\ &= 7(25) + 7(34) + 8(32) + 20(32) = 1309. \end{aligned}$$

These counts are no more than a special case of the previous counts of C and D pairs from a two-way table. The expression for the Somer's coefficient is

$$\gamma_0 = \frac{C - D}{C + D + t_x}.$$

Therefore, this measure of association between *yes*/*no* answers and reported amount of exercise is

$$\gamma_0 = \frac{1390 - 2467}{1390 + 2467 + 1309} = -0.209.$$

The negative value results from the arbitrary coding of the binary *yes*/*no* variable as 1 or 0.

The γ_0-coefficient is related to the Mann-Whitney probability denoted $\hat{P}$, also sometimes called a ridit value (Chapter 8). Specifically, the Mann-Whitney probability $\hat{P} = \frac{1}{2}(1 - \gamma_0)$, and, for the questionnaire data, the value $\hat{P} = \frac{1}{2}(1 + 0.209) = 0.604$. Thus, the estimated probability that a random person who replied "no" exercises less than a random person who replied "yes" is $\hat{P} = 0.604$. A measure of the strength of an association between the *yes*/*no* responses and exercise is $\gamma_0 = 0.209$.

Chi-Square–Based Measures of Association

The chi-square test statistic typically used to identify an association between two categorical variables increases as the table number of rows and columns increases. Therefore, chi-square summary statistics calculated from a table are not effective measures of association; that is, the number of rows and columns primarily determine the chi-square value regardless of the degree of association. Modified chi-square values have been suggested to measure the association between two categorical variables from a table of counts. Five of these measures illustrate:

1. Phi correlation (denoted ϕ)
2. Contingency coefficient (denoted C)
3. Cramer's coefficient (denoted V)
4. Tshuprow's coefficient (denoted T) and
5. Yule's coefficient (denoted Q).

Table 5.10 *Data: Coronary Heart Disease Occurrence by A/B Behavior Type* ($n = 3154$)

	chd	No *chd*	Total
Type A	178	1411	1589
Type B	79	1486	1565
Total	257	2897	3154

These chi-square–based coefficients arise from different strategies to produce commensurate measures of association. To illustrate, these five coefficients are applied to a 2×2 table of $n = 3154$ men classified by *A*/*B*-behavior type and present or absence of a coronary event from an eight-year study (Table 5.10).

The estimated odds ratio is $\hat{or} = 2.373$, and the chi-square test statistic from these *chd* data is $X^2 = 39.80$ (p-value < 0.001).

Specifically five chi-square measures of association are the following:

Phi coefficient:

$$\phi = \sqrt{\frac{X^2}{n}} \quad \text{and} \quad \phi = \sqrt{\frac{39.80}{3154}} = 0.112,$$

Contingency coefficient:

$$C = \sqrt{\frac{X^2}{X^2 + n}} \quad \text{and} \quad C = \sqrt{\frac{39.80}{39.80 + 3154}} = 0.112,$$

Cramer's coefficient:

$$V = \sqrt{\frac{X^2}{n \times \text{minimum}[r-1, c-1]}} \quad \text{and} \quad V = \sqrt{\frac{39.80}{3154}} = 0.112,$$

Tshuprow's coefficient:

$$T = \sqrt{\frac{X^2}{n \times \sqrt{(r-1)(c-1)}}} \quad \text{and} \quad T = \sqrt{\frac{39.80}{3154}} = 0.112,$$

and

Yules's coefficient:

$$Q = \frac{ad - bc}{ad + bc} = \frac{\hat{or} - 1}{\hat{or} + 1} \quad \text{and} \quad Q = \frac{178(1486) - 1411(79)}{178(1486) + 1411(79)} = \frac{2.373 - 1}{2.373 + 1} = 0.407$$

where r represents the number of rows and c represents the number of columns of a $r \times c$ table.

Note the ϕ-coefficient and Yule coefficients apply only to binary variables classified into 2×2 tables. Furthermore, the ϕ-coefficient, C-coefficient, V-coefficient, and T-coefficient yield the same value for a 2×2 table. The C-coefficient, V-coefficient, and T-coefficient also apply to $r \times c$ two-way tables. When the table is square, then $V = T$, otherwise $T < V$.

Table 5.11 *Data: Smoking Exposure (Rows = 3 Levels) and Socioeconomic Status (Columns = 5 Levels) for n = 2522 White Males*

	Socioeconomic status					
	ses_1	ses_2	ses_3	ses_4	ses_5	Total
Smokers	78	124	149	221	343	915
Nonsmokers	102	153	220	352	481	1308
Past smokers	103	81	32	48	35	299
Total	283	358	401	621	859	2522

Yule's coefficient Q is remarkable because it was suggested in the early history of statistics (circa 1900) as a measure of association.

These chi-square based statistics are commensurate indications of association. They are not estimates in the sense that they randomly vary from an underlying parameter. Thus, they have only comparative interpretations and are not often used.

Proportional Reduction in Error Criterion

A strategy central to many statistical techniques consists of a comparison of a summary value calculated under specified conditions to the same summary value calculated after modifying the conditions. For example, to evaluate the influence of age on a mortality rate, comparison of an unadjusted rate ignoring age to the same rate adjusted to account for age identifies the extent of influence of age (Chapter 9). A measure of association that employs this strategy, proposed by statisticians Goodman and Kruskal, is called the *proportional reduction in error criterion*.

This nonparametric measure of association does not require substantial theory. It is fundamentally a comparison between a summary measure of association calculated under one set of conditions to the same summary measure calculated under more restricted conditions. The summary value is a measure of error incurred in classifying observations in an $r \times c$ table.

For an example, consider a 3×5 table created from $N = 2522$ men classified by smoking exposure and socioeconomic status (Table 5.11). A single probability (denoted $\hat{P}_0$) ignoring smoking exposure is compared to a weighted average of probabilities accounting for smoking exposure (denoted $\hat{P}$).

An unadjusted measure of classification error is the probability of misclassifying a study subject based only on socioeconomic status, ignoring smoking exposure. The best estimate (highest probability) of a correct classification is $859/2522 = 0.341$ (ses_5).

Therefore, the corresponding estimate of the probability of misclassification of subject's socioeconomic status is $\hat{P}_0 = 1 - 859/2522 = 1 - 0.341 = 0.659$ ignoring smoking exposure (Table 5.11, last row).

To estimate the same value taking smoking exposure into account generates three estimated misclassification probabilities, one for each smoking category (rows). They are the following:

for smokers: $\hat{P}_1 = 1 - \dfrac{343}{915} = 0.625$,

for nonsmokers: $\hat{P}_2 = 1 - \dfrac{481}{1308} = 0.632$, and

for past smokers: $\hat{P}_3 = 1 - \dfrac{103}{299} = 0.656$.

Combining these probabilities yields an adjusted probability of misclassification of a subject's socioeconomic status accounting for smoking exposure. The weighted average for the row-specific probabilities

$$\begin{aligned}\hat{P} &= \frac{1}{2522}\left\{915\left[1 - \frac{343}{915}\right] + 1308\left[1 - \frac{481}{1308}\right] + 299\left[1 - \frac{103}{299}\right]\right\} \\ &= \frac{(915 + 1308 + 299) - (343 + 481 + 103)}{2522} = \frac{1595}{2522} = 0.632\end{aligned}$$

produces a single summary estimate of misclassification. The weights applied to each estimated probability $\hat{P}_i$ are the total counts from their respective rows of the table.

The estimated proportional reduction in error becomes

$$\lambda = \frac{\hat{P}_0 - \hat{P}}{\hat{P}_0} = \frac{0.659 - 0.632}{0.659} = 0.041,$$

sometimes called the λ-coefficient. Directly, from the *smoking*/*ses* table, the proportional reduction in the error criterion is

$$\lambda = \frac{\hat{P}_0 - \hat{P}}{\hat{P}_0} = \frac{(1 - 859/2522) - (1 - 927/2522)}{1 - 859/2522} = \frac{927 - 859}{2522 - 859} = \frac{68}{1663} = 0.041.$$

In general notation, an expression to estimate the λ-coefficient is

$$\lambda = \frac{\sum \max(n_{ij}) - \max(n_{\cdot j})}{N - \max(n_{\cdot j})}$$

where n_{ij} represents specific cell counts, $n_{\cdot j}$ represents the sums of the column counts, and $\sum \max(n_{ij})$ represents the sum of the maximum count contained in each row. The total number of observations is represented by N. Again, the previous estimated λ-coefficient in this slightly more general form is

$$\hat{\lambda} = \frac{(343 + 481 + 103) - 859}{2522 - 859} = \frac{68}{1663} = 0.041.$$

A parallel λ-coefficient can be calculated from the columns of an $r \times c$ table. Furthermore, it is occasionally suggested that a coefficient created by averaging these row- and column-generated λ-coefficients creates a useful nonparametric summary measure of association from a two-way table of counts.

Applications

6

The 2 × 2 Table

Properties of a 2 × 2 table are primarily properties of the probabilities of the joint occurrence of two binary variables. For concrete terminology, these two binary variables are called presence (denoted D) and absence (denoted $\bar{D}$) of a disease and presence (denoted E) and absent (denoted $\bar{E}$) of an exposure. A 2 × 2 table displays the four joint probabilities (Table 6.1).

Three fundamental relationships among the four joint probabilities flow from a 2 × 2 table:

1. $P(DE) + P(\bar{D}E) = P(E)$

The row and column joint probabilities add to a single *marginal probability*. For example, the probability of the presence of disease and exposure $P(DE)$ and the probability of the absence of a disease and again the presence of exposure $P(\bar{D}E)$ add to the marginal probability of exposure $P(E)$. As displayed, all four marginal probabilities are the sum of two joint probabilities (Table 6.1).

2. $P(DE) = P(D|E)P(E)$

The joint probability of disease and exposure can be expressed as the probability that disease occurs among exposed individuals multiplied by the probability an individual is exposed. The probability of disease (D) is said to be conditional on occurrence of exposure (E), which is another way of saying that only a single row or column in a 2 × 2 table is relevant. The conditional probability of disease among exposed individuals is written as $P(D \mid E)$ and reads "the probability of D given E." For example, conditional on exposure, the 2 × 2 table becomes the following one-way table:

Row 1	$P(DE)$	$P(\bar{D}E)$	$P(E)$

Therefore, the expression for the probability of disease conditional on exposure is then $P(D|E) = P(DE)/P(E)$ or $P(DE) = P(D|E)\,P(E)$.

3. $P(D|E) = P(D)$

The equality of probabilities $P(D)$ and $P(D|E)$ occurs only when disease D is unrelated to exposure E. That is, the probability of the occurrence of disease (D) is the same whether exposure (E) occurs or not ($\bar{E}$). Thus, another expression of the same property is

Table 6.1 *Notation: Joint Distribution of Two Binary Variables Indicating Presence or Absence of Disease* $(D/\bar{D})$ *and Presence or Absence of Exposure* $(E/\bar{E})$

	Disease		
Exposure	D	$\bar{D}$	Total
E	$P(DE)$	$P(\bar{D}E)$	$P(E)$
$\bar{E}$	$p(D\bar{E})$	$p(\bar{D}\bar{E})$	$P(\bar{E})$
Total	$P(D)$	$P(\bar{D})$	1.0

$P(D|\bar{E}) = P(D)$. A direct result of either relationship is that the joint probability $P(DE)$ then equals $P(D)\,P(E)$. In symbols,

$$P(D|E) = P(D) \quad \text{and} \quad P(D|E)P(E) = P(D)P(E), \quad \text{then} \quad P(DE) = P(D)P(E).$$

Under these conditions, the occurrences of disease D and exposure E are said to be *independent* or more precisely *stochastically independent*.

Occasionally, conditional probabilities $P(D|E)$ and $P(E|D)$ are confused. An article published in the *Journal of the American Medical Association* entitled "Major Risk Factors as Antecedents of Fatal and Nonfatal Coronary Heart Disease Events" (2003) provides an example. The article contains an investigation of the "50% myth." In this context, the "50% myth" refers to an occasionally made statement that more than 50% of individuals who die from coronary heart disease lack a major risk factor. Using coronary heart disease (*chd*) data from a Chicago area–based health plan, the authors reported that about 90% of the individuals who died from a coronary event had one or more of 11 major risk factors (denoted one^+). In symbols, they found $P(one^+$ risk factors | *chd* death$) > 0.50$. In addition, they found similar probabilities in various groups and in other studies. These conditional probabilities, however, do not reflect risk. They reflect the frequency of the risk factors among *chd* deaths. For example, the vast majority of individuals who die from a *chd* event are right-handed. The probability is also about 0.90 but being right-handed is not related to the risk of heart disease or, in symbols, P(right handed | *chd* death) = P(right-handed). It is simply the frequency of right-handed individuals among the observed coronary heart disease deaths. If living in the Chicago area were considered a risk factor, then for the *chd*-data P(Chicago-area resident | *chd* death) = 1.0. The probability of interest and relevant to the "50% myth" is P(*chd* death | one^+ risk factors). The more relevant question is: What is the risk of a *chd* event among individuals with one of more risk factors?

Consider $n = 3153$ men at high risk for a coronary event observed for about eight years to study *chd* risk factors. All participants can be classified by presence (D) or absence ($\bar{D}$) of a coronary heart disease event and by one or more major risk factors present (one^+) or absence of risk factors (*none*), creating a 2 × 2 table (Table 6.2). Unlike the Chicago study, individuals without a *chd* event are included. Like the Chicago study, the conditional probability $P(one^+$ risk factors | *chd*$) = 246/257 = 0.957$.

This probability has little to do with heart disease risk and is close to 100% because, as noted, high-risk men were deliberately selected for the study. Clearly more than 50%

Table 6.2 *Data: Presence or Absence of Risk Factors*[a] *and Coronary Heart Disease Events from Eight-Year Follow-up Study of n = 3153 High-Risk Men*

	chd	No *chd*	Total
One$^+$	246	2373	2619
None	11	523	534
Total	257	2896	3153

[a]Risk factors: height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, smoking, and behavior type.

of individuals who had a coronary event have one or more major risk factors, but this observation is not relevant to the 50% myth. A relevant probability is $P(chd|one^+$ risk factors$) = 246/2619 = 0.095$.

The Analysis of a 2 × 2 Table

The general notation for a 2 × 2 table is not consistent. The following discussion uses a popular version (Table 6.3). From a study of risk factors and pregnancy outcome, exceptionally small mothers (<45 kg) classified by ethnic group (African American or white) and by birth weight of their newborn infant ($lbwt \leq$ birth weight of 2500 g or $\overline{lbwt} >$ birth weight of 2500 g) create a typical 2 × 2 table (Table 6.4).

Among many techniques to evaluate an association described by a 2 × 2 table, the most fundamental is the comparison of the proportions of low-birth-weight infants observed in each ethnic group. The statistical question becomes: does the estimated conditional probability $\hat{p}_1 = \hat{P}(lbwt|\text{African American}) = 61/163 = 0.374$ systematically differ from the estimated conditional probability $\hat{p}_2 = \hat{P}(lbwt|\text{white}) = 52/248 = 0.210$?

To start, the conjecture is made that the observed difference in the likelihood of a low-birth-weight infant is only due to random variation. That is, both estimated probabilities are treated as random deviations from the same underlying value (denoted p) or, in symbols, $p_1 = p_2 = p$. The probability of a low-birth-weight infant estimated from the 2 × 2 table ignoring ethnic status is then $\hat{p} = 113/411 = 0.275$ (Table 6.4, last row). Using the estimate $\hat{p}$, the variance of the distribution of an observed difference $\hat{p}_1 - \hat{p}_2$ is estimated from the expression

$$variance(\hat{p}_1 - \hat{p}_2) = \hat{v} = \frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2} = \hat{p}(1-\hat{p})\left[\frac{1}{n_1} + \frac{1}{n_2}\right].$$

Table 6.3 *Notation:* 2 × 2 *Table Containing the Counts of n Joint Occurrences of Two Binary Variables Again Labeled D and E*

	D	$\bar{D}$	Total
E	a	b	n_1
$\bar{E}$	c	d	n_2
Total	m_1	m_2	n

Table 6.4 *Data:* 2 × 2 *Table of Counts of the Joint Occurrence of Ethnic Status and Birth Weight (Ibwt)*[a] among ($n = 411$) *Exceptionally Small Mothers* (<45 kg)

	Birth Weight		
	lbwt	$\overline{lbwt}$	Total
African American	$a = 61$	$b = 102$	$n_1 = 163$
White	$c = 52$	$d = 196$	$n_2 = 248$
Total	$m_1 = 113$	$m_2 = 298$	$n = 411$

[a] *lbwt* = Infant birth weight equal to or less than 2500 g.

The origin of this variance is the binomial distribution (Chapters 4 and 27). Specifically, for the example data, the estimated variance is

$$\hat{v} = 0.275(1 - 0.275)\left[\frac{1}{163} + \frac{1}{248}\right] = 0.00202$$

where $\hat{p} = 0.275$. The value of the test statistic

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{v}}} = \frac{(0.374 - 0.210) - 0}{\sqrt{0.00202}} = 3.655$$

has an approximate standard normal distribution when only random differences in low-birth-weight infants exist between African American and white infants. The associated p-value of $P(|Z| > 3.655 | p_1 = p_2 = p)$ is less than 0.001.

The identical result occurs from a comparison of the estimated conditional probabilities $P_1 = P(\text{African American} \mid lbwt)$ and $P_2 = P(\text{African American} \mid lbwt)$. The columns of the 2 × 2 table of birth weight counts yield estimates $\hat{P}_1 = 61/113 = 0.540$ and $\hat{P}_2 = 102/298 = 0.342$. An estimate of the variance of the observed difference $\hat{P}_1 - \hat{P}_2$ is

$$variance(\hat{P}_1 - \hat{P}_2) = \hat{V} = \frac{\hat{P}(1 - \hat{P})}{m_1} + \frac{\hat{P}(1 - \hat{P})}{m_2} = \hat{P}(1 - \hat{P})\left[\frac{1}{m_1} + \frac{1}{m_2}\right]$$

where $\hat{P} = 163/411 = 0.397$, ignoring infant birth weight status (Table 6.4, last column). Specifically, the estimate becomes

$$\hat{V} = 0.397(1 - 0.397)\left[\frac{1}{163} + \frac{1}{248}\right] = 0.00292.$$

The value of the test statistic is, again,

$$z = \frac{(\hat{P}_1 - \hat{P}_2) - 0}{\sqrt{\hat{V}}} = \frac{(0.540 - 0.342) - 0}{\sqrt{0.00292}} = 3.655.$$

Thus, comparison of two proportions from a 2 × 2 table produces identical results whether the compared conditional probabilities are estimated from the table rows or columns.

Another approach to the comparison of probabilities estimated from two groups employs the Pearson chi-square test statistic (Chapter 1). The first step is to create a theoretical 2 × 2 table. The single estimated probability of a low-birth-weight infant ignoring ethnicity

Table 6.5 *Conjecture:* 2 × 2 *Table of Theoretical Counts of Newborn Infants That Perfectly Conform to the Hypothesis* $p_1 = p_2 = p$[a]

	Birth weights		
	lbwt	$\overline{lbwt}$	Total
African American	$\hat{a} = pn_1 = 0.275(163) = 44.82$	$\hat{b} = (1-p)n_1 = 0.725(163) = 118.18$	163
White	$\hat{c} = pn_2 = 0.275(248) = 68.18$	$\hat{d} = (1-p)n_2 = 0.725(248) = 179.81$	248
Total	113	298	411

[a] $p_1 = p_2 = p = 163/411 = 0.275$.

$\hat{p} = 113/411 = 0.275$ produces a theoretical table that perfectly conforms to the conjecture of no difference exists between African American and white mothers in the likelihood of a low-birth-weight infant (Table 6.5). Specifically, these identical probabilities are $p_1 = 44.82/163 = p_2 = 68.18/248 = \hat{p} = 113/411 = 0.275$ (Table 6.5). A chi-square statistic to evaluate the conjecture that no difference exists between ethnic groups ($p_1 = p_2 = p$) is simply a comparison of four observed values to four theoretical values (Chapter 1). The chi-square test statistic is

$$X^2 = \sum (o_i - e_i)^2/e_i = 13.360 \qquad \text{where} \quad i = 1, 2, 3 \text{ and } 4.$$

The notation o_i represents an observed cell frequency from a 2 × 2 table (Table 6.4), and e_i represents the corresponding theoretical cell frequency (Table 6.5). The approximate chi-square distributed summary statistic X^2 (degrees of freedom = 1) yields a p-value of $P(X^2 > 13.360|\text{no difference}) < 0.001$.

A chi-square expression in a more explicit form (Table 6.4 versus Table 6.5)

$$\begin{aligned} X^2 &= (a-\hat{a})^2/\hat{a} + (b-\hat{b})^2/\hat{b} + (c-\hat{c})^2/\hat{c} + (d-\hat{d})^2/\hat{d} \\ &= (61-44.82)^2/44.82 + (102-118.18)^2/118.18 \\ &\quad + (52-68.18)^2/68.18 + (196-179.81)^2/179.81 = 13.360 \end{aligned}$$

is the same chi-square comparison of observed-to-postulated counts. A short-cut expression to calculate the identical chi-square test statistic is

$$X^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

where n represents the total number of observations (Table 6.1). For the example data (Table 6.4), the chi-square test statistic is again

$$X^2 = \frac{411[61(196) - 102(52)]^2}{(163)(248)(113)(298)} = 13.360.$$

The comparison of estimated conditional probabilities and the chi-square comparison of counts always yields identical results because $X^2 = z^2$ and, for the example, necessarily $(3.655)^2 = 13.360$ (Chapter 1).

It is important to note that two-sample comparison and chi-square approaches to evaluating an association in a 2 × 2 table are approximate. For tables containing one or more small

Table 6.6 *Notation: Six Possible Measures of Relationships within 2 × 2 Table and Expressions for Their Estimation*

	Measures	Symbols	Estimates[a]
Difference in probabilities (rows)[b]	$P(D\|E) - P(D\|\bar{E})$	$\hat{p}_1 - \hat{p}_2$	$\frac{a}{a+b} - \frac{c}{c+d}$
Difference in probabilities (columns)[b]	$P(E\|D) - P(E\|\bar{D})$	$\hat{P}_1 - \hat{P}_2$	$\frac{a}{a+c} - \frac{b}{b+d}$
Relative risk ratio[b]	$\frac{P(D\|E)}{P(D\|E)}$	$\hat{rr}$	$\frac{a/(a+b)}{c/(c+d)}$
Odds ratio (rows)	$\frac{P(D\|E)/P(\bar{D}\|E)}{P(D\|\bar{E})/P(\bar{D}\|\bar{E})}$	$\hat{or}$	$\frac{a/b}{c/d} = \frac{ad}{bc}$
Odds ratio (columns)	$\frac{P(E\|D)/P(\bar{E}\|D)}{P(E\|\bar{D})/P(\bar{E}\|\bar{D})}$	$\hat{or}$	$\frac{a/c}{b/d} = \frac{ad}{bc}$
Attributable risk	$\frac{P(D) - P(D\|\bar{E})}{P(D)}$	$\hat{ar}$	$\frac{ad - bc}{(a+c)(c+d)}$

[a]Notation from Table 6.1.
[b]As noted, both differences yield the identical test statistic.

cell frequencies techniques exist to improve the accuracy of this approximate analysis (to be discussed). In addition, extremely small p-values encountered as part of a computer analysis (such as p-value $= 4.2 \times 10^{-8}$), are equally approximate, and reporting a computer-calculated extremely small probability gives a false sense of accuracy. The best that can be said for these tiny computer-generated probabilities is that they are small or, perhaps very small and are more realistically reported as simply less than 0.001.

Measures of Association in a 2 × 2 Table

A variety of statistical measures describe relationships within a 2 × 2 table. The choice of measure primarily depends on the descriptive properties desired of the summary value because a chi-square analysis is typically adequate to evaluate the influence of random variation. In fact, it can be argued that it is theoretically the best statistic to identify an association. A proportion is a natural measure when two groups are compared. Nevertheless, persuasive reasons exist to describe the relationships within a 2 × 2 table with other measures of association such as an odds ratio or relative risk ratio or attributable risk percentage (Chapter 20).

Table 6.6 contains formal definitions of six common measures of association and expressions for estimation from counts contained in a 2 × 2 table. For a generic measure of association calculated to summarize a relationship within a 2 × 2 table, denoted g and its estimate $\hat{g}$, a statistical test and construction of an approximate confidence interval follow familiar patterns (Chapter 2). The test statistic

$$z = \frac{\hat{g} - g_0}{\sqrt{variance(\hat{g})}}$$

Table 6.7 *Estimates: Expressions to Estimate the Variance of the Distributions of Eight Measures of Association*

	Symbols	Variance estimates	*s. e.*[b]
Difference in proportions (rows)	$\hat{p}_1 - \hat{p}_2$	$\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$	0.046
Difference in proportions (columns)	$\hat{P}_1 - \hat{P}_2$	$\frac{\hat{P}_1(1-\hat{P}_1)}{m_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{m_2}$	0.054
Relative risk ratio	$\hat{rr}$	$\hat{rr}^2\left[\frac{1}{a} - \frac{1}{n_1} + \frac{1}{c} - \frac{1}{n_2}\right]$	0.285[c]
Logarithm of the relative risk ratio ratio[a]	$log(\hat{rr})$	$\frac{1}{a} - \frac{1}{n_1} + \frac{1}{c} - \frac{1}{n_2}$	0.160[c]
Odds ratio	$\hat{or}$	$\hat{or}^2\left[\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right]$	0.507
Logarithm of the odds ratio[a]	$log(\hat{or})$	$\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$	0.225
Attributable risk	$\hat{ar}$	$(1-\hat{ar})^2\left[\frac{b + \hat{ar}(a+d)}{nc}\right]$	0.067
Logarithm of the attributable risk[a]	$log(1-\hat{ar})$	$\frac{b + \hat{ar}(a+d)}{nc}$	0.087

[a]Transformed to increase the accuracy of a normal distribution approximation.
[b]Standard errors for the eight measures of association (birth-weight data: Table 6.4).
[c]Another version of the estimated variance excludes the terms $1/n_1$ and $1/n_2$.

usually has at least an approximate standard normal distribution when the value estimated by $\hat{g}$ is g_0. An approximate normal distribution generated 95% confidence interval, based on the estimate $\hat{g}$,

$$\hat{A} = lower\ bound = \hat{g} - 1.960\sqrt{variance(\hat{g})}$$

and

$$\hat{B} = upper\ bound = \hat{g} + 1.960\sqrt{variance(\hat{g})},$$

is also usually a useful description of the influence of random variation on the accuracy and precision of the estimated value, particularly when moderate sample sizes exist in all cells of the table (>20). Expressions to estimate the variances of the distributions of the estimates of g (Table 6.6) are given in Table 6.7 along with example estimates from the birth weight data (Table 6.4).

Table 6.8 contains the estimated values and their approximate 95% confidence intervals for each measure of association again applied to the low-birth-weight data (Table 6.4).

Although these summary variables appear unrelated, they are in fact different expressions of the same value, namely, $ad - bc$. This quantity is the essential underlying measure of the strength of association between two binary variables classified into a 2 × 2 table. When $ad = bc$, all differences are zero and all ratios are one.

Table 6.8 *Estimates: Approximate 95% Confidence Intervals for Eight Measures of Association Applied to Ethnic Group Status and Birth Weight Data (Table 6.4)*

	Symbols	Estimates	Lower	Upper
Difference in proportions (rows)[a]	$\hat{p}_1 - \hat{p}_2$	0.165	0.075	0.254
Difference in proportions (columns)[a]	$\hat{P}_1 - \hat{P}_2$	0.198	0.091	0.304
Relative risk ratio	$\hat{rr}$	1.785	1.306	2.440
Logarithm of the relative risk ratio	$log(\hat{rr})$	0.579	0.267	0.892
Odds ratio	$\hat{or}$	2.254	1.451	3.502
Logarithm of the odds ratio	$log(\hat{or})$	0.813	0.372	1.253
Attributable risk	$\hat{ar}$	0.237	0.095	0.357
Logarithm of the attributable risk	$log(1 - \hat{ar})$	−0.271	−0.442	−0.100

[a]Identical to the previous chi-square analysis, $X^2 = 13.360$.

Odds Ratio and Relative Risk Ratio

Central to describing binary variables classified into a 2 × 2 table are the odds ratio and the relative risk ratio measures of association. Unlike other measures of association, these two summary values can have similar roles in describing a relationship within a 2 × 2 table, particularly in the case of assessing epidemiologic data.

Expressions for the variance of an estimated value often reveal important properties of the estimate. Estimated variances of the odds ratio, relative risk ratio, and logarithms of these measures of an association are functions of the reciprocal counts from the four values contained in a 2 × 2 table (Table 6.7). For example, the variance of the logarithm of an estimated odds ratio is

$$variance[log(\hat{or})] = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \quad \text{(Chapter 27)}.$$

This expression identifies the fact that small cell frequencies have dominant influences on the precision of the estimate regardless of the other counts in the table.

For example, consider data from a study of female military veterans (Table 6.9) contrasting breast cancer risk between women who served in Vietnam ($n_1 = 3392$) to women who did not served in Vietnam ($n_2 = 3038$). The estimated odds ratio is $\hat{or} = 1.219$, and the estimated relative risk ratio is $\hat{rr} = 1.208$.

The estimated variance of the distribution of the logarithm of the estimated odds ratio is $variance[\log(\hat{or})] = 0.0145$ and is largely determined by the two values $a = 170$ and $c = 126$ where $1/a + 1/c = 1/170 + 1/126 = 0.0138$. Therefore, the effective sample size is more

Table 6.9 *Data: Breast Cancer (D) Diagnosed among Military Women Who Served in Vietnam (E) and Who Did Not Serve in Vietnam ($\bar{E}$) during the War Years 1965–1973*

	D	$\bar{D}$	Total
E	170	3222	3392
$\bar{E}$	126	2912	3038
Total	296	6134	6430

realistically about $170 + 126 \approx 296$ rather than 6430. The precision of the estimated relative risk ratio is similarly dominated by cells with small numbers of observations (Table 6.7). The large influences from small cell frequencies is the nature of these two measures of association, and awareness of this property is an important element in their interpretation.

The relative risk ratio, as the name suggests, estimates risk from a 2×2 table. The odds ratio always measures association from a 2×2 table. When an outcome under study is rare, the estimated odds ratio becomes approximately equal to the estimated relative risk ratio. The odds ratio can then be treated as an approximate measure of risk. Simply, when the two values calculated from the same table are close to equal, the odds ratio and the relative risk ratio have similar properties. Comparison of the odds ratio and the relative risk ratio measures calculated from the Vietnam data (Table 6.9) illustrate. Because breast cancer is a rare disease, the odds ratio can be treated as an approximate measure of risk.

The odds ratio and the relative risk ratio statistics are similar when $a \ll b$ and $c \ll d$ making $a + b \approx b$ and $c + d \approx d$; then

$$\hat{rr} = \frac{a/(a+b)}{c/(c+d)} \approx \frac{a/b}{c/d} = \hat{or}.$$

From the breast cancer data,

$$\hat{rr} = \frac{170/3392}{126/3038} = 1.208 \approx \hat{or} = \frac{170/3222}{126/2912} = 1.219.$$

Parenthetically, as illustrated by the example, when $\hat{rr}$ is greater than 1.0, the estimate $\hat{or}$ is always greater than $\hat{rr}$, and, similarly, when $\hat{rr}$ is less than 1.0, the estimate $\hat{or}$ is always less than $\hat{rr}$.

An additive scale is second nature. For example, the symmetric role of a value b in $a + b$ and $a - b$ is clear and intuitive. The value $-b$ simply represents the same difference but in the opposite direction on the additive scale. The symmetric relationship for values expressed as a ratio is less intuitive. For the ratio $r = a/b$, the symmetric value is $1/r = b/a$. Like the additive scale, the reciprocal indicates the same difference but in the opposite direction. For example, the ratio of male to female Hodgkin's disease mortality rates is $37.8/18.4 = 2.05$. The same relationship is equally described as the ratio of female to male mortality rates as $18.4/37.8 = 1/2.05 = 0.49$. That is, males have twice the female risk or females have half the male risk.

Ratios are compared on a ratio scale. This rule is sometimes forgotten. A ratio such as 0.5 is occasionally expressed as a percentage decrease or

$$\text{percentage decrease} = \frac{1.0 - 0.5}{1.0} \times 100 = 50\%.$$

The arithmetic is correct but the resulting value is not relevant. The ratio 0.5 indicates a decrease by a factor of two (it is said to be a "200% decrease"). In a recent paper entitled "Screening and Prostate-Cancer Mortality in a Randomized European Study" published in the *New England of Medicine* (2009), a conclusion stated, "The rate ratio for death from prostate cancer in the screening group, as compared to the control group, was 0.80 . . . PSA-based testing reduced the rate of death from prostate cancer by 20%." The correct decrease $1/0.80 = 1.25$ relative to the controls. That is, screening reduced the risk of death by a factor of 1.25 among individuals screened compared to controls. When a ratio is in the neighborhood

of 1.0, using an additive scale to describe a ratio produces a similar but technically incorrect value.

An advertisement and a brochure by the March of Dimes read as follows:

> Debbie and Rich Hedding of Pittsford Vt. were devastated when they lost two babies to neural tube defects (NTDs). When Debbie read about folic acid in a March of Dines brochure, she was astonished when learn about the role of folic acid in preventing NTDs; "I was in tears by the time I finished reading the material. I haven't been taking folic acid nor had I been told about it. I couldn't believe that I could have reduced the risk of recurrence by 70 percent. I immediately began taking folic acid and telling every women I could about it."

The odds ratio $\hat{or} = 0.28$ (*Lancet*, 1998) was the source of the value "70 percent" where $(1.0 - 0.28)/1.0 \times 100$ is approximately 70%. If Debbie had done the calculation correctly, she would have been upset to a much greater degree. The risk is $1/0.28 = 3.6$ times less among women using folic acid supplementation.

A Correction to Improve the Normal Distribution Approximation

A comparison of proportions from a 2 × 2 table, like many statistical techniques, frequently employs a normal distribution to estimate approximate significance levels (p-values) (Chapter 4). The accuracy of the approximation is easily improved, which is especially important when at least one category of a 2 × 2 table contains a small number of observations.

A simple example conveys the essence of a commonly used strategy to improve the accuracy of a normally distributed test statistic. Consider a variable labeled X that has the discrete probability distribution:

X	0	1	2	3	4	Total
Probabilities	1/16	4/16	6/16	4/16	1/16	1.0

The exact probability $P(X \geq c)$ can be directly calculated for a chosen value represented by c or calculated from a normal distribution used to approximate the discrete distribution probabilities (Figure 6.1). For the example distribution, the exact probability directly calculated for $c = 2.0$ is $P(X \geq 2.0) = 6/16 + 4/16 + 1/16 = 11/16 = 0.688$. The example probability distribution of X has a mean value $\mu = 2.0$ and a variance $\sigma^2 = 1.0$ (Chapter 27). Using the normal distribution and these two parameter values, an approximate calculation yields

$$z = \frac{c - \mu}{\sigma} = \frac{2.0 - 2.0}{1.0} = 0 \quad \text{and} \quad P(X \geq 2) = P(Z \geq 0) = 0.5.$$

A much improved estimate is achieved by the mechanical process of subtracting half the discrete distribution interval length from the value of c; that is, instead of applying the normal distribution to $P(X > c)$, the expression used is $P(X \geq c - \frac{1}{2}$ [length of the interval]) or, for the example, then $P(X \geq 1.5)$. Continuing the example, again for $c = 2.0$,

$$z_c = \frac{(c - 0.5) - \mu}{\sigma} = \frac{(2.0 - 0.5 - 2.0)}{1.0} = \frac{1.5 - 2.0}{1.0} = -0.5.$$

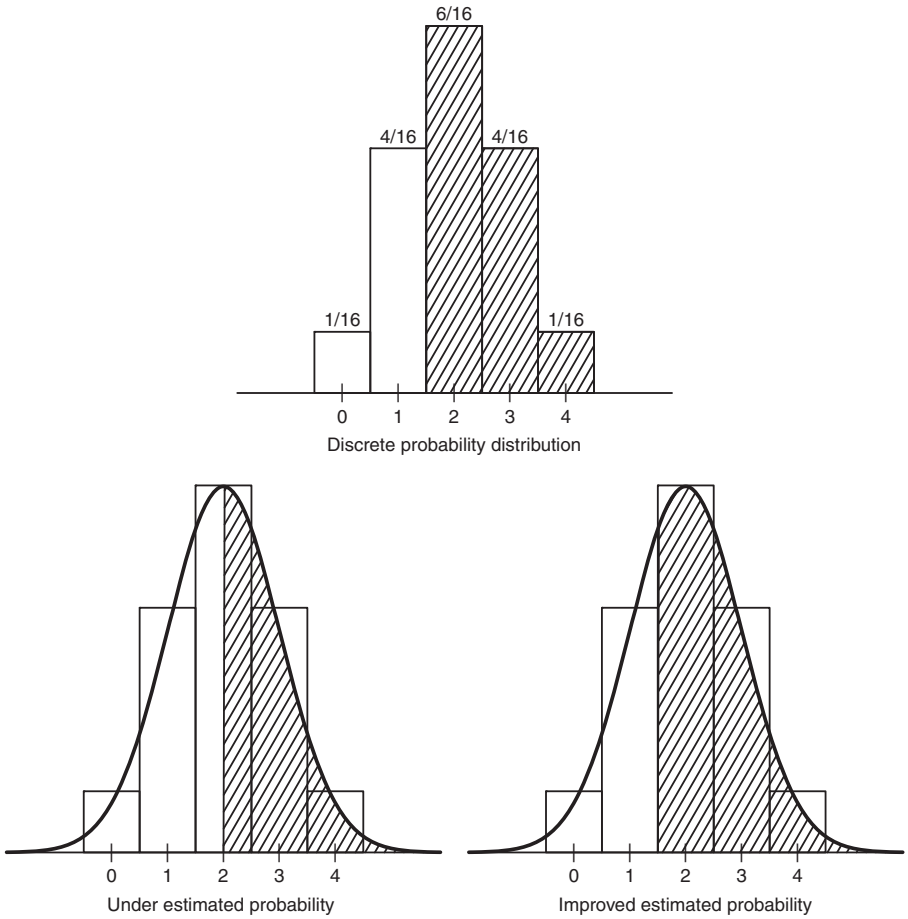


Figure 6.1 Continuous Normal Distribution as Approximation of a Discrete Probability Distribution

The exact probability of 0.688 is now more accurately approximated by the probability $P(X \geq 1.5) = P(Z \geq -0.5) = 0.691$ derived from the continuous normal distribution. The correction factor simply makes the approximate probability of the normal distribution (area) more closely correspond to the exact probability (Figure 6.1, far right).

In the important case of binomial probabilities, the correction factor is again half the interval length (Chapter 4). Specifically, for a binomial distributed value with parameters n and p, the normal distribution approximation used to evaluate an estimate $\hat{p}$ is improved by using where the interval length is $1/n$ ("+" for the left tail and "−" for the right tail probabilities). The continuous standard normal cumulative probability $P(Z \leq z_c)$ is then a more accurate estimate of the exact discrete binomial cumulative probability. As noted, this correction is most important when the sample size is small (Chapter 2). For larger sample sizes, the term $\frac{1}{2}(1/n)$ has negligible influence.

$$z_c = \frac{\hat{p} - p \pm \frac{1}{2}\left[\frac{1}{n}\right]}{\sqrt{p(1-p)/n}}$$

Table 6.10 *Data: A 2 × 2 Table Description of the Joint Occurrence of Case/Control Status and Presence or Absence of B-Type Allele from a Study of a Rare Subtype of Leukemia*

	Alleles		
	B	$\bar{B}$	Total
Cases	2	23	25
Controls	6	22	28
Total	8	45	53

The application to a 2 × 2 table follows the same pattern. The correction factor is

$$\text{correction factor} = \frac{1}{2}\left[\frac{1}{n_1} + \frac{1}{n_2}\right]$$

where again n_1 and n_2 are sample sizes associated with estimated probabilities $\hat{p}_1$ and $\hat{p}_2$. Therefore, for the normal distribution approximation used to make a two-sample comparison of $\hat{p}_1$and $\hat{p}_2$, the test statistic with improved accuracy becomes

$$z_c = \frac{|\hat{p}_1 - \hat{p}_2| \pm \frac{1}{2}\left[\frac{1}{n_1} + \frac{1}{n_2}\right] - 0}{\sqrt{variance(\hat{p}_1 - \hat{p}_2)}}$$

After a series of algebraic manipulations, the short-cut chi-square test statistic employing the correction is

$$X_c^2 = \frac{n(|ad - bc| - n/2)^2}{(a+b)(c+d)(a+c)(b+d)}.$$

The two corrected test statistics $z_c^2 = X_c^2$ are identical.

From a study of childhood leukemia, children are classified by case/control status and by the present or absence of a specific single nucleotide polymorphic allele (labeled B and $\bar{B}$; Table 6.10).

The uncorrected chi-square value is $X^2 = 1.858$ yielding a p-value of 0.173, and the more accurate corrected value is $X_c^2 = 0.958$ yielding a p-value of 0.328.

The Hypergeometric Probability Distribution

The famous statistician R. A. Fisher was introduced to a lady who said she could tell the difference between a cup of tea when tea was added to the milk or the milk was added to the tea. A skeptical R.A. Fisher proposed an experiment. Four cups tea were prepared with milk added to the tea and four cups with tea added to the milk. Upon tasting each cup, the tea-tasting lady identified all eight cups correctly. Two possibilities exist. She cannot tell one cup of tea from another and was lucky, or she has a rare ability to discriminate an extremely subtle difference in tea preparation. Fisher suggested that to evaluate these two possibilities it is useful to know the probability that she identified the eight cups of tea correctly by chance alone. The probability is $(4/8)\ (3/7)\ (2/6)\ (1/5) = 1/70 = 0.014$.

Table 6.11 *Notation: All Possible Outcomes from the Lady Tasting Tea Experiment (Symbol* a *Represents Number Correct) Displayed in* 2×2 *Table* ($a = 0, 1, 2, 3,$ *and* 4)

	Tea/milk	Milk/tea	Total
"Tea/milk"	A	$4 - a$	4
"Milk/tea"	$4 - a$	a	4
Total	4	4	8

The probability 0.014 is a single value from a *hypergeometric probability distribution*. The probability of each outcome of the tea-tasting experiment can be calculated from the expression

$$P(\text{number correct} = a) = \frac{\binom{4}{a}\binom{4}{4-a}}{\binom{8}{4}}$$

where a represents the number of correct determinations. Five probabilities create the probability distribution when the observed outcomes are due to chance alone (Tables 6.11 and 6.12).

In general, for a 2×2 table with fixed marginal values describing two unrelated binary variables, the same hypergeometric probabilities are given by either of two expressions

$$P(\text{outcome} = a) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} \quad \text{(rows)}$$

or

$$P(\text{outcome} = a) = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{n}{a+b}} \quad \text{(columns)},$$

Table 6.12 *Calculations: All Possible Outcomes from the Lady Tasting Tea Experiment – Hypergeometric Probability Distribution (Five* 2×2 *Tables)*

	2×2 tables				
	a	$4 - a$	$4 - a$	a	Probabilities
None correct	0	4	4	0	0.014
One correct	1	3	3	1	0.229
Two correct	2	2	2	2	0.514
Three correct	3	1	1	3	0.229
Four correct	4	0	0	4	0.014

Table 6.13 *Notation: The Outcomes[a] of Keno Displayed in a 2 × 2 Table*

	Casino selects	Not selected	
Player selects	X	$8 - x$	8
Not selected	$20 - x$	$52 + x$	72
Total	20	60	80

[a] x represents the number of values chosen by the player and selected by the casino.

where $a + b + c + d = n$ possible outcomes (Table 6.3). Thus, estimated hypergeometric probabilities are identical whether calculated from the rows or from the columns of a 2×2 table. Like all discrete probabilities distributions, the sum of all possible hypergeometric probabilities is 1.0.

A more extensive example of a hypergeometric distribution comes from the casino gambling game called Keno. A player chooses eight numbers from 1 to 80, and the casino selects 20 random numbers from 1 to 80. The player wins if five or more of the eight chosen numbers are among the 20 random casino-selected numbers. A large jackpot goes to a player whose eight chosen numbers are among the 20 casino selected numbers. Table 6.13 describes the details of the Keno game in terms of possible outcomes classified into a 2×2 table.

The likelihood of each Keno game outcome is described by a hypergeometric probability because no relationship exists between the player's choice of eight numbers and the casino's random selection of 20 numbers, and marginal frequencies 8 and 20 are fixed. The probabilities of all nine possible Keno outcomes create a hypergeometric probability distribution (Table 6.14). It is extremely unlikely that a player chooses eight matching numbers (0.000004), and, in addition, it is also somewhat unlikely that no matching numbers are chosen (0.088). The probability of winning anything is $P(X \geq 5) = 0.018303 + 0.002367 + 0.000160 + 0.000004 = 0.021$.

Furthermore, the expected number of matching values is $EX = \sum p_x x = 2.0$.

Table 6.14 *Computations: Probabilities for the Nine Possible Keno Outcomes, a Hypergeometric Distribution*

Matches	a	b	c	D	Probability
None	0	8	20	52	0.088266
One	1	7	19	53	0.266464
Two	2	6	18	54	0.328146
Three	3	5	17	55	0.214786
Four	4	4	16	56	0.081504
Five	5	3	15	57	0.018303
Six	6	2	14	58	0.002367
Seven	7	1	13	59	0.000160
Eight	8	0	12	60	0.000004

Table 6.15 *Data: Lung Cancer Cases and Controls Classified by Country of Birth (n = 28 Hispanic Nonsmoking Men)*

	Cases	Controls	
U.S.-born	5	3	8
Not U.S.-born	3	17	20
Total	8	20	28

Fisher's Exact Test

Fisher's exact test occasionally applied to evaluate an association in a 2×2 table is a direct application of the hypergeometric probability distribution, which is particularly useful when the sample size is small. When two binary variables summarized in a 2×2 table are not related and the marginal frequencies are fixed values, the probabilities associated with each possible outcome create a hypergeometric probability distribution.

A small sample consisting of $n = 28$ Hispanic nonsmoking men from a case/control study of lung cancer illustrates. Cases of lung cancer and controls are classified by U.S.-born and not U.S.-born subjects (Table 6.15).

All nine possible probabilities calculated under the conjecture that country of birth and the risk of lung cancer are unrelated create a hypergeometric probability distribution (Table 6.16).

The probability associated with the observed outcome of five U.S.-born lung cancer cases is $P(cases = 5) = 0.0205$ when case/control status is unrelated to country of birth (Table 6.16). The probability of a more extreme result, like the tea-tasting experiment, measures the likelihood that the observed result is entirely due to chance and almost always is called a *p*-value in a statistical analysis context. This probability directly calculated from a hypergeometric probability distribution is usually referred to as *Fisher's exact test*. Specifically, when case/control status is unrelated to country of birth, the probability of five or more U.S.-born lung cancer cases is $p\text{-value} = P(cases \geq 5) = 0.0205 + 0.0017 + 0.0001 = 0.0223$ (Table 6.16).

Table 6.16 *Computations: Hypergeometric Probability Distribution Associated with Cases of Lung Cancer and Country of Birth from Case/Control Study of n = 28 Hispanic Nonsmoking Men*

Cases	*a*	*b*	*c*	*d*	Probability
None	0	8	8	12	0.0405
One	1	7	7	13	0.1995
Two	2	6	6	14	0.3492
Three	3	5	5	15	0.2793
Four	4	4	4	16	0.1091
Five	5	3	3	17	0.0205
Six	6	2	2	18	0.0017
Seven	7	1	1	19	0.0001
Eight	8	0	0	20	0.0000

Table 6.17 *Notation: Description of Two Binary Variables with Same Distribution (Labeled X and X′) Creating Four Outcomes*

	$X = 1$	$X = 0$	Total
$X' = 1$	$p^2 + pqr$	$pq - pqr$	p
$X' = 0$	$pq - pqr$	$q^2 + pqr$	q
Total	p	q	1.0

For comparison, the corresponding approximate chi-square statistics are: unadjusted chi-square $X^2 = 6.318$ (p-value $= 0.012$) and adjusted chi-square $X_c^2 = 4.204$ (p-value $= 0.040$).

As noted, the Pearson chi-square test is not likely accurate for tables with small numbers of observations. In this instance, Fisher's exact test is frequently applied to 2 × 2 tables containing one or more small cell frequencies. Hypergeometric probabilities, however, are formally exact only when the marginal frequencies are fixed, which is rarely the case for sampled data. Furthermore, debate exists over the best method to analyze a 2 × 2 table containing small cell frequencies. Alternative statistical techniques exact for small sample sizes (no approximations) exist that employ computer algorithms or nonparametric or bootstrap evaluations (Chapters 11 and 15).

Correlation in a 2 × 2 Table

A 2 × 2 table efficiently summarizes counts of pairs of values each sampled from the same binary distribution (denoted X and X') characterized as *identically distributed* (Chapter 26). A specific statistical structure describes the extent of association within these pairs. The analysis of identically distributed binary variables is not fundamentally different in principle from the general description of a 2 × 2 table. Nevertheless, a close look at such a 2 × 2 table provides additional insight into its properties.

Identically distributed binary variables classified into a 2 × 2 table occur in a number of contexts. Three examples are the following:

Twins: From twin births data, the sex of one twin is indicated by a binary variable ($X =$ male $= 1$ or $X =$ female $= 0$), and the sex of the other twin is indicated by a second binary variable ($X' =$ male $= 1$ or $X' =$ female $= 0$), and the association within twin pairs yields an estimate of zygosity (monozygotic or dizygotic).

Genetics: From genotype data, the presence or absence of one allele represented by a binary variable (X), and the present or absence of a second allele represented by another binary variable (X') yields an estimate of the extent of association within pairs of alleles (admixed or independent).

Inbreeding: From genetic data containing matings between relatives, pairs of binary variables (X and X') are identical because they originate from relatives or are randomly identical and inherited from a general gene pool yields an estimate of the degree of inbreeding (familial or random).

Several expressions for the relationship between two binary variables exist to describe association (Chapter 26). Table 6.17 displays one of these relationships emphasizing the deviation

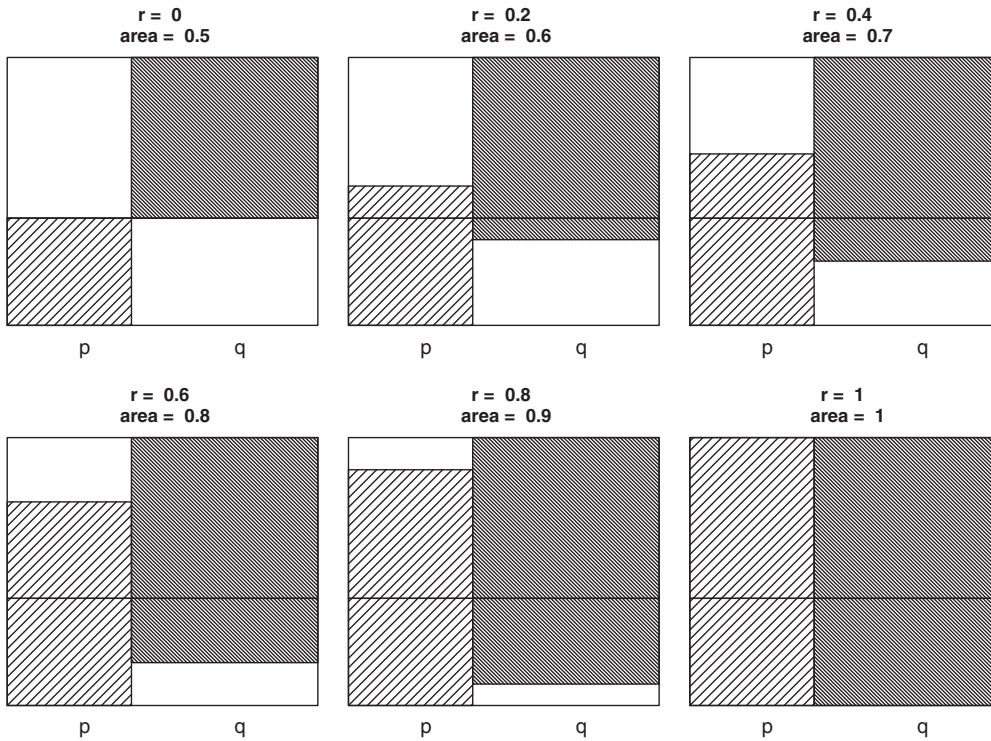


Figure 6.2 Schematic Representation of Role of Correlation Coefficient r ($p = 0.4$)

from random association where $p = P(X = 1) = P(X' = 1)$ and $q = P(X = 0) = P(X' = 0)$ and r measures the correlation within pairs.

Specifically, the term pqr directly indicates the degree of deviation from purely random association (Chapter 27). Thus, when $r = 0$, the binary variables are exactly unrelated and the counted pairs occur entirely by chance alone. When $r = 1$, all pairs are either (1, 1) or (0, 0), said to be concordant. In addition, the parameter represented by r is the usual Pearson product-moment correlation coefficient directly calculated from the binary data that make up pairs consisting of zero or one observation (Chapters 5 and 21).

A schematic representation (Figure 6.2) of the role of the correlation coefficient r is displayed for $p = 0.4$ and six selected values of r ($0 \leq r \leq 1$). Association is indicated by the area of the upper left and lower right rectangles ($area = pqr$). This area increases as r increases.

From another point of view, note that $p^2 + pqr = pr + p^2(1 - r)$ and similarly $q^2 + pqr = qr + q^2(1 - r)$. Thus, the value r represents the proportion of nonrandom pairs identical for the outcome among concordant pairs. For example, the correlation r is the proportion of concordant and identical twin pairs (monozygotic twins), and $1 - r$ is the proportion of concordant and randomly associated twin pairs (dizygotic twins). Thus, binary data can made up of three kinds of pairs: concordant pairs that are always identical with frequency r, concordant pairs that are randomly identical with frequency $(1 - r)(p^2 + q^2)$, and pairs that are not concordant with frequency $2pq(1 - r)$. Table 6.18 displays the notation for data classified by three kinds of pairs.

Table 6.18 *Notation: Distribution of Data Produced by Identically Distributed Binary Variables X and X′*

	XX	XX'	$X'X'$	Total
	(1, 1)	(0, 1) or (1, 0)	(0, 0)	
Counts	a	b	c	N

Estimates of the parameter values p and r (Tables 6.17 and 6.18) are

$$\hat{P}(X=1)=\hat{P}(X'=1)=\hat{P}=\frac{2a+b}{2n} \quad \text{and} \quad \textit{correlation}(X,X')=\hat{r}=1-\frac{b}{2n\hat{p}\hat{q}}.$$

The estimate $\hat{p}=1-\hat{q}$ is the total count of the number of values where X equals 1 $(2a+b)$ divided by the total number of observed values $(2n)$ (Chapter 23). The correlation coefficient r is estimated by equating observed data values to theoretical parameter values, called the *method of moments estimation* (Chapter 27). For example, one version is

$$\frac{b}{n}=2pq-2pqr, \quad \text{therefore} \quad \hat{r}=1-\frac{b}{2n\hat{p}\hat{q}}$$

and another is

$$\frac{a}{n}=p^2+pqr, \quad \text{therefore} \quad \hat{r}=\frac{\frac{a}{n}-\hat{p}^2}{\hat{p}\hat{q}}.$$

The estimated value $\hat{r}$ is the same maximum likelihood estimate (Chapter 27) for both cases.

Again, the essential measure of strength of association in a 2 × 2 table is the underlying relationship $ad-bc$. For pairs of binary data, this relationship simply relates to the estimated correlation coefficient r where

$$ad-bc=ad-b^2/4=n^2[(\hat{p}^2+\hat{p}\hat{q}\hat{r})(\hat{q}^2+\hat{p}\hat{q}\hat{r})-(\hat{p}\hat{q}+\hat{p}\hat{q}\hat{r})^2]=n^2\hat{p}\hat{q}\hat{r}.$$

Therefore, an alternative expression for the estimation of r is

$$\hat{r}=\frac{ad-b^2/4}{n^2\hat{p}\hat{q}}=\frac{ad-b^2/4}{(a+b/2)(c+b/2)}.$$

To illustrate the estimation of the correlation coefficient from identically distributed binary data, twin infants are classified by the sex of the pairs (male, male = MM, female, male = MF and female, female = FF) and by the presence or absence of a cardiac birth defect (Table 6.19).

Table 6.19 *Data: Distribution of Twin Pairs by Sex with and without Observed Cardiac Birth Defects* ($n=54{,}299$, *California*, 1983–2003)

	Twins – birth defects			
	MM	*MF*	*FF*	Total
Birth defects	$a=168$	$b=53$	$c=110$	331
No birth defects	$a=18{,}687$	$b=16{,}093$	$c=19{,}188$	53,968

Table 6.20 *Data: Duffy Genotype Frequencies from Sample of n = 489 African Americans*

	Genotypes – duffy red blood cell types			
	Fy^aFy^a	Fy^aFy^b	Fy^bFy^b	Total
Observed frequency no association ($r = 0$)	$a = 8$	$b = 72$	$c = 409$	489
	$\hat{a} = 3.959$	$\hat{b} = 80.082$	$\hat{c} = 404.959$	489

The estimated probability of a male twin is $\hat{p} = 0.495$, and the estimated probability of a monozygotic pair is $\hat{r} = 0.404$ among twins without a birth defect. The same estimates from twin pairs with a birth defect are $\hat{p} = 0.588$ and $\hat{r} = 0.670$ (Table 6.19). The probability of monozygotic twins (nonrandom same sex pairs) is estimated by $\hat{r}$. Furthermore, the variances of the distributions of $\hat{p}$ and $\hat{r}$ are estimated from the expressions

$$variance(\hat{p}) = \frac{\hat{p}\hat{q}(1+\hat{r})}{2n} \quad \text{and}$$

$$variance(\hat{r}) = (1-\hat{r})\frac{1-2\hat{p}\hat{q}(1-\hat{r})-(1-4\hat{p}\hat{q})(1-\hat{r})^2}{2n\hat{p}\hat{q}}.$$

Specifically, these estimates are $variance(\hat{p}) = 0.00122$ and $variance(\hat{r}) = 0.00172$ for twin pairs with birth defects (Table 6.19, first row).

For the $n = 331$ twin pairs, a confidence interval indicates the influence from random variation, particularly for the key measure of association, the estimated correlation coefficient $\hat{r}$. The bounds of an approximate 95% confidence interval, based on the estimated proportion of monozygotic twins $\hat{r} = 0.670$, are

$$\hat{A} = lower\ bound = 0.670 - 1.960(0.042) = 0.588$$

and

$$\hat{B} = upper\ bound = 0.670 + 1.960(0.042) = 0.751.$$

The estimated confidence interval clearly indicates that the probability of monozygotic twins among the twins with birth defects ($\hat{r} = 0.670$) is almost certainly systematically larger than the same value calculated from twins without birth defects where ($\hat{r} = 0.404$). The 95% confidence interval (0.588, 0.751) clearly excludes the estimate $\hat{r} = 0.404$ (Chapter 2).

Another example of analysis of identically distributed binary variables is an assessment of the association between pairs of alleles within genotypes. Consider a description of red cell antigens creating genotypes by simple Mendelian inheritance called the *Duffy system* (Chapters 25 and 27). The frequencies from a sample of $n = 489$ African Americans produce an estimate of the association between alleles in the sampled population (Table 6.20).

The estimated Fy^a-allele frequency is $\hat{p} = 0.090$ and the estimated correlation between the alleles is $\hat{r} = 0.101$. The genotype frequencies estimated as if no association exists (as if $r = 0$) are given in Table 6.20. For example, the estimated number of the genotypes (Fy^a, Fy^a) is $\hat{a} = n\hat{p}^2 = 489(0.090)^2 = 3.959$. A typical chi-square comparison of observed to theoretical counts directly applies. The general expression of a chi-square assessment of association measured by the estimated correlation coefficient $\hat{r}$ is particularly simple in the case of these

two binary variables and is

$$X^2 = n\hat{r}^2.$$

More specifically, when $r = 0$, the estimated variance of $\hat{r}$ is *variance* $(\hat{r}) = 1/n$. This variance is the previous expression of the variance of $\hat{r}$ when $r = 0$. For a 2×2 table, the chi-square test statistic is then

$$X^2 = z^2 = \left[\frac{\hat{r} - 0}{\sqrt{1/n}}\right]^2 = n\hat{r}^2.$$

Therefore, a measure of the evidence of nonrandom assortment between alleles generated from the Duffy antigen system data (conjecture: correlation $= r = 0$) is given by

$$\text{chi-square test statistic} = X^2 = n\hat{r}^2 = 489(0.101)^2 = 4.980.$$

The test statistic X^2, as noted, has an approximate chi-square distribution (degrees of freedom $= 1$) when no association exists. The p-value associated with this conjecture is $P(X_2 \geq 4.980 | r = 0) = 0.026$. Thus, the sample of genetic data indicates a likely presence of admixture disequilibrium among African Americans (Chapter 25). Technically, the analysis yields evidence that the within-genotype correlation between pairs of alleles, estimated by $\hat{r} = 0.101$, is not likely a random deviation from zero ($r = 0$). The Pearson chi-square comparison of the observed values (a, b, and c) to the theoretical values ($\hat{a}$, $\hat{b}$, and $\hat{c}$) yields the identical result (Table 6.20).

A 2 × 2 Table with a Structural Zero

Another special case of a 2×2 table occurs when collected data are classified into three of four possible cells of a table and the fourth cell is empty because membership in one of the categories is not possible, called a *structural zero*. A random zero sooner or later disappears when the sample size is large enough. A structural zero remains zero regardless of the sample size. For example, when twin pairs are classified by sex as concordant or discordant pairs and by monozygotic and dizygotic status, no discordant monozygotic twin pairs are possible.

Consider a more general situation. It was important to estimate the number of vaccinated infants in a county in Iowa. Two sources of data are available to estimate this number. The first source (source 1) is routinely collected county medical records. The second source (source 2) is a special county-wide survey of parents with newborn infants. When two lists are used to determine the total number of vaccinated infants, the number of infants on neither list is not possibly known, producing a structural zero. The notation for two-list data displayed in a 2×2 table with a structural zero is given in Table 6.21 (Chapter 9).

To estimate the total number of vaccinated infants (denoted N), the assumption is made that the two sources of data are independent (Table 6.22). The number of infants not included on either list (denoted d) then can be estimated. Specifically, this estimate is

$$\begin{aligned}\hat{d} &= N(1 - \hat{p}_1)(1 - \hat{p}_2) = \left[\frac{N\hat{p}_1\hat{p}_2}{N\hat{p}_1\hat{p}_2}\right] \times N(1 - \hat{p}_1)(1 - \hat{p}_2) \\ &= \frac{[N\hat{p}_1(1 - \hat{p}_2)]\,[N(1 - \hat{p}_1)\hat{p}_2]}{N\hat{p}_1\hat{p}_2} = \frac{bc}{a}.\end{aligned}$$

Table 6.21 *Notation: Two-List Counts to Estimate Population Size N*

		Source 2		
		Present	Absent	Total
Source 1	Present	a	b	$a+b$
	Absent	c	$d=?$	?
	Total	$a+c$	?	?

The estimated population size becomes

$$\hat{N} = a + b + c + \hat{d} = a + b + c + \frac{bc}{a} = \frac{(a+b)(a+c)}{a}.$$

Specifically, for the data from Iowa (Table 6.23), the number of infants included on both lists is $a = 33$, on only the list from the community survey (source 1) it is $b = 44$, and on only the medical records (source 2) it is $c = 14$.

These data yield an estimate of $\hat{d} = bc/a = 44(14)/33 = 18.667$ uncounted infants. The county total is then estimated as $\hat{N} = [33 + 44 + 14] + 18.667 = 109.667$. The same estimate in a different form produces the identical count of

$$\hat{N} = \frac{(a+b)(a+c)}{a} = \frac{(44+33)(33+14)}{33} = \frac{(77)(47)}{33} = 109.667.$$

An alternative approach yields the same estimate of population size N using regression techniques (Chapter 9).

Assessing the Accuracy of a Diagnostic Test

Another special case of a 2×2 table occurs when binary data are collected to evaluate the performance of a test used to identify a specific binary outcome such as the presence or absence of a disease (Table 6.24).

The language, notation and issues are different in the context of assessing classification accuracy but the structure and properties of the 2×2 table are essentially the same as previously discussed.

Table 6.22 *Notation: Model of Two Independent Data Sources to Estimate Population Size Where p_1 Represents Probability of Inclusion from Source 1 and p_2 Represents Independent Probability of Inclusion from Source 2*

Two-list model				
		Source 1		
		Present	Absent	Total
Source 2	Present	$Np_1\,p_2$	$Np_1(1-p_2)$	Np_1
	Absent	$N(1-p_1)\,p_2$	$N(1-p_1)(1-p_2)$	$N(1-p_1)$
	Total	Np_2	$N(1-p_2)$	N

Table 6.23 *Data: Counts to Estimate Number of Vaccinated Infants in a County in Iowa from Two Sources*

		Source 2		
		Present	Absent	Total
Source 1	Present	$a = 33$	$b = 44$	$a + b = 77$
	Absent	$c = 14$	$d = ?$	?
	Total	$a + c = 47$	?	?

Three conditional probabilities describe the issues central to the application of a 2 × 2 table to assess classification accuracy and are given special names:

1. Probability $P(T^+|D)$ is called *sensitivity*

Sensitivity is defined as the conditional probability that a test identifies an individual as positive for a disease among those who have the disease (Table 6.24, column 1), estimated by $a/(a + c)$.

2. Probability $P(T^-|\bar{D})$ is called *specificity*

Specificity is defined as the conditional probability that a test identifies an individual as negative for a disease among those who do not have the disease (Table 6.24, column 2), estimated by $d/(b + d)$.

3. Probability $P(\bar{D}|T^+)$ is called *false positive*

False positive is defined as the conditional probability that an individual who does not have the disease is among those who tested positive (Table 6.24, row 1), estimated by $b/(a + b)$.

For example, when athletes are tested for illegal drug use after a competition, high sensitivity means that drug users will likely be correctly identified. Similarly, high specificity means that nondrug users will likely be correctly identified. A false positive determination occurs when an athlete who is a nonuser is among those with a positive test. These conditional probabilities are also commonly used to describe classification accuracy in terms of a geometric curve (Chapter 22).

Table 6.24 *Notation: 2 × 2 Table Describing Joint Distribution of a Positive Test* (T^+) *and a Negative Test* (T^-) *to Identify Presence* (D) *or Absence* ($\bar{D}$) *of a Disease*

	Disease		
Test	D	$\bar{D}$	Total
T^+	$a = P(DT^+)$	$b = P(\bar{D}T^+)$	$P(T^+)$
T^-	$c = P(DT^-)$	$d = P(\bar{D}T^-)$	$P(T^-)$
Total	$P(D)$	$P(\bar{D})$	1.0

Table 6.25 *Computation: Probabilities of a False Positive Test* $= P(\bar{D}|T^+)$ *with Sensitivity* $= 0.95 = P(T^+|D)$ *and Specificity* $= 0.90 = P(T^-|\bar{D})$ *for Five Selected Probabilities of Disease,* $P(D)$

$P(D)$	$P(T^+)$	$P(\bar{D}\|T^+)$
0.50	0.525	0.095
0.10	0.185	0.486
0.05	0.143	0.667
0.005	0.104	0.954
0.0005	0.100	0.995
0	0.100	1.000

The following relationship among the probabilities from a 2×2 table is the first step in describing an important property of classification accuracy based on binary variables. For two events labeled A and B, the same joint probability $P(AB)$, expressed two ways, is

$$P(AB) = P(A|B)P(B) \quad \text{or} \quad P(BA) = P(B|A)P(A).$$

Therefore, because $P(AB) = P(BA)$, then

$$P(A|B)P(B) = P(B|A)P(A) \quad \text{or} \quad P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

The last expression, called *Bayes' theorem* (after Sir Thomas Bayes, b. 1701), describes the relationship between conditional probabilities $P(A|B)$ and $P(B|A)$. Therefore, applying Bayes' theorem, a false positive probability consists of three components where

$$\begin{aligned}\text{false positive probability} = P(\bar{D}|T^+) &= \frac{P(T^+|\bar{D})P(\bar{D})}{P(T^+)} \\ &= \frac{[1 - P(\text{specificity})]P(\text{disease is absent})}{P(\text{positive test})}.\end{aligned}$$

To illustrate an often unappreciated property of a false positive determination consider a hypothetical test (T) with sensitivity of $P(T^+|D) = 0.95$ and specificity of $P(T^-|\bar{D}) = 0.90$ for a disease (D). Then, Bayes' theorem yields the relationship

$$\text{false positive probability} = P(\bar{D}|T^+) = \frac{0.10P(\bar{D})}{P(T^+)} = \frac{0.10P(\bar{D})}{0.95P(D) + 0.10P(\bar{D})}$$

where the probability of a positive test is

$$P(\text{positive test}) = P(T^+) = P(T^+D) + (T^+\bar{D}) = P(T^+|D)P(D) + P(T^+|\bar{D})P(\bar{D}).$$

Sensitivity and specificity are obvious properties of an accurate classification. Less obvious is the fact that performance of a binary test, measured in terms of the likelihood of a false positive result, is strongly influenced by the frequency of disease. As disease prevalence decreases, the probability of a false positive determination dramatically increases. Table 6.25 contains examples of false positive probabilities for five decreasing levels of disease prevalence where sensitivity $= P(T^+|D) = 0.95$ and specificity $= P(T^-|\bar{D}) = 0.90$. For example,

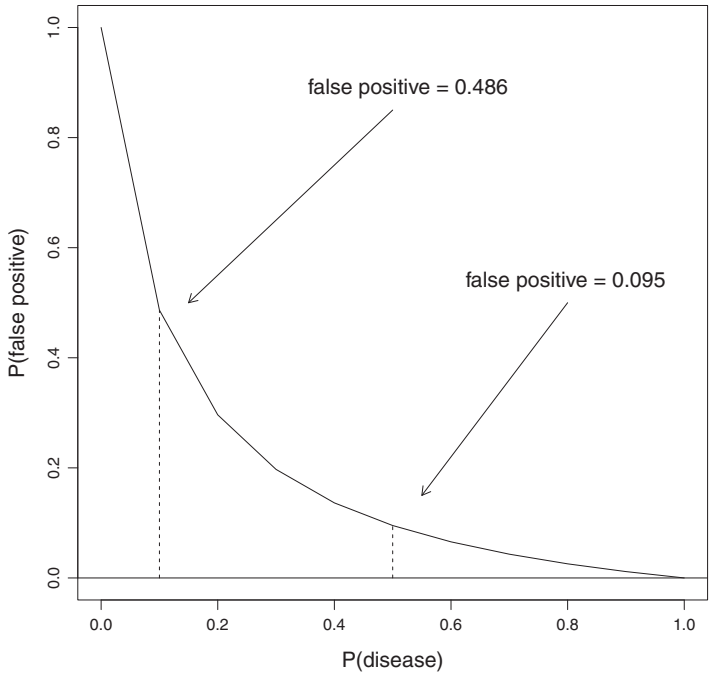


Figure 6.3 Probability of False Positive as Function of Probability of Disease

the probability of a false positive test is close to 0.5 when disease prevalence is $P(D) = 0.1$ (Figure 6.3). At the extreme, if a disease is not present, all positive results are false.

When an outcome is rare, such as HIV-positive individuals, a positive test is treated as evidence that more extensive and sometimes more expensive tests are necessary because of the likelihood of false positive results.

7

Linear Bivariate Regression Model

When famous physicists are photographed the background is frequently a blackboard covered with mathematics. Mathematics is the language of physics. Even written descriptions are typically insufficient to describe precisely or even to understand extremely complex and technical concepts of modern physics. String theory requires 10-dimensional geometry to describe the properties of subatomic particles. Clearly visual or verbal descriptions are a small part of a complete understanding. What the language of mathematics lacks in poetry, it makes up with precision and clarity.

To a lesser degree, the description of statistical concepts similarly requires unambiguous and unequivocal terminology. The language and terms used in statistics are also used in nonstatistical settings. For example, the terms independence, interaction, and confounding have unequivocal meaning in the context of statistical models but are used in everyday language in a much broader sense, frequently leading to confusion and misinterpretation of the statistical meaning. The following discussion, like physics, employs mathematics and symbols to define, explain, and discuss a number of basic concepts important to statistical data analysis. A linear bivariate regression model provides a definitive mathematical and statistical context. It is simple mathematically and sufficiently sophisticated statistically to provide an explicit and accurate context to unambiguously define and explore the properties of a variety of general statistical concepts. Specifically, the topics are additivity, regression model coefficients, correlation, adjustment, interaction, confounding, and collinearity.

The Bivariate Linear Regression Model

An expression for a bivariate linear regression model is

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \varepsilon_i \quad i = 1, 2, \ldots, n = \text{number of observations.}$$

This is usually referred to as a linear model because it consists of a sum of terms, but in this specific case the terms themselves are linear ($b_i\, x_i$). Perhaps this model should be called a "linear-linear" model. A visual representation of a bivariate linear model is displayed in Figure 7.1.

Like many statistical models, a bivariate linear model has deterministic and stochastic parts. The deterministic part is geometrically a plane ($b_0 + b_1\, x_1 + b_2\, x_2$; Figure 7.1). Coefficient b_1 represents the slope in the x_1 direction, and coefficient b_2 represents the slope of the same plane in the x_2 direction. Data points (y, x_1, x_2) are located above and below the plane (circles; Figure 7.1). The failure of these points to lie exactly on the plane (dots) depicts the influence of random variation. The probability distribution of these distances,

Table 7.1 *Description: Case Study Variables (n = 610 Mother-Infant Pairs)*

Variables	Units	Minimums	Medians	Maximums	Mean Values	*s. d.*[a]
bwt	Pounds	4.41	7.69	11.03	7.67	1.10
wt	Pounds	73.06	119.90	255.00	124.86	29.38
ht	Inches	58.85	65.57	72.72	64.53	2.26
gest	Weeks	28	39	42	39	1.51

[a]Estimated standard deviations.

denoted ε, is the stochastic part of the model. A bivariate linear model is the sum of deterministic and stochastic parts. The variables represented by x have several of names including independent variables, explanatory variables, and predictor variables. The variable represented by y has corresponding names, dependent variable, outcome variable, and response variable.

The properties of a typical bivariate linear model are illustrated by a case study of infant birth weight. A sample of $n = 610$ mothers and their infants provides four variables: the dependent variable infant birth weight (bwt_i) and three independent variables: maternal pre-pregnancy weight (wt_i), maternal height (ht_i), and infant gestational age ($gest_i$). These mothers are self-reported as white and are a small sample from the University of California, San Francisco Perinatal Data Base. A few descriptive statistics provide a sense of the data (Table 7.1).

A specific bivariate linear model applied to summarize and describe the influences from maternal height and weight on birth weight from the mother-infant data is

$$bwt_i = b_0 + b_i wt_i + b_2 ht_i + \varepsilon_i$$

and provides a concrete case study of the statistical issues that arise in the application of regression techniques that are also important issues in a large number of other kinds of statistical analyses.

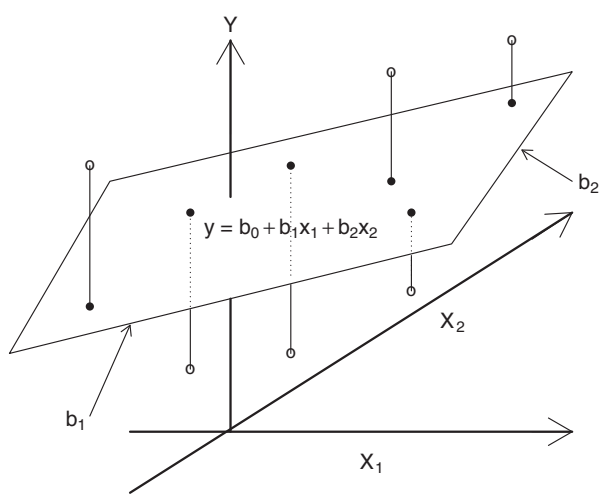


Figure 7.1 A Visual Representation of a Bivariate Linear Regression Model

Table 7.2 *Correlation Array: Estimated Pairwise Product-Moment Correlation Coefficients: Infant Birth Weight* (bwt)*, Maternal Prepregnancy Weight* (wt)*, Maternal Height* (ht)*, and Gestational Age* ($gest$)

	bwt	wt	ht	$gest$
bwt	1.000	0.126	0.167	0.452
wt	–	1.000	0.329	−0.034
ht	–	–	1.000	0.018
$gest$	–	–	–	1.000

A regression analysis ideally begins with estimation of a correlation array containing the product-moment correlation coefficients from all possible pairs of data variables (Table 7.2) (Chapter 5). These pairwise correlations, as will be described, are basic components of the sophisticated summary values at the center of a regression analysis.

Estimation of the bivariate model parameters from the case study data is the starting point for a rigorous and detailed discussion of a variety of general statistical topics in the specific context of a regression analysis (Table 7.3).

The bivariate model estimate of infant birth weight (denoted $b\hat{w}t$) based on the maternal weight and height is

$$\hat{y}_i = b\hat{w}t_i = 3.316 + 0.003wt_i + 0.062ht_i \quad i = 1, 2, \ldots, n = 610.$$

Also important, the stochastic contributions (ε) are estimated by the differences between observed and estimated birth weights or

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = bwt_i - b\hat{w}t_i = bwt_i - [\hat{b}_0 + \hat{b}_1 wt_i + \hat{b}_2 ht_i],$$

called *residual values*.

Two indispensable estimated variances are associated with the dependent variable y. The deviations $y_i - \bar{y}$ reflect the variation of the dependent variable entirely ignoring the independent variables. This measure of variability does not involve the regression analysis and is estimated by the usual expression

$$variance(y) = S_Y^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2 \quad i = 1,\ 2, \ldots, n.$$

The variability of the observations y is measured relative to the single estimated mean value $\bar{y}$.

Table 7.3 *Estimation: Bivariate Linear Regression Model Relating Infant Birth Weight* (y) *to Maternal Pre-pregnancy Weight* (x_1) *and Maternal Height* (x_2) – $n = 610$ *Mother-Infant Pairs*

	Estimates	*s. e.*[a]	z Values	p-values
b_0 = intercept	3.31638	–	–	–
b_1 (wt)	0.00299	0.0016	1.888	0.059
b_2 (ht)	0.06160	0.0184	3.349	<0.001

[a]Estimated standard errors.

A second variance based on the regression model accounting for the influences of the independent variables x_1 and x_2 is estimated by the expression

$$variance(y) = S^2_{Y|x_1,x_2} = \frac{1}{n-2}\sum (y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum \hat{\varepsilon}_i^2 \quad i = 1,\ 2,\ \ldots,\ n$$

summarizing the variation among the residual values $y_i - \hat{y}_i = \hat{\varepsilon}_i$. This estimated variance again reflects the variability of the dependent variable y but relative to the estimated x_1/x_2 regression plane (distances $\hat{\varepsilon}$). In addition, most linear regression models require the knowledge or the assumption that the stochastic contributions have a normal distributions with a variance represented by $\sigma^2_{Y|x_1,x_2}$ and estimated by $S^2_{Y|x_1,x_2}$. For the birth weight data, the estimated variance is $S^2_{Y|x_1,x_2} = 1.182$. At first glance, the variance might appear small. The magnitude of the variance of the birth weight variable, however, is a function of measurement units, in this case pounds. Like most statistical techniques the choice of measurement units does not affect the analytic results.

Additivity

An *additive model* is characterized by the property that the sum of individual estimated influences is exactly equal to the joint effect of a multivariate dependent variable. In less formal language, "the sum of the parts equals the whole." In more concrete statistical language, the influence of the x_1 variable on the outcome variable y is the same regardless of the value of the variable x_2 and vice versa. This fundamental property results from the structure of an additive regression model. Specifically, in symbols, for additive bivariate linear models

$$y_i = b_0 + b_1x_{1i} + b_2x_{2i} \quad \text{and} \quad y_i' = b_0 + b_1\,(x_{1i} + 1) + b_2x_{2i}$$

then

$$y_i' - y_i = [b_0 + b_1\,(x_{1i} + 1) + b_2x_{2i}] - [b_0 + b_1x_{1i} + b_2x_{2i}] = b_1.$$

Thus, additivity dictates that a one unit change in the independent variable x_1 produces a b_1 unit change in the dependent variable y, regardless of the value of the independent variable x_2. Variable x_1 is frequently said to have an influence "independent of variable x_2," which is not quite true. In fact, the model dictates that variable x_1 has an independent influence. Thus, the separate estimated influences of x_1 and x_2 combine to exactly reflect the bivariate influence of the dependent variable y. The fundamental feature of additive regression models in general is a description of a single multivariate observation as a series of separate measures of the influences of its component parts.

Coefficients

Headlines in a British newspaper read, "Tall Men Earn More than Shorter Colleagues." Citing a study published in *Men's Health* (an Australian journal), the news story states that "research shows that it appears that size matters." The article goes on to report that an extra two inches of height is associated with an average increase in income of 15%. In statistical terms, the researchers are likely reporting a regression analysis–estimated coefficient of

$\hat{b} = 0.075$ or an increase in height of one inch is associated with an increase in wages of 7.5%.

The interpretation of an estimated regression coefficient is not that simple. For example, does a man six feet tall who weighs 280 pounds outearn, on average, a man who is two inches shorter but weighs 180 pounds? More technically the question becomes: Is the linear height-income relationship influenced by weight? In addition, tall men are usually heavier than their shorter colleagues. It is certainly possible that size is the determining factor and not height alone or, perhaps, the difference in earnings relates entirely to differences in weight, indirectly reflected by differences in height. Other variables are undoubtedly involved such as the ages of the individuals sampled. It is likely that the reported observation comes from an estimated regression equation constructed from a number of additive independent variables. Furthermore, the estimated regression coefficient measuring the influence of height is subject to sampling variation. Thus, the interpretation of the comment "size matters" is not straight forward and requires considerable care.

A bivariate regression analysis provides an accurate and detailed description of an additive model, and, in addition, three simple linear regression models produce an unambiguous and detailed description of the origins and properties of the bivariate model regression coefficients.

A complete understanding depends on a number of technical details. The first regression model is

$$bwt_i = A_1 + B_1 ht_i.$$

Thus, an estimated infant birth weight based on a regression model and only maternal height is $\hat{bwt}_i = 2.952 + 0.073\ ht_i$. The differences between an observed birth weight bwt_i and the model-estimated birth weights $\hat{bwt}_i$, determined entirely by maternal height, generates a series of residual values $res_{bwt.ht} = bwt - \hat{bwt}$. These residual values reflect infant birth weight with the influence from maternal height "removed." More technically, as the term "removed" suggests, the residual values measure variation in birth weight not attributable to maternal height. For example, if a residual value is zero, then maternal height perfectly predicts infant birth weight. Otherwise, the residual values reflect the degree that birth weight is not "explained" by maternal height. The correlation between these residual values and maternal height is zero or, in symbols, $correlation(res_{bwt.ht}, ht) = 0$. The words "removed" and "explained" are used in the narrow statistical sense that a simple linear regression model (a straight line) accurately reflects the extent that infant birth weight is influenced by maternal height.

The second regression model

$$wt_i = A_2 + B_2 ht_i$$

produces an estimate of maternal pre-pregnancy weight based again entirely on maternal height in terms of model estimates, $\hat{wt}_i = -121.551 + 3.869\, ht_i$. A second set of residual values $res_{wt.ht} = wt - \hat{wt}$ are estimates of maternal weight also with the influence from maternal height "removed." Again, the correlation between residual values and maternal heights is zero or, in symbols, $correlation\ (res_{wt.ht}, ht) = 0$.

A third regression model is then applied to these two sets of residual values that are height "removed" infant birth weights and height "removed" maternal weights. A regression model

describing the relationship between these two residual values is

$$bwt - residual\ values = res_{bwt.ht} = a + b(res_{wt.ht}).$$

The estimated coefficient $\hat{b}$ is now said to measure the influence of maternal weight on infant birth weight that is not "explained" by maternal height. The estimated model coefficients are $\hat{a} = 0$ and $\hat{b} = 0.00299$. The estimated intercept $\hat{a}$ is always exactly zero. The coefficient $\hat{b}$ estimated from the "height removed" residual values and the coefficient $\hat{b}_1$ estimated from the bivariate linear model analysis (Table 7.3) are algebraically identical ($\hat{b} = \hat{b}_1$). The estimated regression coefficients from the bivariate linear model, therefore, separately indicate the influence of each independent variable statistically isolated from the influences of the other independent variable in the sense just described; that is, using three simple linear relationships, the influence from maternal height is "removed," and the regression coefficient $\hat{b}_1$ then measures the "pure" influence of maternal weight on infant birth weight. In the same way, the estimated coefficient $\hat{b}_2$ separately measures the "pure" influence of the maternal height on infant birth weight with influence from the maternal weight "removed." Three regression models or a single bivariate model, therefore, produce the same measures of the additive ("separate") influences of maternal height and weight.

A more complete reporting of the increase in earnings associated with increased height quoted in the newspaper article might read as follows:

> When the analytic results from an artificial linear and additive regression model are used to compare two individuals, the estimated coefficient measuring the influence of height indicates that an estimated increase of two inches is associated with 15% average increase in earnings when isolated from all other relevant influences such as age, ethnicity, and kind of employment. Furthermore, the accuracy of the estimate "15% increase" depends on the accuracy a linear and additive model to represent the income-height relationship that can be substantially influence by random variation.

This more complete description does not make good headlines. Media reports rarely mention the fact that such a single summary value extracted from a multivariable analysis results from an application of a statistical and almost always additive regression model. Furthermore, newspaper articles rarely give even a hint of the influence of sample size, sampling variation, or adequacy of an additive linear model to summarize the relationship between independent and dependent variables.

Although estimated coefficients adjusted for the influences of other variables are the central purpose of a multivariable analysis, their estimated values are not automatically important or easily interpreted. A physician is not likely to say "your cholesterol level is over 300 mg/dL but don't worry because after statistical adjustment with a regression model including your other risk factors, its influence is considerably reduced."

Multiple Correlation Coefficient

It is critically important to assess the accuracy of an additive regression model, for which we have a number of techniques (Chapter 12). A natural and readily available measure of accuracy is the correlation between estimated and observed values. The correspondence between observed values (y) and regression model estimates of the same value ($\hat{y}$) is summarized by

a single product-moment correlation coefficient called the *multiple correlation coefficient* (Chapters 5 and 12). The value typically reported is the squared coefficient. A multiple correlation coefficient is estimated directly from the observed and estimated values. For example, the birth weight analysis yields a squared estimated $correlation(y, \hat{y}) = r^2_{y\hat{y}} = (0.183)^2 = 0.034$.

The same value calculated from three product-moment correlation coefficients is

$$r^2_{y\hat{y}} = \frac{r^2_{y1} + r^2_{y2} - 2r_{y1}r_{y2}r_{12}}{1 - r^2_{12}}.$$

This not very useful expression does show that when x_1 is not correlated with $x_2 (r_{12} = 0)$, then $r^2_{y\hat{y}} = r^2_{y1} + r^2_{y2}$. Thus, the additive influences of uncorrelated variables x_1 and x_2 on the dependent variable y are separate and proportional to their respective squared correlation coefficients.

The squared correlation coefficient $r^2_{y\hat{y}}$ also summarizes the difference between sums of squared deviations of the data values $(y - \bar{y})$ and the residuals values $(y - \hat{y})$ where

$$r^2_{y\hat{y}} = \frac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \approx \frac{S^2_Y - S^2_{Y|x_1,x_2}}{S^2_Y} \quad i = 1, 2, \ldots, n.$$

The squared correlation coefficient $r^2_{y,\hat{y}}$ is essentially a comparison of two estimated variances, S^2_Y and $S^2_{Y|x_1,x_2}$. Thus, a multiple correlation coefficient indicates the proportional reduction in variation associated with the independent variables x_1 and x_2 included in a regression model relative to the maximum possible variation associated with the dependent value (y values). The squared correlation coefficient is an elegant summary of the magnitude of joint influence of the dependent variables in terms of a single statistical measure between 0 and 1. Geometrically, the coefficient $r^2_{y,\hat{y}}$ measures the reduction in variability of the dependent y relative to a horizontal line $(\bar{y})$ versus measuring variability relative to an x_1/x_2 estimated plane $(\hat{y})$. Such a reduction measures the joint influence of the independent variables x_1 and x_2.

A correlation coefficient that measures the strength of the relationship between two variables after the elimination of the influences of a third variable is sometimes a useful product of a regression analysis. From the example birth weight bivariate analysis, the residual values $res_{bwt.ht}$ and $res_{wt.ht}$ are constructed so that the height influence *ht* is "removed." Correlation between these two sets of residual values is then an estimate of the remaining association between variables birth weight and maternal weight. From the birth weight data, this correlation is $correlation(res_{bwt.ht}, res_{wt.ht}) = 0.076$, called a *partial correlation coefficient*.

Although calculated directly from the n pairs of residual values, a bit of algebra produces a general expression for the estimated partial correlation coefficient as

$$correlation(res_{y.j}, res_{i.j}) = r_{yi.j} = \frac{r_{yi} - r_{yj}r_{ij}}{\sqrt{(1 - r^2_{ij})(1 - r^2_{yj})}}$$

in terms of pairwise correlation coefficients (Table 7.2).

This expression identifies the consequences of "removing" the redundancy between the independent variables with a bivariate regression model. Some examples are the following:

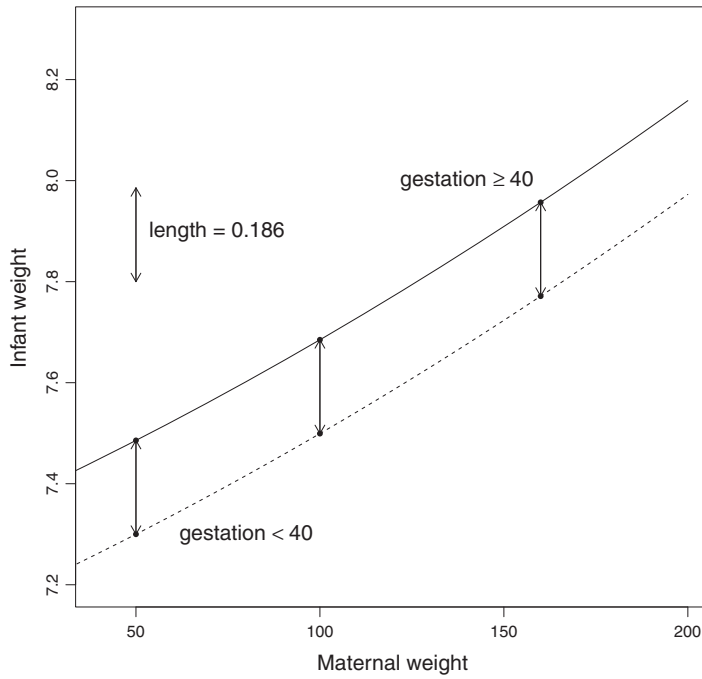


Figure 7.2 Regression Model – Two Levels of Gestational Age (<40 and ≥40 Weeks) Adjusted for Influence of Maternal Weight

1. For $r_{y2} = 0$, then the partial correlation coefficient $r_{y1.2} = \frac{r_{y1}}{\sqrt{1-r_{12}^2}}$ indicates that an influence from x_2 remains when x_2 is related to x_1 and x_1 is related to y.
2. When $r_{12} = r_{y2} = 0$, only then is the correlation between y and x_1 unaffected by x_2 or, in symbols, $r_{y1.2} = r_{y1}$.
3. When $r_{y2}r_{12} > 0$, the difference between $r_{y1.2}$ and r_{y1} reflects the degree of "pure" influence associated with x_2. For the example, the correlation coefficient between birth weight and maternal weight with height "removed" is $r_{y1.2} = 0.076$ and not removed is $r_{y1} = 0.126$.

Adjustment

Birth weight data classified into two gestational age categories illustrate the application of statistical adjustment using an additive linear regression model. When F_i represents a binary indicator variable (gestational age < 40 weeks, $F_i = 0$, and gestational age ≥ 40 weeks, $F_i = 1$), a bivariate regression model relating infant birth weight (bwt) to both gestational age (F) and maternal prepregnancy weight (wt) is

$$bwt_i = b_0 + b_1 F_i + b_2 wt_i + b_3 wt_i^2.$$

The model dictates additive influences on birth weight from gestational age and the two maternal weight variables producing an additive nonlinear (quadratic) influence. The binary variable F causes the relationship between infant birth weight and maternal weight to have identical quadratic influences in both gestation groups (Figure 7.2). The additive

Table 7.4 *Estimation: Linear Regression Model Coefficients Relating Infant Birth Weight* (y) *to Maternal Pre-pregnancy Weight* (wt_i) *for Two Levels of Gestational Age (gest* < 40 *Weeks and gest* ≥ 40*) for* $n = 610$ *Mother-Infant Pairs*

	Estimates	*s. e.*[a]	z statistics	p-values
b_0 (*intercept*)	7.126	–	–	–
b_1 (*gest*)	0.186	0.0096	1.924	0.055
b_2 (wt)	0.003	0.0097	3.347	–
b_3 (wt^2)	5×10^{-6}	3.4×10^{-5}	0.147	<0.001

[a]Estimated standard errors.

model coefficient b_1, therefore, describes the magnitude of the difference associated with gestational age (<40 and ≥40) and is said to be adjusted for maternal weight. Specifically, for $F_i = 1$, the model is

$$bwt_i' = b_0 + b_1 + b_2 wt_i + b_3 wt_i^2$$

and for $F_i = 0$, the model becomes

$$bwt_i = b_0 + b_2 wt_i + b_3 wt_i^2$$

making the difference $bwt_i' - bwt_i = b_1$. To repeat, because the model is additive, the influence of gestational age measured by b_1 is not affected by maternal weight (Table 7.4). Geometrically, the difference between the two gestational age categories is represented by the constant distance between two quadratic curves created to be identical for all maternal weights (Figure 7.2). The model-estimated coefficient $\hat{b}_1$ summarizes this difference with a single estimated value described in a variety of ways. To list a few: The estimated difference between the two gestational age categories is the same for all maternal weights or is independent of maternal weight or accounts for maternal weight or is adjusted for maternal weight or treats maternal weight as constant or controls for maternal weight or partitions the influences of gestation and maternal weight and removes the influence of maternal weight.

These descriptions refer to the statistical property that an additive model produces a single constant measure of the difference between categories with the influences of the other variables in the model "removed." From the birth weight model, the adjusted estimate of the influence of gestational age is the distance between two quadratic curves. Specifically, the estimated distance is $\hat{b}_1 = 0.186$ (Figure 7.2). In other words, the measure of influence of the categorical variable is created to be identical for all maternal weights. In this sense, the influence of maternal weight is "removed" as an issue in the interpretation of the influence of gestational age. It is the same for all maternal weights. It is important to note that when an additive model fails to reflect accurately the relationships within the data, the measures of the individual influences estimated from the model coefficients equally fail.

Interaction

A statistical interaction is the failure of a relationship among estimates or summary values or most any statistical measure to be the same for all values of a third variable. The direct

consequence is that a summary value, at best, becomes difficult to interpret and is generally misleading. In terms of a regression model, interaction is another term for nonadditivity. For example, consider the additive bivariate model

$$y_i = a + bF_i + cx_i$$

where F_i indicates a binary classification. For $F_i = 0$, the model is $y_i = a + cx_i$, and when $F_i = 1$, the model becomes $y_i' = a + b + cx_i$. The relationship between independent variable y and dependent variable x is summarized by a straight line for both values of F, and, furthermore, the two straight lines have the same slope (slope $= c$) within both categories. Because the relationship between variables x and y is the same for both levels of F, it is said that "no interation exists," which is again not quite true. The model dictates that no interaction exists. In other words, an additive model is created so that the difference $y_i' - y_i = b$ is constant, or, more to the point, the difference is the same for all values of the variable x. The presence or absence of an interaction is a property of the model and not the data.

The same bivariate linear regression model with an interaction term added is

$$y_i = a + bF_i + cx_i + d(F_i \times x_i).$$

In this nonadditive case, when $F_i = 0$, the model again is $y_i = a + cx_i$, and when $F_i = 1$, the model becomes $y_i' = (a + b) + (c + d)x_i$. The relationship between variables x and y is described by two different straight lines. The intercepts are a and $a + b$, and the slopes are c and $c + d$. The difference $y_i' - y_i = b + dx_i$ is not constant for all values of x. The variables F and x are said to interact. The value of the regression coefficient b is influenced by the variable $x(d \neq 0)$ because the summary lines are not parallel. Thus, the single regression coefficient b fails to usefully summarize the relationship between binary variable F and dependent variable y because the relationship is different for all values of x. Again, interaction is a property of the model and not the data.

Another version of a bivariate linear regression model including an interaction term is

$$y_i = b_0 + b_1x_{1i} + b_2x_{2i} + b_3(x_{1i} \times x_{2i}).$$

When $b_3 \neq 0$, again the difference from a one-unit increase in the dependent variable x_1 does not produce a simple and constant measure of change in the independent variable y, measured by a single regression coefficient b_1. Specifically, for a nonadditive bivariate model,

$$\begin{aligned} y_i &= b_0 + b_1x_{1i} + b_2x_{2i} + b_3(x_{1i} \times x_{2i}) \\ &= b_0 + (b_1 + b_3x_{2i})x_{1i} + b_2x_{2i}. \end{aligned}$$

The influence from the variable x_1 depends on the value of x_2 because the coefficient describing the influence x_1 depends on the value x_2. In symbols, the model measure of influence of x_1 is $(b_1 + b_3x_2)$. The presence of an interaction again means that a single regression coefficient is not a useful summary of the influence of the dependent variable x_1. For example, confidence intervals and statistical tests based on the estimate $\hat{b}_1$ are meaningless.

From the example birth weight data, a nonadditive model produces four estimated coefficients (Table 7.5). An increase of one pound in maternal weight produces an estimate of a $\hat{b}_1 + \hat{b}_3x_2$ pound increase in infant birth weight for a maternal height of x_2. From the

Table 7.5 *Estimation: Nonadditive Linear Bivariate Regression Model Relating Infant Birth Weight* (y) *to Maternal Pre-pregnancy Weight* (x_1) *and Height* (x_2)

Coefficients		Estimates	*s. e.*	*p*-value
Intercept	b_0	13.134	–	–
wt	b_1	−0.0752	0.016	–*
ht	b_2	−0.0909	0.074	–*
$ht \times wt$	b_3	0.0012	0.005	0.144

*Does not reflect a separate influence.

interaction model, for a women 63 inches tall, the estimated increase in infant birth weight per pound of maternal weight would be $-0.075 + 0.0012(63) = 0.0$ pounds, demonstrating that the single estimated coefficient $\hat{b}_1 = -0.075$ is meaningless as a single summary of the influence of maternal weight. Statistical software frequently display meaningless statistical evaluations of single coefficients estimated from nonadditive models.

Confounder Bias

A dictionary definition of confounding states: mixed up with something else so that the individual elements become difficult to distinguish. A definition from an epidemiology textbook states, "The common theme with regard to confounding by a variable is: the association between an exposure and a given outcome is induced, strengthened, weakened or eliminated by a third variable." Based on the bivariate linear model, the definition of confounding is rigorously defined, simply described, and entirely unambiguous. Although wider applications of the concept of confounding are useful, an explicit description in the context of a bivariate regression model again serves as a concrete foundation for understanding these also important but often less obvious situations.

For the bivariate linear regression case, the statistical question of confounding is simply: Does the independent variable x_2 have a useful role in the model? In other words, does the variable x_2 influence the relationship between independent variable x_1 and dependent variable y? This question is relevant only in the context of an additive model. If an interaction exists between dependent variables x_1 and x_2, the question of the role of x_2 is answered. It is necessary to include the variable x_2 in the model to accurately describe the influence of x_1. The question of a confounding influence disappears.

When a bivariate model is additive, the degree of confounding is measured by the difference between two estimated regression coefficients. Consider the two linear additive regression models:

$$Y_i = B_0 + B_1 x_{1i}$$

and

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i}.$$

A measure of the extent of confounding is the difference $\hat{B}_1 - \hat{b}_1$. This definition refers to estimated values and not to the underlying model parameters B_1 and b_1. Thus, confounding is

a property of the data sampled. For example, tests of significance and confidence intervals are not relevant to the question of the confounding influence. The difference $\hat{B}_1 - \hat{b}_1$ measures bias.

The least squares estimate of the simple linear regression coefficient B_1 from a sample of n pairs of observations (Y, x_1) is

$$\hat{B}_1 = \frac{S_{Y1}}{S_1^2}.$$

Similarly, the least squares estimate of the bivariate linear regression coefficient b_1 from a sample of n sets of observations (y, x_1, x_2) is

$$\hat{b}_1 = \frac{S_{y1}S_2^2 - S_{y2}S_{12}}{S_1^2S_1^2 - S_{12}^2}.$$

The difference between these two estimates

$$\hat{B}_1 - \hat{b}_1 = \hat{b}_2 r_{12}\left[\frac{S_2}{S_1}\right]$$

directly measures the confounding bias incurred by not including x_2 in the model.

The influence of the confounding variable x_2 depends on the model-estimated correlation coefficient r_{12} and regression coefficient $\hat{b}_2$. Direct consequences are the following:

1. No confounding bias exists when x_1 and x_2 are uncorrelated ($r_{12} = 0$).
2. No confounding bias exists when $\hat{b}_2 = 0$.
3. When $r_{y2} = 0$, confounding bias remains as long as $r_{12} \neq 0$. In symbols, the measure of confounding bias becomes $\hat{B}_1 - \hat{b}_1 = \frac{\hat{B}_1}{1-r_{12}^2}$.

To summarize, for independent variable x_2 to cause a confounding bias, it must be both correlated with x_1 ($r_{12} \neq 0$) and directly related to dependent variable y ($\hat{b}_2 \neq 0$).

The question of whether a confounding variable should or should not be included in the model does not have a rigorous answer. It is occasionally suggested that to be considered a confounding influence, the change in $\hat{B}_1$ relative to $\hat{b}_1$ should be more than 10%. The value 10% is a conventional value with no theoretical basis. The question of confounding concerns bias. Confounding is a property of the sampled data, and a choice has to be made to include or not include the confounding variable in the model regardless of the sources of the confounding influence.

Situations arise where nonconfounding variables are usefully included in an additive regression model. Such a situation occurs when a variable unrelated or nearly unrelated to the other independent variables in the regression equation strongly relates to the dependent variable. As would be expected, including such a variable in the model reduces the variance of the estimated values $\hat{y}$ which reduces the variances of the estimated regression coefficients, creating a more powerful statistical tool.

For the birth weight data, the lack of correlation between gestational age (*gest*) and maternal height ($r_{gest,ht} = 0.02$) and weight ($r_{gest,wt} = -0.03$) means that gestational age has only a minor confounding influence (Table 7.2). Nevertheless, the correlation of gestational age with infant birth weight is $r_{gest,bwt} = 0.452$ and including gestational age in the regression model substantially increases the precision of the model-estimated coefficients. As a result,

the inclusion of gestational age reduces the p-value associated with the estimated maternal weight coefficient $\hat{b}_1$, for the example analysis, from 0.059 (gestational age excluded) to 0.001 (gestational age included). As expected, the squared multiple correlation coefficient $r^2_{y\hat{y}}$ increases, specifically from 0.034 to 0.238. As demonstrated, it is sometimes advantageous to ignore a small and often inconsequential bias to achieve substantial gains in precision and interpretation (Chapter 10).

Collinearity

The definition of the term collinear is as follows: two values are collinear when they lie on a straight line. Collinear variables are not an issue in statistical analysis. One variable can be used and other collinear variables ignored. Collinear variables have a correlation coefficient of exactly 1.0, and, therefore, a single variable perfectly represents the influence of any other collinear variable or variables. The analysis is not affected by their exclusion.

Statistical issues arise in a regression analysis, however, when independent variables are nearly collinear. An extremely high correlation between independent variables is not a frequent occurrence, but occasionally such variables are a desired part of a regression analysis. For example, in a study of risk of hypertension among textile workers in China, age of the worker and time employed were both thought to make important contributions to the analysis. Collinearity became an issue because employer policy was to hire new workers at age 19 or occasionally at age 20, causing variables age and years worked to be close to collinear because age essentially equals years worked plus 19 (correlation > 0.9).

The issues surrounding collinearity are clearly illustrated again with a bivariate linear regression model. When a bivariate plane is estimated from sampled data, it is important that the observed values cover a wide range for both independent variables x_1 and x_2 to produce precise estimates of the model coefficients. Technically this property is referred to as the "area of support" for the estimated plane. A rough analogy is the placement of the legs of a table. When the legs are placed at four corners of a table top, they produce a stable surface. As the legs are moved toward the center of the table, the table becomes increasingly unstable. Estimated regression coefficients have an analogous property of becoming increasingly unstable as the support for the plane diminishes. The principle is simple. The smaller the range of either observed independent variable, the smaller the support, the less precise the estimated coefficients.

A statistical example of the relationship between support and precision arises from the analysis of the simple regression model $y_i = a + bx_i$. An estimate of the variance of the estimate of the coefficient b is

$$variance(\hat{b}) = \frac{S^2_{Y|x}}{\sum (x_i - \bar{x})^2}.$$

The degree of support is directly determined by the spread ("area of support") of the x-values measured by the sum of squared deviations $\sum (x_i - \bar{x})^2$, which largely dictates the degree of precision of the estimated regression coefficient ($\hat{b}$).

A parallel summary measure of support from bivariate data is the correlation between independent variables (r_{12}). Figure 7.3 displays the x_1/x_2 area of support for a set of observations. Figure 7.4 again displays the x_1/x_2 area of support but over a reduced area (larger

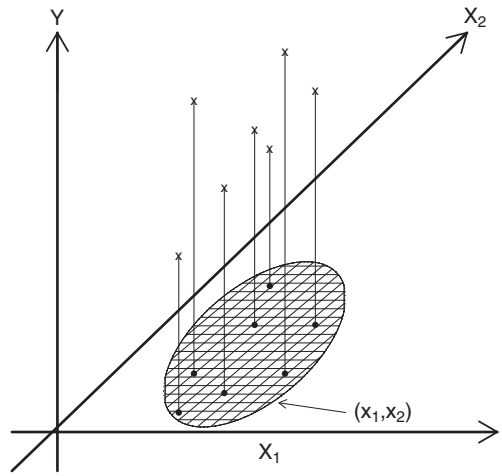


Figure 7.3 "Support" for Estimation of Regression Coefficients for an Estimated Plane

correlation coefficient r_{12}). The correlation coefficient of $r_{12} = 0$ yields the maximum spread of both variables x_1 and x_2. At the other extreme, when the correlation is $r_{12} = 1$, bivariate support entirely disappears and estimation of the bivariate regression model coefficients is no longer possible. Technically, the variances of the coefficients become infinite.

From an analytic point of view, the expression for the estimated variance of a bivariate regression coefficient $\hat{b}_i$ is

$$variance(\hat{b}_1) = \frac{S^2_{Y|x_1\,x_2}}{(n-1)\,S_i^2\,\left(1-r_{12}^2\right)}.$$

The expression pinpoints the relationship between the area of support, measured by $(1 - r_{12}^2)$, and precision of the estimated regression coefficient. The factor $1/(1 - r_{12}^2)$

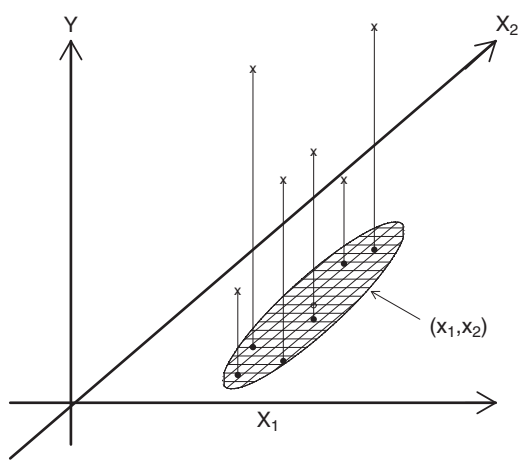


Figure 7.4 "Support" Producing a Less Precise Estimation of Regression Coefficients for an Estimated Plane Due to Increased x_1/x_2 Correlation

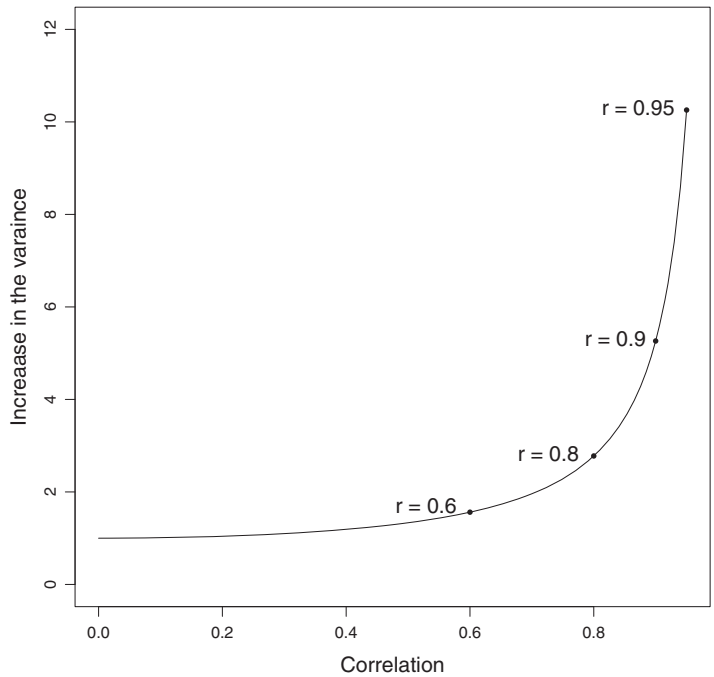


Figure 7.5 Increase in Variance of Estimated Regression Coefficients Due to Correlation between Dependent Variables x_1 and $x_2(r_{12})$ in a Bivariate Linear Regression Analysis

increases the variance of the estimate $\hat{b}_i$ proportional to increase in the squared value of the estimated correlation coefficient between independent variables x_1 and x_2 and becomes a substantial influence when a correlation is close to 1.0 (Figure 7.5). For example, when $r_{12} = 0.7$, the variances of the estimated coefficients $\hat{b}_1$ and $\hat{b}_2$ close to double and when $r_{12} = 0.95$, the estimated variances increase by a factor of close to 10 relative to the maximum area of support that occurs at $r_{12} = 0$.

For most data, particularly data concerning human disease or mortality, correlation coefficients larger than 0.7 are rare. For the case study example, the correlation between maternal height and pre-pregnancy weight is high for most human data. This estimated height-weight correlation of $r_{12} = 0.329$, however, only moderately increases the variance of an estimated regression coefficients by a factor of $1/(1-0.329^2) = 1.12$.

8

The 2 × *k* Table

The Pearson chi-square statistic is the usual choice to assess an association between two categorical variables described by data classified into in a two-way table (Chapters 1 and 6). One such table that contains the counts of 165 individuals who responded to the survey question "Do you think exercise increases the length of life (yes/no)?" classified by their reported physical activity is displayed in Table 8.1.

When the columns of such a table have a logical sequence and in a second table they do not, the chi-square test statistic is identical ($X^2 = 6.997$ – degree of freedom = 3). Statistical techniques exist that capitalize on the ordering of both numeric and nonnumeric (ordinal) values to produce a more sensitive approach to statistical analysis of a $2 \times k$ table.

Table 8.2 contains counts of subjects from a coronary heart disease study (*chd*) classified by the reported number of cigarettes smoked per day.

The smoking exposure categories have a definite numeric ordering, and a typical chi-square analysis includes an evaluation based on the slope of a linear response ("test of trend"). The fact that the categories are constructed from meaningful numeric quantities does not guarantee that the relationship is linear. It is possible, for example, that an increase in smoking among light smokers has a different influence on the risk of a *chd* event than the same increase in heavy smokers. An alternative statistical strategy between completely ignoring an ordered classification and assuming that a linear response accurately reflects a response begins with the Wilcoxon rank sum test that applies to both ordinal (Table 8.1) and numeric (Table 8.2) data classified into a $2 \times k$ table.

Wilcoxon (Mann-Whitney) Rank Sum Test

Qualitative or quantitative variables that can be ordered without relying on a numeric scale are called *ordinal variables*. Regardless of the origins and properties of observed data, the Wilcoxon procedure starts by ranking the observations producing specific ordinal values. The properties of the distribution that produced the original data are no longer relevant. The data simply become a set of integers (ranks = 1, 2, 3, . . . , N = number of observations).

A principal application of a Wilcoxon rank sum analysis is identification of differences between two samples of data providing an alternative to Student's *t*-test. The Wilcoxon rank sum test is slightly less efficient (lower statistical power; Chapter 15) than the corresponding *t*-test but does not require knowledge or the assumption that numeric data consist of values sampled from two normal distributions with the same variance. In addition, the Wilcoxon test produces exact results for small sample sizes.

Table 8.1 *Data: To Illustrate an Important Property of the Chi-Square Test for Independence Applied to a Two-Way Table*

	None	Occasional	Moderate	Strenuous	Total
Yes	7	7	8	20	42
No	25	34	32	32	123
Total	32	41	40	52	165

To compare two groups (designated as groups 1 and 2 for convenience), the Wilcoxon test statistic is the sum of ranked values (denoted R_i) of n_1 observations from group 1 or n_2 observations from group 2 where all $N = n_1 + n_2$ observations are ranked from the smallest to the largest without regard to group membership. When only random differences exist between compared groups, the sum of ranks (denoted $W = \Sigma R_i$) from each group is likely to be close to proportional to the number of observations in the group. That is, the sum of the ranks of the values in group 1 is likely to be in the neighborhood of $n_1(N + 1)/2$. The quantity $(N + 1)/2$ is the mean rank. In symbols, the expected value of the distribution of W is $EW = n_1(N + 1)/2$ where n_1 represents the number of observations in group 1. When the mean values of two sampled groups systematically differ, the sum of the ranks of the observations from group 1 is likely to be substantially smaller or larger than the value $n_1(N + 1)/2$. The sum of the ranks from group 2 has the same properties and could equally be used to identify systematic differences between groups.

For an example, consider the following artificial data:

$$\text{Group 1} : 4, 35, 21, 28, 66 \quad n_1 = 5 \text{ and}$$
$$\text{Group 2} : 10, 42, 71, 77, 90 \quad n_2 = 5$$

making $N = n_1 + n_2 = 5 + 5 = 10$ the total number of observations. The sum of the ranks associated with group 1 is $W = \Sigma\, R_i = 1 + 3 + 4 + 5 + 7 = 20$ (Figure 8.1, circles). When no systematic differences exist between groups 1 and 2, the test statistic W calculated from group 1 differs from the expected value $EW = n_1(N + 1)/2 = 5(11)/2 = 27.5$ by chance alone.

A seemingly different approach to the same two-sample comparison is the *Mann-Whitney test.* The Mann-Whitney test statistic is created by counting the number of observations in one

Table 8.2 *Data: Coronary Heart Disease (Present/Absent) Classified by Four Levels of Reported Cigarette Smoking Exposure (Chapter 9)*

	Cigarettes per day				
	0	1–20	21–30	>30	Total
chd	98	70	50	39	257
No *chd*	1554	735	355	253	2897
Total	1652	805	405	292	3154

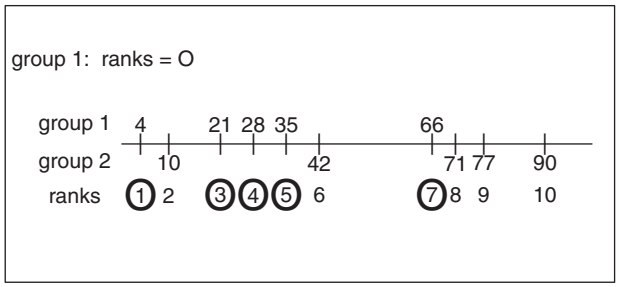


Figure 8.1 Visual Display of Wilcoxon Test Statistic (W) Based on Ranks Used to Measure Difference between Two Samples of Data

group that are smaller than each observation in the other group (denoted u_i for the ith observation). Like the Wilcoxon strategy, the Mann-Whitney counts are scale-free, and analysis, therefore, does not depend on the properties of the population distributions sampled. Such statistical measures are called *nonparametric* or *distribution-free*.

For the example data, two observations in group 2 (values 10 and 42) are less than the fifth largest value in group 1 (value 66) yielding the Mann-Whitney count of $u_5 = 2$. The Mann-Whitney test statistic is the sum of these u_i values (denoted $U = \Sigma u_i$ where $i = 1, 2, \ldots, n_1$). For the example data, the $n = 10$ observations yield a Mann-Whitney statistic of $U = \Sigma u_i = 0 + 1 + 1 + 1 + 2 = 5$ among $n_1(n_2) = 5(5) = 25$ possible pairs. Like the Wilcoxon test statistic, the count U reflects differences between the compared values from each group and likely becomes extreme when the two groups systematically differ.

A computational table suggests one way to envision the statistic U. A table is constructed so the rows are ordered values from group 1, the columns are ordered values from group 2, and the cells of the table are assigned a value 1 when a value in group 2 is less than the value in group 1 and a 0 otherwise. Such a table constructed from the example data is the following:

Group 1		4	21	28	35	66
	10	0	1	1	1	1
	42	0	0	0	0	1
Group 2	71	0	0	0	0	0
	77	0	0	0	0	0
	90	0	0	0	0	0
u_i		0	1	1	1	2

The Mann-Whitney value U is the total number of 1's among the $n_1 n_2$ possible pairs in the table. For the example, the value is $U = 5$ among the $5(5) = 25$ possible pairs.

The Wilcoxon and Mann-Whitney test statistics (W or U) produce the identical statistical analysis because the test statistics W and U differ by a constant value determined only by sample sizes of the groups compared. Specifically, for group 1, $W = U + n_1(n_1 + 1)/2$ where n_1 is again the number of observations in group 1. For the example, when $W = 20$, then

$U = \Sigma\, u_i = 5$ because W and U differ by $n_1(n_1 + 1)/2 = 5(6)/2 = 15$, a nonrandom value determined entirely by sample size $n_1 = 5$.

The rank of the ith observation from group 1 is the sum of two values. It is the rank of the observation within group 1 plus the number of observations less than that observation in group 2, namely, u_i. In symbols, the rank of the ith observation in group 1 is $R_i = i + u_i$. The Wilcoxon test statistic W is the sum of these ranks:

$$W = \sum R_i = \sum (i + u_i) = \sum i + \sum u_i = \frac{n_1(n_1 + 1)}{2} + U$$

where $\Sigma i = n_1(n_1 + 1)/2$ (Chapter 27). Specifically, for the example data, a computational table for group 1 illustrates:

Group 1	R_i	i	u_i
4	1	1	0
21	3	2	1
28	4	3	1
35	5	4	1
66	7	5	2
Sum	20	15	5

Therefore, $U = W - n_1(n_1 + 1)/2 = \Sigma\, R_i - n_1(n_1 + 1)/2 = 20 - 5(6)/2 = 20 - 15 = \Sigma\, u_i = 5$ (last row).

To statistically describe differences between two groups, the test statistic U has a useful and intuitive interpretation. Among the $n_1 n_2$ total number of possible pairs of observations from two samples, the value U is the total number of values from group 2 less than each value from group 1. The proportion $\hat{P} = U/(n_1 n_2)$, therefore, becomes an estimate of the probability that a randomly selected value from group 2 is less than a randomly selected value from group 1. For the example, this estimate is $\hat{P} = U/(n_1 n_2) = 5/25 = 0.2$. Figure 8.2 schematically displays the logic of the Mann-Whitney test statistic $\hat{P}$.

A Mann-Whitney estimated probability $\hat{P}$ near 0.5 provides little evidence of a difference between compared groups; that is, a randomly chosen value from group 1 is equally likely to be smaller or larger than a randomly chosen value from group 2. As the estimate $\hat{P}$ increasingly differs from 0.5, the likelihood increases that the compared groups systematically differ (Figure 8.2). Furthermore, computation and interpretation of $\hat{P}$ remain the same when the difference between the compared groups is measured by a variable that is not numeric but has a meaningful order (an ordinal variable).

Approximate statistical tests based on the normal distribution exist to evaluate the influence of sampling variation on the equivalent test statistics W, U, and $\hat{P}$ when neither sample size n_1 and n_2 is small (both samples greater than 10 observations). The values analyzed are ordinal (distribution-free), but a nonparametric test statistic has an approximate normal distribution. For comparisons between groups when the sample size is small in one or both groups, special tables or computer applications produce an exact analysis (no approximations).

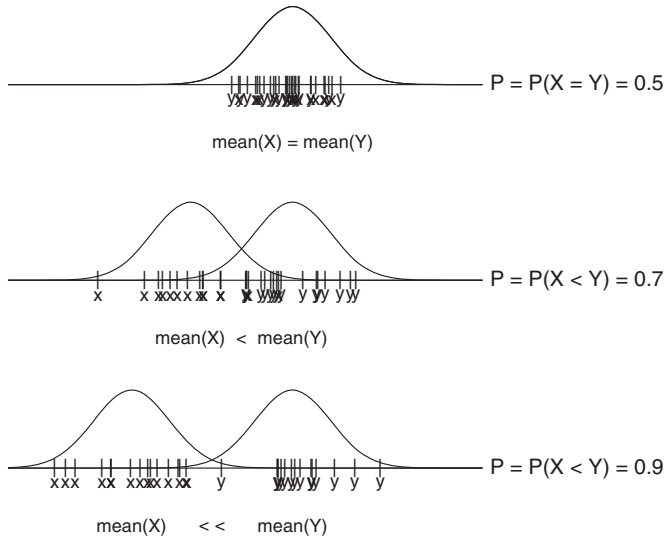


Figure 8.2 Visual Display of the Mann-Whitney Test Statistic as a Measure of the Difference Observed between Two Samples of Data

Specifically, when only random differences exist between the two groups ($P = 0.5$), the Mann-Whitney estimated probability $\hat{P}$ has an approximate normal distribution with a mean value of 0.5 and an exact and approximate variance of

$$variance(\hat{P}) = \frac{1}{12n_1n_2}[N+1] \approx \frac{1}{12n_1} + \frac{1}{12n_2}$$

regardless of the original distributions of the sampled observations (Chapter 27). The test statistic

$$z = \frac{\hat{P} - 0.5}{\sqrt{\frac{1}{12n_1} + \frac{1}{12n_2}}}$$

then has an approximate standard normal distribution when $P = 0.5$. This test statistic addresses the statistical question: Do systematic differences likely exist between the two sampled groups? The nonparametric statistical tests based on ranks W or counts U produce the same results (same p-value). A Mann-Whitney test statistic $\hat{P}$ also plays an important role in evaluating the accuracy of classification strategies (Chapter 22).

A Mann-Whitney assessment of a difference between two samples is symmetric in the sense that the same test statistic can be constructed based on the probability that a randomly selected observation from group 2 is larger than a randomly selected observation from group 1. For the example, this value is $\hat{Q} = 20/25 = 0.8 = 1 - \hat{P}$. However, no difference occurs in the assessment because the deviations $|\hat{P} - 0.5|$ and $|\hat{Q} - 0.5|$ are equal.

Cholesterol levels and behavior-type data for the 40 heaviest study participants (greater than 225 lb) from a coronary heart disease study illustrate the Mann-Whitney (Wilcoxon) test (Table 8.3).

The mean cholesterol level from behavior type A men is 245.05 and from type B men is 210.30, making the mean difference 34.75. The statistical question becomes: Is the observed

Table 8.3 *Data: Cholesterol Levels (*chol*) of* 40 *Men at High Risk of Coronary Disease Who Weigh More than* 225 *Pounds Classified by A/B Behavior Type*

	chol	A/B		*chol*	A/B		*chol*	A/B		*chol*	A/B		*chol*	A/B
1	344	B	9	246	B	17	224	A	25	242	B	33	252	A
2	233	A	10	224	B	18	239	A	26	252	B	34	202	A
3	291	A	11	212	B	19	239	A	27	153	B	35	218	A
4	312	A	12	188	B	20	254	A	28	183	B	36	202	B
5	185	B	13	250	B	21	169	B	29	234	A	37	212	A
6	250	A	14	197	A	22	226	B	30	137	B	38	325	A
7	263	B	15	148	B	23	175	B	31	181	A	39	194	B
8	246	A	16	268	A	24	276	A	32	248	A	40	213	B

mean difference likely due to chance alone? The two-sample Student's t-test yields a test statistic $t = 2.562$ with an associated p-value of $P(|t| \geq 2.562|$ *no difference*$) = 0.015$.

As described, the parallel nonparametric Wilcoxon rank test begins by combining both groups and replacing each of the $N = 40$ observed cholesterol values with its rank, making the analysis distribution-free. The test statistic becomes the sum of the ranks associated with one of the two groups (again denoted W). For the behavior-type data, the sum of the ranks of the cholesterol levels from the $n_1 = 20$ type A men is $W = \Sigma\ R_i = 503$. When no systematic difference exists between the mean cholesterol levels of type A and type B individuals, the expected sum of the ranks W for type A group individuals is $EW = n_1\ (N + 1)/2 = 20(41)/2 = 20(20.5) = 410$. The variance of the test statistic W is given by the expression

$$variance(W) = \frac{n_1 n_2 (N + 1)}{12} = \frac{20(20)(41)}{12} = 1366.67$$

where n_1 is the number of type A individuals and n_2 is the number of type B individuals ($N = n_1 + n_2 = 20 + 20 = 40$). The specific test statistic

$$z = \frac{W - n_1(N + 1)/2}{\sqrt{variance(W)}} = \frac{503 - 410}{\sqrt{1366.67}} = 2.516$$

has an approximate standard normal distribution when only random differences exist between type A and type B individuals. The p-value is $P(|W| \geq 503 \mid$ *no difference*$) = P(|Z| \geq 2.516 \mid$ *no difference*$) = 0.0119$. The parallel t-test produces almost the identical p-value of 0.0148 ($t = 2.562$).

The Mann-Whitney test statistic ($\hat{P}$) applied to the same data equally reflects differences in cholesterol levels and has the further interpretation as the probability that a randomly selected type B individual will have a lower cholesterol level than a randomly selected individual from the type A group. This probability is estimated by

$$\hat{P} = \frac{U}{n_1 n_2} = \frac{293}{20(20)} = 0.733$$

where

$$U = W - \frac{n_1(n_1 + 1)}{2} = 503 - 210 = 293.$$

Table 8.4 *Data: Responses* ($n = 165$) *to a Question on Life Expectancy and Reported Physical Activity (an Ordinal Variable)*

	Reported exercise				
	None	Occasional	Moderate	Strenuous	Total
Yes	7	7	8	20	42
No	25	34	32	32	123
Total	32	41	40	52	165

The probability estimated by $\hat{P}$ is

$$P(\text{cholesterol value of a random type B} < \textit{cholesterol value of a random type A}) = 0.733.$$

As noted, the Mann-Whitney approach based on the measure $\hat{P}$ yields the identical analytic result as the Wilcoxon rank test. Specifically, the test statistic is

$$z = \frac{\hat{P} - 0.5}{\sqrt{variance(\hat{P})}} = \frac{0.733 - 0.5}{\sqrt{0.00854}} = 2.516$$

and, as before, the p-value is 0.0119. Note that the $variance(\hat{P}) = variance(W)/(n_1 n_2)^2 = 1366.67/(20 \times 20)^2 = 0.00854$ because W and U differ by a constant value (Chapter 27).

Nonparametric Analysis of a 2 × *k* Table

The Mann-Whitney estimated probability $\hat{P}$ applied to a $2 \times k$ table effectively identifies differences between two groups and, in addition, applies to assessing differences created by ordinal variables. Consider again the answers to the question from a health attitudes survey: Do you think exercise increases the length of life? The yes or no responses from a sample of 165 individuals surveyed are also classified by the amount of physical exercise (none, occasional, moderate, and strenuous – an ordinal variable) reported by the respondents (Table 8.4 – repeat of Table 8.1).

The Mann-Whitney test statistic $\hat{P}$ is an estimate of the probability that a random person selected from group 2 (the no group) reports a lower level of exercise than a random person selected from group 1 (the yes group). This estimate is, as before, the count of the total number of individuals in the no group who reported lower levels of exercise than each member of the yes group divided by the total number of possible pairs. Individuals reporting the same level of exercise (same exercise category) are assumed to be uniformly distributed (Chapter 1). Consequently, within a category, half of the no individuals are assumed to exercise less than the yes individuals. In the context of a $2 \times k$ table, the Mann-Whitney test statistic $\hat{P}$ is occasionally called a *ridit value.*

Mechanically, for each level of physical activity, the Mann-Whitney u_i values are the total number of individuals from the no group in categories to the left of each of the individuals in the yes group plus one-half the number of the individuals in the same category ("tied" values). For example, in the moderate physical activity category, $25 + 34 = 59$ individuals (none and occasional) in the no group reported less physical activity than each of the eight individuals in the yes group (counts to the left). In addition, there are 32 individuals within the same category of physical activity. Therefore, $25 + 34 + \frac{1}{2}(32) = 75$ individuals in the no

group have a level of physical activity less than each of the eight individuals in the yes group who reported moderate levels of exercise. The total count is then $8(75) = 600$ individuals; that is, the Mann-Whitney u_i count is $u_3 = 8(75) = 600$ individuals with lower levels of exercise than the members of the moderate exercise group. The u_i counts for each of the four physical activity categories are (Table 8.4) the following:

$$\text{None}: u_1 = 7 \times \left[\frac{1}{2}(25)\right] = 87.5$$

$$\text{Occasional}: u_2 = 7 \times \left[25 + \frac{1}{2}(34)\right] = 294$$

$$\text{Moderate}: u_3 = 8 \times \left[25 + 34 + \frac{1}{2}(32)\right] = 600 \text{ and}$$

$$\text{Strenuous}: u_4 = 20 \times \left[25 + 34 + 32 + \frac{1}{2}(32)\right] = 2140.$$

Thus, the Mann-Whitney test statistic, as before, is the sum of these u_i values divided by the total number of pairs compared or

$$\hat{P} = \frac{U}{n_1 n_2} = \frac{\sum u_i}{n_1 n_2} = \frac{87.5 + 294 + 600 + 2140}{42(123)} = \frac{3121.5}{5166} = 0.604$$

where there are a total of $n_1(n_2) = 42(123) = 5166$ pairs of comparisons among the 165 surveyed individuals. Thus, the estimated probability that a randomly selected person from the no group has a lower level of physical activity than a randomly selected person from the yes group is 0.604.

In general notation, the Mann-Whitney probability $\hat{P} = \frac{\sum u_i}{n_1 n_2}$ estimated from a $2 \times k$ table is

$$\hat{p} = \frac{n_{11}\left[\frac{1}{2}n_{21}\right] + n_{12}\left[n_{21} + \frac{1}{2}n_{22}\right] + n_{13}\left[n_{21} + n_{22} + \frac{1}{2}n_{23}\right] + n_{14}\left[n_{21} + n_{22} + n_{23} + \frac{1}{2}n_{24}\right] + \cdots}{n_1 n_2}.$$

where, for the ith category,

$$u_i = n_{1i}\left[n_{21} + n_{22} + n_{23} + \cdots + \frac{1}{2}n_{2i}\right] \quad i = 1, 2, \ldots, k.$$

The Mann-Whitney test statistic $\hat{P}$ calculated from a $2 \times k$ table of counts has the same statistical properties already described. For the health survey data, the estimated probability of $\hat{P}$ is 0.604. The probability $P = 0.5$ is the expected probability when the level of reported physical exercise is exactly unrelated to the answer of the yes/no exercise question. The estimated variance of the approximate normal distribution of the estimate $\hat{P}$ is

$$variance(\hat{P}) = \frac{1}{(n_1 n_2)^2} variance(W) \approx \frac{1}{12n_1} + \frac{1}{12n_2} = \frac{1}{12(42)} + \frac{1}{12(123)} = 0.00266$$

making the test statistic

$$z = \frac{0.604 - 0.5}{\sqrt{0.00266}} = 2.020.$$

Table 8.5 *Notation: Cell Frequencies, Marginal Sums and Probabilities for a 2 × k Table*

	$X=1$	$X=2$	$X=3$	$X=4$		$X=k$	Total ($n_i.$)
$Y=1$	n_{11}	n_{12}	n_{13}	n_{14}	...	n_{1k}	$n_{1.}$
$Y=0$	n_{21}	n_{22}	n_{23}	n_{24}	...	n_{2k}	$n_{2.}$
Total ($n_{.j}$)	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	...	$n_{.k}$	n
p_j	p_1	p_2	p_3	p_4	...	p_k	1.0

The p-value is 0.043 and, in detail, is the probability P(more extreme $\hat{P}|P = 0.5) = P(|\hat{P}| \geq 0.604 \mid P = 0.5) = P(|Z| \geq 2.020 \mid \text{no association}) = 0.043$. The estimate $\hat{P} = 0.604$ is, therefore, unlikely to differ from 0.5 entirely because of random variation.

The normal distribution test statistic is understated because the estimate of the variance is biased (too large). This bias results from not accounting for the reduction in variability due to the "tied" observations (same category) that is unavoidable in a table. Corrections exist for this usually slight and conservative bias. As noted, the Mann-Whitney probability $\hat{P}$ is related to the γ-coefficient measuring association in a $2 \times k$ table. Specifically, the γ-coefficient $= \gamma = 2\hat{P} - 1$, therefore, $\gamma = 2(0.604) - 1 = 0.208$ (Chapter 5).

A Chi-Square Analysis of a 2 × *k* Table

A binary categorical variable (denoted Y) and a k-level numeric variable (denoted X) provide an opportunity to explore specific patterns of an association from data contained in a $2 \times k$ table, a pattern sometimes called a "dose response." Table 8.5 contains the notation for such a table.

Table 8.6 displays a binary variable, presence ($Y = chd = 1$) and absence ($Y =$ no $chd = 0$) of a coronary heart disease event and a meaningful numeric categorical variable, namely, seven categories ($k = 7$) of body weights of high-risk men (coded $X = 1, 2, \ldots, 7$). It is well established that coronary heart disease risk is associated with increasing body weight. The weight-risk pattern of association has not been as fully explored. A technique called a "test for trend" is frequently used to identify patterns in a $2 \times k$ table but applies only to categorical variables that have a linear numeric response. Such a test is more accurately referred to as a "test for linear trend."

It is important to note that any equally spaced values of X produce the identical analytic results. That is, for constant values A and B, the values $X' = AX + B$ yield the same results

Table 8.6 *Data: Coronary Heart Disease Events Classified by Seven Categories of Body Weight from a Study of High-Risk Men* ($n = 3153$) *– a 2 × 7 Table*

	Body weights categories (X)							
	$X=1$	$X=2$	$X=3$	$X=4$	$X=5$	$X=6$	$X=7$	Total
$chd = 1$	25	24	20	46	45	50	47	257
No $chd = 0$	417	426	327	493	350	398	485	2896
Total ($n_{.j}$)	442	450	347	539	395	448	532	3153
$\hat{p}_j$	0.057	0.053	0.058	0.085	0.114	0.112	0.088	0.082

(p-value) as the value X. For the example, the weight categories could be seven 30 pound intervals, and the analytic results would be unchanged (Table 8.6).

A test for linear trend begins by estimating a straight line from the data to represent the outcome-risk relationship. The estimated slope of this line (denoted $\hat{b}$) is compared to the horizontal line $b = 0$ as a measure of association. The comparison of the summary value $\hat{b}$ to the theoretical value $b = 0$ (no association) is a comparison of two straight lines. Therefore, to be a useful comparison, the probabilities p_j must have at least an approximately linear relationship to the categorical variable X (Table 8.6, last row). It is critical for any analysis based on a statistical summary that the summary value accurately characterizes the relationship under study. A linear summary of a nonlinear relationship likely produces misleading results.

A partitioning of the usual Pearson chi-square test statistic applied to a two-way table produces a useful evaluation of the utility of an estimated straight line as a summary measure of association. To describe this partition the following notation and expressions from a 2 × k table are used (Table 8.5):

Number of observed values in the ijth cell of the table: n_{ij}
Number of observed values in the jth column of the table: $n_{.j}$
Number of observed values in the ith row of the table: $n_{i.}$
Total number of observations in the table: $n = \Sigma\, n_{.j} = n_{1.} + n_{2.}$
Data estimated probability for the jth row category: $\hat{p}_j = n_{1j}/n_{.j}$
Linear model estimated probability for the jth row category: $\hat{p}_j$ and
Estimated proportion of observations $Y = 1$ in the table: $\hat{P} = \Sigma\, n_{1j}/n = n_{1.}/n$.

The partitioning of a chi-square test statistic starts with the simple relationship

$$\hat{P}_j - \hat{P} = (\hat{P}_j - \tilde{P}_j) + (\tilde{P}_j - \hat{P}) \quad j = 1, 2, \ldots, k = \text{number of categories}$$

where $\tilde{p}_j$ represents a linear model estimated probability for the jth category (details to be described). Squaring the elements of the partition, summing over the k categories, and dividing by $\hat{P}(1 - \hat{P})$ produces three approximate chi-square distributed summary statistics. The usual chi-square statistic

$$X^2 = \frac{\sum n_j(\hat{p}_j - \hat{P})^2}{\hat{P}(1 - \hat{P})}, \quad j = 1, 2, \ldots, k$$

measures the difference between the estimated probabilities and a horizontal line ($\hat{P}$). In addition, the chi-square statistic

$$X_L^2 = \frac{\sum n_j(\tilde{p}_j - \hat{P})^2}{\hat{P}(1 - \hat{P})}, \quad j = 1, 2, \ldots, k$$

measures the difference between the data estimated line and a horizontal line (L for linear), and

$$X_{NL}^2 = \frac{\sum n_j(\hat{p}_j - \tilde{p}_j)^2}{\hat{P}(1 - \hat{P})}, \quad j = 1, 2, \ldots, k$$

measures the difference between the probabilities estimated from the data and the corresponding probabilities from an estimated straight line (*NL* for nonlinearity). All three chi-square statistics measure different sources of variation among the estimated table probabilities; that is, these three chi-square statistics reflect the extent of specific sources of variation relative to the total variation $P(1 - P)$ (Chapter 1). Furthermore, the partition guarantees $X^2 = X_L^2 + X_{NL}^2$. The degrees of freedom associated with these chi-square test statistics follow the same pattern or *total* = *linear* + *nonlinear* or $df = k - 1 = 1 + (k - 2)$. The chi-square statistic X_{NL}^2 is an evaluation of the accuracy of a straight line as a summary measure of association, in symbols, $\tilde{p}_j$ versus $\hat{p}_j$. The chi-square statistic X_L^2 is a summary assessment of the difference between the data estimated straight line and the horizontal line $b = 0$ as a measure of the strength of association, in symbols, $\tilde{p}_j$ versus $\hat{P}$.

The line summarizing the x/y relationship is the typical ordinary least squares estimated regression line. Least squares estimation produces a line totally without conditions. It is simply a weighted average (Chapter 3). For data classified into a $2 \times k$ table, it is possible to succinctly calculate the summary values necessary to estimate the slope b:

$$\text{Mean values: } \bar{x}_1 = \frac{\sum n_{1j}x_j}{n_{1.}}, \bar{x}_2 = \frac{\sum n_{2j}x_j}{n_{2.}} \quad \text{and} \quad \bar{x} = \frac{n_{1.}\bar{x}_1 + n_{2.}\bar{x}_2}{n}$$

$$\text{Sum of squares: } SS_{xx} = \sum n_{.j}(x_j - \bar{x})^2, SS_{yy} = \frac{n_{1.}n_{2.}}{n} \quad \text{and} \quad SS_{xy} = (\bar{x}_1 - \bar{x}_2)SS_{yy}$$

(Chapter 5).

These expressions yield the least squares estimated line that becomes the source of the linear model estimated probabilities, denoted $\tilde{p}_j$.

Specifically, for the weight/*chd* data (Table 8.6),

$$\textit{slope} = \hat{b} = \frac{SS_{xy}}{SS_{xx}} = \frac{117.406}{13068.730} = 0.009 \quad \text{and}$$

$$\textit{intercept} = \hat{a} = \hat{P} - \hat{b}\bar{x} = 0.082 - 0.009(3.100) = 0.054[3].$$

The straight line estimated probabilities are $\tilde{p}_j = 0.054 + 0.009\, x_j$ (Figure 8.3, dots).

The observed probabilities ($\hat{p}_j$) and estimated probabilities ($\tilde{p}_j$) associated within the seven weight categories x_j are the following:

	Body weight categories (X)							
	$X = 1$	$X = 2$	$X = 3$	$X = 5$	$X = 5$	$X = 6$	$X = 7$	Mean values
$\hat{p}_j$	0.057	0.053	0.058	0.085	0.114	0.112	0.088	0.082
$\tilde{p}_j$	0.054	0.063	0.072	0.081	0.090	0.099	0.108	0.082

It is helpful to note that $\tilde{p}_j - \hat{P} = \hat{b}(x_j - \bar{x})$, and, therefore, the chi-square statistic X_L^2 directly reflects the influence of the slope of the line summarizing the risk-outcome relationship. For example, when the estimated line is exactly horizontal ($\hat{b} = 0$), then $X_L^2 = 0$.

The resulting partitioned chi-square test statistics are $X^2 = 22.490$ (degrees of freedom = 6), $X_L^2 = 14.088$ (degrees of freedom = 1), and $X_{NL}^2 = 8.402$ (degrees of freedom = 5). The respective p-values are <0.001, <0.001, and 0.135. Both the plot and moderately small

Table 8.7 *Data: Numbers of Cases (Childhood Cancer) and Controls* (Y) *Classified by Number of Maternal X-rays during Pregnancy* (X) *from the Oxford Survey of Childhood Cancer* ($n = 16{,}226$)

	Maternal x-ray exposures (X)						
	$X = 0$	$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X \geq 5$	Total
Case $= Y = 1$	7332	287	199	96	59	65	8038
Control $= Y = 0$	7673	239	154	65	28	29	8188
Total (n_j)	15,005	526	353	161	87	94	16,226
$\hat{p}_j$	0.489	0.546	0.564	0.596	0.678	0.691	0.495

p-value of 0.135 suggest that a linear relationship is marginally successful as a summary of the weight-risk relationship; that is, the chi-square analysis indicates strong evidence of a nonrandom association ($X^2 = 22.490$) but marginal evidence of a linear relationship between body weight and the probability of a coronary disease event (Table 8.6), making the assumption of linearity suspect.

Another Example: Childhood Cancer and X-Ray Exposure

Data from a case/control study (Oxford Survey of Childhood Cancer) provide another example of an analysis of a $2 \times k$ table using a linear regression model to assess the association between a binary variable and a k-level categorical variable (Table 8.7 and Figure 8.4). These data were collected to study the influence of prenatal maternal x-ray exposure on the risk of childhood cancer.

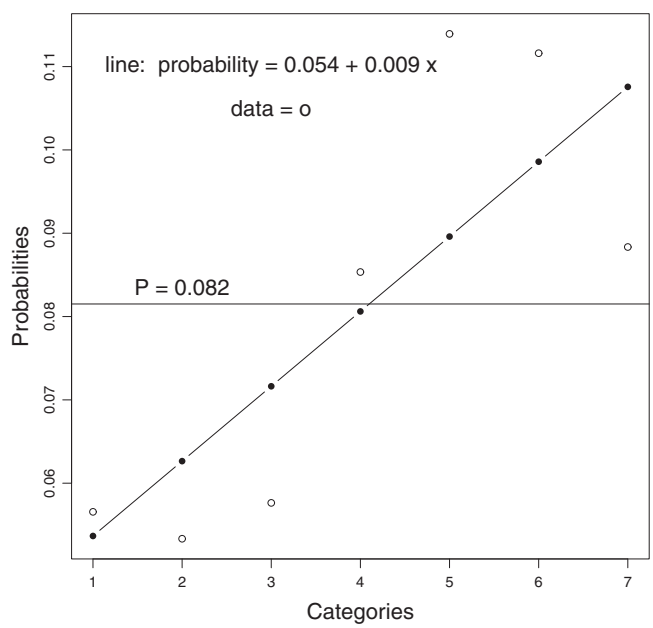


Figure 8.3 Data and Estimated Line: Estimated (Dots) and Observed (Circles) Probabilities of a Coronary Event for Seven Levels of Body Weight from High-Risk Study Subjects ($n = 3153$)

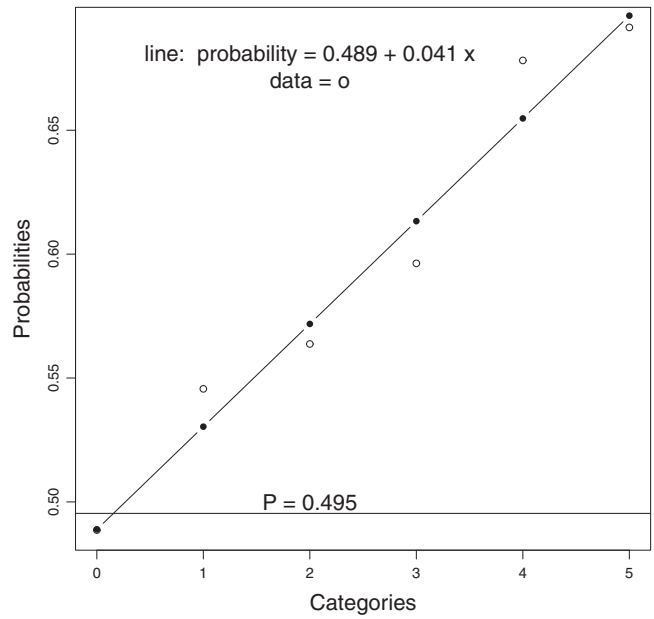


Figure 8.4 Six Categories of Case/Control Data Describing X-Ray Exposure from a Study of Childhood Cancer ($n = 16{,}226$)

The same chi-square partition applied to the weight-*chd* data (Table 8.6) applies to the assessment of risk from maternal x-rays potentially providing a summary of response in terms of a straight line. The previous computational expressions produce a least squares estimated straight line and linear model estimated probabilities $\tilde{p}_j$ for each of six x-ray exposure levels. Specifically, for the x-ray data,

$$slope = \hat{b} = \frac{SS_{xy}}{SS_{xx}} = \frac{279.208}{6733.580} = 0.041$$

with

$$intercept = \hat{a} = \hat{P} - \hat{b}\bar{x} = 0.459 - 0.041(0.156) = 0.489.$$

The estimated line is $\tilde{p}_j = 0.489 + 0.041x_j$ (Figure 8.4). The six observed ($\hat{p}_j$) and estimated probabilities ($\tilde{p}_j$) are the following:

	Maternal x-ray exposures (X)						
	$X = 0$	$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X \geq 5$	Mean values
$\hat{p}_j$	0.489	0.546	0.564	0.596	0.678	0.691	0.495
$\tilde{p}_j$	0.489	0.530	0.572	0.613	0.655	0.696	0.495

The resulting partitioned chi-squares test statistics are $X^2 = 47.286$ (degrees of freedom = 5), $X^2_L = 46.313$ (degrees of freedom = 1), and $X^2_{NL} = 0.972$ (degrees of freedom = 4). The respective p-values are <0.001, <0.001, and 0.912. Both the plot and the extremely small chi-square value X^2_{NL} (large p-value = 0.912) indicate that linear model is a highly

Table 8.8 *Summary: Partitioned Chi-Square Statistic – Two Examples*

			Body weight			X-ray exposure		
	Chi-square statistics	df^a	df^a	X^2	p-values	df^a	X^2	p-values
Linear	$X_L^2 = \dfrac{\sum n_{.j}(\tilde{p}_j - \hat{P})^2}{\hat{P}(1-\hat{P})}$	1	1	14.088	<0.001	1	46.313	<0.001
Nonlinear	$X_{NL}^2 = \dfrac{\sum n_{.j}(\hat{p}_j - \tilde{p}_j)^2}{\hat{P}(1-\hat{P})}$	$k-2$	5	8.402	0.135	4	0.972	0.914
Overall	$X^2 = \dfrac{\sum n_{.j}(\tilde{p}_j - \hat{P})^2}{\hat{P}(1-\hat{P})}$	$k-1$	6	22.490	<0.001	5	47.286	<0.001

[a] df = degrees of freedom and $j = 1, 2, \ldots, k$ = number of categories.

successful summary of the x-ray-risk relationship. More specifically, the linear model estimated probabilities $\tilde{p}_i$ essentially replicate the data estimated probabilities $\hat{p}_i$ (Figure 8.4). The analysis, therefore, leaves little doubt that a linear relationship accurately describes increasing risk of childhood cancer associated with increasing x-ray exposure. Identification of this relationship from these data brought a worldwide halt to use of maternal x-rays as a diagnostic tool during pregnancy (1950). This obviously important contribution to medical practice is primarily due to British epidemiologist Alice Stewart (b. 1906).

Difference in mean number of maternal x-rays between cases and controls equally reflects the association between x-ray exposure and childhood cancer. The mean number of x-rays among mothers of cases is $\bar{x}_{cases} = 0.191$ and among mothers of controls is $\bar{x}_{controls} = 0.122$. The larger value among cases is consistent with the previous chi-square analysis, indicating increased childhood cancer risk associated with x-ray exposure. In fact, a formal statistical evaluation of the mean difference exactly duplicates the previous chi-square result.

To assess the influence from random variation on the observed mean difference $\bar{x}_{cases} - \bar{x}_{controls} = 0.069$, an expression for the variance of the distribution of the mean difference is required. An estimate calculated from data contained a $2 \times k$ table is

$$\hat{V} = variance(\bar{x}_{cases} - \bar{x}_{controls}) = \frac{S_X^2}{n}\left[\frac{1}{n_{.1}} + \frac{1}{n_{.2}}\right].$$

The variance of the number of x-rays is estimated under the assumption that the variability associated with both case ($Y = 1$) and control ($Y = 0$) distributions of x-ray exposures is identical. A two-sample test statistic becomes

$$z = \frac{(\bar{x}_{cases} - \bar{x}_{controls}) - 0}{\sqrt{\hat{V}}}$$

and has an approximate standard normal distribution when no case/control differences exist. Specifically, from the x-ray data, the value of the test statistic is

$$z = \frac{0.191 - 0.122}{\sqrt{\dfrac{6733.58}{16226}\left[\dfrac{1}{8038} + \dfrac{1}{8188}\right]}} = \frac{0.191 - 0.122}{0.0101} = 6.805.$$

This apparent different analytic approach is identical to the previously generated chi-square value ($X_L^2 = 46.313 = z^2 = [6.805]^2$). A little algebra shows that these two approaches are the same for any $2 \times k$ table. Therefore, whether the assessment of association is made among x-ray categories (columns) with a chi-square statistic or a comparison between mean number of x-rays (rows), the answers are identical. The body weight and x-ray analyses are summarized in Table 8.8.

9

The Loglinear Poisson Regression Model

Regression analyses come in a variety of forms. They can be parametric, nonparametric, or even semiparametric. A parametric analysis starts with choosing a model. Two fundamental considerations are the kind of outcome to be analyzed and summary statistics desired. When the outcome to be analyzed is a continuous value the model often chosen is a *linear model* (Chapter 7). The model coefficients measure change in the dependent variable for a one-unit change in an independent variable. When the outcome to be analyzed is a 0 or 1 binary value, the model often chosen is a *logistic model.* The model coefficients measure change in the log-odds for a one-unit change in an independent variable. When the data to be analyzed are counts, probabilities, or rates, the model often chosen is a *loglinear Poisson model.* The model coefficients reflect the change in a logarithm of a count for a one-unit change in the independent variable.

A specific expression for a loglinear additive Poisson model is

$$log(y) = log(\text{count}) = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

where y represents a count. Like all additive models, when

$$log(y') = a + b_1x_1 + b_2x_2 + \cdots + b_i(x_i + 1) + \cdots + b_kx_k$$

and

$$log(y) = a + b_1x_1 + b_2\mathrm{x}_2 + \cdots + b_ix_i + \cdots + b_kx_k$$

then, for a one-unit change in the variable x_i,

$$log(y') - log(y) = b_i \quad \text{or} \quad \frac{y'}{y} = e^{b_i}.$$

Thus, the model coefficients measure influence in terms of a ratio. Also important, like all additive models, the influence from each independent variable is summarized by a single ratio unaffected by the values of other variables in the model. The separation of the influence of a multivariate observation into multiplicative components is a principal feature of an additive loglinear analysis.

A Simple Poisson Regression Model

A simple linear regression model

$$y_x = a + bx$$

Table 9.1 *Data: Number of Cases, Number of Births and Rates of Autism per* 10,000 *Live Births by Year (California – 1987 to 1995)*

Years	Cases	Births	Rates[a]
1987	202	504,853	4.00
1988	255	534,174	4.77
1989	333	570,976	5.83
1990	442	613,336	7.21
1991	571	610,701	9.35
1992	716	602,269	11.89
1993	803	585,761	13.71
1994	996	568,263	17.53
1995	1323	577,113	22.92

[a]Rate of autism cases per 10,000 live births.

is called simple because the dependent variable y is related only to a single independent variable x. The coefficient b measures the linear change in the dependent variable y for a one-unit change in the independent variable x, summarizing the relationship between y and x as a straight line.

A simple Poisson regression model is

$$y_x = ae^{bx}$$

where the coefficient b measures the multiplicative change in the variable y for a one-unit change in the variable x, that is, the dependent value $y_{x+1} = ae^{b(x+1)} = ab^x \times e^b = y_x \times e^b$. For example, if $b = 0.3$, then $e^{0.3} = 1.35$ and a one-unit change in the variable x produces a change of 1.35 times the original value y_x or a ratio of $y_{x+1}/y_x = e^b = 1.35$. When the dependent variable y_x is transformed to $log(y_x)$, the Poisson regression model becomes the simple loglinear model

$$log(y_x) = log(a) + bx.$$

When y_x represents a count, a probability, or a rate, the coefficient b then measures the linear change in logarithm of the dependent variable y from a one-unit change in the independent variable x.

The loglinear Poisson model has a variety of useful properties. As always, linear relationships are intuitive and familiar. Logarithms transform multiplicative relationships into additive relationships that are often accurately assessed using a normal probability distribution. The model assumptions and estimation of parameters closely resemble those of a linear regression model yielding a familiar pattern of evaluation. The accuracy of a loglinear model is directly evaluated with a chi-square statistic. Finally, an additive loglinear model provides estimated ratio measures of the separate influences from each independent variable, obtained by exponentiating the model coefficients.

Data from the state of California indicate a remarkable increase in the risk of autism for the period 1987 to 1995 (Table 9.1). Rates and log rates of autism are displayed in Figure 9.1 (dots = observed rates and log rates).

Table 9.2 *Data/Estimates: Model Estimated and Observed Rates per* 10,000 *Live Births and Number of Cases of Autism by Year (California – 1987 to 1995)*

Years	Data rates[a]*	Model rates[a]	Data cases	Model cases
1987	4.00	3.81	202	192.29
1988	4.77	4.75	255	253.80
1989	5.83	5.93	333	338.41
1990	7.21	7.39	442	453.46
1991	9.35	9.22	571	563.24
1992	11.89	11.50	716	692.91
1993	13.71	14.35	803	840.67
1994	17.53	17.90	996	1017.36
1995	22.92	22.33	1323	1288.87

[a]Rate of autism cases per 10,000 live births.

A simple loglinear Poisson regression model estimated from these autism data is

$$log(rate) = \hat{a} + \hat{b}(\text{year}) = -447.2 + 0.221(\text{year}).$$

For example, the estimate $log(rate_{1987}) = -447.2 + 0.221(1987) = -7.873$. Therefore, the model-estimated rate of autism for the year 1987 is $rate_{1987} = e^{-7.873} \times 10{,}000 = 3.809$ cases per 10,000 live births. The observed rates and number of cases contrasted to model-estimated rates and number of cases show an exceptionally close correspondence between data and model for all years (Table 9.2 and Figure 9.1). For example, over the eight years 1987 to 1995, the rate of autism increased by a model-estimated factor of 1.247 ($e^{0.221}$) each year. Thus, the model-estimated rate ratio for the eight-year period is

$$\text{rate ratio} = \hat{rr} = \frac{\hat{r}_{1995}}{\hat{r}_{1987}} = (e^{\hat{b}})^8 = (e^{0.221})^8 = (1.247)^8 = 5.864.$$

Alternatively, the identical estimated rate ratio is

$$\hat{rr} = \frac{\hat{r}_{1995}}{\hat{r}_{1987}} = \frac{22.333}{3.809} = 5.864$$

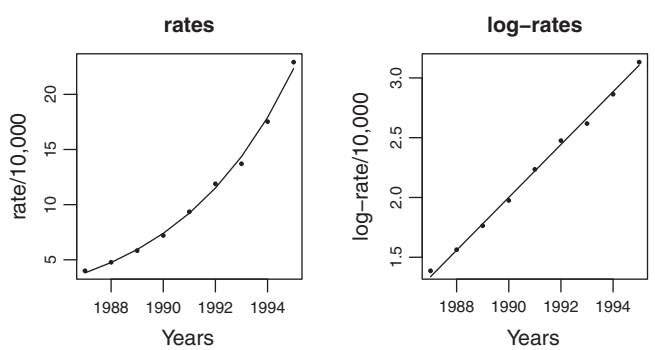


Figure 9.1 Rates and Log Rates of Autism per 10,000 Live Births (California – 1987 to 1995)

Table 9.3 *Data: Rates, Cases, and Ratios from Stomach Cancer Mortality Data by Age and Sex per* 10,0000 *Persons At-Risk (California – 1999–2008)*

	Males			Females			
Ages	Deaths	Populations	Rates[a]	Deaths	Populations	Rates[a]	Rate ratios
45–54	84	2,014,272	4.17	52	1,967,034	2.64	0.634
55–64	136	1,417,097	9.60	69	1,474,064	4.68	0.488
65–74	131	775,192	16.90	77	891,864	8.63	0.511
75–84	154	466,018	33.05	121	642,748	18.83	0.570
≥85	76	161,990	46.92	92	325,940	28.23	0.602
Total	581	4,834,569	12.02	411	5,301,650	7.75	0.645

[a] R_i^{male} and R_i^{female} are rates per 100,000 persons at-risk (ith age strata).

calculated directly from the model-estimated rates (Table 9.2 – second column). The observed rate ratio is $22.924/4.001 = 5.729$.

It is clear from the plots that the loglinear model is an accurate representation of the trend in autism risk. Nevertheless, a formal assessment of model accuracy ("fit") is achieved by a direct comparison of observed number of cases to model-estimated number of cases (column 4 (o_i) versus column 5 (e_i); see Table 9.2). In general, like a linear regression model (Chapters 7 and 12), the model-estimated counts from a loglinear Poisson model directly compared to the corresponding observed counts produce a straightforward and simply interpreted evaluation of model accuracy (Chapter 12). The usual comparison $X^2 = \Sigma(o_i - e_i)^2/e_i = 4.790$ has a chi-square distribution (degrees of freedom = 7) when the differences between observed and model-estimated numbers of autism cases are due entirely to random variation. The associated p-value is 0.686. In general, the degrees of freedom for a chi-square comparison of model generated estimates to observed values are the total number of observations minus the number of model parameters estimated. Thus, the nine autism rates summarized with a two-parameter model produce the degrees of freedom $= 9 - 2 = 7$.

Poisson Regression Model: Analysis of Vital Statistics Data

A rate ratio is a statistic frequently used to compare rates between two groups (Chapter 3). For example, the male to female age-specific stomach cancer mortality rates, denoted R_i^{male} and R_i^{female}, can be compared as a ratio (Table 9.3).

Selecting the female age-specific counts as a "standard population" (denoted P_i, column 6), two summary rates are (Chapters 3 and 16)

$$\text{female rate} = \bar{R}_{female} = \sum P_i R_i^{female} \Big/ \sum P_i = 7.752 \text{ per } 100{,}000 \text{ persons at-risk}$$

and

$$\text{adjusted male rate} = \bar{R}_{male} = \sum P_i R_i^{male} \Big/ \sum P_i = 13.949 \text{ per } 100{,}000 \text{ persons at-risk}$$

Table 9.4 *Model Results: Loglinear Model Analysis of Stomach Cancer Mortality by Sex and Age (California – 1999–2008)*

	Coefficients	Estimates	*s. e.*	*p*-values
Intercept	$\hat{a}$	−12.704	–	–
Age	$\hat{b}_1$	0.061	0.002	<0.001
Sex	$\hat{b}_2$	−0.599	0.065	<0.001
$-2 \times$ log-likelihood = 9.047				

The age-adjusted male rate $\bar{R}_{male}$ is the rate expected among males as if their age distribution is exactly the same as the female age distribution, P_i. The estimated summary rate ratio is

$$rate\ ratio = standard\ mortality\ ratio = \frac{\bar{R}_{female}}{\bar{R}_{male}} = \frac{7.752}{13.949} = 0.556.$$

Note that $\bar{R}_{male}/\bar{R}_{female} = 1/0.556 = 1.799$ is the identical summary of stomach cancer mortality cancer risk (Chapter 6). A single ratio accurately summarizes the female-male difference in mortality risk when the underlying age-specific rate ratio estimated is a single value. That is, the female-male age-specific ratios differ only because of random variation (Table 9.3, last column).

The loglinear Poisson regression model can be extended to incorporate other issues making the estimated rate ratio a more sophisticated contrast of cancer risk. Of critical importance, the Poisson model allows an evaluation of the homogeneity of the age-specific risk ratios. When each age-specific ratio does not estimate the same underlying value, a model approach likely indicates that a single summary value is misleading.

An additive loglinear Poisson regression model applied to estimate an age-adjusted female-male rate ratio is

$$log(rate) = a + b_1\, age + b_2\, sex$$

where variable $sex = \{\text{male} = 1 \text{ and female} = 0\}$ and variable age is assigned values $age = \{45, 55, 65, 75, \text{and } 85 \text{ years}\}$. Additivity of the model produces the female-male rate ratio that is the same for all age groups. Table 9.4 contains the regression analysis results from the stomach cancer mortality data (Table 9.3).

The model-estimated age-adjusted rate ratio summarizing the female-male risk is $e^{\hat{b}_2} = e^{-0.599} = 0.549$ (model-free = 0.556). The all-important assessment of the accuracy of the additive model produces a chi-square distributed test statistic of $X^2 = 8.947$ (degrees of freedom = 10 − 3 = 7) with an associated p-value = 0.255, based on observed (o_i) and model-estimated (e_i) number of deaths. Specifically, the direct comparison of counts of deaths is the following:

	Males					Females				
Observed (o_i)	84	136	131	154	76	52	69	77	121	92
Estimated (e_i)	96.8	125.7	127.0	141.0	90.5	51.9	71.9	80.3	106.9	100.1

Table 9.5 *Data: Mortality from Stomach Cancer among White and African American Young Adults Ages* 25 *to* 44 *for Females and Males (California – 1999–2008)*

	Ages	Deaths	Populations	Rates[a]
White male	25–34	44	16,569,463	2.66
	35–44	162	17,170,525	9.43
White female	25–34	44	16,652,229	2.64
	35–44	135	18,283,791	7.38
African American male	25–34	22	2,829,377	7.78
	35–44	44	2,644,142	16.64
African American female	25–34	12	2,981,382	4.02
	35–44	58	2,975,071	19.50

[a]Rate of death per 1,000,000 persons at-risk.

Unlike the model-free estimate, the direct evaluation of the model accuracy provides some assurance of the homogeneity of the rate ratios.

Rate Ratios Adjusted for Several Variables

The loglinear Poisson model is directly extended to a multivariable description of risk. Stomach cancer mortality data in young adults (ages 25–44) classified by age, sex, and ethnicity illustrate (Table 9.5).

A relevant loglinear Poisson additive model is

$$log(rate) = a + b_1\, age + b_2\, sex + b_3\, ethnicity.$$

The variable *age* is binary (age 25–34 = 0 and age 35–44 = 1), the variable sex is binary (male = 1 and female = 0), and the variable *ethnicity* is binary (white = 0 and African-American = 1). Assuming age-specific numbers of deaths have at least approximate Poisson distributions, the additive model parameters provide estimates of the influences of three measures of cancer risk (Table 9.6). For another point of view, the Poisson model describes data distributed into a $2 \times 2 \times 2$ table, and analysis identifies the separate influences of the three binary variables on stomach cancer mortality.

A first issue is the accuracy of an additive model.

Table 9.6 *Model Results: Estimates of the Influences (Coefficients) on the Rate of Stomach Cancer Mortality among Young Adults from Age, Sex, and Ethnicity (California – 1999–2008)*

	Coefficients	Estimates	*s. e.*	*p*-values
Intercept	$\hat{a}$	−12.764	–	–
Age	$\hat{b}_1$	1.148	0.103	<0.001
Sex	$\hat{b}_2$	−0.152	0.088	0.084
Ethnicity	$\hat{b}_3$	0.780	0.100	<0.001
$-2 \times$ log-likelihood = 5.629				

Table 9.7 *Results: The Model Estimated and Observed Number of Deaths and Mortality Rates*

	Ages	Data deaths	Model deaths	Data rates[a]	Model rates[a]
White male	25–34	44	47.4	2.7	2.9
	35–44	162	154.9	9.4	9.0
White female	25–34	44	40.9	2.6	2.5
	35–44	135	141.8	7.4	7.8
African American male	25–34	22	17.7	7.8	6.2
	35–44	44	52.0	16.6	19.7
African American female	25–34	12	16.0	4.0	5.4
	35–44	58	50.3	19.5	16.9

[a]Rates per 1,000,000 persons at-risk.

The data and model-estimated values again allow a typical chi-square comparison between observed and model-estimated numbers of deaths (Table 9.7). The approximate chi-square distributed test statistic $X^2 = 5.698$ (degrees of freedom $= 8 - 4 = 4$) yields a p-value $=$ 0.231. As usual, the log-likelihood value of 5.629 is similar (Chapter 10).

The specific rate ratios estimated from the model, $\hat{rr}_i = e^{b_i}$ ($i = 1, 2,$ and 3), summarize the separate influences from age, sex, and ethnicity on stomach cancer mortality. The estimated variance of the distribution of the logarithm of each estimated rate ratio is given by the expression

$$variance(\hat{rr}) = \frac{1}{d_1} + \frac{1}{d_2}$$

where d_1 represents the number of deaths in one group and d_2 in the other (Chapter 27). Essentially the same variance estimates are produced as part of the maximum likelihood estimation of the loglinear regression model coefficients (Table 9.6).

The estimated ratio ratios and their approximate 95% confidence intervals for the three risk variables are the following:

$$\text{age: rate ratio} = \hat{rr}_{age} = e^{1.148} = 3.153$$

$$variance[log(\hat{rr})] = \frac{1}{122} + \frac{1}{399} = 0.011, \text{ then}$$

$$95\% \text{ confidence interval} = e^{1.148 \pm 1.960(0.103)} \rightarrow (2.574, 3.862);$$

$$\text{sex: rate ratio} = \hat{rr}_{sex} = e^{-0.152} = 0.859$$

$$variance[log(\hat{rr})] = \frac{1}{272} + \frac{1}{249} = 0.008, \text{ then}$$

$$95\% \text{ confidence interval} = e^{-0.152 \pm 1.960(0.088)} \rightarrow (0.724, 1.021);$$

$$\text{ethnicity: rate ratio} = \hat{rr}_{ethnicity} = e^{0.780} = 2.181$$

$$variance[log(\hat{rr})] = \frac{1}{136} + \frac{1}{385} = 0.010, \text{ then}$$

$$95\% \text{ confidence interval} = e^{0.780 \pm 1.960(0.100)} \rightarrow (1.794, 2.653).$$

Table 9.8 *Data: Counts and Probabilities of Coronary Heart Disease Events at Four Levels of Smoking Exposure among High-Risk White Males* ($n = 3154$) *(Chapter 8)*

	Cigarettes per day				
	0	1–20	21–30	>30	Total
chd	98	70	50	39	257
No *chd*	1554	735	355	253	2897
n_x	1652	805	405	292	3154
p_x	0.059	0.087	0.123	0.134	0.081

Poisson model-estimated rate ratios and multivariable analysis of stomach cancer mortality data illustrate the general strategy of applying Poisson models to effectively analyze and describe data contained in multilevel tables with the added benefit of an accompanying goodness-of-fit assessment of homogeneity (additivity) of the summary values (Chapter 10).

A Test for Linear Trend

Data classified into numeric categories of a $2 \times k$ table are frequently used to explore the pattern of response (Chapter 8). Data describing the risk from reported smoking exposure and the occurrence of a coronary heart disease event (*chd*) illustrate an application of a Poisson regression model to identify a specific dose-response relationship (Table 9.8). More technically, the relationship is an exposure-response relationship.

The statistical question is: Do constant and multiplicative increasing probabilities of a coronary event (denoted p_x) accurately describe the risk from each cigarette smoked (denoted x)? This question translates into an assessment of the linear pattern of the logarithms of the probabilities of a *chd* event. In symbols, the statistical question is: do the values $\log(p_{x+1}) - log(p_x) = b$ or $p_{x+1}/p_x = e^b$ accurately summarize the pattern of smoking on the risk of a *chd* event with a single value? A loglinear model to explore the utility of summarizing smoking influence as a constant multiplicative risk is

$$log(p_x) = a + bx$$

where $x = \{0, 10, 20, \text{and } 30 \text{ cigarettes per day}\}$ represent four levels of exposure.

Thus, the model estimates $\hat{a} = -2.785$ and $\hat{b} = 0.0294$ describe the linear increase in the logarithm of the variable p_x per cigarette smoked (Table 9.9). For example, the Poisson model parameters produce the estimated risk ratio associated with smoking of 10 cigarettes per day of

$$\hat{rr} = \frac{P(chd|x+10)}{P(chd|x)} = \frac{e^{\hat{a}+\hat{b}(x+10)}}{e^{\hat{a}+\hat{b}x}} = e^{\hat{b}(10)} = e^{0.0294(10)} = 1.341.$$

The model-estimated logarithms of the probabilities of a *chd* event $log(p_x)$ are required to have an exactly linear increasing pattern of risk associated with increasing cigarette exposure, which appears to be a reasonable summary of the observed exposure-response relationship (Figure 9.2 and Table 9.10).

Table 9.9 *Model Results: The Estimates of the Poisson Model Parameters for the* 2×4 *Table of Coronary Events by Smoking Exposure*

	Estimates	*s. e.*	*p*-value
$\hat{a}$	−2.785	–	–
$\hat{b}$	0.0294	0.0056	<0.001
$-2 \times$ log-likelihood = 1.345			

Table 9.10 *Model Results: Estimates from Coronary Heart Disease Data at Four Levels of Smoking Exposures for High-Risk White Males* ($n = 3154$)

	Cigarettes per day[a]				
	0	1–20	21–30	>30	Model estimates ($x = 10$)
Estimate: $log(\hat{p}_x)$	−2.785	−2.492	−2.198	−1.904	Difference = 0.294
Estimate: $\tilde{p}_x$	0.062	0.083	0.111	0.149	Ratio = 1.341
Observed: $\hat{p}_x$	0.059	0.087	0.123	0.134	–

[a]Exposures = {0, 10, 20, and 30} cigarettes per day and $\hat{p}_x = -2.785 + 0.0294$.

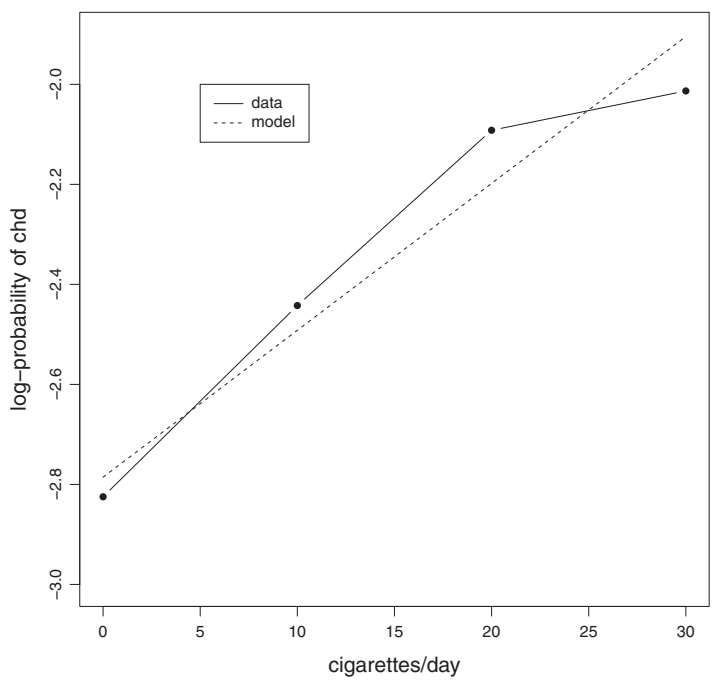


Figure 9.2 Coronary Risk and Exposure: Model Estimate (Dashed Line) and Data (Dots) from Coronary Heart Disease Events at Four Levels of Smoking Exposure for High-Risk White Males ($n = 3154$)

Table 9.11 *Data: A* 3 × 4 *Table of the Counts*[a] *of Newborn Infants with Birth Defects Classified by Maternal Ethnicity and Vitamin Use* ($n = 411$)

	White	African american	Hispanic	Asian	Total
Always	5 (55.4)	14 (14.6)	22 (21.0)	39 (34.1)	125
During	84 (72.6)	18 (19.2)	30 (27.5)	32 (44.7)	164
Never	48 (54.0)	16 (14.2)	17 (20.5)	41 (33.2)	122
Total	182	48	69	112	411

[a]Values expected (parentheses) when vitamin use and ethnicity are unrelated.

As always, despite close correspondence between model-estimated and observed values as well as a moderately large p-value = 0.497 (chi-square value = 1.395 degrees of freedom = 2) it remains a subject matter decision whether estimated model values are a sufficiently accurate reflection of the smoking-*chd* relationship. Failure to statistically identify nonadditivity is only evidence of the adequacy of an additive model and is not a guarantee of an underlying additive relationship.

A Graphical Display of a Table

Plots of logarithms of table frequencies provide frequently useful visual descriptions of the relationships among the categorical variables used to create a table. Additivity translates geometrically into parallel lines (Chapter 10). Therefore, departures from parallel indicate the degree and pattern of association. From a more technical point of view, such plots can display the failure of the relationships among the rows of a table to be the same for all columns and vice versa. In model terms, a plot displays the magnitude and type of interaction (nonadditivity) between categorical variables (Chapters 7 and 10).

In a study of diet and risk of a birth defect, it was important to describe the relationship between vitamin use and ethnicity. Data that address this issue are from 411 mothers whose newborn infant has a birth defect. Ethnicity is self-reported and use of vitamin supplements is classified as always used, used only during pregnancy, and never used (Table 9.11).

A typical Pearson chi-square evaluation of the association between vitamin use and ethnicity produces a test statistic of $X^2 = 10.271$ (degrees of freedom = 6) with associated p-value of 0.114 (Chapter 10). Such a statistical assessment gives no hint of the properties or patterns of association. To further explore the association of vitamin use and ethnicity, a plot displays the table of observed log frequencies, and including additive model-estimated log frequencies from a Poisson additive regression model provides a detailed visual comparison.

Figure 9.3 is a plot of the logarithms of the cell frequencies (Table 9.12, solid lines) and the same *log*-frequencies estimated from the additive model (Table 9.12, parallel dashed lines).

The plot identifies that the nonadditivity causing a modest association between vitamin use and ethnicity is largely due to a different pattern of vitamin use among Asian mothers. Otherwise, the ethnicity categories show little evidence of an association (solid versus dashed lines).

Table 9.12 *Data: A* 3 × 4 *Table of Logarithm of Counts*[a] *of Newborn Infants with Birth Defects by Maternal Ethnicity and Vitamin Use* ($n = 411$)

	Ethnicity			
	White	African american	Hispanic	Asian
Always	3.91 (4.01)	2.64 (2.68)	3.09 (3.04)	3.66 (3.53)
During	4.43 (4.29)	2.89 (2.95)	3.40 (3.32)	3.47 (3.80)
Never	3.87 (3.99)	2.77 (2.66)	2.83 (3.02)	3.71 (3.50)

[a]Values expected (parentheses) when vitamin use and ethnicity are unrelated.

Implementation of Poisson Models Applied to Tabular Data

Usually the explanation of statistical software systems and computer code is best left to manuals, computer help routines, or Google queries. In the case of loglinear models, three basic issues become evident from the construction of a computer routine. Popular computer applications (SAS, STATA, and R) produce the identical model-estimated values and likelihood values. For example, log-likelihood value from the three computer software systems is $G^2 = 8.839$ estimated from a small set of artificial data (Table 9.13 and 9.14).

It should be noted that computer-estimated model coefficients are not always identical. There are several ways to estimate model parameters that best represent the data, and statistical software occasionally apply different approaches. The choice has no consequences because different computational strategies produce identical statistical results.

The computer software requires the number of counts in each cell of the table and a method to identify the cell (Table 9.14).

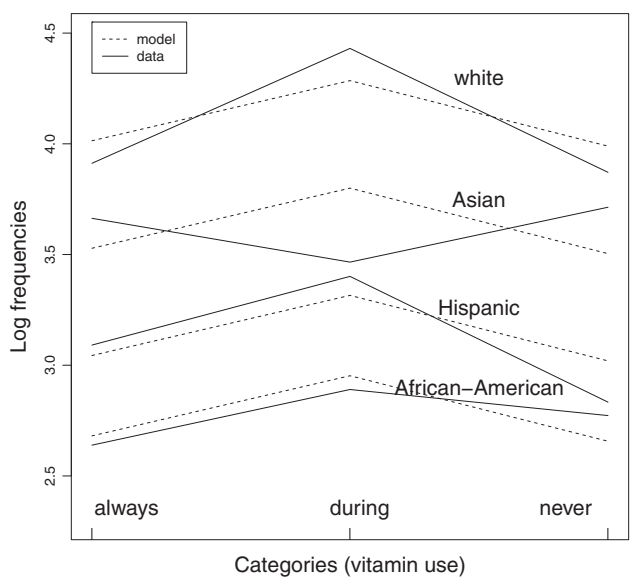


Figure 9.3 *Log*-Frequencies of Birth Defects from the 3 × 4 Table of Vitamin Use by Ethnicity ($n = 411$)

Table 9.13 *Example: "Data" to Illustrate Computer Application*

	A_0	A_1	A_2
Table 0			
B_0	1	6	5
B_1	4	3	8
Table 1			
B_0	2	12	12
B_1	7	6	16

Table 9.14 *Computer Code: Three Major Statistical Analysis Systems Used to Apply Loglinear Models to Tables of Categorical Data (Table 9.13)*

STATA

```
Clear
*data
input rr cc ll n
0 0 0 1
1 0 0 4
0 1 0 6
1 1 0 3
0 2 0 5
1 2 0 8
0 0 1 2
1 0 1 7
0 1 1 1 2
1 1 1 6
0 2 1 1 2
1 2 1 1 6
End

xi:poisson n i.rr i.cc i.ll
predict N

list n N
```

SAS

```
data temp;
input rr cc ll n;
datalines;
0 0 0 1
1 0 0 4
0 1 0 6
1 1 0 3
0 2 0 5
1 2 0 8
0 0 1 2
1 0 1 7
0 1 1 1 2
1 1 1 6
0 2 1 1 2
1 2 1 1 6
run;

proc genmod data = temp; class rr cc ll;
model n = rr cc ll/dist = poisson
link = log;
output out = pred pred = pred_n;
run;
proc print data = pred;
run;
```

R

```
#data
n  ← c(1,4,6,3,5,8,2,7,12,6,12,16)
rr ← factor(c(0,1,0,1,0,1,0,1,0,1,0,1))
cc ← factor(c(0,0,1,1,2,2,0,0,1,1,2,2))
ll ← factor(c(0,0,0,0,0,0,1,1,1,1,1,1))
f  ← glm(n~rr+cc+ll, family = poisson)
summary (f)
cbind(n, fitted(f))
```

Table 9.15 *Notation: An Additive Loglinear Additive Model of Incomplete Data Generated from Two Sources*[a] *of Counts (Denoted S_1 and S_2)*

Frequency	S_2	*Not* S_2
S_1	n_{11}	n_{12}
Not S_1	n_{21}	$n_{22} = ?$
Log frequency	S_2	*not* S_2
S_1	l_{11}	l_{12}
Not S_1	l_{21}	$l_{22} = ?$
Model	S_2	*not* S_2
S_1	μ	$\mu + a$
Not S_1	$\mu + b$	$[\mu + a + b]$

[a] S_i represents identified by source i and *not* S_i represents not identified by source i.

For example, the cell count of three located in column 2, row 2 in table 0 is coded (1, 1, 0). Therefore, no specific order is required of the table counts. Of more importance, it is not required that data from all cells be available.

Situations arise where specific combinations of the categorical variables never produce a count, called a *structural zero* (Chapter 6). For example, the U.S. Census Bureau does not report data when the number of individuals surveyed is extremely small. This suppression of data protects confidentially. Another example occurs in tables of male and female cancer mortality rates. Females do not die of prostate cancer, and males do not die of ovarian cancer. In a table of cancer mortality rates and sex, these cell frequencies are necessarily zero. One of the properties of a loglinear Poisson model applied to tabular data is that theory, estimates, and likelihood statistics are the same in principle for incomplete data. Thus, when an observation are not available, the corresponding computer code is not included. The only important change in theory and analysis is the degrees of freedom are reduced by the number of cells in the table with unavailable counts.

Analysis of Tables with Incomplete Data

As noted, occasionally data in a table are not complete. Analyzing the logarithms of the cell frequencies of such a table has conceptual and statistical advantages. The simplest case is a 2×2 table with a single "missing" cell (Table 9.15). Estimates from data sampled from two sources (denoted S_1 and S_2) necessarily are based on observations recorded by at least one source (Chapter 6). Clearly the data not reported from either sources are not available. The loglinear additive model expression for the "missing value" is $\mu + a + b$ and is estimated by $\hat{l}_{22} = l_{12} + l_{21} - l_{11}$ because an additive model (Table 9.15) dictates that

$$l_{22} = l_{12} + l_{21} - l_{11} \rightarrow (\mu + a) + (\mu + b) - \mu = \mu + a + b.$$

Thus, three observed cell frequencies are sufficient to estimate the model parameters μ, a, and b. An estimate from the loglinear model of the unavailable cell count n_{22} (in brackets)

Table 9.16 *Data: Counts and Logarithms of Counts to Estimate Total Number of Vaccinated Infants in a County in Iowa from Two Independent Sources (Denoted S_1 and S_2) (Chapter 6)*

Counts:		
	S_2	*not* S_2
S_1	33	44
not S_1	14	?
Logarithms of the counts:		
Model	S_2	*not* S_2
S_1	3.497	3.787
not S_1	2.639	?

becomes

$$log(\hat{n}_{22}) = \hat{l}_{22} = \mathrm{l}_{12} + \mathrm{l}_{21} - l_{11} \text{ and the estimated count is } \hat{n}_{22} = e^{\hat{l}_{22}} = \frac{e^{l_{12}} e^{l_{21}}}{e^{l_{11}}}.$$

The previous data from the Iowa survey of vaccinated children illustrate a loglinear additive model estimate of population size from two reported sources (Table 9.16) (Chapter 4). Directly from the additive loglinear model (Table 9.15), the estimate of the "missing" count is

$$\hat{l}_{22} = l_{12} + l_{21} - l_{11} = log(44) + log(14) - log(33) = 3.787 + 2.639 - 3.497 = 2.927$$

and

$$\hat{n}_{22} = e^{\hat{l}_{22}} = e^{2.927} = 18.667 \quad \text{or} \quad \hat{n}_{22} = 14(44)/33 = 18.667 \quad \text{(Table 9.16)}.$$

The estimated total number of vaccinated children is then $\hat{N} = [33 + 44 + 14] + 18.667 = 109.667$ (Chapter 6).

Another Illustration of the Analysis of an Incomplete Table

Using slightly more extensive incomplete data, again a loglinear Poisson model produces an estimate of population size. As part of a research project concerning the risk of a birth defect, it was important to estimate the number of cases of spina bifida among African American children in New York State. Three sources of data are available to identify cases: birth certificates, death certificates, and state rehabilitation records. A direct count of cases from these three sources is biased (too small) because cases not identified by least one source are not included (a structural zero). A loglinear model again provides a statistical tool to estimate the "missing" count. This estimate added to the total observed cases again provides an estimate of the total number of affected children (Table 9.17).

As usual, key to a useful estimate is the choice of a model. Specifically, four choices are the following:

Additive: all three sources are independent

1. $log(n_{ijk}) = a + b_1x_1 + b_2x_2 + b_3x_3$

Table 9.17 *Data: Counts of Spina Bifida Cases among African American Children (New York State) from Three Sources*[a] *to Estimate Total Number of Cases*

S_3	S_2	*Not* S_2
S_1	$n_{111} = 8$	$n_{121} = 13$
not S_1	$n_{211} = 1$	$n_{221} = 3$
not S_3	S_2	*not* S_2
S_1	$n_{112} = 1$	$n_{122} = 8$
not S_1	$n_{212} = 3$	$n_{222} = ?$

[a] S_i represents identified by source i, and *not* S_i represents not identified by source i.

where coefficient x_i represents a binary variable coded presence = 1 or absence = 0 from the ith data source,

Interaction: source 1 and source 2 are associated

2. $log(n_{ijk}) = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4(x_1 \times x_2)$

Interaction: source 1 and source 3 are associated

3. $log(n_{ijk}) = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4(x_1 \times x_3)$

Interaction: source 2 and source 3 are associated

4. $log(n_{ijk}) = a + b_1x_1 + b_2x_3 + b_3x_3 + b_4(x_2 \times x_3)$.

The loglinear model-estimated coefficients, as usual, produce estimated cell frequencies. The focus of these four models is on the estimate of the "missing" cell frequency n_{222} (Tables 9.17 and 9.18). From model 1, for example, making the estimated total number of cases $\hat{N} = [8 + 13 + 1 + 3 + 1 + 8 + 3] + 15.782 = 52.782$. The estimates from the other three models follow the same pattern.

$$\hat{n}_{222} = e^{\hat{a}+\hat{b}_1+\hat{b}_2+\hat{b}_3} = e^{0.416+1.118+1.118+0.106} = e^{2.759} = 15.782$$

Table 9.18 *Summary: Four Loglinear Poisson Models and Their Estimated Coefficients Yielding Estimates of Total Number of Spina Bifida Cases (denoted* $\hat{N}$*)*

	Estimated model coefficients					Estimates			
Model	$\hat{a}$	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\hat{b}_4$	$\hat{l}_{222}$	$\hat{n}_{222}$[a]	$\hat{N}$	p-values[b]
1	0.416	1.118	1.118	0.106	–	2.759	15.78	52.78	0.110
2	0.000	1.705	1.705	0.000	−0.085	2.565	13.00	50.00	0.103
3	0.439	0.575	1.569	−0.811	1.876	3.648	38.40	75.40	0.806
4	0.163	0.875	1.658	0.811	−1.533	1.974	7.200	44.20	0.324

[a] $\hat{n}_{222} = e^{\hat{l}_{222}}$.
[b] Results from a chi-square goodness-of-fit test.

Table 9.19 *Data: Counts of Mother/Daughter Pairs Classified by Less than High School Education, High School Education Only, and Beyond High School*

		Mothers			
		$<HS$	HS	$>HS$	Total
Daughters	$<HS$	$n_{11} = 84$	$n_{12} = 100$	$n_{13} = 67$	$n_{1\cdot} = 251$
	HS	$n_{21} = 142$	$n_{22} = 106$	$n_{23} = 77$	$n_{2\cdot} = 325$
	$>HS$	$n_{31} = 43$	$n_{32} = 48$	$n_{33} = 92$	$n_{3\cdot} = 183$
	Total	$n_{\cdot 1} = 269$	$n_{\cdot 2} = 254$	$n_{\cdot 3} = 236$	$n = 759$

The simplest assumption, but likely the least realistic, is that the sources of ascertainment are independent. The choice would then be model 1 ($\hat{N} = 52.8$). The most accurate model ("best fit") would be model 3 ($\hat{N} = 75.4$). In fact, no unequivocal method exists to select the "correct" model without further information.

Quasi-independence: Analysis of an Incomplete Table

Quasi-independence is a term for independence among categorical variables when data are not available for specific cells within a table. For example, in a 2×2 table of matched pairs data, only discordant pairs are relevant because concordant pairs are matched for the risk factor and do not play a role in the analysis (Chapter 10). Nevertheless, important questions are addressed from an analysis based on the remaining discordant pairs.

To illustrate the application of a loglinear analysis to identify possible quasi-independence, data describing mother-daughter educational status provide an example (Table 9.19). The mother-daughter pairs are classified by three levels of attained education: less than high school ($<HS$), high school only (HS), and beyond high school ($>HS$).

Two different statistical assessments of the mother-daughter association produce exactly the same results. When educational levels are not associated, then estimated cell frequencies are completely determined by marginal probabilities (Chapter 10). For example, the joint probability

$$P(mother < HS \text{ and } daughter > HS) = P(mother < HS)P(daughter > HS)$$

is estimated by $\hat{n}_{31} = n(\frac{n_{3\cdot}}{n})(\frac{n_{\cdot 1}}{n})$. A specific estimated count is then $\hat{n}_{31} = 759(\frac{183}{759})(\frac{269}{759}) = 64.858$. The same pattern of estimation applied to the other eight cells produces nine theoretical values as if mother-daughter educational levels are exactly independent (Table 9.20).

An additive loglinear Poisson model applied to the same data (Table 9.19) produces identical results. The additive model is

$$log(n_{ij}) = l_{ij} = a + r_i + c_j$$

and the estimated coefficients are contained in Table 9.21.

From the Poisson model, for example, $l_{31} = \hat{a} + \hat{r}_3 + \hat{c}_1 = 4.488 - 0.316 + 0.0 = 4.172$ again making $\hat{n}_{31} = e^{4.172} = 64.858$. The chi-square test statistic to formally compare the data (Table 9.19) to the model-estimated cell frequencies (Table 9.20) yields a chi-square test statistic of $X^2 = 48.806$ (degrees of freedom $= 4$) with an associated p-value less than

Table 9.20 *Estimates: Counts of Mother/Daughter Pairs Classified by Less than High School Education, High School Education Only, and Beyond High School as If Educational Levels Are Exactly Independent*

		Mothers			
		$<HS$	HS	$>HS$	Total
Daughters	$<HS$	$n_{11} = 88.96$	$n_{12} = 84.00$	$n_{13} = 78.05$	$n_{1.} = 251$
	HS	$n_{21} = 115.18$	$n_{22} = 108.76$	$n_{23} = 101.05$	$n_{2.} = 325$
	$>HS$	$n_{31} = 64.86$	$n_{32} = 61.24$	$n_{33} = 56.90$	$n_{3.} = 183$
	Total	$n_{.1} = 269$	$n_{.2} = 254$	$n_{.3} = 236$	$n = 759$

0.001. It is usual that mothers and daughters generally have similar educational levels, so the observed association is expected.

An interesting question concerns the pattern of educational status associated with daughters whose educational status differs from their mothers. The question becomes: Does a similar association exist among the six mother-daughter pairs that differ in educational status? (Table 9.19). The analysis is then limited to the six counts of frequencies where mother-daughter pairs differ in educational status. The three dominant mother-daughter pairs that do not differ in educational status are excluded from the analysis. The additive loglinear model coefficients estimated from this incomplete data (discordant mother-daughter pairs) are given in Table 9.22.

As before, the coefficients from the additive loglinear model provide estimates of the six cell values as if no association exists between mother and daughter educational status (Table 9.23). For example, for mothers with less than high school education ($<HS$) and daughters with more than high school education ($>HS$), the estimated log count is

$$\hat{l}_{31} = \hat{a} + \hat{r}_3 + \hat{c}_1 = 4.689 - 0.848 + 0.0 = 3.842$$

producing the estimated cell count

$$\hat{n}_{31} = e^{\hat{l}_{31}} = e^{3.842} = 46.614.$$

Table 9.21 *Results: Estimated Model Coefficients from an Additive Loglinear Poisson Model Analysis of Mother/Daughter Education Data (Table 9.19)*

Coefficients	Estimates	*s. e.*
$\hat{a}$	4.488	–
r_1	0.0	–
$\hat{r}_2$	0.258	0.084
$\hat{r}_3$	−0.316	0.097
c_1	0.0	–
$\hat{c}_2$	−0.057	0.087
$\hat{c}_3$	−0.131	0.089
$-2 \times$ log-likelihood $= 46.592$		

Table 9.22 Results: Model-Estimated Coefficients Resulting for Additive Loglinear Model Analysis of Quasi-independence for Three Educational Levels[a]

Coefficients	Estimates	*s. e.*
$\hat{a}$	4.690	–
r_1	0.0	–
$\hat{r}_2$	0.240	0.139
$\hat{r}_3$	−0.848	0.143
c_1	0.0	–
$\hat{c}_2$	−0.049	0.149
$\hat{c}_3$	−0.540	0.124
$-2 \times$ log-likelihood $= 1.162$		

[a]Pairs concordant for the mother/daughter attained education status are ignored.

The chi-square comparison of the six model-estimated to the six observed cell counts produces the chi-square test statistic $X^2 = 1.163$ (degrees of freedom $= 1$) yielding a p-value of 0.281. The degrees of freedom are the number of cells in the table (6) minus the number of estimated parameters in the model (5). Conditional on differing educational attainment, no strong evidence of a mother-daughter association is apparent.

A Case Study – Adjusted Mortality Rates: Black/White Infant Mortality

A weighted average and a Poisson regression model are useful statistical tools to rigorously address the question: Do African America newborn infants experience lower perinatal mortality than white infants after accounting for their generally lower birth weight? A perinatal mortality rate is defined by the expression

$$\text{perinatal mortality rate} = \frac{\text{infant deaths } (< 28 \text{ days}) + \text{fetal deaths}}{\text{live births} + \text{fetal deaths}}.$$

A perinatal "rate" is not a rate but rather an estimate of the probability of a perinatal death traditionally called a rate.

Table 9.23 *Estimates: Counts of Mother-Daughter Discordant Pairs for Educational Status Classified by Less than High School Education, High School Education Only, and Beyond High School as if Educational Levels Are Exactly Unrelated (Observed Counts in Parentheses)*

	Mothers			
Daughters	*<HS*	*HS*	*>HS*	Total
<HS	–	$n_{12} = 103.61$ (100)	$n_{13} = 63.39$ (67)	167
HS	$n_{21} = 138.39$ (142)	–	$n_{23} = 80.61$ (77)	219
>HS	$n_{31} = 46.61$ (43)	$n_{32} = 44.39$ (48)	–	91
Total	185	148	144	477

Table 9.24 *Data: Perinatal Deaths, Births, Rates*/1000 *(California – 1988) and Rate Ratio (Black/White) for* 35 *Birth Weight Categories (Grams)*

	White infants			Black infants			
Weights	Deaths	Births	Rate	Deaths	Births	Rate	Ratio
800–900	173	322	537.27	65	131	496.18	0.924
901–1000	148	337	439.17	40	122	327.87	0.747
1001–1100	134	398	336.68	30	131	229.01	0.680
1101–1200	106	381	278.22	29	137	211.68	0.761
1201–1300	103	444	231.98	21	143	146.85	0.633
1301–1400	86	427	201.41	19	143	132.87	0.660
1401–1500	108	597	180.90	19	165	115.15	0.637
1501–1600	85	560	151.79	20	167	119.76	0.789
1601–1700	84	682	123.17	24	219	109.59	0.890
1701–1800	86	722	119.11	12	194	61.86	0.519
1801–1900	100	935	106.95	26	298	87.25	0.816
1901–2000	81	978	82.82	15	299	50.17	0.606
2001–2100	74	1589	46.57	21	420	50.00	1.074
2101–2200	87	1714	50.76	10	453	22.08	0.435
2201–2300	82	2322	35.31	14	603	23.22	0.657
2301–2400	80	2885	27.73	12	763	15.73	0.567
2401–2500	80	4149	19.28	13	977	13.31	0.690
2501–2600	77	4916	15.66	14	1189	11.77	0.752
2601–2700	93	7455	12.47	10	1654	6.05	0.485
2701–2800	93	8855	10.50	17	1796	9.47	0.901
2801–2900	100	14,197	7.04	11	2545	4.32	0.614
2901–3000	86	17,903	4.80	9	2947	3.05	0.636
3001–3100	92	19,969	4.61	12	2851	4.21	0.914
3101–3200	90	27,068	3.32	9	3557	2.53	0.761
3201–3300	96	29,107	3.30	9	3324	2.71	0.821
3301–3400	79	35,627	2.22	11	3577	3.08	1.387
3401–3500	67	32,926	2.03	1	3119	0.32	0.158
3501–3600	69	36,360	1.90	9	2952	3.05	1.607
3601–3700	58	30,612	1.89	7	2250	3.11	1.642
3701–3800	59	32,119	1.84	2	2176	0.92	0.500
3801–3900	40	24,004	1.67	3	1573	1.91	1.145
3901–4000	35	23,217	1.51	1	1348	0.74	0.492
4001–4100	30	16,232	1.85	3	909	3.30	1.786
4101–4200	19	14,233	1.33	1	735	1.36	1.019
4201–4300	17	9,781	1.74	1	489	2.04	1.177
Total	2897	404,023	7.17	520	44,356	11.72	1.635

All African American and white live births plus fetal deaths reported during the year 1988 on California birth and death certificates allow an analysis of ethnicity-specific perinatal mortality risk. During 1988, a total of 448,379 births and fetal deaths occurred (44,356 African American and 404,023 white infants). Birth certificates contain the mother's ethnicity (reported by the mother) and her infant's birth weight (recorded in grams). The data are tabulated by ethnicity and classified into 100 gram birth weight intervals (less than 800 grams, 801 to 900 grams, . . . , 4201 to 4300 grams; Table 9.24).

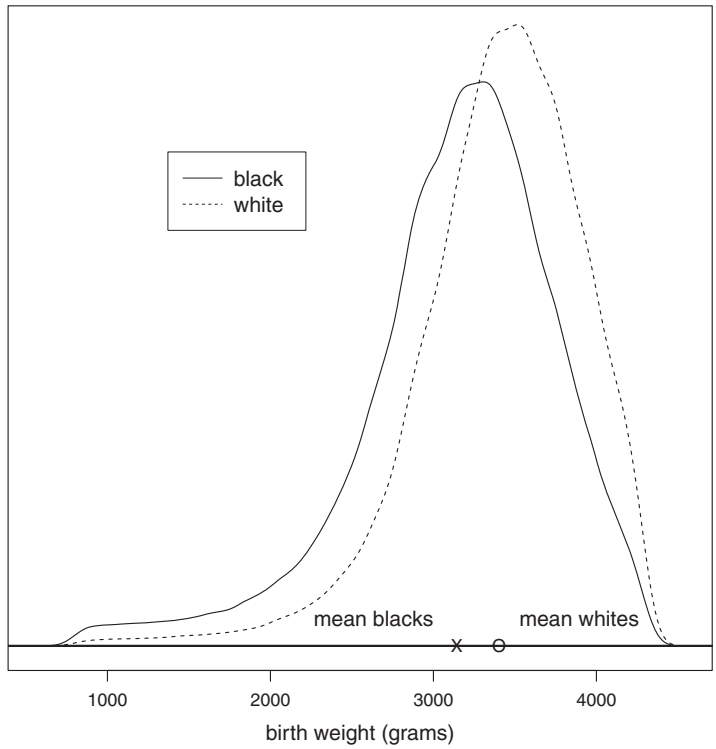


Figure 9.4 Distributions of Birth Weights of White and African American Infants (California – 1988)

The 70 ethnicity- and weight-specific categories contain the number of perinatal deaths, the number of live births plus fetal deaths (labeled "births"), and perinatal mortality rates for white and black infants. The overall perinatal mortality rate among black infants is $(520/44{,}356) \times 1000 = 11.723$ deaths per 1000 births and among white infants $(2897/404{,}023) \times 1000 = 7.170$ deaths per 1000 births. Thus, the overall black-white perinatal mortality rate ratio is $11.723/7.170 = 1.635$, sometimes call the *crude ratio* because it is not adjusted for possible influencing factors (Table 9.24, last row). This case study is similarly described in *Statistical Tools for Epidemiologic Research* (Oxford University Press, 2001).

Black newborn infants weigh considerably less, on average, than white newborn infants (mean values: $\bar{x}_{black} = 3143.2$ grams and $\bar{x}_{white} = 3403.4$ grams), and for all infants, lower birth weight is associated with higher mortality. The two estimated birth weight distributions are displayed in Figure 9.4.

An important question is: Would the differences in perinatal mortality persist if the birth weight distributions were the same for black and white infants? More technically the question becomes: What is the black-white perinatal mortality rate ratio when statistically adjusted for the influence of differing birth weight distributions? This question can be addressed with several statistical strategies. Three approaches are described in the following: weight-specific ratios, a model-free weighted average, and a Poisson loglinear regression model.

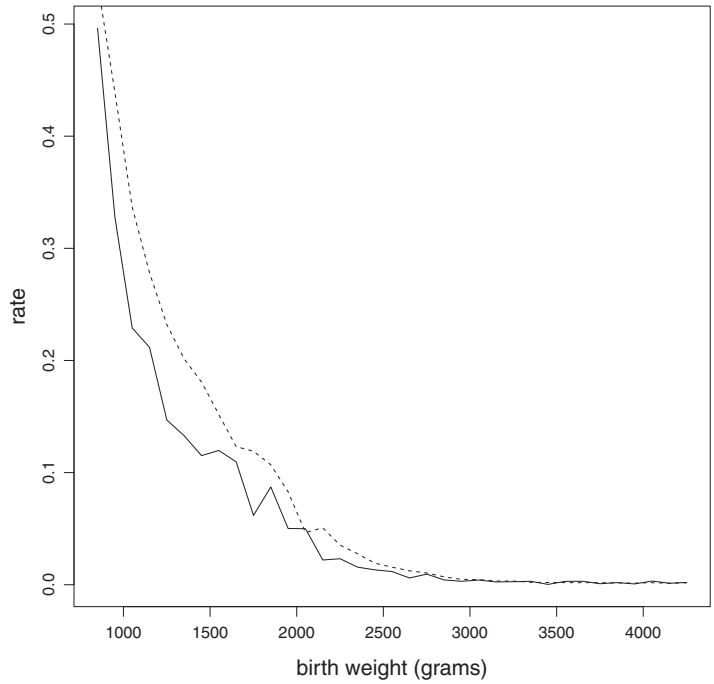


Figure 9.5 Data: Perinatal Mortality Rates by Birth Weight for African American and White Infants, 1988

First Approach: Weight-Specific Comparisons

An easily applied contrast of perinatal mortality risk essentially unaffected by differences in birth weight distributions is accomplished by directly comparing black to white mortality rates within each of 35 birth weight categories (strata). Differences between black and white birth weights within a 100 gram interval are small and negligible (Table 9.24). In terms of the weight-specific rate ratios (*black rate*-white *rate*) 27 out of 35 perinatal mortality rate ratios are less than one (*black rate* < *white rate*). These differences are not due to differences in birth weight distributions because, as noted, within each of the 35 strata the black-white mean birth weights are practically the same (balanced). The fact that about 80% of weight-specific rate ratios show black infants with lower perinatal mortality is a persuasive indication that, after equalizing for birth weight differences, black infants have consistently lower perinatal mortality. In other words, when infants with essentially the same birth weights are compared, the black infants are more likely to survive, particularly black infants with low and extremely low birth weights. Figure 9.5 displays the black and white perinatal mortality rates by birth weight. Clearly, the black perinatal mortality rates (solid line) are generally below the white rates (dashed line) for most birth weights.

A more detailed view of these strata-specific differences is achieved by comparing the logarithms of the rates (Figure 9.6). A plot of the logarithms of the perinatal mortality rates by birth weight over a reduced range creates a clearer visual picture of the black-white mortality patterns, especially for extremely high birth weights. Contrasting log rates necessarily produces the same number of observed black-white mortality ratios less than 1.0

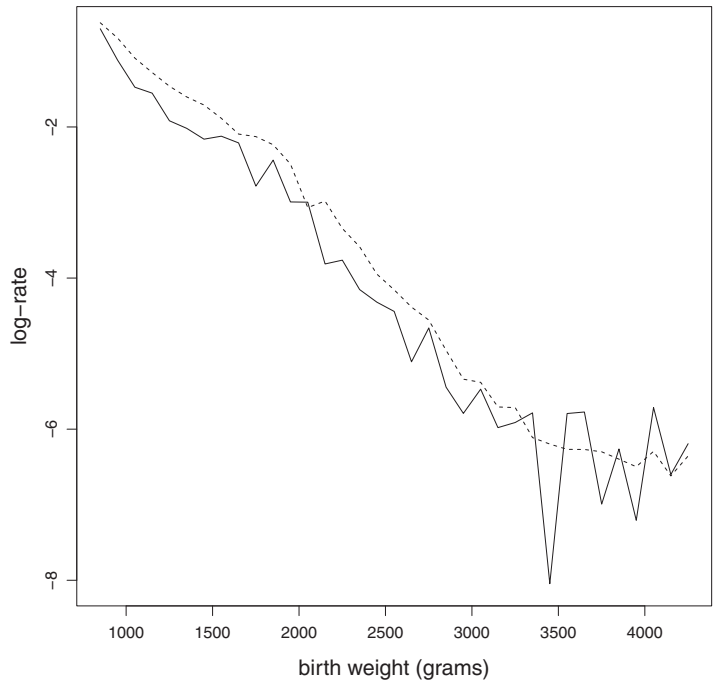


Figure 9.6 Perinatal Mortality Log Rates by Birth Weight for African American versus White Infants, 1988

(27 lower black *log*-rates out of 35) but on a more visually sensitive scale. For example, on the log scale, black perinatal mortality rates clearly fluctuate above the white rates among infants only in the neighborhood of 4000 gram birth weights where these rates are highly unstable due to extremely small numbers, particularly black perinatal deaths (a total of 5 black and 2133 white deaths). This feature of the relationship between white and black mortality rates is invisible when the rates are directly plotted (Figure 9.5).

The last column in Table 9.24 contains the weight-specific white-black perinatal rate ratios for the 35 birth weight categories. To accurately summarize these values with a single rate ratio it is necessary that the underlying ratio be the same for all birth-weight strata (homogeneous).

The issue of a constant rate ratio is directly displayed with a graphical approach. To start, a line describing these rate ratios is estimated as if the ratios are constant (dashed line) and a line constructed directly from the 35 pairs of rates themselves (solid line) are plotted (Figure 9.7). Comparison of these two summary lines gives a direct visual indication of the extent of homogeneity among the observed rate ratios additivity.

A line representing a constant rate ratio is estimated from the simplest possible regression model where the intercept is required to be zero or, in symbols,

$$\textit{rate-ratio} = \frac{r_i^{(black)}}{r_i^{(white)}} = b \quad \text{or} \quad r_i^{(black)} = b r_i^{(white)} \quad i = 1, 2, \ldots, 35.$$

The notation $r_i^{(ethnicity)}$ represents the ith ethnicity- and weight-specific perinatal mortality rate. A weighted average produces the ordinary least-squares estimated slope b (denoted $\hat{b}$)

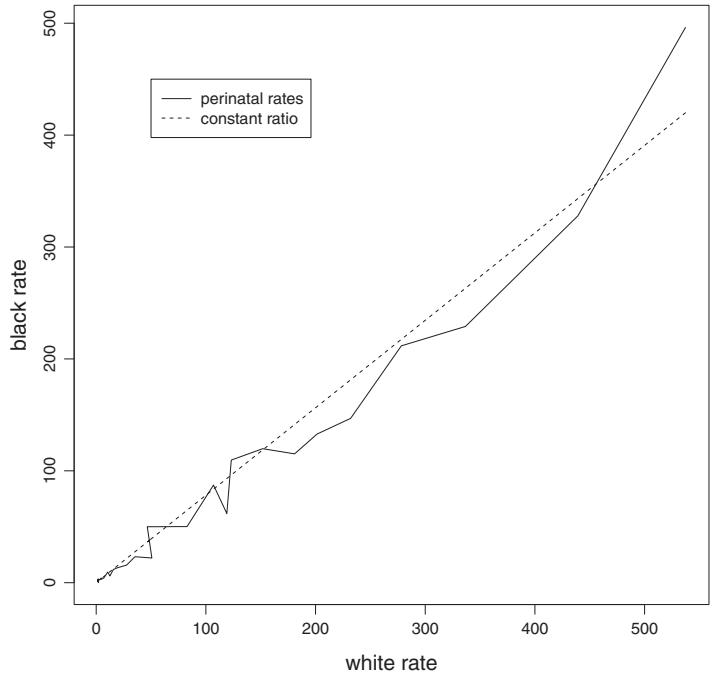


Figure 9.7 White and Black Perinatal Mortality Rates (Solid Line) and the Line Estimated as if Rate Ratio Is Constant for All Birth Weights (Dashed Line)

as if the rate ratios are exactly constant (Chapter 3). Figure 9.7 displays the 35 pairs of white and black perinatal mortality rates ($r_i{}^{(white)}$, $r_i{}^{(black)}$) connected by a line (solid line) as well as the estimated straight line (dashed line).

For the California perinatal mortality data, the estimated slope from the simple linear regression model is $\hat{b} = 0.782$ and characterizes a constant perinatal mortality rate ratio as if the 35 differences among the observed ratios ($r_i^{(black)}/r_i^{(white)}$) are strictly random (Table 9.24, last column).

Second Approach: A Model-Free Summary

The 35 separate ethnicity- and weight-specific rate ratios clearly reflect the black-white differences in perinatal mortality but are not a parsimonious summary description of overall risk. A weighted average of the weight-specific rate ratios, however, provides a single comprehensive summary ratio. Unlike the 35 individual strata-specific comparisons, the influence of sampling variation on this single summary ratio can be directly estimated, allowing a rigorous statistical evaluation.

Parallel to applying a number of statistical summaries, such as an estimated odds ratio or relative risk, it is statistically advantageous to consider the logarithms of the rate ratios rather than the ratios themselves (Chapter 6). Three reasons for this frequently used statistical strategy are the following:

(1) As noted, a reduced range makes a graphic display clearer by increasing the visibility of differences, especially in the extreme ranges

(2) Log-rate ratios convert multiplicative comparisons to an additive scale, making differences easier to describe and to statistically manipulate [ratios a/b become differences $log(a) - log(b)$] and
(3) As previously noted, *log*-rates tend to have more symmetric and, therefore, more normal-like distributions producing a simple and intuitive assessment of their statistical properties.

Again, the symbol $r_i^{(ethnicity)}$ denotes the perinatal ethnicity-specific mortality rate from the *i*th birth weight category. Then, the estimated weight-specific rate ratio contrasting black to white perinatal mortality rates is denoted

$$\hat{R}_i = \frac{r_i^{(black)}}{r_i^{(white)}}$$

for the *i*th weight category (Table 9.24, last column). A weighted average of the logarithms of these 35 estimated rate ratios yields a single summary perinatal mortality *log*-rate ratio of

$$\overline{log\text{-R}} = \frac{\sum w_i log(\hat{R}_i)}{\sum w_i}$$

where the weights w_i are chosen to be the reciprocal values of the estimated variances of $log(\hat{R}_i)$ (Chapter 3). In symbols, the weights are $w_i = 1/S^2_{log(\hat{R}_i)}$. Specifically, the estimated variances of the distributions of these 35 weight-specific estimated *log*-rate ratios are

$$S^2_{log(\hat{R}_i)} = \frac{1}{d_i^{(black)}} + \frac{1}{d_i^{(white)}}$$

where $d_i^{(ethnicity)}$ represents the number of ethnicity-specific perinatal deaths in the *i*th birth-weight strata (Chapter 27).

For example, for the birth-weight category 2501 to 2600 grams ($i = 18$), the rate ratio is

$$\hat{R}_{18} = \frac{\hat{r}_{18}^{(black)}}{\hat{r}_{18}^{(white)}} = \frac{14/1189}{77/4916} = \frac{11.775}{15.663} = 0.752$$

and the *log*-rate ratio becomes

$$\begin{aligned} log(\hat{R}_{18}) &= log\left[\hat{r}_{18}^{(black)}\right] - log\left[\hat{r}_{18}^{(white)}\right] = log(11.775) - log(15.663) \\ &= log(0.752) = -0.285. \end{aligned}$$

The associated estimated variance is

$$S^2_{log(\hat{R}_{18})} = \frac{1}{d_{18}^{(black)}} + \frac{1}{d_{18}^{(white)}} = \frac{1}{14} + \frac{1}{77} = 0.0.0012.$$

The weighted average of the logarithms of the 35 rate ratios is $\overline{log\text{-}R} = -0.261$ making the estimated summary ratio $\bar{R} = e^{\overline{log\text{-}R}} = e^{-0.261} = 0.770$. That is, an estimate of the underlying ratio of black-white perinatal mortality rates is $\bar{R} = 0.770$. Thus, the black perinatal infant mortality rate is, on average, about three-quarters of the white rate accounting for the influence of the lower birth weights among African American infants. Again a single weighted average is an effective summary of black-white perinatal mortality only when differences among the rate ratios from each of the 35 weight-specific strata are random fluctuations from a single underlying value.

The estimated variance of a weighted average, when the weights are the reciprocal of the estimated variances of the quantities averaged (Chapter 3), is

$$variance(\overline{log\text{-}R}) = \frac{1}{\sum w_i}.$$

For the black-white perinatal data, the estimated variance becomes

$$variance(log\text{-}rate\ ratio) = S^2_{\overline{log\text{-}R}} = \frac{1}{427.350} = 0.00234$$

making the standard error $S_{\overline{log\text{-}R}} = \sqrt{0.00234} = 0.048$.

A normal distribution approximation of the distribution of the logarithm of the summary rate ratio allows the construction of confidence interval in the usual way. An approximate 95% confidence interval is

$$\overline{log\text{-}R} \pm 1.960 S_{\overline{log\text{-}R}} = -0.261 \pm 1.960(0.048) \rightarrow (-0.356, -0.166)$$

based on the normal distribution and the estimate $\overline{log\text{-}R} = -0.261$. Because the estimated summary rate ratio is $\bar{R} = e^{\overline{log\text{-}R}} = 0.770$, the associated approximate 95% confidence interval becomes

$$(e^{lower\ bound}, e^{upper\ bound}) = (e^{-0.356}, e^{-0.166}) = (0.701, 0.847).$$

Like all 95% confidence intervals, the probability that the interval (0.701, 0.847) contains the value estimated by $\bar{R} = 0.770$ is approximately 0.95, providing a sense of accuracy and precision. The estimated confidence interval provides definite assurance that the lower weight-adjusted mortality observed among African Americans is not likely due entirely to random variation (ratio < 1.0).

Third Approach: Poisson Regression Model

Another approach to summarizing black-white differences in perinatal mortality is achieved with an application of an additive Poisson regression model. Such a model postulates a specific relationship between birth weight and ethnicity making it possible to estimate, evaluate, and display the separate influences on the pattern of perinatal mortality.

For the California birth/death certificate data, a potentially useful additive Poisson regression model is

$$log(r_{ij}) = a + bF_j + P(bwt_i) \quad j = 1, 2 \quad \text{and} \quad i = 1, 2, \ldots, 35$$

where $P(x) = c_1x + c_2x^2 + c_3x^3 + c_4x^4 + c_5x^5$. The weight-specific perinatal mortality rate is denoted r_{ij} for the jth ethnicity within the ith birth-weight strata. The variable F represents a binary indicator variable where $F_j = 1$ indicates African American and $F_j = 0$ indicates white mother-infant pairs. A detailed description of the relationship between birth weight and perinatal mortality is not a primary interest and is pragmatically represented by a rather complex fifth-degree polynomial. Other functions could equally represent the additive influence of infant birth weight. The primary purpose of any choice is to characterize the mortality pattern associated with birth weight as part of a clear and focused description of the difference between black-white perinatal mortality rates.

Table 9.25 *Model Estimates: Coefficients[a] from the Poisson Regression Model Describing Black/White Perinatal Mortality Racial Differences by Birth Weight*

	Coefficient	Estimate	*s. e.*
Intercept	a	−3.994	–
Ethnicity	b	−0.290	0.048
Linear	c_1	−16.415	0.194
Quadratic	c_2	1.558	0.195
Cubic	c_3	2.035	0.186
Quartic	c_4	0.371	0.180
Quintic	c_5	−0.739	0.159
Loglikelihood = −27.261			

[a]These coefficients result from a specialized numerical/computer technique that allows for the extreme range of the independent birth-weight variables (bwt_i) but does not affect the analytic results.

Geometrically, the polynomial function $P(bwt)$ and the additive model allow the black and white *log*-rates to be accurately represented by exactly the same relationship within their ethnicity categories: that is, two identical curves, $P(x)$. The constant distance between these two curves measures the black-white perinatal mortality difference, adjusted for birth weight (Chapter 7). Thus, the model difference in log rates (model parameter denoted b) directly summarizes the black-white rate ratio as if the perinatal mortality risk between infants is identical for all birth weights (no interaction/homogeneous). In addition, the numbers of deaths in each of the 70 ethnicity- and weight-specific categories is postulated to have Poisson distributions. The additive regression model and the "Poisson assumption" make it possible to estimate the seven model coefficients and their standard errors (Table 9.25). Of central importance, the analysis produces a single weight-adjusted estimate of the black-white difference in perinatal mortality experience, namely, the estimated coefficient $\hat{b}$.

The Poisson model and its maximum likelihood estimated coefficients directly yield an estimate of the *log*-rate $[log(\hat{r}_{ij})]$, which, in turn, yields an estimate of the perinatal mortality rate $(\hat{r}_{ij})$, leading to model-estimated number of perinatal deaths in each ethnicity- and weight-specific strata.

For example, the white infant *log*-rate ($F = 0$) for the weight category 2501–2600 grams ($i = 18$) is estimated by

$$\text{estimated } log\text{-rate} = log\left(\hat{r}_{18}^{(white)}\right) = -3.994 - 0.290(0) + \hat{P}(2550) = -4.152.$$

The model perinatal mortality rate is then $\hat{r}_{18}^{(white)} = e^{-4.152} \times 1000 = 15.732$ deaths per 1000 births. The number of model-estimated perinatal deaths is the estimated rate multiplied by the number of births or $\hat{d}_{18}^{(white)} = 0.0157(4916) = 77.3$. The corresponding observed quantities are $r_{18}^{(white)} = (77/4916) \times 1000 = 15.663$ deaths per 1000 births (perinatal mortality rate) and $d_{18}^{(white)} = 77$ perinatal deaths (Table 9.24).

Following the typical goodness-of-fit pattern, the 70 ethnicity- and weight-specific model-estimated numbers of perinatal deaths ($\hat{d}_i$) are compared to the 70 corresponding observed numbers of deaths (d_i). A chi-square comparison produces the test statistic

$$X^2 = \sum \frac{(d_i - \hat{d}_i)^2}{\hat{d}_i} = 56.216 \quad i = 1, 2, \ldots, 70 \quad i = 1, 2, \ldots, 70$$

with 63 degrees of freedom. Once again, the degrees of freedom are the number of observations (strata) in the table minus the number of model parameters or $70 - 7 = 63$ yielding the p-value $P(X^2 \geq 56.216 \mid \textit{model fits}) = 0.715$. As usual, the phrase "model fits" means that all differences between observed and estimated numbers of deaths are due only to random variation. A p-value of 0.715 indicates a close correspondence between the model-estimated and the observed numbers of deaths, implying that the single estimated coefficient ($\hat{b}$) from the additive model accurately summarizes the black-white perinatal mortality differences.

The generally excellent fit of the estimated Poisson regression model allows ethnicity-birth-weight relationships to be separated into two components, one reflecting the influences of ethnicity and the other reflecting the influence of birth weight on the likelihood of a perinatal mortality death. That is, no apparent evidence exists of an ethnicity–birth weight interaction. More specifically, a constant coefficient $\hat{b}$ accurately reflects a constant black-white racial difference in risk. In terms of the model parameters, for the ith birth-weight category, the separate influence of ethnicity measured by the estimated *log*-rate difference is

$$log(\hat{R}_i) = log(\hat{r}_i^{(black)}) - log(\hat{r}_i^{(white)}) = [\hat{a} + \hat{b}(1) + \hat{P}(bwt_i)] - [\hat{a} + \hat{b}(0) + \hat{P}(bwt_i)] = \hat{b}$$

and is the same for all weight categories (no interaction). Therefore, the model-estimated single rate ratio $\hat{R} = e^{\hat{b}} = e^{-0.290} = 0.749$ is likely an accurate measure of black-white perinatal mortality differences. As before, the estimated ratio indicates the black perinatal mortality rate is about 75% of the white rate after accounting for the birth-weight differences. In statistical language, the estimated rate ratio $\hat{R}$ is said to be adjusted for the influence from the differing white and African American birth-weight distributions (Chapter 7).

Because, like all estimates, the estimate $\hat{b}$ is subject to sampling variation, an assessment of this influence is frequently an important consideration. The test statistic

$$z = \frac{\hat{b} - 0}{S_{\hat{b}}} = \frac{-0.290 - 0}{0.048} = -6.009$$

has an approximate standard normal distribution when $b = 0$ (rate ratio $= R = 1$). The estimated standard error $S_{\hat{b}} = 0.048$ is calculated as part of the maximum likelihood model estimation process (Table 9.25). The maximum likelihood estimate $\hat{b} = -0.290$ and, therefore, the approximately normally distributed z-value yields a p-value of $P(|\hat{b}| \geq 0.290 | b = 0) = P(|Z| \geq 6.009 | b = 0) < 0.001$. The extremely small p-value again supplies substantial evidence of a systematic difference in perinatal mortality between white and African American infants, as suspected from the previous two model-free analyses.

The influence of sampling variation on the estimated summary rate ratio is additionally described by a confidence interval. An approximate 95% confidence interval is

$$\hat{b} \pm 1.960 S_{\hat{b}} = -0.290 \pm 1.960(0.048) \rightarrow (-0.384, -0.195)$$

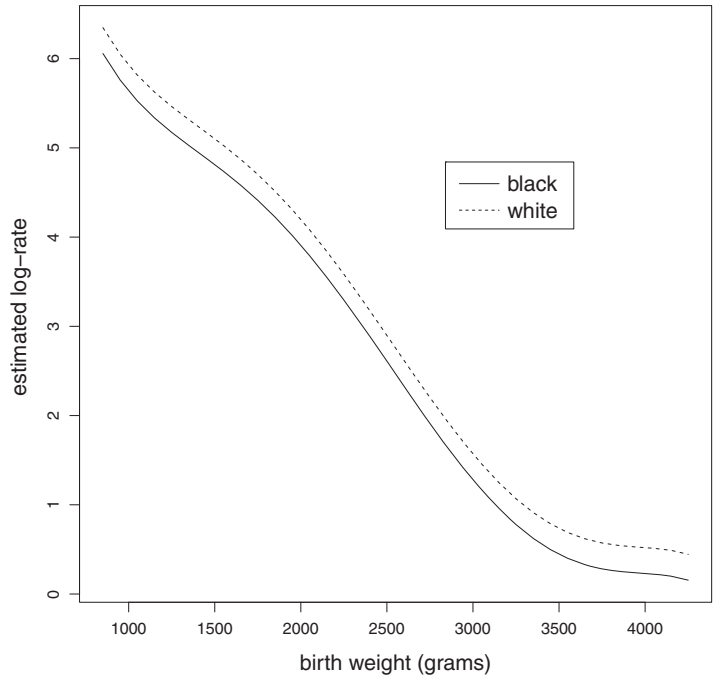


Figure 9.8 Estimated Perinatal Black-White Mortality Log Rates by Birth Weight

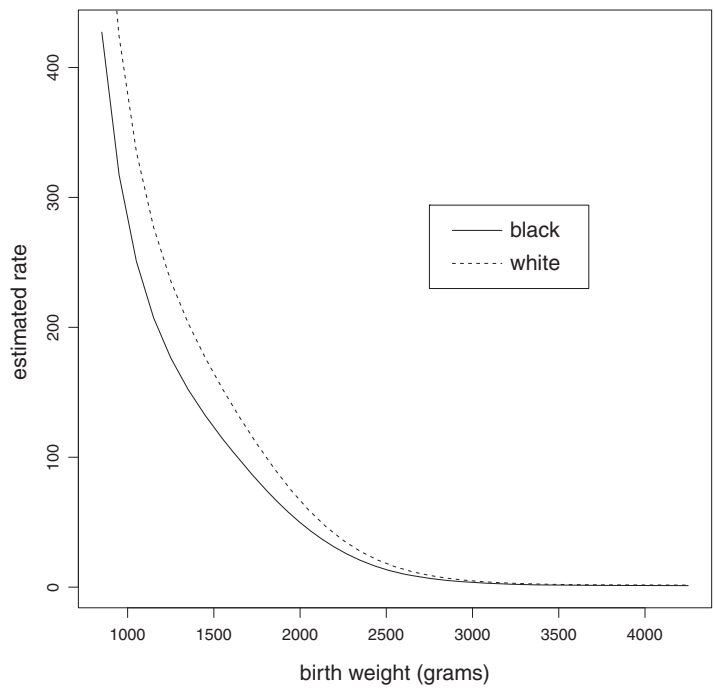


Figure 9.9 Estimated Black-White Mortality Rates by Birth Weight

based on the normal distribution and model estimate $\hat{b} = -0.290$. Because the summary rate ratio $\hat{R} = 0.749$ is estimated by $\hat{R} = e^{\hat{b}} = e^{-0.290}$, the confidence interval bounds for the underlying single rate ratio R become

$$(\mathrm{e}^{lower\ bound}, \mathrm{e}^{upper\ bound}) = (\mathrm{e}^{-0.384}, e^{-0.196}) = (0.681, 0.823).$$

This model-based confidence interval only slightly differs from the previous model-free (weighted average) approximate 95% confidence interval interval (0.701, 0.847). The primary difference between the two approaches is that the Poisson model yields an evaluation of the homogeneity of the strata-specific rate ratios.

Figure 9.8 visually displays the geometry of the *log*-rates calculated from the additive model. As dictated by the choice of an additive model, the distance (on a log scale) between the identical polynomial curves describes the model birth-weight/mortality pattern for black and white infants as constant for all birth weights, $|\hat{b}| = 0.290$. Also, because $\hat{b}$ is less zero, the model-estimated black perinatal mortality log rates are less than the model-estimated white log-rates for all birth weights, again dictated by the model. Figure 9.9 displays the same estimated log rates exponentiated to produce a more natural picture of model-estimated perinatal mortality rates.

A clear and simple summary of black-white perinatal mortality risk is produced by the additive Poisson loglinear model. Furthermore, the assessment of the model ("goodness-of-fit") indicates that this single summary is likely an accurate reflection of the underlying difference in black-white risk. Particularly noteworthy is the extreme confounding influence of birth weight indicated by an adjusted rate ratio (black-white) of 0.749 and, without adjustment, a crude rate ratio (black-white) of $11.723/7.170 = 1.635$.

10

Two-Way and Three-Way Tables and Their Analysis

A table is a summary statistic, arguably the most important summary statistic. Like most summaries, it is a convenient and useful way to condense data to reflect more clearly the issues the data were collected to explore. Approaches to the analysis of tabular data frequently appear to require specialized statistical tools, but, in fact, many of these techniques are specific applications of familiar methods. One of these statistical tools is the likelihood statistic, typically associated with the analysis of data with statistical models (Chapters 9 and 27). Before describing the ways that tables provide insight into the relationships among categorical variables, a working understanding of likelihood comparisons is essential. The full theory and notation are extensive, but the following examples provide an accessible introduction.

A fad of a number of years ago was to collect data to determine if the probability of dying was influenced by an individual's birthday. In its simplest form, individuals were classified into two categories; died during the six months before or during the six months after their birthday. Perhaps these analyses were motivated by the large amounts of public available data from a wide variety of individuals (famous people, royalty, presidents, movie stars, and so on). A random sample ($n = 348$) of birth and death dates from a volume of *Who's Who* produces such data (Table 10.1).

To assess the influence of random variation on these before and after data, the estimated probability of death before a birthday (denoted $\hat{p}$) is compared to the probability expected when no association exists, namely, $p_0 = 0.5$. This comparison can be made using either estimated probabilities or a table of counts; both give the same analytic result (Chapter 6). The normal distribution approximation of a binomial distribution provides a statistical comparison of an estimated probability ($\hat{p} = 0.425$) to theoretical probability ($p_0 = 0.5$) where

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\hat{p} - 0.5}{\sqrt{\frac{1}{4n}}} = \frac{0.425 - 0.5}{\sqrt{\frac{1}{4(348)}}} = -2.787$$

and yields the p-value of $P(|Z| \leq -2.787 \mid p_0 = 0.5) = 0.0053$ (Chapters 1, 4, and 6). Based on a comparison of counts using a chi-square–distributed test statistic (degrees of freedom = 1), the

$$\text{Pearson chi-square statistic} = X^2 = \frac{(148 - 174)^2}{174} + \frac{(200 - 174)^2}{174} = 7.770$$

yields the identical p-value $= P(X^2 \geq 7.770|\textit{no association}) = P(X^2 \geq (-2.787)^2|\textit{no association}) = 0.0053$ [1].

Table 10.1 *Data: Tabulation of Deaths before and after a Birthday* ($n = 348$)

	Deaths		
	Before	After	Total
Observed	148	200	348
Theory	174	174	348
$\hat{p}_i$	0.425	0.575	1.0
p_0	0.5	0.5	1.0

A likelihood approach to the same comparison, as might be expected, starts with a likelihood value calculated from the observed value $\hat{p} = 0.425$ (Table 10.1). For the before/after birthday data, the likelihood statistic is

$$\text{likelihood value} = \binom{348}{148}(\hat{p})^{148}(1-\hat{p})^{200} = \binom{348}{148}(0.425)^{148}(1-0.425)^{200}.$$

This likelihood value is simply the estimated binomial probability of the occurrence of the observed value of 148 deaths before a birthday among $n = 348$ sampled individuals ($\hat{p} = 148/348 = 0.425$) (Chapter 4). When no association exists between a birthday and the day of death, the likelihood value is calculated as if the probability of death before or after birthday is the same or $p_0 = 0.5$. The data estimate $\hat{p}$ is replaced by the theoretical value p_0, and likelihood value becomes

$$\text{likelihood value} = \binom{348}{148}(p_0)^{148}(1-p_0)^{200} = \binom{348}{148}(0.5)^{148}(0.5)^{200}.$$

Likelihood values are frequently tiny, sometimes extremely tiny, probabilities. Nevertheless, the comparison of likelihood values reflects differences between data and theory. It would be expected that the likelihood value based on data-generated probability $\hat{p} = 0.425$ is larger than the likelihood value based on the theoretical conjecture $p_0 = 0.5$. The statistical question becomes: is the observed increase sufficiently larger to suggest a nonrandom influence?

For several reasons the logarithm of likelihood values are used to make this comparison. For example, sums are more intuitive and mathematically tractable than products. The description of the details and specific requirements of log-likelihood comparisons are more completely described (Chapter 27). The log-likelihood value calculated with the observed probability $\hat{p} = 0.425$ is

$$\text{log-likelihood value} = L = K + 148\,log(0.425) + 200\,log(0.575) = -237.316,$$

where $K = log[\binom{348}{148}]$. A conjecture of no before/after association creates the parallel log-likelihood value calculated with the theoretical probability of $p_0 = 0.5$ and is

$$\text{log-likelihood value} = L_0 = K + 148\,log(0.5) + 200\,log(0.5) = -241.215.$$

The comparison of these two log-likelihood values (denoted G^2) is

$$\begin{aligned} G^2 &= 2[(148\,log(0.425) + 200\,log(0.575)) - (148\,log(0.5) + 200\,log(0.5))] \\ &= 2[L - L_0] = 2\lceil -237.316 - (-241.215)\rceil = 7.799. \end{aligned}$$

Table 10.2 *Data: Tabulation of Deaths before and after a Birthday by Month* ($n = 348$)

	Month before death						Month after death						
	6	5	4	3	2	1	1	2	3	4	5	6	
Counts (n_i)	25	23	25	25	27	23	42	40	32	35	30	21	348
p_i	0.072	0.066	0.072	0.072	0.078	0.066	0.121	0.115	0.092	0.101	0.086	0.060	1.0
p_0	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	1.0

More succinctly, the same likelihood test statistic is expressed as

$$G^2 = 2[148\, log(0.425/0.5) + 200\, log(0.575/0.5)] = 7.799.$$

Notice that the term in the likelihood expression $K = \binom{348}{148}$ disappears from the comparison of the likelihood values. Similar constant terms appear in all likelihood statistics. These values do not influence likelihood comparisons and are ignored in the following discussion (Chapter 27).

In general, the expression for a log-likelihood comparison of two sets of probabilities ($\hat{p}$ versus p_i) is

$$G^2 = 2\sum n_i\, log(\hat{p}_i/p_i) \qquad i = 1, 2, \ldots, k = \text{number of compared probabilities},$$

where estimates $\hat{p}_i$ are based on n_i observations and probabilities p_i are based on theory. Formally, the statistic G^2 is called the likelihood ratio test statistic. When $\hat{p}_i = p_i$ for all comparisons, this perfect correspondence produces the likelihood value $G^2 = 0$.

At heart of a log-likelihood comparison is the test statistic G^2 that has an approximate chi-square distribution with degrees of freedom equal to the difference in the number of estimates used to generate each of the compared probability distributions. For the example comparison (Table 10.1), the degrees of freedom are 1, and the p-value is $P(G^2 \geq 7.799 \mid \text{no association}) = 0.005$. As always, a small p-value indicates that the observed value $\hat{p} = 0.425$ is not likely a random deviation from theoretical probability $p_0 = 0.5$.

Both Pearson goodness-of-fit chi-square statistic X^2 and log-likelihood G^2 test statistics have approximate chi-square distributions when the observed variation is strictly random. Both test statistics have the same degrees of freedom, and both are a comparison of data to theoretical values. It is not surprising that these two test statistics usually produce close to the same value when applied to the same data. For the birthday data, the two test statistics are $X^2 = 7.770$ and $G^2 = 7.799$.

A more extensive comparison of observed and theoretical probabilities of death before and after a birthday is presented by month (Table 10.2).

The previous $n = 348$ observations (Table 10.1) are distributed into a table with $k = 12$ categories instead of two, but the likelihood analysis changes little in principle. The 12 data-estimated probabilities ($\hat{p}_i = n_i/n$) are compared with 12 theoretical probabilities ($p_0 = 1/12 = 0.083$). The general likelihood comparison is again

$$G^2 = 2\sum n_i\, log(\hat{p}_i/p_0)$$

and, for the before-after birthday example data, becomes

$$G^2 = 2[25\,log(0.072/0.083) + 23\,log(0.066/0.083) + \cdots + 21\,log(0.060/0.083)]$$
$$= 17.332.$$

The value of test statistic G^2 has an approximate chi-square distribution ($k - 1 = 11$ degrees of freedom) when the underlying probability of death is the same for each month ($p_0 = 1/12 = 0.083$). The associated p-value is $P(G^2 \geq 17.332 \mid p_0 = 1/12) = 0.098$. The corresponding Pearson chi-square test statistic is $X^2 = 18.069$ (p-value $= 0.080$).

Note that the analyses of the 2×2 table of birthday data and the 2×12 table constructed from the same data produce rather different results (p-value $= 0.005$ and 0.098). Construction of a table frequently requires selection of more or less arbitrary categories. Furthermore, as illustrated, the choice of size or number of categories can considerably influence the analytic results. As they say in a legal world, "buyer beware," because the analysis of a table and accompanying inference frequently are influenced by the choices that determine how the table is constructed.

Likelihood comparisons used to measure the correspondence between data and theoretical values have a number of remarkable features. Likelihood statistics apply to both discrete (as above) and continuous (next section) data. Likelihood statistics summarizing continuous data are not always simply calculated, but their interpretation is not different from the discrete case. Log-likelihood values are frequently used to produce estimates of model parameters (Chapter 27). An inescapable fact is that the more model parameters added to represent relationships within sampled data, the more accurate the model representation becomes. Adding parameters to a model always creates a log-likelihood value closer to zero (perfect fit). Log-likelihood statistics faithfully reflect such an increase between model-estimated values and data. Thus, likelihood comparisons serve as an excellent statistical tool directly reflecting the consequences of including or excluding model parameters to explore the delicate balance between model simplicity and accuracy. Thus, the magnitude of differences between likelihood values is central to the process of selecting an effective model. The difference between Pearson chi-square test statistics can increase or decrease when a variable is added to a model, and, therefore, comparisons between chi-square statistics are not always a reliable tool for model selection.

The following simple example hints at the use of likelihood statistics as an important part of a model selection strategy. Table 10.3 contains artificial data to illustrate the log-likelihood approach to contrasting the utility of three models as possible summaries of the relationship within a table.

The general expression of the likelihood statistic associated with the example table is

$$\text{likelihood} = K \times \hat{p}_1^{n_1} \times \hat{p}_2^{n_2} \times \hat{p}_3^{n_3} \times \hat{p}_4^{n_4},$$

and the corresponding log-likelihood statistic becomes

$$\text{log-likelihood value} = L = log(K) + \sum n_i\, log(\hat{p}_i) \quad i = 1, 2, 3, \text{ and } 4,$$

Table 10.3 *Example: Three Possible Models to Represent Relationship within a Table of Example Data*

	Probability distributions					
	1	2	3	4	5	Total
			Data			
Counts (n_i)	16	12	20	28	24	100
Probabilities (p_i)	0.16	0.12	0.20	28	0.24	1.0
			Models			
0. Uniform ($p_i^{[0]}$)	0.20	0.20	0.20	0.20	0.20	1.0
1. Linear ($p_i^{[1]}$)	0.14	0.17	0.20	0.23	0.26	1.0
2. Threshold ($p_i^{[2]}$)	0.14	0.14	0.20	0.26	0.26	1.0

where again $\hat{p}_i = n_i/n$ and $n = \sum n_i$ is the total number of observations (Table 10.3, first two rows). Specifically for the data-estimated probabilities $\hat{p}_i$, the log-likelihood value is

$$\begin{aligned}\text{log-likelihood value} = L &= log(K) + 16\,log(0.16) + 12\,log(0.12) + \cdots + 24\,log(0.24)\\ &= -156.847.\end{aligned}$$

For the simplest uniform model (model 0 – parameters = 1), the corresponding theoretical likelihood value is

$$\text{likelihood} = K \times \left(p_1^{[0]}\right)^{n_1} \times \left(p_2^{[0]}\right)^{n_2} \times \left(p_3^{[0]}\right)^{n_3} \times \left(p_4^{[0]}\right)^{n_4},$$

and the log-likelihood value becomes

$$\text{log-likelihood} = L_0 = log(K) + \sum n_i\, log\left(p_i^{[0]}\right).$$

The specific value is

$$\begin{aligned}\text{log-likelihood value} = L_0 &= log(K) + 16\,log(0.20) + 12\,log(0.20) + 28\,log(0.20)\\ &\quad + 24\,log(0.20) = log(K) + \sum n_i\, log(0.20) = -160.944.\end{aligned}$$

The log-likelihood comparison of the data-generated probabilities follows, as before,

$$\text{log-likelihood chi-square statistic} = G_0^2 = 2[L - L_0] = 2[-156.847 - (-160.944)] = 8.193.$$

For a linear and slightly more complicated model (model 1 – parameters = 2), the log-likelihood value is

$$\text{log-likelihood} = L_1 = log(K) + \sum n_i\, log\left(p_i^{[1]}\right) = -158.391,$$

and the log-likelihood comparison of the data-generated probabilities ($\hat{p}_i$) to the second set of theoretical probabilities ($p_i^{[1]}$) is

$$\text{log-likelihood chi-square statistic} = G_1^2 = 2[L - L_1] = 2[-156.847 - (-158.391)] = 3.087.$$

For the more extensive threshold model (model 2 – parameters = 3), the log-likelihood value is

$$\text{log-likelihood} = L_2 = log(K) + \sum n_i\, log\left(p_i^{[2]}\right) = -157.288,$$

Table 10.4 *Data: Number of Winter Colds* (n_{ij})[a] *Reported during Clinical Trial of the Effectiveness of Large Doses of Vitamin C (Placebo versus Vitamin C – n = 237)*

	Counts of winter colds			
	No colds	1–4	5 or more	Total
Placebo	42	87	29	158
Vitamin C	27	48	4	79
Total	69	135	33	237

[a]Where $i = 1, 2$ = the number of rows and $j = 1, 2, 3$ = the number of columns.

and the log-likelihood comparison to the data-generated probabilities is

$$\text{log-likelihood chi-square statistic} = G_2^2 = 2[L - L_2] = 2[-156.847 - (-157.288)] = 0.881.$$

Note that when the theoretical values are identical to the observed values, the chi-square statistic is zero ($G^2 = 0$), and the model is called saturated.

To summarize, the three model log-likelihood chi-square statistics are

$$\text{log-likelihood chi-square statistic} = G_k^2 = 2[L - L_k] = 2\sum n_i \, log\left(\hat{p}_i / p_i^{[k]}\right) \quad k = 0, 1, \text{and } 2.$$

As the selected models increase in complexity, the corresponding reduction in likelihood test statistics reflects the increased accuracy of each model to summarize the data. For the example, the three likelihood values are again

$$\begin{aligned} G_0^2 &= 2\sum n_i \, log\left(\hat{p}_i / p_i^{[0]}\right) = 8.193, \\ G_1^2 &= 2\sum n_i \, log\left(\hat{p}_i / p_i^{[1]}\right) = 3.087, \text{ and} \\ G_2^2 &= 2\sum n_i \, log\left(\hat{p}_i / p_i^{[2]}\right) = 0.881. \end{aligned}$$

Another example illustrates a log-likelihood assessment of data contained in an $r \times c$ two-way table. The data are from a Canadian clinical trial conducted to assess the effectiveness of large doses of vitamin C as a way to reduce the frequency of winter colds (Table 10.4).

Table 10.5 contains the observed probability distribution.

Table 10.5 *Data: Proportions of Winter Colds* ($\hat{p}_{ij}$) *Reported during Clinical Trial of the Effectiveness of Large Doses of Vitamin C*

	Data estimated probabilities – $\hat{p}_{ij}$			
	No colds	1–4	5 or more	Total
Placebo	0.177	0.367	0.122	0.667
Vitamin C	0.114	0.203	0.017	0.333
Total	0.291	0.570	0.139	1.000

Table 10.6 *Theoretical Values: Probabilities of Winter Colds (p_{ij}) Calculated Exactly as if No Treatment Effect Exists from Vitamin C*

	Model probabilities – p_{ij}			
	No colds	1–4	5 or more	Total
Placebo	0.194	0.380	0.093	0.667
Vitamin C	0.097	0.190	0.046	0.333
Total	0.291	0.570	0.139	1.000

A natural comparison is between data-generated probabilities ($\hat{p}_{ij}$) and theoretical probabilities (p_{ij}) calculated as if vitamin C and the frequency of winter colds are not associated. These probabilities can be calculated in a variety of ways. Perhaps the simplest is to note that when no association exists, then for vitamin C (Table 10.6)

$$\begin{aligned}\text{probability} &= P(\text{vitamin C} \mid \text{no colds}) = P(\text{vitamin C}|1-4 \text{ colds}) \\ &= P(\text{vitamin C}|5^{+}\text{colds}) = P(\text{vitamin C}) = \tfrac{79}{237} = 0.333.\end{aligned}$$

For the example, the probability $p_{11} = P(\text{placebo and no colds}) = P(\text{placebo})\,P(\text{no colds}) = 0.667(0.291) = 0.194$, and the probability $p_{21} = P(\text{vitamin C and no colds}) = P(\text{vitamin C})\,P(\text{no colds}) = 0.333(0.291) = 0.097$ (Table 10.6, column 1).

In general, the likelihood value from a two-way table of observations $\hat{p}_{ij}$ is

$$\text{likelihood value} = K \prod\prod \hat{p}_{ij}^{n_{ij}} \qquad \text{where } n_{ij} = \text{a cell count}$$

for $i = 1, 2, \ldots, r =$ number of rows, $j = 1, 2, \ldots, c =$ number of columns. Then, specifically for the clinical trial data (Table 10.5), the log-likelihood value becomes

$$\begin{aligned}\text{log-likelihood} = L &= log(K) + \sum\sum n_{ij}\, log(\hat{p}_{ij}) = 42\, log(0.177) + \cdots + 4\, log(0.017) \\ &= -372.412.\end{aligned}$$

The theory-generated log-likelihood value, based on the conjecture that vitamin C use is exactly unrelated to the observed number of colds, is

$$\begin{aligned}\text{log-likelihood} = L_0 &= log(K) + \sum\sum n_{ij}\, log(p_{ij}) = 42\, log(0.194) + \cdots + 4\, log(0.046) \\ &= -377.034,\end{aligned}$$

where values p_{ij} are theoretical probabilities created as if exactly no association exists (Table 10.6). The likelihood comparison of data- and theory-generated probability distributions (Table 10.5 versus Table 10.6) produces the likelihood ratio test statistic

$$G^2 = 2[L - L_0] = 2[-372.412 - (-377.034)] = 9.244,$$

or, directly, the identical test statistic is

$$\begin{aligned}G^2 &= 2\sum\sum n_{ij}\, log(\hat{p}_{ij}/p_{ij}) \\ &= 2[42\, log(0.177/0.194) + \cdots + 4\, log(0.017/0.046)] = 9.244\end{aligned}$$

Table 10.7 *Notation: Representation of Additive Model for a Two-Way Table* – 4 × 3 *Table of Observations Denoted* y_{ij}

	Column variable		
Row variable	1	2	3
1	$y_{11} = a + r_1 + c_1$	$y_{12} = a + r_1 + c_2$	$y_{13} = a + r_1 + c_3$
2	$y_{21} = a + r_2 + c_1$	$y_{22} = a + r_2 + c_2$	$y_{23} = a + r_2 + c_3$
3	$y_{31} = a + r_3 + c_1$	$y_{32} = a + r_3 + c_2$	$y_{33} = a + r_3 + c_3$
4	$y_{41} = a + r_4 + c_1$	$y_{42} = a + r_4 + c_2$	$y_{43} = a + r_4 + c_3$

for $i = 1, 2$ and $j = 1, 2$, and 3. The degrees of freedom are two resulting in a p-value of $P(G^2 \geq 9.244 \mid \text{no association}) = 0.010$. The traditional Pearson chi-square comparison yields $X^2 = 8.025$ and a p-value 0.018 (Chapter 8).

Analysis of Tables – Continuous Data

Data classified into a two-way table create a fundamental question: Is the categorical variable used to classify observations into table rows related to the categorical variable used to classify the same observations into table columns? Typically statistical analysis of a two-way table starts with postulating that the categorical classifications are unrelated. The data are then analyzed for evidence that this conjecture is not true (Chapter 9).

Unrelated in the context of a statistical analysis of a two-way table translates into measuring the extent of nonadditivity with a statistical model.

For data classified into an $r \times c$ table, an additive model is

$$y_{ij} = a + r_i + c_j,$$

where again $i = 1, 2, \ldots, r =$ number of rows, $j = 1, 2, \ldots, c =$ number of columns, and the observed value in the ijth cell is denoted y_{ij}. Mechanically the model requires that the difference between any two rows is the same for all columns, and the difference between any two columns is the same for all rows. Table 10.7 illustrates additivity within a table with four rows ($r = 4$) and three columns ($c = 3$) containing continuous values denoted y_{ij}.

As always, the defining property of additivity is that influences from one variable are not influenced by any other variable; that is, contributions to the observed value y_{ij} from the row parameters r_i are the same for all levels of a column variable, and vice versa (Table 10.7). Additive influences in less precise language are referred to as independent, not associated, separate, uncorrelated, or unrelated.

Two-way tables can be distinguished by the kind of values classified. The values in a table can be measured quantities (continuous variables) such as mean values or rates.

Also, values in a table can be counts (discrete variables) such as the number of deaths or the number of low-birth-weight infants. Although analyses of both kinds of data are similar, they sufficiently differ to be usefully discussed separately.

An analysis of continuous data begins with the assumption of additivity between the categorical variables, and, in addition, the observed values classified into the table (again denoted y_{ij}) are known or assumed to have at least approximate normal distributions.

Table 10.8 *Data: Rates of Infants with Birth Weights Less than 2500 Grams per 1000 Live Births Classified by Parity and Maternal Education* (*< HS = Less than Grade* 12, *HS = Grade* 12 *and > HS = beyond Grade* 12)

	Maternal education		
Parity	<HS	HS	>HS
0	26.35	25.36	22.72
1	25.57	27.49	26.43
2	26.92	28.10	23.58
≥3	28.17	26.84	26.10

Note: HS denotes completed a high school education.

For example, the rates of low-birth-weight infants (weights less than 2500 grams per 1000 live births) classified by parity (four levels) and maternal education (three levels) are assumed to have at least approximate normal distributions with the same variance (Table 10.8).

Applying an additive model to these data to analyze and describe the influences of maternal education and parity on an infant's birth weight produces estimated row and column parameters (Table 10.9) (Chapter 9). Twelve model-estimated rates calculated from the six model parameter estimates $\hat{a}, \hat{r}_i$, and $\hat{c}_j$ perfectly conform to the additive model (Table 10.10).

The degrees of freedom associated with an analysis of a two-way table are the number of cells (observations) in the table ($r \times c$) reduced by the number of parameters estimated to establish the model-estimated values. The r rows require $r - 1$ estimates because it is not necessary to estimate one value because the r row estimated values must add to the observed row total. Similarly, the c columns require $c - 1$ estimated values. The degrees of freedom are the difference between the number of cells in the table (observations) and the number of estimated values (parameters) or $df = rc - [(r - 1) + (c - 1) + 1] = (r - 1)(c - 1)$ including the single estimated model constant term a. Computer software to estimate additive

Table 10.9 *Model: Computer-Estimated Parameters from Additive Model (Birth Weight Classified by Parity and Education)*

Parameters	Estimates	*s. e.*
$\hat{a}$	25.427	–
$\hat{r}_1$	0.0	–
$\hat{r}_2$	1.689	1.042
$\hat{r}_3$	1.390	1.042
$\hat{r}_4$	2.227	1.042
$\hat{c}_1$	0.0	–
$\hat{c}_2$	0.195	0.902
$\hat{c}_3$	−2.045	0.902

Table 10.10 *Estimation: Exactly Additive Rates of Low-Birth-Weight Infants per 1000 Live Births Based on Model-Estimated Parameters (Table 10.9)*

	Maternal education		
Parity	<HS	HS	>HS
0	25.427	$25.427 + 0.195 = 25.62$	$25.427 - 2.045 = 23.38$
1	$25.427 + 1.687 = 27.11$	$25.427 + 1.687 + 0.195 = 27.31$	$25.427 + 1.687 - 2.045 = 25.07$
2	$25.427 + 1.390 = 26.82$	$25.427 + 1.390 + 0.195 = 27.01$	$25.427 + 1.390 - 2.045 = 24.77$
≥3	$25.427 + 2.227 = 27.65$	$25.427 + 2.227 + 0.195 = 27.85$	$25.427 + 2.227 - 2.045 = 25.61$

model parameters is one way to calculate model-generated values (Tables 10.9 and 10.10). An easily applied and intuitive method is described next.

An alternative to generating parameter estimates to construct a formal additive model is a direct and model-free process given the name *mean polish*, a term originated by statistician J. W. Tukey (b. 1915), that provides an natural and simple analysis of a two-way $r \times c$ table. The first step is again to assume additivity among the categorical variables. As a direct result, the additive row and column influences can be estimated separately based on the row and column mean values.

Continuing the low-birth-weight example, the row and column mean values are displayed in Table 10.11.

Subtracting the estimated row variable mean values (parity) from each corresponding row value produces a "new" table that describes only the column variable influences (education) and lack of fit of an additive relationship. The additive row influences are "removed." Table 10.12 contains the results. Repeating the process, similarly subtracting the column mean values (education) from each remaining column value "removes" the additive column influences (Table 10.12, bottom row). The resulting table contains the difference between values calculated under the conjecture of additivity and the observed data, leaving only values that result from nonadditivity, the residual values (Table 10.13).

When categorical variables have exactly additive influence, all residual values are zero. Thus, subtracting the residual values from the original data produces exactly additive estimated values (Table 10.14). Furthermore, these mean polish estimated values are identical

Table 10.11 *Mean Polish: Data and Mean Values of the Infant Birth-Weight Data (Repeated Table 10.8)*

	Education			
Parity	<HS	HS	>HS	Mean values
0	26.35	25.36	22.72	24.81
1	25.57	27.49	26.43	26.50
2	26.92	28.10	23.58	26.20
≥3	28.17	28.17	26.10	27.04
Mean values	26.75	26.95	24.71	26.14

Table 10.12 *Mean Polish: Row Influence Removed Leaving the Column Influences on the Low-Birth-Weight Data*

	Education			
Parity	<HS	HS	>HS	Mean values
0	1.54	0.55	−2.09	0.0
1	−0.93	0.99	−0.07	0.0
2	0.72	1.90	−2.62	0.0
≥3	1.13	−0.20	−0.94	0.0
Mean values	0.616	0.81	−1.43	0.0

to the values generated from the computer-estimated parameters from an additive model analysis. For example, the model-estimated rate of low-birth-weight infants for mothers with high school education and infants with parity 1 is $\hat{y}_{22} = 25.427 + 1.687 + 0.195 = 27.308$ (Tables 10.10 and 10.14). Note that the differences between column estimates are the same for any pair of rows, and the differences between the row estimates are the same for any pair of columns, a required property of additive estimates (Table 10.14). These two analyses (model or mean polish) are a partitioning of the observed values into two parts, additive and residual components (observation = additive + residual). The residual values reflect either random variation only or random variation plus systematic influences caused by an association between the categorical variables (nonadditivity). For the parity/education data all 12 residuals values appear small and random relative to the observed values (Table 10.13).

A statistical analysis of a table can take a number of forms. One simple and effective approach is to explore the pattern of the residual values (Chapter 9). When the residual values are random, small, and inconsequential, they support the conjecture that the two categorical variables are unrelated. Conversely, large residual values, particularly when they are distributed in nonrandom patterns, are not only an indication of an association; they potentially provide insight into the properties of the apparent association between the categorical variables. Thus, residual values are key to revealing a pattern or lack of a pattern in an $r \times c$ table.

Table 10.13 *Mean Polish: Column Influences Removed Leaving the Residual Values from the Low-Birth-Weight Data*

	Education			
Parity	<HS	HS	>HS	Mean values
0	0.92	−0.26	−0.66	0.0
1	−1.54	0.18	1.36	0.0
2	0.10	1.09	−1.19	0.0
≥3	0.52	−1.01	0.49	0.0
Mean values	0.0	0.0	0.0	0.0

Note: No row or column additive influences.

Table 10.14 *Mean Polish: Additive Model Estimates – Estimated Residual Values Subtracted from Observed Values (Identical to Table 10.10)*

	Education		
Parity	<HS	HS	>HS
0	25.43	25.62	23.38
1	27.11	27.31	25.07
2	26.82	27.01	24.77
≥3	27.65	27.85	25.61

A 6 × 6 table of mean birth weights (kilograms) classified by maternal age (columns – six levels) and infant parity (rows – six levels) illustrates (Table 10.15). Applying either the computer-estimated-model approach or Tukey's mean polish produces values estimated under the conjecture of additivity of the categorical variables (age and parity). Thus, the estimated birth weights are exactly additive. As before, the differences between these additive model values and the observed data yield a table of residual values. The magnitude of these residual values depends on the measurement units. Standardization, for example, dividing by the standard deviation of the residual values, yields a table of readily interpretable residual values (Table 10.16).

A property of residual values is that the row and column mean values are exactly zero. The variance of these residual values is estimated by

$$\textit{estimated variance}(\text{residual}) = \hat{v} = \frac{1}{df}\sum\sum r_{ij}^2 \quad i = 1, 2, \ldots, r \text{ and } j = 1, 2, \ldots, c,$$

where r_{ij} denotes the residual value from the ijth cell of the table. When an additive model accurately represents the observed values and the original data have at least an approximate normal distribution, the standardized residual values $(r_{ij}/\sqrt{\hat{v}})$ have a standard normal distribution (mean value = 0 and variance = 1.0). Table 10.16 displays the $r \times c = 6 \times 6 = 36$

Table 10.15 *Data: Mean Birth Weights (Kilograms) of Newborn Infants Classified by Parity and Maternal Age (New York State Vital Records Data)*

	Maternal age					
Parity	<20	20–25	25–30	30–35	35–40	>40
0	3.281	3.305	3.291	3.258	3.225	3.202
1	3.278	3.354	3.371	3.356	3.322	3.303
2	3.280	3.360	3.397	3.390	3.374	3.335
3	3.220	3.345	3.405	3.422	3.412	3.392
4	3.202	3.332	3.399	3.434	3.435	3.417
≥5	3.333	3.315	3.398	3.443	3.467	3.482

Table 10.16 *Analysis: Standardized Residual Values from Infant Birth Weights Classified by Parity and Maternal Age*

	Maternal age						
Parity	<20	20–25	25–30	30–35	35–40	>40	Total
0	2.10	1.17	0.04	−0.77	−1.21	−1.33	0.0
1	0.61	0.74	0.24	−0.21	−0.67	−0.71	0.0
2	0.14	0.35	0.25	−0.03	−0.13	−0.57	0.0
3	−1.29	−0.16	0.21	0.41	0.44	0.39	0.0
4	−1.74	−0.51	0.01	0.58	0.83	0.82	0.0
≥5	0.19	−1.59	−0.75	0.02	0.74	1.40	0.0
Total	0.0	0.0	0.0	0.0	0.0	0.0	0.0

standardized residual values from the analysis of the birth-weight data (Table 10.15) where the estimated variance is $\hat{v} = 0.00241$. Large and small values become apparent.

A visual version of the distribution of the residual values is created by representing positive residual values with a plus sign and negative residual values with a minus sign (Table 10.17). A clear pattern emerges from the birth-weight residual values. The negative residual values are associated with young mothers with high-parity infants and with older mothers with low-parity infants. Necessarily, the positive values are then associated with older mothers with high-parity infants and with younger mothers with low-parity infants. Thus, the pattern of residual values suggests decreasing birth weight with increasing maternal age for low-parity infants and increasing birth weights with increasing maternal age for high-parity infants.

The description of the properties of a two-way table containing continuous observations using an additive model is a special case of a wide range of statistical techniques called the analysis of variance (Chapter 13). Entire books are devoted to this statistical technique.

Matched Pairs – The Categorical Variable Case

Matched pairs analysis is a statistical strategy frequently involving categorical data classified into a two-way table (Chapter 15). These tables summarize counts of collected pairs, not

Table 10.17 *Analysis: A Representation of the Standardized Residual Values for Birth Weights of Newborn Infants Classified by Parity and Maternal Age (+ and −) SSP*

	Maternal age					
Parity	<20	20–25	25–30	30–35	35–40	>40
0	+	+	+	−	−	−
1	+	+	+	−	−	−
2	+	+	+	−	−	−
3	−	−	+	+	+	+
4	−	−	+	+	+	+
≥5	+	−	−	+	+	+

Table 10.18 *Data: Counts of Matched Pairs* ($n = 502$) *Classified by Case/Control and Socioeconomic Status (SES)*

		Controls		
		High SES	Low SES	Total
Cases	High SES	$a = 43$	$b = 225$	268
	Low SES	$c = 132$	$d = 102$	234
	Total	175	327	502

single observations. A 2×2 table made up of counts of four kinds of matched pairs summarizes the joint occurrence of two binary variables. For example, the binary variables, case/control status, and high and low socioeconomic status create a table of matched pairs (Table 10.18). These case/control pairs are part of a study of childhood leukemia (cases) matched for sex, ethnicity, and age (Chapter 15). The matching control subjects are selected from birth certificate records of healthy children who match the sex, ethnicity, and age of the case.

A table of matched data consists of pairs concordant (a and d) or discordant (b and c) for the risk factor. Concordant pairs do not provide information about the relationship between the risk factor and case/control status. They are "matched" for the risk factor; that is, any observed difference within concordant pairs cannot be attributed to a risk factor because both members of the pair have the risk factor or both do not have the risk factor. Not only are the concordant pairs matched for selected factors, they are also "matched" for the risk factor. The data relevant to identifying an association between outcome (case/control status) and risk factor (SES status) consist of the remaining discordant pairs. The analysis of matched pairs data, therefore, is based on only discordant pairs. For the example data, there are $b = 225$ and $c = 132$ discordant pairs (Table 10.18). The total number of discordant pairs $n = b + c = 357$ becomes the sample analyzed.

The evidence of an association between case/control status and a risk factor is measured by a comparison of the counts of the two kinds of discordant pairs (denoted b and c). When case/control status is unrelated to the risk factor, counts b and c differ by chance alone. More formally, probabilities P(risk factor present | case) and P(risk factor present | control) are equal. From another perspective, the statistical analysis is an assessment of symmetry of the 2×2 matched pairs table. When exactly no association exists, the matched pairs table is symmetric. That is, the counts of discordant pairs are equal or $b = c$.

A test statistic applied to assess the equality of n discordant pairs using a typical normal distribution to approximate binomial probabilities is

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}},$$

where $\hat{p} = b/n = 0.630$ estimates the probability that a case processes the risk factor among the discordant pairs. When $p_0 = 0.5$ (no case/control difference), then only random

Table 10.19 *Data: Counts of Case/Control Matched Pairs (n_{ij}) Classified by Level of Socioeconomic Status (SES) Distributed into a 5 × 5 Table ($n = 502$) from a Study of Childhood Leukemia*

		Controls					
		ses_1	ses_2	ses_3	ses_4	ses_5	Total
Cases	ses_1	(11)	20	12	11	28	82
	ses_2	15	(20)	11	14	37	97
	ses_3	8	12	(12)	13	33	78
	ses_4	8	9	8	(10)	46	81
	ses_5	6	16	12	38	(92)	164
	Total	48	77	55	86	236	502

Note: ses_i represents the ith socioeconomic level.

differences exist between counts of the two kinds of discordant pairs. Because $\hat{p} = b/(b + c)$ and $p_0 = 0.5$, the test statistic z simplifies to

$$X^2 = z^2 = \frac{(b - c)^2}{b + c},$$

frequently called McNemar's test and has an approximate chi-square distribution when no association exists within pairs. For the matched pairs case/control leukemia data where $b = 225$ and $c = 132$, the chi-square–distributed test statistic (degrees of freedom = 1)

$$X^2 = \frac{(225 - 132)^2}{225 + 132} = 24.227$$

generates a p-value less than 0.001.

This same logic and pattern of analysis applies to any categorical variable used to classify matched case/control pairs data into a table. For a k-level categorical variable, a $k \times k$ table results. The matched pairs data (Table 10.18) distributed into five levels ($k = 5$) of increasing socioeconomic status are displayed in a 5 × 5 table (Table 10.19).

The analysis of the relationship between case/control status and socioeconomic levels again translates into assessing table symmetry. Like the 2 × 2 table, a symmetric matched pairs table occurs when exactly no association exists between case/control status and socioeconomic levels. Thus, theoretical counts of corresponding symmetric discordant pairs are equal, or, in symbols, the cell counts are $n_{ij} = n_{ji}$. Also, like the 2 × 2 table, concordant pairs are "matched" for the risk factor (*ses*). For example, from Table 10.19, the count $n_{13} = 12$ and the corresponding symmetric count $n_{31} = 8$ would then differ by chance alone; that is, the observed number of pairs would randomly differ from the theoretical value given by $n_{13} = n_{31} = (8 + 12)/2 = 10$. In general, when case/control status is unrelated to the classification of the counts n_{ij}, the expected cell frequencies are

$$\text{expected value case/control frequency} = e_{ij} = \frac{n_{ij} + n_{ji}}{2}.$$

Using the data in Table 10.19, these theoretical counts of pairs discordant for SES produce the corresponding symmetric case/control array (Table 10.20).

Table 10.20 *Theory: Case/Control Status and Socioeconomic Levels Matched Pairs Calculated as if Exactly No Within-Pair Association Exists* ($n = 502$)

		Controls					
		ses_1	ses_2	ses_3	ses$_4$	ses_5	Total
Cases	ses_1	(11)	17.5	10.0	9.5	17.0	65
	ses_2	17.5	(20)	11.5	11.5	26.5	87
	ses_3	10.0	11.5	(12)	10.5	22.5	66.5
	ses_4	9.5	11.5	10.5	(10)	42.0	83.5
	ses_5	17.0	26.5	22.5	42.0	(92)	200
	Total	65	87	66.5	83.5	200	502

Pearson's chi-square test statistic to evaluate the differences between observed (Table 10.19) and expected counts (Table 10.20) is

$$X^2 = \sum\sum (n_{ij} - e_{ij})^2/e_{ij},$$

where $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, k$. Similar to the 2×2 case/control analysis, the expression for the chi-square test statistic becomes

$$X^2 = \sum\sum \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}} \qquad \text{for all pairs } i < j = 1, 2, \ldots, k(k-1)/2,$$

where $i < j$ indicates counts of pairs above the diagonal table values (n_{ii} – Table 10.20); that is, the chi-square statistic is the sum of McNemar-like statistics. The degrees of freedom are equal to number of these values, namely, $df = k(k-1)/2$. In addition, as noted, the concordant pairs clearly do not influence the test statistic (diagonal values: $n_{ii} - n_{ii} = 0$) making the effective sample size again $n = 357$ discordant pairs. For the example of matched pairs data, the degrees of freedom are $df = 5(4)/2 = 10$, and the corresponding chi-square test statistic $X^2 = 37.427$ yields again a p-value less than 0.001.

Analysis of Tables – Count Data

Commonly tables contain counts. A typical analysis employs a multiplicative measure of association between categorical variables; that is, for category levels of two independent variables represented by A_i and B_j, the corresponding cell frequency is determined by the product of probabilities $P(A_i \text{ and } B_j) = P(A_i)P(B_j)$. Therefore, when categorical variables are unrelated, the marginal probabilities entirely determine the cell probabilities and a table is unnecessary. Consider the counts from a study of coronary heart disease risk where one categorical variable indicates the presence or absence of a coronary event (two levels) and a second variable indicates reported smoking exposure (four levels) contained in the $r \times c = 2 \times 4 = 8$ cells of a table (Table 10.21).

Table 10.21 *Data: Counts of Coronary Events (chd) from High-Risk Men (Type-A Behavior with High Blood Pressure) by Four Levels of Smoking Exposure* ($n = 388$)

	Cigarettes smoked per day				
	0	1–20	20–30	>30	Total
chd	29	21	7	12	69
No *chd*	155	76	45	43	319
Total	184	97	52	55	388

When these categorical variables are independent, then the expected number of coronary events among nonsmokers is

$$n_{11} = nP(A_1 \text{ and } B_1) = nP(A_1)P(B_1) = 388\left[\frac{69}{388}\right]\left[\frac{184}{388}\right] = 32.72$$

among the $n = 388$ study subjects (Table 10.22). Thus, the marginal frequencies 69 and 184 determine the theoretical frequency of *chd* events among the nonsmokers (32.72). Applying the same calculation to each cell produces a table of calculated counts as if a *chd* event is exactly unrelated to smoking exposure (Table 10.22).

Logarithms of independent cell counts transform multiplicative relationships into additive relationships. In symbols, products of probabilities become differences in the logarithms of probabilities; that is,

$$\text{nonadditivity} = log[P(A_i \text{ and } B_j)] - [log(A_i) + log(B_j)]$$

measures the association between variables A_i and B_j. Therefore, parallel to the continuous data case, an additive model (no association) among categorical variables again becomes

$$log(n_{ij}) = a + r_i + c_j.$$

Thus, additivity of *log*-values of table counts is equivalent to multiplicative independence of categorical variables.

Again, parametric assumptions produce a rigorous analysis of a two-way table of counts based on an additive model. A Poisson probability distribution is a frequent choice as a description of the distribution of the observed cell counts (Chapter 9). A loglinear regression

Table 10.22 *Results: Estimated Counts of Coronary Heart Disease Events among High-Risk Men as if* chd *Events Are Exactly Unrelated to Smoking Exposure* ($n = 388$)

	Levels of smoking exposure				
	0	1–20	20–30	>30	Total
chd	32.72	17.25	9.25	9.78	69
No *chd*	151.28	79.75	42.75	45.22	319
Total	184	97	52	55	388

Table 10.23 *Model: Estimated Model Parameters from the Additive Loglinear Model of the Relationship between* chd *Events and Smoking Exposure* ($n = 388$)

Parameters	Estimates	*s. e.*
$\hat{a}$	3.488	–
r_1	0.0	–
$\hat{r}_2$	1.531	0.133
c_1	0.0	–
$\hat{c}_2$	−0.640	0.126
$\hat{c}_3$	−1.264	0.157
$\hat{c}_4$	−1.208	0.154

model and this parametric assumption applied to coronary heart disease and smoking data produce five model-estimated parameters a, r_i, and c_j (Table 10.23).

The estimated *log*-counts from this additive loglinear Poisson model, when exponentiated, produce cell frequencies calculated as if coronary events and smoking exposure are statistically independent. The estimated values are identical to the values calculated from the marginal probabilities. For example, two estimated frequencies are

$$\hat{n}_{22} = e^{\hat{a}+\hat{r}_2+\hat{c}_2} = e^{3.488+1.531-0.640} = e^{4.379} = 79.750 \text{ (model) or } 388(319/388)(184/388) = 79.750 \text{ (table).}$$

These *chd*-data produce a distribution of estimated probabilities of the coronary events (denoted $\hat{p}_{ij}$ – Table 10.24), and the loglinear model estimates produce a distribution of the theoretical and additive probabilities (denoted $\hat{P}_{ij}$ – Table 10.24).

The log-likelihood chi-square statistic (degrees of freedom = 3) provides a direct comparison of these two probability distributions. Specifically, the test statistic to assess the conjecture of additivity

$$G^2 = 2\sum\sum n_{ij}\, log(\hat{p}_{ij}/\hat{P}_{ij}) = 2.765 \quad i = 1, 2, \quad \text{and} \quad j = 1, 2, 3, \quad \text{and } 4$$

yields the corresponding p-value of 0.429. The application of the Pearson chi-square test statistic to the same estimates, as usual, produces the expected similar result

$$X^2 = \sum (n_{ij} - n\hat{P}_{ij})^2/n\hat{P}_{ij} = 2.783$$

that yields a p-value of 0.426.

Table 10.24 *Estimates: Data* ($\hat{p}_{ij}$) *and Model Estimated* ($\hat{P}_{ij}$) *Probabilities of a* chd *Event for Four Levels of Smoking*

	Levels of smoking exposure (*chd*)				Levels of smoking exposure (no *chd*)				
	0	1–20	20–30	>30	0	1–20	20–30	>30	Total
n_{ij}	29	21	7	12	155	76	45	43	388
$\hat{p}_{ij}$	0.075	0.054	0.018	0.031	0.399	0.196	0.116	0.111	1.0
$\hat{P}_{ij}$	0.084	0.044	0.024	0.025	0.390	0.206	0.110	0.117	1.0

Table 10.25 *Estimates: Nonparametric Median Polish Analysis Applied of Coronary Events among High-Risk Men Classified by Four Levels of Smoking Exposure Produce Additive* Log-*Counts and Independent Counts (in Parentheses)*

	Median polish			
	0	1–20	20–30	>30
chd	3.55 (34.93)	3.04 (20.81)	2.22 (9.25)	2.47 (11.84)
No *chd*	4.86 (128.68)	4.34 (76.68)	3.62 (34.07)	3.78 (43.60)

An alternative nonparametric analysis is created with a minor modification of the mean polish *analysis*. Applied to the logarithms of table counts, row influences are estimated and subtracted from the table counts to "remove" the additive row influences. To be truly a nonparametric technique, however, the median row values and not the mean row values are used. Then the column influences are estimated by their column median values and are subtracted from the newly created table values. As before, the table now contains residual values. Unlike the mean polish, the median polish can require this process to be repeated several times until the estimates converge to stable values. Subtracting the stable residual values from the observed values and exponentiating these median polish log-estimates again produce theoretical counts as if the row and column categorical variables are exactly unrelated, without a formal model or parametric assumptions; that is, unrelated means additive influences. For the example *chd*-data, estimated values are displayed in Table 10.25.

For an analysis of a two-way table, as the example illustrates, parametric and nonparametric methods produce similar but not identical results. The median polish of the *log*-counts also yields estimated values that are statistically independent (values in the parentheses). Using these estimated values, plots, chi-square evaluations and residual value analyses directly apply. Furthermore, a nonparametric analysis is distribution-free and minimally affected by extreme observations.

Alternatively, a parametric model analysis easily generalizes to more extensive tables (more than two categorical variables). Of more importance, a model-based approach provides a particularly efficient and valuable interpretation of the difference between observed data and additive model estimates.

For a two-way table, the loglinear model can be extended to include interaction terms. The expression for the nonadditive model becomes

$$log(n_{ij}) = a + r_i + c_j + [r_i \times c_j].$$

This nonadditive model of cell counts requires a total of $(r-1)(c-1)$ parameters which equals to the degrees of freedom. As noted, such a model is said to be saturated, and model-estimated cell counts are identical to observed cell counts ($\hat{n}_{ij} = n_{ij}$). The value of the chi-square test statistic is zero. Therefore, for any value derived from estimated parameters of a saturated model, the identical value can be calculated directly for the data. From another point of view, the interaction terms directly measure the extent of nonadditivity among the

Table 10.26 *Description: A General Introductory Characterization of Four Kinds of Three-Way Tables*

Relationships	Properties	Symbols	Examples[a]
Complete independence	No associations	$A + B + C$	$AB = AC = BC = 0$
Joint independence	One association	$A + B + C + AB$	$AC = BC = 0$
Conditional independence	Two associations	$A + B + C + AB + AC$	$BC = 0$
Additive associations	Three associations	$A + B + C + AB + AC + BC$	$ABC = 0$

[a]If two variables are independent, a three-way association does not occur.

categorical variables. Thus interaction effects directly measure the lack of fit (accuracy) of the additive model. To summarize,

$$\text{interaction effects} = \text{data} - \text{additive effects} = \text{lack of fit}.$$

Therefore, small and negligible interaction effects are necessary for an accurate additive model; that is, nonadditivity measures association. The model-defined terms interaction and additivity are synonymous with the less technical terms association and independence when applied to categorical variables in a table of counts. Evaluation of interaction influences, from a model approach, furthermore, are particularly useful in the analysis of tables constructed from more than two categorical variables (see next sections).

Three-Way Tables

An analysis of a three-way table is essentially a more intensive application of the techniques applied to a two-way table and, in addition, introduces the path to analysis of higher-dimensional tables. Three categorical variables (labeled A, B, and C for convenience) produce four kinds of three-way tables (Table 10.26) and eight specific loglinear models.

Along the same lines as a two-way table, log-likelihood comparisons are an ideal statistical tool to explore relative accuracy of various models to identify relationships within a table constructed from three or more categorical variables. As usual, the analysis is a search for a simple but effective description of the relationships within collected data.

To contrast statistical models using log-likelihood statistics calculated from a loglinear model, three requirements exist. The data used to estimate the parameters of compared models must be exactly the same. This requirement is particularly important because the magnitude of a log-likelihood value directly depends on the number of observations. The models compared must be *nested*. Nested means that a simpler model is created by eliminating one or more variables from a more extensive model. Therefore, the reduced model (nested model) requires fewer parameters. Last, the difference between log-likelihood values has a chi-square distribution when the two models differ by chance alone. The difference in log-likelihood values then results only from the increased ability of the more extensive model to capitalize on random variation to improve the correspondence between model estimates and data.

The Analysis of the Three-Way Table

To start, three binary variables, again designated A, B, and C, tabulated into $2 \times 2 \times 2$ tables are displayed at the end of this section illustrating four kinds of loglinear models applied to the analysis of three-way tables in detail (Table 10.26). The "data" created to conform perfectly to the model requirements demonstrate clearly the details and properties of each analysis (residuals = 0). The analysis of higher dimensional tables follows much the same patterns. Two case studies conclude the chapter.

Complete Independence

Complete independence is another name for additivity of three categorical variables (in symbols, $AB = 0$, $AC = 0$, and $BC = 0$). An expression for the additive loglinear model is

$$log(n_{ijk}) = \mu + a_i + b_j + c_k.$$

For this model and the following models, the indexes i (rows), j (columns), and k (layers) identify the combinations of the model parameters used to estimate the table counts (denoted $\hat{n}_{ijk}$) where 0 indicates the absence and 1 indicates the present of a parameter (Display 1.0). For the example, the cell frequency n_{001} represents the absence of parameter a_i, the absence of parameter b_j and the presence of parameter c_k. For example, the cell frequency n_{001} (upper left corner of the C_1-table – Display 1.0). For model parameter estimates $\hat{\mu} = 2.773$ and $\hat{c} = 0.693$, the cell frequency n_{001} is

$$\hat{n}_{001} = e^{log(\hat{n}_{001})} = e^{\hat{\mu}+\hat{c}} = e^{2.773+0.693} = 32.$$

Three marginal probabilities $P(A) = 0.8$, $P(B) = 0.8$ and $P(C) = 0.333$ determine the probabilities for all levels of the variables A, B, and C (Display 1). For example, the probability $P(AB \mid C_1) = P(AB) = P(A)P(B) = (0.80)(0.80) = 0.64$, thus $\hat{n}_{001} = 50(0.64) = 32$. Furthermore, tables formed by adding any pair of three categorical variables (in the display adding tables C_0 and C_1) produces a table with values determined entirely by the other two variables (A and B in the example) and remain independent. This property described in terms of the odds ratios is

$$\text{complete independence: } or_{AB|C_0} = or_{AB|C_1} = or_{AB|C_0+C_1} = 1.0.$$

The variable pairs AC and BC have the identical property.

As noted, the defining property of complete independence is that a table is unnecessary. The probability of any combination of the categorical variables A, B, and C is a product of the corresponding marginal probabilities. A completely additive model plays a minor role in a model selection process. It is the model that produces the maximum possible log-likelihood comparison because it is the simplest possible nested model (fewest parameters) that can be constructed to represent the relationships among the categorical variables. Each parameter added to the complete independence model produces estimated cell frequencies that more closely resemble the data values and produce a log-likelihood value closer to zero (perfect fit).

Joint Independence

The term *joint independence* refers to relationships within a three-way table when one of the three categorical variables is unrelated to the other two. A loglinear model description of a three-way table where variable C is unrelated to both variables A and B is

$$log(n_{ijk}) = \mu + a_i + b_j + c_k + [a_i \times b_j].$$

The detailed case of joint independence of binary variable C (in symbols, $AB \neq 0$ and $AC = BC = 0$) is illustrated in Display 2. The variable C does not influence the AB-relationship. Like a two-way table, the interaction parameter (denoted $a_i \times b_j$) measures the extent of nonadditivity (association) between variables A and B. The fundamental feature of joint independence in a three-way table is that one variable is not relevant to the analysis of the other two variables. For the example, because variable C is unrelated of variables A and B, sometimes called a nonconfounding variable, the table created by combining the C_0-table and the C_1-table exhibits exactly the same relationships found in each subtable. From a mechanical point view, subtables C_0 and C_1 are proportional to the collapsed table, or, in symbols,

$$f \times [(C_0 + C_1)\text{-table}] = C_0\text{-table and } (1 - f) \times [(C_0 + C_1)\text{-table}] = C_1\text{-table}.$$

The value f for the example tables is 0.25 (Display 2). Because the observations in each subtable are proportional, multiplicative summary measures such as ratios or odds ratios, remain the same in the combined table. These measures become more precise and yield a simpler description of the AB-relationship because they are estimated from larger numbers of observations (smaller variance) contained in a single table. In terms of the odds ratio (Display 2), this property is illustrated by

$$\text{joint independence: } or_{AB|C_0} = or_{AB|C_1} = or_{AB|C_0+C_1} = or_{AB},$$

and, for the example, the odds ratio in all three tables is $or_{AB} = 4.0$.

Conditional Independence

When only one pair of categorical variables is independent, a loglinear model requires two interaction terms. An expression for this nonadditive model applied to a three-way table is

$$log(n_{ijk}) = \mu + a_i + b_j + c_k + [a_i \times c_k] + [b_j \times c_k],$$

where variables A and B are independent (in symbols, $AB = 0$, $AC \neq 0$, and $BC \neq 0$). A detailed illustration describes "data" when only variables A and B are unrelated (Display 3).

The existence of one pair of independent categorical variables allows subtables to be usefully created that again produce more precise and simpler descriptions of the role of each of two categorical variables. These combined tables can be created only from one or

the other of the two independent variables. Specifically, when variable A is independent of variable B, then the table formed by collapsing over variable A provides an unbiased analysis of the BC-relationship (Display 4). The table formed by collapsing over variable B similarly provides an unbiased analysis of the AC-relationship (Display 4). In terms of odds ratios, this property is illustrated by

$$or_{BC|A_0} = or_{BC|A_1} = or_{BC|A_0+A_1} = 4.0 \quad \text{and } or_{AC|B_0} = or_{AC|B_1} = or_{AC|B_0+B_1} = 0.75.$$

No Pairwise Independence

Continuing the previous pattern, when all three categorical variables are associated, the loglinear model requires three interaction terms (in symbols, $AB \neq 0$, $AC \neq 0$, and $BC \neq 0$). For a three-way table (Display 5), the expression for the loglinear model becomes

$$log(n_{ijk}) = \mu + a_i + b_j + c_k + [a_i \times b_j] + [a_i \times c_k] + [b_j \times c_k].$$

When no pairwise independence exists, it is not possible to collapse a three-way table to create unbiased assessments of any of the three pairwise relationships. Nevertheless, the interaction terms in the model do not interact; that is, additivity allows an assessment of a specific interaction free from influence from the other categorical variable. In other words, each of the three pairwise interaction terms in the model provides an estimated measure of association uninfluenced by the third variable. Thus, three separate pairwise associations can be routinely estimated and assessed. Specifically, again in terms of odds ratios,

$$or_{AB|C_0} = or_{AB|C_1} = or_{AB} = e^d, or_{AC|B_0} = or_{AC|B_1} = or_{AC} = e^f \text{ and}$$
$$or_{BC|A_0} = or_{BC|A_1} = or_{BC} = e^e.$$

From the illustration (Display 5), these three odds ratios are $or_{AB} = e^d = e^{1.386} = 4.0$ (unaffected by variable C), $or_{AC} = e^f = e^{0.405} = 1.5$ (unaffected by variable B) and $or_{BC} = e^e = e^{-0.288} = 0.75$ (unaffected by variable A) calculated from the interaction model coefficients d, f, and e.

The loglinear model analysis of a table is primarily an analysis of interactions. The noninteraction terms in the model, sometimes called main effects *terms*, reflect the frequency of the categorical variables sampled. Although these terms have a necessary technical role, the main effects do not influence the analysis of the pairwise relationships among the categorical variables used to construct the table. As always, when interaction terms are a component of a model, main effect terms are not simply interpreted.

The essential loglinear model properties are the following:

Association	Summary	Analysis
None	No table	Collapses to three probabilities
One pair	One summary table	Collapses over the independent variable (nonconfounding)
Two pairs	Two summary tables	Collapses over either variable of the independent pair
Three pairs	No summary table	Separate estimates of the three pairwise associations

Log-Linear Models for Four 2 × 2 × 2 Three-Way Tables

The following examples display four loglinear models applied to the analysis of three-way tables created from three binary variables (labeled A, B, and C) illustrated with artificial data that perfectly conform to the model requirements (no random variation or all residual values = 0).

Display 1. Example: Variables *A*, *B*, and *C* Pairwise Independent (Complete Independence – *AB* = *AC* = *BC* = 0)

C_0	A_0	A_1	Total
B_0	16	4	20
B_1	4	1	5
Total	20	5	25

C_1	A_0	A_1	Total
B_0	32	8	40
B_1	8	2	10
Total	40	10	50

Table $C_0 + C_1$

$C_0 + C_1$	A_0	A_1	Total
B_0	48	12	60
B_1	12	3	15
Total	60	15	75

Model and Data

Frequency	n_{000}	n_{100}	n_{010}	n_{110}
Counts	16	4	4	1
Model	μ	$\mu + a$	$\mu + b$	$\mu + a + b$

Frequency	n_{001}	n_{011}	n_{101}	n_{111}
Counts	32	8	8	2
Model	$\mu + c$	$\mu + a + c$	$\mu + b + c$	$\mu + a + b + c$

Model Values

Parameters	μ	a	b	c
Estimates	2.773	−1.386	−1.386	0.693
$e^{\text{estimates}}$	16	0.25	0.25	2.0

Example of Completely Independent Variables

additive scale: $log(n_{111}) = 2.773 - 1.386 - 1.386 + 0.693 = 0.693$

multiplicative scale: $n_{111} = 16(0.25)(0.25)(2) = 2$

Example Probabilities

When $P(A_0) = 0.8$, $P(B_0) = 0.8$ and $P(C_0) = 0.5$, then
$P(A_0B_0|C_0) = P(A_0B_0|C_1) = P(A_0\mathrm{B}_0|C_0 + C_1) = P(A_0B_0) = P(A_0)P(B_0) = 0.8(0.8)$
$= 0.64$.

Display 2. Example: Variables A and B Are Both Independent of Variable C (Joint Independence – $AC = BC = 0$)

Table C_0

C_0	A_0	A_1	Total
B_0	2	1	3
B_1	4	8	12
Total	6	9	15

Table C_1

C_1	A_0	A_1	Total
B_0	6	3	9
B_1	12	24	36
Total	18	24	45

Table $C_0 + C_1$

$C_0 + C_1$	A_0	A_1	Total
B_0	8	4	12
B_1	16	32	48
Total	24	36	60

Model and Data

Frequency	n_{000}	n_{100}	n_{010}	n_{110}
Counts	2	1	4	8
Model	μ	$\mu + a$	$\mu + b$	$\mu + a + b + d$

Frequency	n_{001}	n_{011}	n_{101}	n_{111}
Counts	6	3	12	24
Model	$\mu + c$	$\mu + a + c$	$\mu + b + c$	$\mu + a + b + c + d$

Model Values

Parameters	M	A	b	c	d
Estimates	0.693	−0.693	0.693	1.099	1.386
$e^{\text{estimates}}$	2	0.5	2.0	3.0	4.0

Example of Jointly Independent Variables

additive scale: $log(n_{111}) = 0.693 - 0.693 + 0.693 + 1.099 + 1.386 = 3.178$

multiplicative scale: $n_{111} = 2(0.5)(2)(3)(4) = 24$

Example Probabilities

$$P(A_0B_0|C_0) = P(A_0B_0|C_1) = P(A_0B_0|C_0 + C_1) = 0.133$$
$$P(A_0|C_0) = P(A_0) = 0.4 \text{ and } P(B_1|C_1) = P(B_1) = 0.8$$

Display 3. Example: Variables A and B Independent and AC and BC Associated (Conditional Independence – $AB = 0$)

Table C_0

C_0	A_0	A_1	Total
B_0	6	4	10
B_1	3	2	5
Total	9	6	15

Table C_1

C_1	A_0	A_1	Total
B_0	2	1	3
B_1	4	2	6
Total	6	3	9

Model and Data

Frequency	n_{000}	n_{100}	n_{010}	n_{110}
Counts	6	4	3	2
Model	μ	$\mu + a$	$\mu + b$	$\mu + a + b$

Frequency	n_{001}	n_{011}	n_{101}	n_{111}
Counts	2	1	4	2
Model	$\mu + c$	$\mu + a + c + f$	$\mu + b + c + d$	$\mu + a + b + c + d + e$

Model Values

Parameters	μ	a	b	C	d	e
Estimates	1.792	−0.405	−0.693	−1.099	1.386	−0.288
$e^{\text{estimates}}$	6	0.667	0.5	0.333	4.0	0.75

Example of Conditional Independent Variables

additive scale: $log(n_{111}) = 1.792 - 0.405 - 0.693 - 1.099 + 1.386 - 0.288 = 0.693$
multiplicative scale: $n_{111} = 6(0.667)(0.5)(0.333)(4)(0.75) = 2$

Example Probabilities

$$P(A_0B_0|C_0) = 0.4,\ P(A_0B_0|C_0 + C_1) = 0.333 \text{ and}$$
$$P(A_0C_0|B_0) = 0.462,\ P(A_0C_0|B_0 + B_1) = 0.391$$

Display 4. Example: Collapsed Tables When Variables *A* and *B* Are Independent – Conditional Independence

Table A_0

A_0	C_0	C_1	Total
B_0	6	2	8
B_1	3	4	7
Total	9	6	15

Table A_1

A_1	C_0	C_1	Total
B_0	4	1	5
B_1	2	2	4
Total	6	3	9

Table $A_0 + A_1$

$A_0 + A_1$	C_0	C_1	Total
B_0	10	3	13
B_1	5	6	11
Total	15	9	24

odds ratios: $or_{BC|A_0} = or_{BC|A_1} = or_{BC|A_0+A_1} = 4.0$ or $e^d = e^{1.386} = 4.0$.

Table B_0

B_0	C_0	C_1	Total
A_0	6	2	8
A_1	4	1	5
Total	10	3	13

Table B_1

B_1	C_0	C_1	Total
A_0	3	4	7
A_1	2	2	4
Total	5	6	11

Table $B_0 + B_1$

$B_0 + B_1$	C_0	C_1	Total
A_0	9	6	15
A_1	6	3	9
Total	15	9	24

odds ratios: $or_{AC|B_0} = or_{AC|B_1} = or_{AC|B_0+B_1} = 0.75$ or $e^e = e^{-0.288} = 0.75$.

Display 5. Example: No Pairs of Variables Are Independent (Three Additive Pairwise Interactions – $AB \neq 0$, $AC \neq 0$, and $AB \neq 0$)

Table C_0

C_0	A_0	A_1	Total
B_0	6	3	9
B_1	2	4	6
Total	8	7	15

Table C_1

C_1	A_0	A_1	Total
B_0	4	3	7
B_1	1	3	4
Total	5	6	11

Model and Data

Frequency	n_{000}	n_{100}	n_{010}	n_{110}
Counts	6	3	2	4
Model	M	$\mu + a$	$\mu + b$	$\mu + a + b + d$

Frequency	n_{001}	n_{011}	n_{101}	n_{111}
Counts	4	3	1	3
Model	$\mu + c$	$\mu + a + c + f$	$\mu + b + c + e$	$\mu + a + b + c + d + e + f$

Model Values

Parameters	μ	a	b	c	d	e	f
Estimates	1.792	−0.693	−1.099	−0.405	1.386	−0.288	0.405
$e^{\text{estimates}}$	6	0.5	0.333	0.667	4.0	0.75	1.5

Example of Additive Relationships

additive scale: $log(n_{111}) = 1.792 - 0.693 - 1.099 - 0.405 + 1.386 - 0.288 + 0.405 = 1.099$

multiplicative scale: $n_{111} = 6(0.5)(0.333)(0.667)(4)(0.75)(1.5) = 3.0$

Table 10.27 *Data:* $2 \times 5 \times 2$ *Table Containing Number of Lung Cancer Cases among Men Classified by Age, Smoking Status, and Five Increasing Socioeconomic Status Categories* ($n = 2228$)

Nonsmokers	ses_1	ses_2	ses_3	ses_4	ses_5	Total
Age $<$ 55	38	34	58	54	88	272
Age $\geq$ 55	40	90	91	167	260	648
Total	78	124	149	221	348	920
Smokers	ses_1	ses_2	ses_3	ses_4	ses_5	Total
Age $<$ 55	45	54	80	77	90	346
Age $\geq$ 55	57	99	140	275	391	962
Total	102	153	220	352	481	1308

Odds Ratio Measures of the Association between Pairs of Variables

$$or_{AB|C_0} = or_{AB|C_1} = or_{AB} = e^d = e^{1.386} = 4.0,$$
$$or_{BC|A_0} = or_{BC|A_1} = or_{BC} = e^e = e^{-0.288} = 0.75,$$
$$or_{AC|B_0} = or_{AC|B_1} = or_{AC} = e^f = e^{0.405} = 1.5.$$

A Case Study – Joint Independence

Lung cancer risk in men has been shown to be inversely associated with socioeconomic status. The following case study is an investigation of whether smoking exposure parallels the incidence pattern observed in lung cancer. The data (Table 10.27) are a sample of $n = 2228$ men classified by smoking exposure (smoker or nonsmoker), age ($<$55 years and $\geq$55 years) and five increasing levels of increasing socioeconomic status (income levels, denoted ses_i).

Selecting a model to describe the relationships among three categorical variables can start with the simplest possible model and evaluate influences created by adding variables or can start with the most extensive model and evaluate influences created by eliminating variables. The easiest results to interpret are produced by the later. A loglinear model with three interaction terms to evaluate the pairwise associations among the three categorical variables (*smk*/*ses*/*age*) is

$$log(n_{ijk}) = \mu + smk_i + ses_j + age_k + [smk_i \times ses_j] + [smk_i \times age_k] + [ses_j \times age_k].$$

The summary model with three pairwise interaction terms provides the best fit possible. A model with three-way interaction terms included would be saturated ($X^2 = 0$) and is not a summary. The estimates of the model parameters, specifically the interaction terms, are the beginning of the search for a simpler model that remains an effective description of the relationships within table (Table 10.28).

The test statistic (G^2) has a chi-square distribution when estimated model and observed counts differ by chance alone. Therefore, the p-value $P(G^2 \geq 5.893 \mid \text{random differences}) = 0.207$ (degrees of freedom $= 4$) suggests that the model with three interaction terms is an adequate summary of the observed data (Table 10.27).

Table 10.28 *Estimates: Parameters from Loglinear Model with Three Interaction Terms Describing* $2 \times 5 \times 2$ *Three-Way Table of Smoking/*ses*/Age Data (Table 10.27)*

	Estimates	*s. e.*	*p*-values
Intercept	3.627	–	–
smk	0.072	0.159	0.0653
ses_2	0.101	0.185	0.585
ses_3	0.446	0.173	0.009
ses_4	0.364	0.171	0.033
ses_5	0.753	0.161	<0.001
age	0.188	0.159	0.237
$smk \times ses_2$	0.611	0.197	0.002
$smk \times ses_3$	0.355	0.184	0.054
$smk \times ses_4$	1.054	0.180	<0.001
$smk \times ses_5$	1.141	0.172	<0.001
$smk \times$ age	0.149	0.097	0.001
$ses_2 \times$ age	−0.080	0.194	0.681
$ses_3 \times$ age	0.109	0.184	0.556
$ses_4 \times$ age	0.163	0.175	0.931
$ses_5 \times$ age	0.0186	0.168	0.912
Deviance[a] = 5.893			

[a]Deviance $= G^2 = -2 \times$ log-likelihood value.

Then, using this model for comparison, the log-likelihood values from three reduced models provide measures to evaluate the smoking/*ses*/age relationships when the specific interaction terms involving age are removed from the model (Table 10.29). The statistical question becomes: Is age a substantial contributor to the model?

Interaction terms (associations) involving the variable age produce little change in the log-likelihood values when either or both of the interaction terms including the variable age are removed from the model relative to the model including three pairwise interactions (Table 10.29). Thus, the influence of the binary age variable appears minimal. A plot of the logarithms of the observed counts (solid line) and the values calculated from the model

Table 10.29 *Results: Three Loglinear Models from Smoking/*ses*/Age Data with Interaction Terms Containing Variable Age Removed*

Interaction terms	Analysis			
$ses \times age = 0$	Deviance	Test[a]	*df*	*p*-value
	9.361	3.467	4	0.483
$smk \times age = 0$	Deviance	Test[a]	*df*	*p*-value
	8.233	2.339	1	0.126
$age \times ses = 0$ and $smk \times age = 0$	Deviance	Test[a]	*df*	*p*-value
	11.964	6.069	5	0.300

[a]Deviance comparison to the model containing all three pairwise interaction terms (Table 10.28).

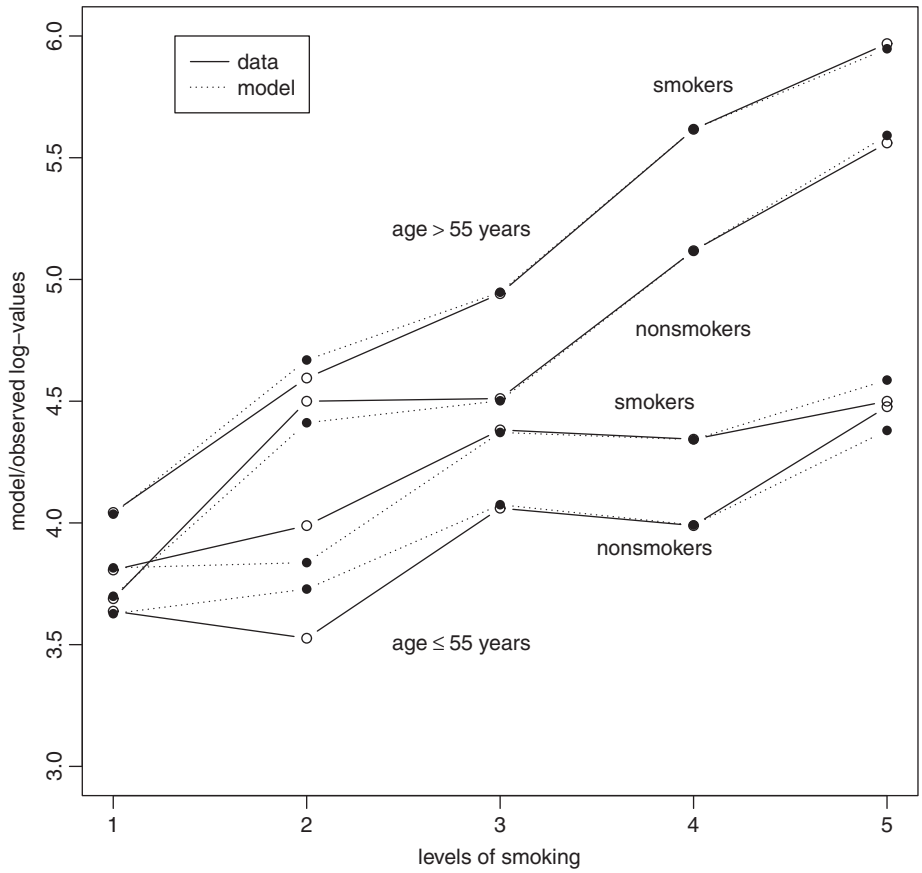


Figure 10.1 Model Values (Dashed Line) and Observed (Solid Line) *Log*-Values for the $2 \times 5 \times 2$ Table of Smoking, Socioeconomic Status by Age Table

containing all three interaction terms further displays evidence that the variable age has a modest influence on the relationship between *ses*-status and smoking exposure (Figure 10.1, dotted versus solid lines).

When a variable such as age is weakly associated with the other two variables in a three-way table, the table that results from combining the data over the levels of that variable produces a two-way table incurring a small but a frequently negligible bias. The collapsed table, as noted, substantially increases the sample size creating a corresponding increase in precision and, in addition, a much simpler description of the smoking and socioeconomic status relationship in a single table (Table 10.30 and Figure 10.2).

From the collapsed 2×5 table, a single estimated summary ratio from the reduced log-linear model characterizes the non-smoker-smoker association with *ses*-status ignoring age (Table 10.30). Thus, a loglinear model produces an estimated geometric distance between smokers and nonsmokers on the *log*-scale that is exactly equal for all five categories of socioeconomic status. Each of the five data estimated non-smoking-smoking ratios are approximately equal to 1.4 (Table 10.30). The model-estimated parameter 0.352 yields the single estimated ratio $e^{0.352} = 1.422$. In model terms, there is clearly no evidence of a

Table 10.30 *Data:* 2 × 5 *Table Describing Individuals Classified by Smoker and Nonsmoker for Five Socioeconomic Categories (Table 10.27 – Marginal Frequencies)*

	Socioeconomic status					
	ses_1	ses_2	ses_3	ses_4	ses_5	Total
Smokers	78 (74.3)	124 (114.4)	149 (152.4)	221 (236.6)	348 (342.3)	920
Nonsmokers	102 (105.7)	153 (162.6)	220 (216.6)	352 (336.4)	481 (486.7)	1308
Total	180	277	369	573	829	2228
Ratios	1.308	1.234	1.477	1.593	1.382	1.422[a]
Probabilities	0.433	0.448	0.404	0.386	0.420	0.413[a]

[a]Summary values from an *ses*/smoking model analysis (not shown).
Note: Values in parentheses are model estimates of the cell frequencies.

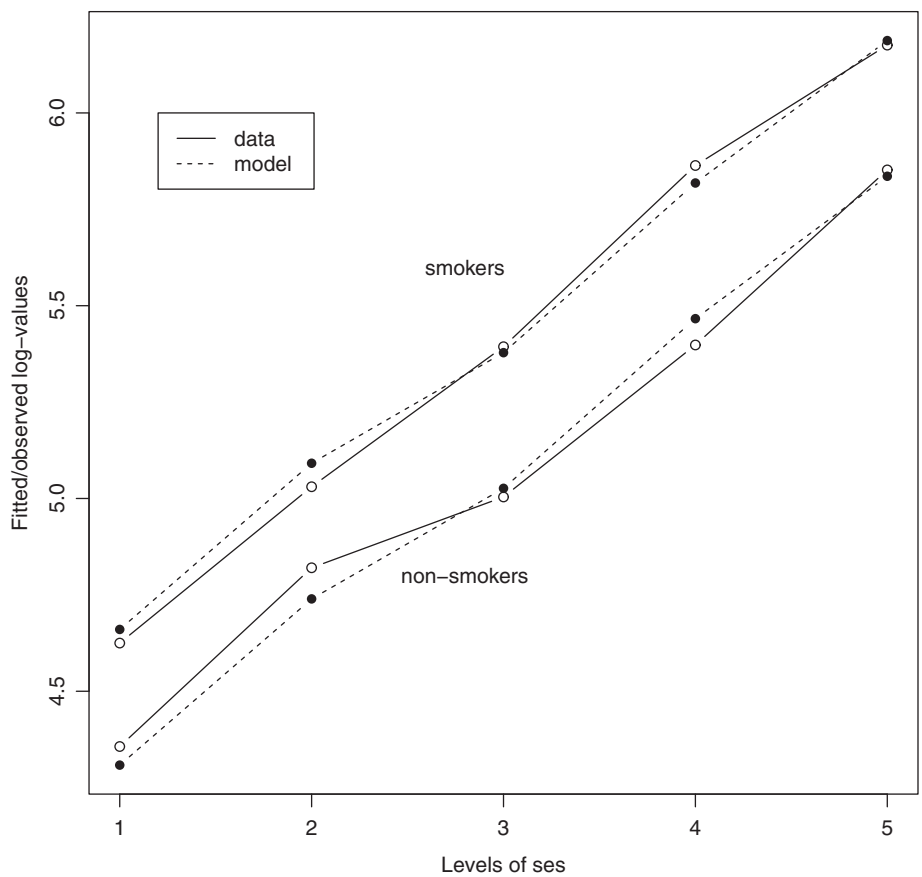


Figure 10.2 Model Values (Dashed Line) and Observed Log-Counts (Solid Line) from 2 × 5 Table of Smoking and Socioeconomic Status

Table 10.31 *Data: Counts of Neural Tube Birth Defects by Case/Control Status, Maternal Weight (bmi < 29 and bmi ≥ 29) and Ethnicity – a 2 × 2 × 4 Table*

White non-hispanics	Case	Control	Total	Foreign-born latina	Case	Control	Total
$bmi < 29$	252	131	383	$bmi < 29$	46	35	81
$bmi \geq 29$	15	21	36	$bmi \geq 29$	9	5	14
Total	267	152	419	Total	55	40	95
U.S.-Born Latina	Case	Control	Total	African American	Case	Control	Total
$bmi < 29$	57	24	81	$bmi < 29$	12	5	17
$bmi \geq 29$	12	8	20	$bmi \geq 29$	1	1	2
Total	69	32	101	Total	13	6	19

smoking-*ses* interaction (Figure 10.2). The close correspondence between the plotted model values (dots) and data values (circles) is apparent. Furthermore, a constant ratio of counts requires the probability of a smoker at each of the five *ses*-levels to be equally constant ($\hat{p} = 0.413$ – Table 10.30). A formal statistical analysis of an association between smoking and socioeconomic status from the collapsed 2×5 table produces the log-likelihood goodness-of-fit test statistic $G^2 = 3.733$ (comparison of model estimates to data counts – Table 10.30) yielding a p-value of 0.444 (degrees of freedom = 4). For a slight decrease in model accuracy, the relationship between socioeconomic status and smoking becomes parsimonious and apparent. To summarize, once age is removed from the model incurring a small bias, the resulting analysis of the reduced model yields no persuasive evidence of a relationship between socioeconomic status and smoking exposure (Figure 10.2).

A Case Study – Conditional Independence in a 2 × 2 × 4 Table

Conditional independence requires that one pair of variables among the three possible pairs is unrelated (independent). A three-way table from a study of neural tube birth defects illustrates (Table 10.31). The three categorical variables are case/control status, ethnicity (white, non-Hispanic, foreign-born Latina, U.S.-born Latina and African American), and pre-pregnancy maternal weight classified as obese ($bmi \geq 29$) and not obese ($bmi < 29$). The three-variable loglinear model with three pairwise interactions terms included is

$$log(n_{ijk}) = \mu + bmi_i + race_j + status_k + [bmi_i \times race_j] + [bmi_i \times status_k] + [race_j \times status_k].$$

All possible reduced loglinear models produce seven log-likelihood test statistics (Table 10.32).

The pairwise association between case/control status and ethnicity appears to have little influence (p-value = 0.412). Therefore, two separate tables created by combining the data over these two variables (ethnicity or case/control status) again produces a more powerful analysis and a simpler description based on two separate two-way tables instead of one three-way table at the cost of incurring a slight bias from excluding a weakly related third variable (ethnicity).

Table 10.32 *Analysis: Results from the Seven Possible Loglinear Models Applied to Three-Way Table Created by Case/Control Counts of Neural Tube Birth Defects Classified by Ethnicity and Obesity (Table 10.31)*

Model: $status \times race = 0$				
$logL$	$logL_0$	Test	df	p-value
−1.875	−3.310	2.871	3	0.412
Model: $bmi \times race = 0$				
$logL$	$logL_0$	Test	df	p-value
−1.875	−7.298	10.846	3	0.013
Model: $status \times bmi = 0$				
$logL$	$logL_0$	Test	df	p-value
−1.875	−4.652	5.555	1	0.018
Model: $status \times race = bmi \times race = 0$				
$logL$	$logL_0$	Test	df	p-value
−1.875	−8.545	13.340	6	0.038
Model: $bmi \times status = status \times race = 0$				
$logL$	$logL_0$	Test	df	p-value
−1.875	−5.899	8.049	4	0.090
Model: $bmi \times race = status \times bmi = 0$				
−1.875	−9.887	16.024	4	0.003
Model: $status \times race = status \times bmi = race \times bmi = 0$				
$logL$	$logL_0$	Test	df	p-value
−1.875	−11.134	18.518	7	0.010

Note: $logL$ = log-likelihood value of the three variable interaction model (all pairwise interactions included), $logL_0$ = log-likelihood value for models with one or more interaction terms removed, and "Test" represents the chi-square test statistic created by the difference between these two log-likelihood statistics.

Combining the four ethnicity categories produces a 2 × 2 table of case/control status and *bmi*-levels (Table 10.33). The estimated odds ratio measuring the association between case/control status and obesity is $\hat{or} = 1.780$ (p-value = 0.022) with an approximate 95% confidence interval of (1.087, 2.917). The collapsed table produces clear and parsimonious evidence of the association between neural tube birth defects and maternal obesity, taking advantage of the knowledge that omitting ethnicity from the model only slightly biases the estimated relationship.

Similarly, a weighted average of the four ethnic-specific odd ratio is 2.018 with some evidence of homogeneity of ethnicity from the loglinear model analysis (Table 10.33). The

Table 10.33 *Data: Counts of Neural Tube Birth Defects by Maternal Body Mass Index (bmi) and Case/Control Status (Combined over Ethnicity)*

	Case	Control	Total
$bmi < 29$	367	195	562
$bmi \geq 29$	37	35	72
Total	404	230	634

Table 10.34 *Data: Counts by Maternal bmi and Ethnicity Status (Combined over Case/Control Status)*

	White	Foreign-born latina	U.S.-born latina	African American	Total
$bmi < 29$	383	81	81	17	562
$bmi \geq 29$	36	14	20	2	72
Total	419	95	101	19	634

cost of this more focused and easily interpreted analysis is again a small bias caused by ignoring the influence of ethnicity.

Somewhat incidentally, the table that results from combining over case/control status (Table 10.34) shows an appreciable association between ethnicity and obesity. The chi-square–distributed likelihood ratio test statistic $G^2 = 10.469$ (degrees of freedom $= 3$) provides substantial evidence that the apparent association is not likely to have occurred by chance alone (p-value $= 0.015$).

11

Bootstrap analysis

A bootstrap is a small loop found at the back of a boot, used to help pull it on. Bootstrap estimation gets its name from the expression "to pull yourself up by your bootstraps." An eighteenth-century German folk tale, *The Surprising Adventures of Baron Munchausen* by Rudolph Erich Raspe (1785), chronicles the Baron saving himself from drowning by lifting himself out a lake by his bootstraps. In a statistical context, a bootstrap estimate is also a readily available, self-contained, efficient method that applies in most situations. Like the Baron's rescue, it is a surprisingly simple method to effortlessly produce important results.

To start, consider the not very realist situation where 2000 mean values each consisting of $n = 25$ observations are estimated from 2000 separate random samples of the numbers 1 to 25. The estimated mean value of the resulting $k = 2000$ sample mean values is $\bar{x} = 13.025$. The estimated variance of these mean values, estimated in the usual way, again from these same 2000 mean values is $S_{\bar{X}}^2 = \Sigma(\bar{x}_i - \bar{x})^2/1999 = 2.165$. The distribution of these mean values is displayed in Figure 11.1. Not surprisingly, the estimated mean value and variance of this estimated distribution are almost identical to the theoretical values of $\mu = 13$ and $\sigma_{\bar{X}}^2 = 2.167$. Using these two estimates and taking advantage of the fact that mean values typically have at least approximate normal distributions (Figure 11.1), a routine approximate 95% confidence interval becomes

$$\bar{x} \pm 1.960 S_{\bar{X}} = 13.025 \pm 1.960(1.472) \rightarrow (10.140, 15.910).$$

In most realistic situations, collected data yield a single sample of n observations that produces a single estimated mean value. Statistical theory, however, provides an estimate of the variance of the distribution of the mean value $\bar{x}$ estimated from a single sample of data. Specifically, the variance is estimated by $S_{\bar{X}}^2 = S_X^2/n$ (Chapter 3). It appears hopeless that a single sample could give any realistic indication of the entire distribution of a sample mean value without additional assumptions or theory or information. Nevertheless, like the Baron, an amazingly simple method exists to estimate not only the mean and its variance but the entire distribution of the estimated mean value from a single sample.

Since it is rarely practical to sample an original population more than once, an alternative is achieved by resampling the sample. The single sample of n observations is sampled k times where each observation is "replaced" after it is sampled so that the next sampled value is selected from exactly the same original observations. The process is called *sampling with replacement.* Each sample contains some original values more than once, and other values will not be included at all. If the values sampled are not "replaced," the 2000 mean values (Figure 11.1) of the 25 sample values would be identical. This sampling process is repeated

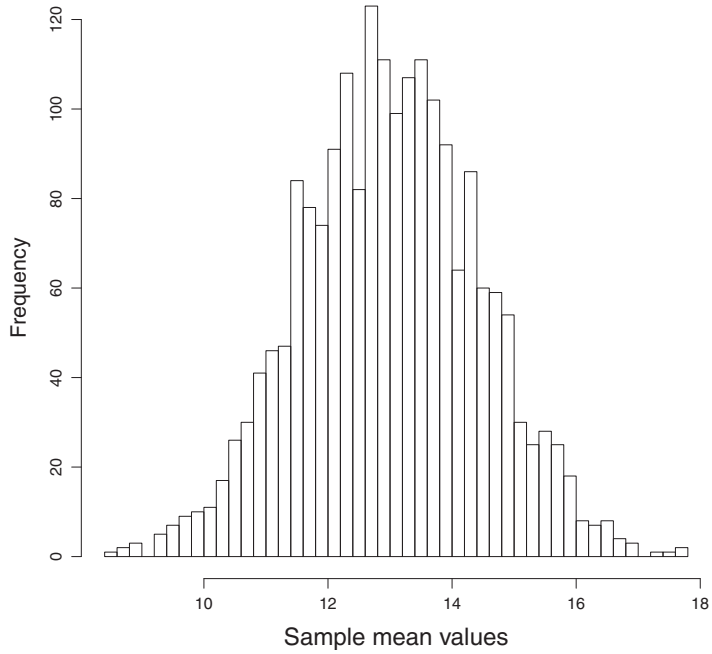


Figure 11.1 Distribution of 2000 Sample Mean Values Created from Independently Sampling the Numbers 1 to 25 Each of Sample Size $n = 25$

a large number of times producing k *replicate* samples of n values ($k = 2000$ samples, for example). Each replicate sample produces an estimated mean value.

Another view of replicate sampling comes from treating the original sampled observations as a "population." A specific random sample of $n = 25$ values from the numbers 1 to 25 is

$$X = \{9, 15, 20, 4, 17, 15, 7, 14, 14, 1, 5, 22, 1, 13, 4, 19, 23, 18, 13, 7, 12, 2, 23, 22, 13\}.$$

This sample becomes the "population" sampled. It has a "fixed" mean value ($\bar{x} = 12.520$) and variance ($S_X^2 = 50.677$). The "population" size is $n = 25$. Each of the k replicate samples is a random selection (with replacement) from this same "population" producing a distribution of estimated mean values.

The differences among these k bootstrap-estimated mean values calculated from the k replicate samples reflect the variability in the original population sampled. Of more importance, these replicate mean values provide an estimate of the entire distribution of the estimated mean value. The first six replicate samples of $k = 2000$, each made up of $n = 25$ values from the original single sample of observations (labeled X), are contained in Table 11.1. In addition, the mean values estimated from each of these first six replicate samples (denoted $\bar{x}_{[i]}$) are $\bar{x}_{[1]} = 13.24$, $\bar{x}_{[2]} = 12.92$, $\bar{x}_{[3]} = 12.20$, $\bar{x}_{[4]} = 12.16$, $\bar{x}_{[5]} = 10.48$ and $\bar{x}_{[6]} = 13.32$. To repeat, the $k = 2000$ replicate mean values differ because of the variability in the original sample of n observations. The distribution of all 2000 replicate mean values calculated from randomly sampled sets of $n = 25$ values from the original observations labeled X is displayed in Figure 11.2.

Table 11.1 *Estimates: First Six Bootstrap Replicate Samples Used to Estimate Distribution of a Mean Value Based on Sampling with Replacement of the Same Sample of* $n = 25$ *Observations* $k = 2000$ *Times from Observations Labeled X*

Six example replicate samples												
Sample 1												
13	15	13	10	19	15	11	11	17	3	19	17	15
10	4	23	12	17	15	10	23	8	13	16	2	
Sample 2												
12	15	24	2	10	12	19	12	8	15	4	22	16
9	9	9	12	3	11	23	12	22	22	15	5	
Sample 3												
2	23	9	22	19	15	22	12	10	13	5	12	8
3	3	12	12	4	22	17	2	9	24	10	15	
Sample 4												
11	13	12	24	3	24	23	23	15	13	9	11	4
9	3	3	19	12	3	19	17	5	15	9	5	
Sample 5												
9	13	9	15	23	8	3	2	3	13	19	19	4
8	8	2	10	12	12	11	3	24	11	2	19	
Sample 6												
19	16	19	22	13	3	19	10	12	13	23	17	19
12	15	3	19	15	2	2	12	4	19	13	12	

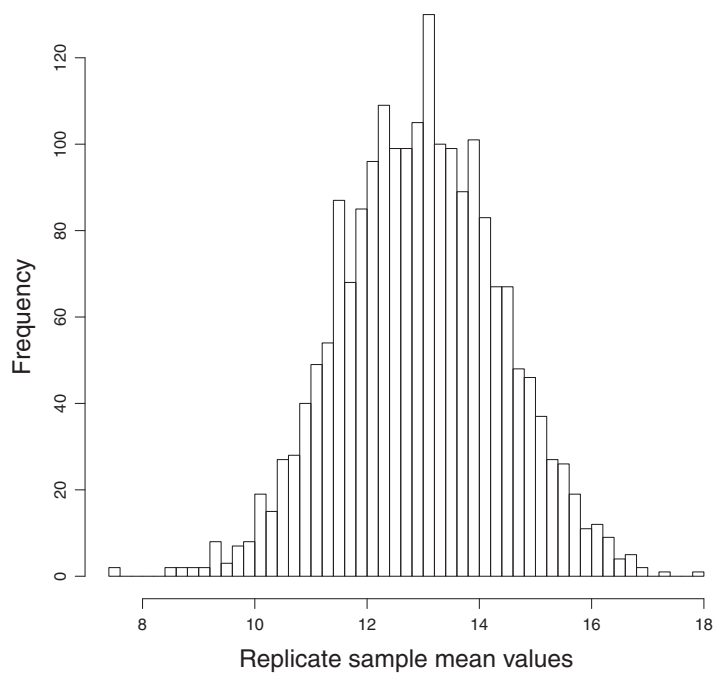


Figure 11.2 Distribution of Sample Mean Values Created by Replicate Sampling a Single Sample of $n = 25$ Observations (Labeled X) $k = 2000$ Times

Thus, replicate samples produce a bootstrap estimate of the distribution of the mean value $\bar{x}$ without assumptions or statistical theory or approximations. All that is required is a computer. The mean value of this estimated distribution is

$$\begin{aligned}\text{bootstrap estimated mean value} = \bar{x}_{[.]} &= \frac{1}{k}\sum \bar{x}_{[i]} \\ &= \frac{1}{2000}(13.24 + 12.92 + 12.20 + \cdots + 11.33) \\ &= 12.972\end{aligned}$$

where $i = 1, 2, \ldots, k = 2000$ replicate mean values directly estimate the "population" mean value, namely, $\bar{x} = 12.515$. In a similar fashion, the bootstrap-estimated variance is also directly calculated from the estimated distribution of the $k = 2000$ bootstrap-estimated mean values. The bootstrap estimate of the variance of the sample mean value is

$$\text{bootstrap-estimated variance} = S^2_{\bar{x}_{[i]}} = \frac{1}{k-1}\sum (\bar{x}_{[i]} - \bar{x}_{[.]})^2 = 2.046.$$

Almost like a folk tale, the estimated distribution is a complete statistical description of the distribution of an estimated mean value $\bar{x}$ (Figure 11.2).

A simple way to compare two distributions is to sort each set of sampled values from the smallest to largest and plot the corresponding pairs (Chapter 12). When the two distributions sampled are identical, these pairs randomly deviate from a straight line with intercept of 0.0 and slope 1.0. A plot of 2000 such pairs (Figure 11.3) demonstrates that the distribution of 2000 mean values each estimated from independent samples of values (Figure 11.1) is essentially the same as the distribution estimated from 2000 replicate mean values each sampled with replacement from the same sample of 25 observations (Figure 11.2). Resampling the population is rarely practical, but resampling the sample is almost always possible and usually produces effective estimates.

A bootstrap-estimated distribution offers two approaches to calculating a confidence interval. Assuming that the $k = 2000$ bootstrap-estimated values have at least an approximate normal distribution, the bootstrap-estimated mean value and variance produce a confidence interval in the usual way. From the example bootstrap-estimated distribution (Figure 11.2), an approximate 95% confidence interval based on the estimated mean value $= \bar{x}_{[.]} = 12.972$, and its estimated variance $= S^2_{\bar{x}_{[i]}} = 2.046$ is $\bar{x}_{[.]} \pm 1.960 S_{\bar{x}\,[i]} = 12.972 \pm 1.960\,(1.430) \rightarrow (10.169, 15.775)$. Note again that the actual mean value of the population sampled is 13.0 and variance is 2.167 (calculated from theory).

Because bootstrap replicate sampling produces an estimate of the entire distribution of the sample mean value, it is a simple matter to select the 2.5th percentile and the 97.5th percentile values to create an assumption-free 95% confidence interval. The bootstrap-estimated distribution alone provides all the necessary information to construct the confidence interval. For the example data, bootstrap distribution percentiles are 2.5th percentile = 10.203 (*rank* = 50) and 97.5th percentile = 15.761 (*rank* = 1950) directly creating an alternative and assumption-free 95% confidence interval of (10.203, 15.761), sometimes called a *nonparametric confidence interval* (Chapter 15).

Bootstrap estimation of the distribution of a sample mean employs a special case of general bootstrap notation. For a statistical summary value denoted g and its estimated value denoted

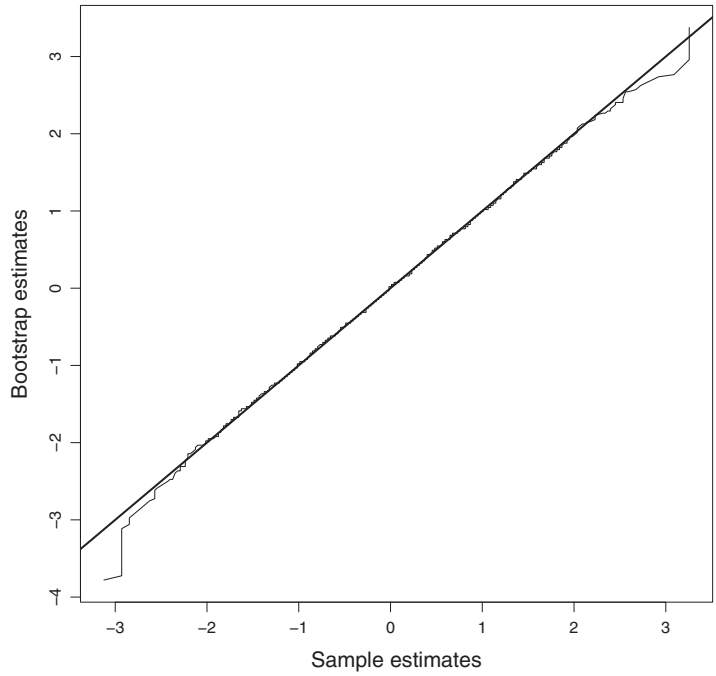


Figure 11.3 Comparison of Two Estimated Distributions of Mean Value: Distribution of Estimated Mean Values from 2000 Independent Random Samples of the Numbers 1 to 25 (Figure 11.1) to Bootstrap-Estimated Distribution Created from 2000 Replicate Sample Mean Values Each Estimated from a Single Random Sample of the Numbers 1 to 25 (Figure 11.2)

$\hat{g}$, the bootstrap-estimated mean value is the mean of the bootstrap replicate estimates,

$$\bar{g}_{[.]} = \frac{1}{k}\sum \hat{g}_{[i]} \quad i = 1, 2, \ldots, k = \text{number of sets of replicate samples}$$

where $\hat{g}_{[i]}$ represents a single bootstrap estimate of the summary statistic g calculated from the ith replicate sample of n observations. The bootstrap-estimated variance of the distribution of the estimated summary value $\hat{g}$ is

$$variance\left(\hat{g}_{[i]}\right) = S^2_{\hat{g}_{[i]}} = \frac{1}{k-1}\sum \left(\hat{g}_{[i]} - \bar{g}_{[.]}\right)^2 \qquad i = 1, 2, \ldots, k,$$

again based on k bootstrap estimates $\hat{g}_{[i]}$ from k replicate samples from the original n sampled observations. These two bootstrap-estimated values are the usual estimates of a sample mean and variance applied to the special case of k estimated values produced by replicate sampling. In addition, bootstrap estimates frequently have approximate normal distributions because bootstrap estimates are summary mean values (Chapter 1).

The choice of the number of replicate samples is not much of an issue because the precision of a bootstrap-estimated value is dictated by the number of observations (n) and variation of the original sampled values. The fundamental purpose of a bootstrap analysis is to accurately describe the properties of the estimated value for any sample size. For example, the high variability of an estimate from a small sample of observations remains high and potentially unstable regardless of the number of replicate samples; that is, the sample size n determines

Table 11.2 *Data: Pairs of before/after Measurements of Thiocyanate Levels from Extremely Heavy Cigarette Smokers Participating in an Intensive Intervention Trial (a Subset of $n = 20$ Subjects from a Large Study)*

	Thiocyanate levels									
	1	2	3	4	5	6	7	8	9	10
Before	131	154	45	89	38	104	127	129	148	181
After	111	13	16	35	23	115	87	120	122	228
Percentage change	15.3	91.6	64.4	60.7	37.8	−10.6	31.5	7.0	17.6	−26.0
	11	12	13	14	15	16	17	18	19	20
Before	122	120	66	122	115	113	163	67	156	191
After	118	121	20	153	139	103	124	37	163	246
Percentage change	3.3	−0.8	69.7	−25.4	−20.9	8.8	23.9	44.8	−4.5	−28.8

the precision of the estimated value, and the k replicate samples describe the properties of the estimate. Modern computers allow calculation of a large number of replicate samples without any particular consequences. A set of 2000 replicate calculations (k) rarely takes more than fractions of a second and almost always produces a clear sense of the entire estimated distribution. To repeat, the number of replicate samples does not affect the precision of data estimated values based on n observations.

Example: Analysis of Paired Data

A small subset of data from an intervention trial focused on heavy cigarette smokers (three or more packs of cigarettes per day) produces an example of a bootstrap analysis (Table 11.2). The observed values are thiocyanate measurements from blood samples that biochemically indicate the amount of recent cigarette smoking exposure. The example observations consist of $n = 20$ pairs of before measurements followed by after measurements of thiocyanate from each participant in a trial to evaluate the effectiveness of an extremely intensive intervention program where every effort was made to reduce smoking.

A typical approach to evaluating differences between before/after data is Student's paired t-test. The difference in mean levels of thiocyanate measured before ($\bar{x}_{before} = 119.05$) and then after ($\bar{x}_{after} = 104.70$) intervention is $\bar{x}_{before} = -\bar{x}_{after} = 14.35$. The resulting t-statistic $T = 1.515$ (degrees of freedom $= 19$) produces a p-value of 0.146. An important concern is that a t-test requires the compared mean values be sampled from at least approximate normal distributions with the same variance.

A more critical and rarely addressed issue is the choice of the way before or after differences are measured. The mean difference between before or after observed values may not be the most effective assessment of the reduction in smoking exposure. It might be easier for individuals with extremely high exposure levels to achieve the same reduction as those individuals with lower levels. Therefore, perhaps a more revealing measure of intervention success would be the relative before or after percentage decrease given by the expression

$$\hat{P} = 100 \times \frac{\text{before value} - \text{after value}}{\text{before value}}.$$

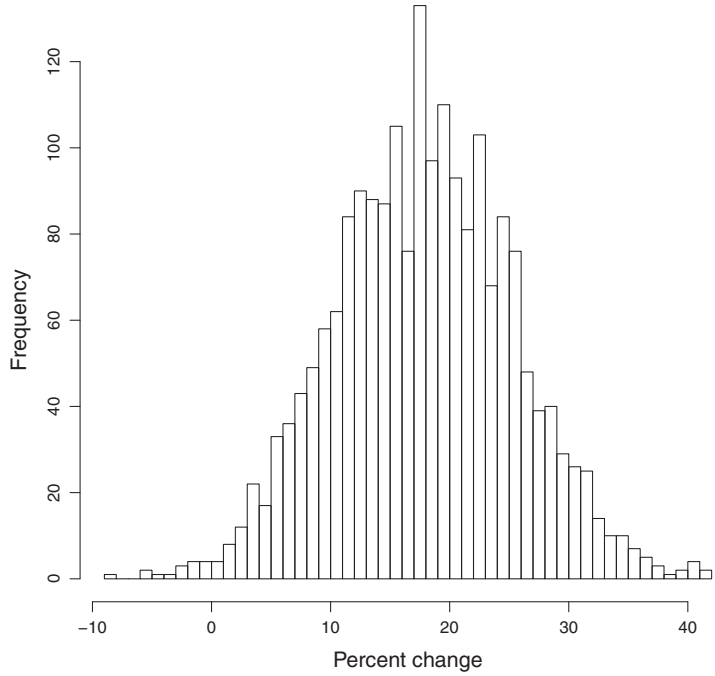


Figure 11.4 Bootstrap-Estimated Distribution of the Mean Percent Decrease from before/after Smoking Reduction Intervention Trial ($k = 2000$ Replicate Samples from $n = 20$ Participants)

The observed percentage decrease (denoted $\hat{P}_i$) ranges from -28.8 to 91.6 estimated from each of the $n = 20$ participants (Table 11.2).

The distribution of the mean value of this potentially more sensitive measure of reduction in thiocyanate levels is directly estimated by resampling the percentage decrease values $\hat{P}_i$ calculated from the 20 study participants (Table 11.2, last row). Bootstrap sampling produces an estimate of the entire distribution of the mean percentage decrease (Figure 11.4). The bootstrap-estimated mean value is $\hat{P}_{[.]} = 17.818$, and the corresponding estimated standard error of the distribution of the mean value is $S_{\hat{P}_{[i]}} = 7.509$, based on $k = 2000$ replicate sample estimates of the before/after percentage decrease.

A bootstrap-estimated distribution provides two approaches to assessing the influence of random variation on an observed mean percentage reduction of thiocyanate levels. First, based on the assumption the estimated mean percentage change has an approximate normal distribution, which is strongly supported by the plot of the distribution of the replicate estimated values (Figure 11.4). The test statistic

$$z = \frac{0 - \bar{P}_{[.]}}{S_{\hat{P}_{[i]}}} = \frac{0 - 17.818}{7.509} = -2.373$$

has an approximate standard normal distribution when the intervention has no influence on smoking behavior. The small associated p-value of $P(Z \leq -2.373 \mid P = 0) = 0.009$ indicates that it is not likely that the observed reduction occurred by chance alone.

An alternative and assumption-free evaluation of the intervention trial results from directly counting the replicate percentage change estimates $\hat{P}_{[i]}$ less than zero (no before or after

Table 11.3 *Data from a Case/Control Study: Distances (Meters) to Nearest High-Voltage Electric Power Line of Residents with Newborn Children with a Birth Defect and Corresponding Randomly Chosen Controls ($n = 10$ Pairs – Small Subset from a Large Study)*

	Case/control distances									
	1	2	3	4	5	6	7	8	9	10
Cases	20	85	40	187	368	144	494	20	55	670
Controls	700	155	167	278	460	333	600	345	125	50

differences). The observed number is 16. An empiric p-value then becomes $16/2000 = 0.008$. Thus, in light of the entire distribution of the estimates $\hat{P}_{[i]}$, the observed bootstrap mean decrease $\bar{P}_{[i]} = 17.818$ is large relative to zero; that is, evidence exists that the intervention strategy substantially shifts the distribution of thiocyanate measurements away from zero.

The two kinds of confidence intervals follow the previous pattern where

$$\text{parametric: } \bar{P}_{[.]} \pm 1.960 S_{\hat{P}_{[i]}} = 17.818 \pm 1.960\,(7.509) \rightarrow (3.100, 32.536)$$

or

$$\text{nonparametric: 2.5th percentile} = 3.374 \quad \text{and} \quad \text{97.5th percentile} = 32.382$$

alternatively creating an assumption-free 95% confidence interval from the smoking intervention trial data.

Example: Evaluation of a Difference between Two Harmonic Mean Values

A small subset of data ($n = 10$ case/control pairs) from a large study of birth defects conducted in France illustrates again the pattern of bootstrap analysis (Table 11.3). Case and control participants were selected from residents in rural France to assess the possible association between birth defects and electromagnetic radiation exposure from the high voltage electric power line nearest to case/control residence. The statistical question addressed by measuring case and control distances to these power lines is: Does evidence exist of a nonrandom case/control influence on the difference in distance to the nearest high voltage power line? The mean distance for cases is $\bar{x}_{cases} = 208.3$ meters and for controls is $\bar{x}_{controls} = 321.3$ meters. A conventional two-sample t-test yields a test statistic of $T = 1.150$ (degrees of freedom $= 18$) with an associated p-value of 0.133.

These data contain large noninformative distances that highly influence a direct comparison of mean case/control distances, particularly the estimated variance. An approach that minimizes this unwanted influence is to measure distance (denoted d_i) by its reciprocal value $(1/d_i)$. This inverse measure emphasizes small distances and deemphasizes large distances (Chapter 21). Therefore, influences from large noninformative distances are essentially eliminated. In symbols, a summary measure of case and control distance becomes

$$\bar{h} = \frac{1}{\frac{1}{n}\sum \frac{1}{d_i}} = \frac{n}{\sum \frac{1}{d_i}} \qquad i = 1, 2, \ldots, n,$$

Table 11.4 *Estimation: First Five Bootstrap Replicate Samples from* $k = 2000$ *Samples from French Birth Defects Data* ($\bar{D}_{[i]}$)

	Example replicate samples									
Replicates	1	2	3	4	5	6	7	8	9	10
Sample 1										
Cases	50	333	278	345	700	345	600	167	333	345
Controls	670	368	670	368	20	670	187	494	40	40
Sample 2										
Cases	167	460	333	167	278	333	125	50	125	167
Controls	144	670	85	85	20	85	144	144	85	20
Sample 3										
Cases	460	460	345	600	155	278	155	460	278	155
Controls	494	40	40	494	20	187	670	187	40	144
Sample 4										
Cases	333	600	125	50	155	50	278	333	125	155
Controls	550	494	40	144	144	187	20	85	40	20
Sample 5										
Cases	600	125	278	278	278	460	167	278	345	333
Controls	550	368	494	368	368	670	20	368	40	187

Table 11.4 *(Continued) Summary Values: First Five Bootstrap-Estimated Case/Control Differences between Harmonic Mean Values (*$K = 2000$ *Replicates)*

	Replicate mean differences				
	1	2	3	4	5
$\bar{D}_{[i]} = \bar{h}_{cases_{[i]}} - \bar{h}_{control_{[i]}}$	125.8	93.1	198.2	70.6	158.8

called the *harmonic mean value.* A harmonic mean value directly reflects "closeness" in meters. For the birth defects data (Table 11.3), the case and control harmonic mean distances are $\bar{h}_{cases} = 57.646$ meters and $\bar{h}_{controls} = 181.13$ meters.

Bootstrap sampling provides an estimate of the distribution of the difference between case/control distances measured by the difference between harmonic mean values (denoted $\bar{D}_{[i]} = \bar{h}_{case[i]} - \bar{h}_{control_{[i]}}$). The sample values and differences in harmonic mean values from the first five replicate samples from $k = 2000$ are presented in Table 11.4.

Bootstrap estimates based on these 2000 replicate harmonic mean differences directly yields a mean value of $\bar{D}_{[.]} = 124.921$. An estimated standard deviation of the distribution from the replicate mean values $\bar{D}_{[i]}$ is $S_{\bar{D}_{[i]}} = 71.556$. The estimated distribution of harmonic mean differences is displayed in Figure 11.5. The graphical description displays persuasive visual evidence that the distribution of estimated harmonic mean differences $\bar{D}_{[i]}$ will be accurately approximated by a normal distribution (solid line).

A more rigorous assessment of bootstrap-estimated distribution requires a computer-generated normal distribution. Specifically, a sample of 2000 normally distributed computer-generated values with the same mean and variance as the bootstrap estimates is sorted and

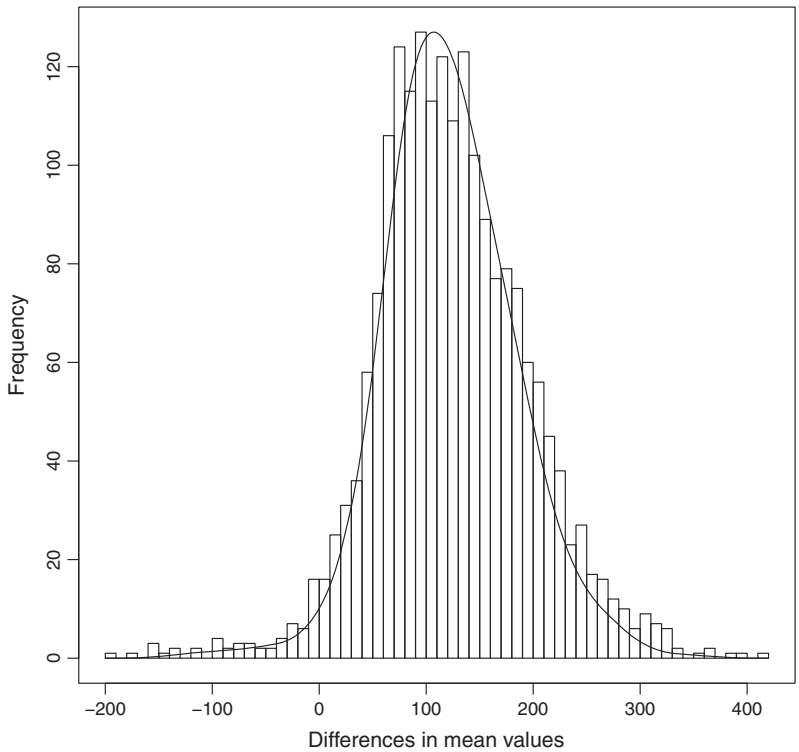


Figure 11.5 Bootstrap Distribution of the Difference between Case/Control Harmonic Mean Values Based on 2000 Replicate Samples

compared to the 2000 sorted bootstrap-estimated values (Chapter 12). As before, these 2000 pairs of mean values randomly deviate from a straight line with intercept = 0.0 and slope = 1.0 when the bootstrap and computer generated estimates have the same normal distribution. Such a plot provides substantial evidence that the estimated differences in harmonic mean values have close to a normal distribution (Figures 11.5 and 11.6).

Therefore, the bootstrap-estimated distribution-based test statistic

$$z = \frac{0 - \bar{D}_{[.]}}{S_{\bar{D}_{[i]}}} = \frac{0 - 124.921}{71.556} = -1.746$$

has an approximate standard normal distribution when electromagnetic radiation exposure has no association with the occurrence of a birth defect. The estimated p-value is $P(\bar{D} \leq 0|$ *no difference*$) = 0.040$. As before, the number of observations less than zero provides an assumption-free estimate of an alternative p-value of $84/2000 = 0.042$.

Two kinds of 95% confidence intervals again follow the previous patterns:

$$\text{parametric: } \bar{D}_{[.]} \pm 1.960 S_{\bar{D}_{[i]}} = 124.921 \pm 1.960\,(71.556) \rightarrow (-15.329, 265.171)$$

and

$$\text{nonparametric: 2.5th percentile} = -15.3 \text{ and 97.5th percentile} = 265.2.$$

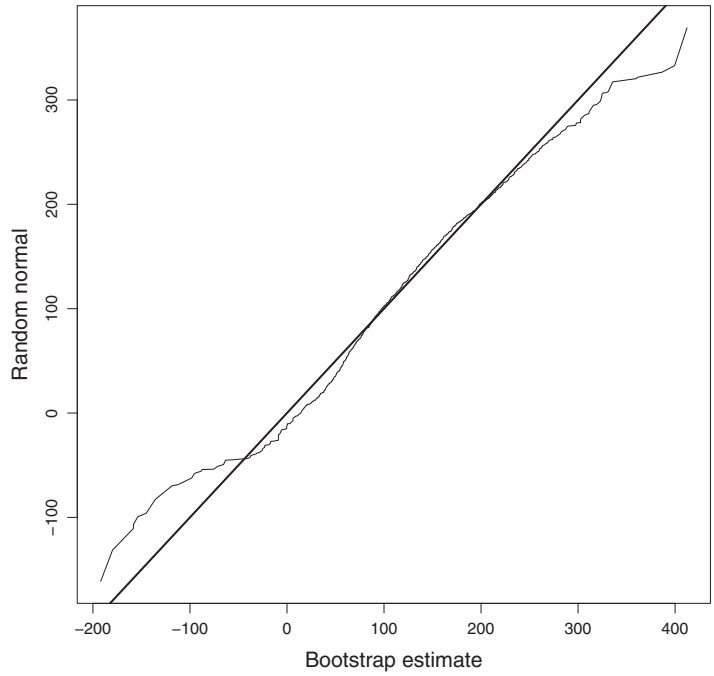


Figure 11.6 Bootstrap Distribution of $k = 2000$ Estimated Mean Differences Compared to 2000 Computer-Generated Normally Distributed Values with Same Mean and Variance

A confidence interval, as usual, gives a sense of the influence on random variation of the estimated difference between two harmonic mean values (precision) and an indication of the likely location of the underlying harmonic mean case/control difference (accuracy).

An important feature of bootstrap estimation is that the choice of a test statistic or summary value is essentially unrestricted. For example, for the birth defects data, the distance measure could have been $1/d_i^2$ or $log(1/d_i)$ or $1/\sqrt{d}_i$ or any other chosen measure of distance. A bootstrap analysis is not in the least constrained to traditional test statistics. Bootstrap evaluations equally apply to usual statistical measures, complicated measures, or specially created statistics without additional statistical theory or assumptions.

An Example of Bootstrap Estimation from Categorical Data

A bootstrap summary statistic and its properties estimated from a table requires the data be "detabled" back into the individual observations. A table is a summary that coalesces observations into a compact and useful display that reduces a sample of unorganized and sometimes extensive number of observations into a succinct description of the relationships among categorical classifications. To sample a table with replacement, it is necessary to convert the table back to the original observations. From a computer point of view, unorganized and extensive numbers of observations are not issues.

The bootstrap approach is introduced with a small artificial table (Table 11.5).

Suppose the summary statistic of interest is

$$\hat{p} = \hat{p}_1 - 2\hat{p}_2 + \hat{p}_3.$$

Table 11.5 *Artificial Data: Illustration of a Bootstrap Estimate of a Summary Statistic Calculated from Categorical Data* ($n = 20$)

Values	1	2	3	Total
x	10	5	5	20
p_i	0.50	0.25	0.25	1.0

From the example data, the observed value is $\hat{p} = 0.5 - 2\,(0.25) + 0.25 = 0.25$ estimated from the table.

The original data contained in the three cells of the table converted back to the 20 original observations are the following:

$$D = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3 \text{ and } 3\}.$$

The data contained in the three table categories are now 20 values labeled D and are randomly sampled with replacement in the same manner as the previous bootstrap replicate estimates. Each sample of 20 values produces a replicate three category table, and each replicate table produces a replicate estimate of the summary value, denoted $\hat{P}_{[i]}$. To illustrate, the first five replicate tables from a sample of $k = 2000$ are the following:

Five example replicate tables															
	1			2			3			4			5		
X	10	6	4	8	8	4	14	2	4	10	4	6	13	1	6

producing bootstrap estimates of $\hat{p}$

$$\hat{p}_{[1]} = 0.10,\ \hat{p}_{[2]} = -0.20,\ \hat{p}_{[3]} = 0.70,\ \hat{p}_{[4]} = 0.40 \text{ and } \hat{p}_{[5]} = 0.85.$$

The replicate values $\hat{p}_{[i]}$ describe the distribution of the estimated summary statistic $\hat{p}$ based on 2000 replicate samples. It is then straightforward to estimate the bootstrap mean and variance. Specifically these estimates are

$$\text{bootstrap mean value} = \bar{p}_{[.]} = \frac{1}{2000}\sum \hat{p}_{[i]} = 0.254$$

and

$$\text{bootstrap variance}(\hat{p}_{[i]}) = \frac{1}{1999}\sum (\hat{p}_{[i]} - \bar{p}_{[.]})^2 = 0.088.$$

For this simple illustration, the variance can be estimated from theory and is $S^2_{\hat{p}_{[i]}} = 0.085$. Needless to say, for more complex and extensive tables, the theory can be difficult, often cumbersome, and, in many cases, not available. Regardless, bootstrap sampling routinely provides the estimation and evaluation of most summary statistics calculated from data classified into a table.

Table 11.6 *Notation: Data Collected to Assess Agreement between Two Raters (p_{ij} Represents Proportion of Observations in ijth Cell of the Table)*

		Rater 1		
		C_1	C_2	Total
Rater 2	C_1	p_{11}	p_{12}	q_1
	C_2	p_{21}	p_{22}	q_2
	total	p_1	p_2	1.0

Kappa Statistic – A Bootstrap-Estimated Confidence Interval

As noted, bootstrap estimation of the distribution of a summary statistic calculated from a table only slightly differs from the general resampling process already described. Sampling with replacement of the original sampled observations is again the key.

A popular measure of agreement is called the *kappa statistic.* A kappa statistic is a frequently used summary value that indicates the degree of agreement between two different and independent observers, often called "raters" in the context of assessing consistency. The kappa statistic (denoted K) is a measure of agreement divided by the probability agreement did not occur by chance. For the case of two categories labeled C_1 and C_2, the notation is presented in Table 11.6.

The components of a kappa statistic are

$$\text{crude agreement} = \hat{P} = p_{11} + p_{22},$$
$$\text{random agreement} = \hat{p} = p_1 q_1 + p_2 q_2,$$
$$\text{adjusted agreement} = \text{crude agreement} - \text{random agreement} = \hat{P} - \hat{p},$$

and the expression for the kappa statistic is

$$\text{estimated kappa statistic} = \hat{K} = \frac{\hat{P} - \hat{p}}{1 - \hat{p}}.$$

A kappa statistic is constructed so that when the raters exactly agree only by chance, the value is zero ($K = 0$). The normalization of the kappa statistic by the estimated probability of nonrandom agreement produces a value such that when the agreement is perfect (all observations are on the diagonal of the agreement or disagreement table) the value of the kappa statistic is one ($K = 1$).

The expression for the variance of the distribution of an estimated kappa statistic ($\hat{K}$) is complicated, extensive, and not presented. As expected, using bootstrap sampling simply produces an estimated distribution, its variance and a 95% confidence interval.

Consider data generated by two independent raters viewing 312 microarray images and classifying subjects by the presence or absence of coronary artery disease (Table 11.7). All arteries sampled are from different individuals and disease status (present/absent) is evaluated by a visual process called color deconvolution. The estimated kappa statistic is

$$\hat{K} = \frac{\hat{P} - \hat{p}}{1 - \hat{p}} = \frac{0.856 - 0.518}{1 - 0.518} = 0.721,$$

Table 11.7 *Coronary Artery Data: Agreement between Pathologic Determinations Given by Two Independent Raters (N = 312 Individuals)*

		Rater 1		
	Counts	C_1	C_2	Total
Rater 2	C_1	105	24	129
	C_2	18	165	183
	Total	123	189	312
or				
		Rater 1		
	Proportions	C_1	C_2	Total
Rater 2	C_1	0.337	0.077	0.413
	C_2	0.058	0.529	0.587
	Total	0.394	0.606	1.00

where $\hat{P} = 0.337 + 0.529 = 0.865$ and $\hat{p} = (0.394)(0.413) + (0.606)(0.587) = 0.518$ (Table 11.7).

The observed data (Table 11.7) are "detabled" into 312 values (1, 2, 3, and 4) and sampled $k = 2000$ times with replacement. Each replicate sample produces a table, and each table produces a bootstrap-estimated kappa statistic $\hat{K}_{[i]}$. Table 11.8 contains the first 10 bootstrap samples from $k = 2000$ replicate tables used to estimate the distribution of the kappa statistic calculated from the coronary artery diagnosis data. The bootstrap estimate of the kappa statistic is $\bar{K}_{[.]} = 0.720$. The bootstrap-estimated variance of the distribution of the estimated kappa statistic is $variance(\hat{K}_{[i]}) = S^2_{\hat{K}_{[i]}} = 0.00163$. In an analysis of agreement, the likelihood the observed kappa value is a random deviation from zero is rarely an issue. A 95% confidence interval, however, is an important part of evaluation and interpretation of the degree of agreement.

Table 11.8 *Example: The First 10 Bootstrap Tables from 2000 Replicate* 2×2 *Tables Used to Estimate Distribution of the Kappa Statistic* $(\hat{K})$

Replicates	$\hat{p}_{11}$	$\hat{p}_{12}$	$\hat{p}_{21}$	$\hat{p}_{22}$	$\hat{K}_{[i]}$
1	0.372	0.071	0.051	0.506	0.752
2	0.340	0.103	0.048	0.510	0.691
3	0.356	0.058	0.080	0.506	0.718
4	0.337	0.064	0.051	0.548	0.758
5	0.365	0.067	0.048	0.519	0.764
6	0.410	0.103	0.061	0.426	0.674
7	0.333	0.064	0.064	0.538	0.732
8	0.346	0.083	0.090	0.481	0.647
9	0.292	0.067	0.051	0.590	0.740
10	0.279	0.096	0.087	0.538	0.608

Table 11.9 *Coronary Heart Disease Analysis: Results from Logistic Model Applied to Describe the Relationship between Body Weight and Probability of a Coronary Event (n = 3153 High-Risk Men)*

Variables	Estimate	*s. e.*	*z*-Value	*p*-value
Constant	−4.215	–	–	–
Weight	0.011	0.0029	3.757	<0.001

Two kinds of estimated confidence intervals again follow the previous pattern. Assuming that the bootstrap distribution of $\hat{K}$ has at least an approximate normal distribution, a 95% confidence interval is

$$\text{parametric: } \bar{K}_{[.]} \pm 1.960 S_{\hat{K}_{[i]}} = 0.720 \pm 1.960\,(0.040) \rightarrow (0.641, 0.800)$$

or

$$\text{nonparametric: 2.5th percentile} \quad \text{and} \quad \text{97.5th percentile} \rightarrow (0.643, 0.789)\,.$$

When the variability of the estimate $\hat{p}$ is ignored, an approximate expression for the variance of the estimate $\hat{K}$ becomes

$$variance(\hat{K}) = variance\left[\frac{\hat{P}-\hat{p}}{1-\hat{p}}\right] \approx \frac{1}{(1-\hat{p})^2}\, variance(\hat{P}) = \frac{1}{(1-\hat{p})^2}\frac{\hat{P}(1-\hat{P})}{n}$$

for n rated subjects (Chapters 4 and 27). For the example estimates $\hat{P} = 0.865$ and $p = 0.518$, the approximate variance based on this expression is close to the bootstrap-estimated value, namely, $variance(\hat{K}) = 0.00161$. Thus, bootstrap estimation can occasionally be used to indicate the consequences of a restrictive assumption.

A Graphic Application of Bootstrap Estimation

Bootstrap sampling enriches graphical presentations in a simple way. Data that describe the influence of body weight on the risk of a coronary heart disease event (*chd*) illustrate. From an eight-year study, occurrences of coronary events were recorded (present/absent) and the weight of each study participant was also part of the collected data. An analysis of $n = 3153$ high-risk men produces an estimate of a logistic curve describing the relationship between weight and the likelihood of a *chd* event (Table 11.9).

Estimated logistic model probabilities of a coronary event then are calculated from the expression

$$\text{estimated probability of } chd \text{ event} = \hat{P}(chd|\text{weight} = wt_i) = \frac{1}{1+e^{-(-4.215+0.011wt_i)}}.$$

A plot of this weight/*chd*-estimated relationship described by a logistic curve is straightforward (Figures 11.7 and 11.8, solid line).

A bootstrap sample of these 3153 (weight/*chd*) pairs of observations generates a replicate logistic curve that differs from the original data estimated curve (Table 11.9) only because of sampling variability. This bootstrap-estimated logistic curve can be added to the plot of

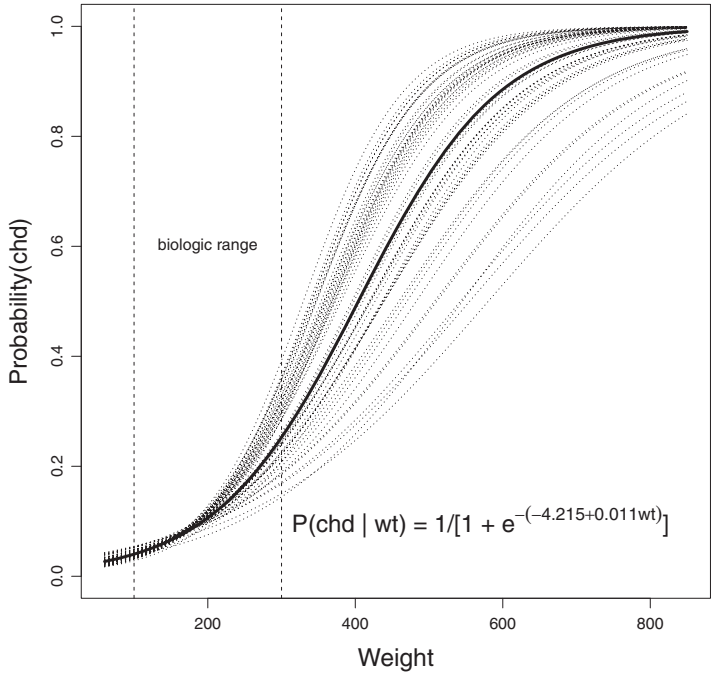


Figure 11.7 Plot of Logistic Model–Estimated Probabilities Relating Risk of a Coronary Event Associated with Increasing Body Weight Using Bootstrap-Estimated Curves to Indicate the Influence of Sampling Variability

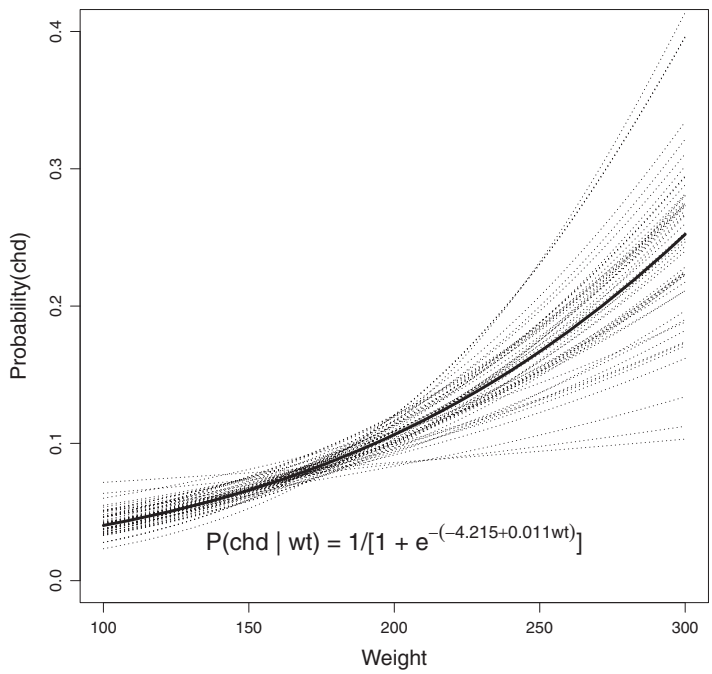


Figure 11.8 Plot of Logistic Model–Estimated Probabilities Relating Risk of a Coronary Event Associated with Increasing Body Weight Using Bootstrap Estimated Curves to Indicate the Influence of Sampling Variability for a Range of Realistic Body Weights (Biologic Range)

original data generated estimate. Additional bootstrap sampling produces additional curves. Adding these estimates to the plot simply displays the influence of sampling variation on the original estimated curve (Figures 11.7 and 11.8, $k = 50$ replicate estimates). The influence of always present sampling variation becomes clearly visible. For the example, it is apparent the sampling variation increases dramatically as the estimated curve deviates from the center of the data (a body weight of about 170 pounds – Figure 11.8).

A Property of Bootstrap Estimation

Bootstrap estimation is referred to as "distribution-free." In other words, knowledge or assumptions about the properties of the sampled population are unnecessary. Situations arise where bootstrap estimation fails to yield useful estimates. Figure 11.9 (top left) displays the distribution of the mean value based on $k = 5000$ independent samples of $n = 25$ observation sampled from a normal distribution. The distribution of the mean values from 5000 bootstrap replicate samples from of a single sample of $n = 25$ observations from the same normal distribution is also displayed (top right). As expected, independent multiple sampling of a population distribution and bootstrap resampling of a single sample produce essentially the same estimate of the distribution of the mean value. It is important to note that for this comparison a different estimated mean value is produced from each of the 5000 replicate samples. The replicate estimated mean values are then said to be "smooth," which, in the context of bootstrap estimation, means that even small differences between the replicate samples produce differences in the bootstrap-estimated mean values.

The second row (Figure 11.9, middle left) displays the distribution of the range (the minimum value subtracted from the maximum value) estimated from the 5000 independent samples of $n = 25$ values. The plot on the right displays a bootstrap estimate of the distribution of the range (middle right). The range is not a "smooth" estimate. Many different replicate samples produce the same value and many other values are not produced at all. In fact, the 5000 replicate samples produced only 36 different ranges. Similarly, bootstrap estimation produced only 18 different median values (bottom right). In both cases, replicate sampling produces only a few samples with different values. Clearly, these bootstrap-estimated distributions are an incomplete picture and any subsequent analysis has no utility. Along with the range and median, examples of other statistical summaries that do not produce useful bootstrap estimates are maximum values, minimum values, percentiles, quantiles, and nearest neighbor distances.

Randomization and Bootstrap Analysis Applied to Two-Sample Data

The comparison of two samples of data is at the center of a number of statistical techniques, bootstrap estimation included (Chapters 8, 13, 17). A statistical strategy called a *randomization test* is particularly appropriate for evaluating differences observed between samples from two independent populations. Table 11.10 contains nine representative observations from a larger data set where the issue to be explored is the relationship between maternal age and weight gained during pregnancy ($n = 187$ mothers). The statistical question is: Does weight gained during pregnancy systematically differ between younger and older mothers?

Table 11.10 *Data (Partial): Maternal Age and Weight Gained (Pounds) during Pregnancy (n = 187 White Mothers)[a] to Illustrate a Two-Sample Analysis Using Randomization and Bootstrap Methods Gain/Age Data*

	1	2	3	4	5	6	7	8	...	187
Age	22	29	31	19	26	22	36	21	...	29
Gain	12.5	27.3	47.2	31.7	21.5	33.0	22.0	21.7	...	44.0
Category	0	0	1	0	0	0	1	0	...	0

[a] $n_0 = 141$ mother's age < 30 years (coded 0) and $n_1 = 46$ mother's age ≥ 30 years (coded $= 1$).

The estimated mean weight gain $\bar{x}_0 = 17.213$ based on $n_0 = 141$ mothers younger than age 30 and $\bar{x}_1 = 19.207$ based on $n_1 = 46$ mothers age 30 and older create an observed mean difference of $\bar{x}_1 - \bar{x}_0 = \bar{d} = 1.994$ pounds. The corresponding two-sample t-test statistic $T = 2.089$ yields a p-value of 0.022 (degrees of freedom $= 185$).

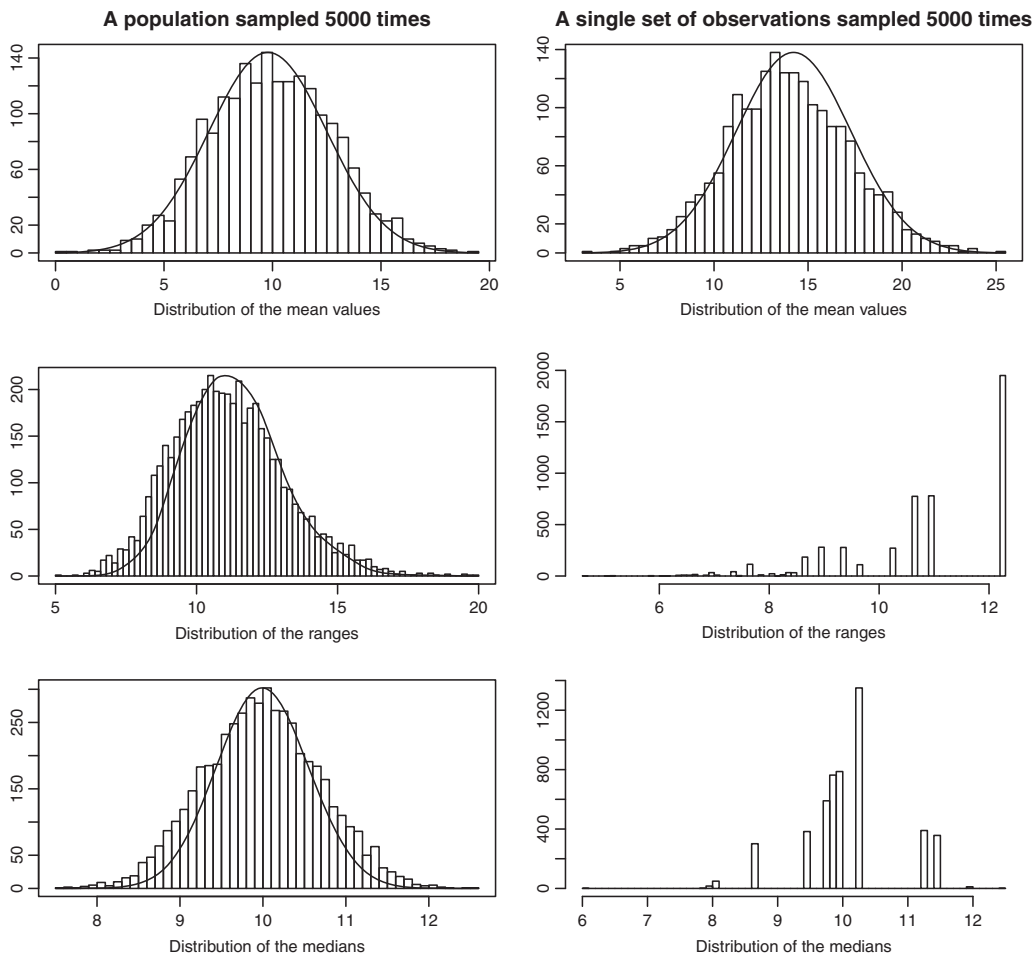


Figure 11.9 Examples of Bootstrap Estimation of a "Smooth" Summary Value (Mean) and Two "Unsmooth" Summary Values (Range and Median)

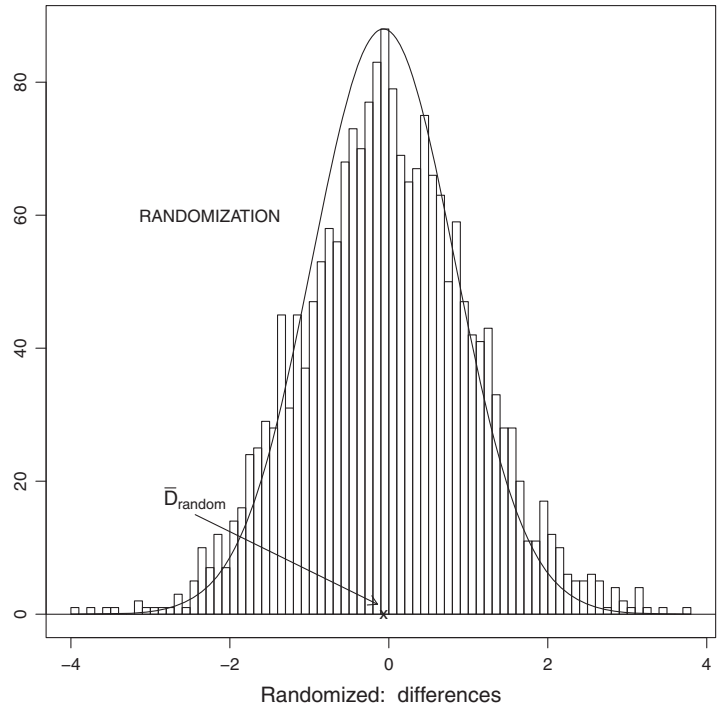


Figure 11.10 Distribution Describing Difference in Mean Differences in Weight Gain during Pregnancy Created from Samples of $n = 187$ Mothers Randomized into Two Groups ($n_0 = 141$ and $n_1 = 46 - k = 2000$)

Randomization Test

At the heart of a randomization test applied to a two-sample comparison is a computer-generated distribution of a summary statistic reflecting differences between the two compared groups as if the two sampled populations are identical. The conjecture that only random differences exist is frequently called the *null hypothesis* and occasionally the "dull" hypothesis. The null distribution is directly estimated by combining of the two samples of data into a single sample of $n = n_0 + n_1$ subjects and then randomly selecting n_0 subjects. The randomly selected observations (n_0 values) and the remaining observations (n_1 values) only differ because of random variation. Of more importance, the difference in mean values calculated from each sample differ only because of random variation. Thus, the underlying mean difference is zero. These mean differences, calculated from 2000 samples randomized into groups of $n_0 = 141$ and $n_1 = 46$ mothers, produce a distribution of 2000 estimated mean differences as if the original groups differ only because of the natural random variation in weight gain (Figure 11.10). From another perspective, the estimated distribution describes the influence of random variation on the differences in mean weight gain as if maternal age has absolutely no influence. All mothers are randomly distributed between the two groups and, other than random variation, no other reason exists for the two sampled mean values to differ.

The estimated mean difference based on 2000 differences between randomized mean values is mean difference $= \bar{D}_{[.]} = \frac{1}{2000}\sum \bar{D}_{[i]} = -0.008$, where $\bar{D}_{[i]}$ represents the ith

difference in weight gain estimated by the two mean values calculated from randomized samples. When it is possible to create exactly all possible differences between randomized mean values, the mean value $\bar{D}_{[i]}$ is exactly zero. The corresponding randomization of weight gain estimated variance based on 2000 replicate values $\bar{D}_{[i]}$ is

$$\text{variance of the difference} = \textit{variance}(\bar{D}_{[i]}) = S^2_{\bar{D}_{[i]}} = \frac{1}{1999}\sum (\bar{D}_{[i]} - \bar{D}_{[.]})^2 = 1.134.$$

An assessment of the original observed mean difference $\bar{d} = 1.994$ follows the usual pattern. The likelihood that this mean difference arose by chance alone provides an assessment of the influence of maternal age. Therefore, assuming the randomization distribution is accurately approximated by a normal distribution (Figure 11.10, solid line), then the test statistic

$$z = \frac{\bar{d} - 0}{S_{\bar{D}_{[i]}}} = \frac{1.994 - 0}{1.065} = 1.872$$

has an approximate standard normal distribution. The associated p-value is $P(\bar{D} \geq 1.994 \mid \text{null distribution}) = 0.031$. A parallel nonparametric estimate of the same p-value is simply the count of randomized mean differences that exceed the data observed mean difference $\overline{d} = 1.994$ in weight gain. Thus, an empiric and assumption-free significance probability is $p\text{-value} = P(\bar{D} \geq 1.994 \mid \text{no difference between groups}) = 65/2000 = 0.033$.

Bootstrap Analysis

A bootstrap approach addressing the same issue of association between weight gain and maternal age begins with previously described resampling the original data with replacement. Each replicate sample produces a mean value from each group, and the corresponding mean difference (denoted $\overline{d}_{[i]}$) is a replicate estimate of the mean difference in weight gain. Unlike the randomization approach, no hypothesis is involved, and each difference between these two bootstrap mean values estimates the underlying difference between the two groups. Like the randomized mean values, these differences are subject only to the variation present in the original sample data. Thus, a series of $k = 2000$ replicate samples produces an estimated distribution of the mean difference in weight gain and its variance (Figure 11.11).

Specifically, the mean difference in weight gain estimated from this bootstrap generated distribution is

$$\text{mean difference} = \bar{d}_{[.]} = \frac{1}{2000}\sum \bar{d}_{[i]} = 1.985$$

with estimated variance

$$\text{variance of the difference} = \textit{variance}(\bar{d}_{[i]}) = S^2_{\bar{d}_{[i]}} = \frac{1}{1999}\sum (\bar{d}_{[i]} - \bar{d}_{[.]})^2 = 0.949.$$

Possible assessments of the observed mean difference are, as before, two kinds of statistical tests and two kinds of confidence intervals based on the properties of the bootstrap-estimated distribution.

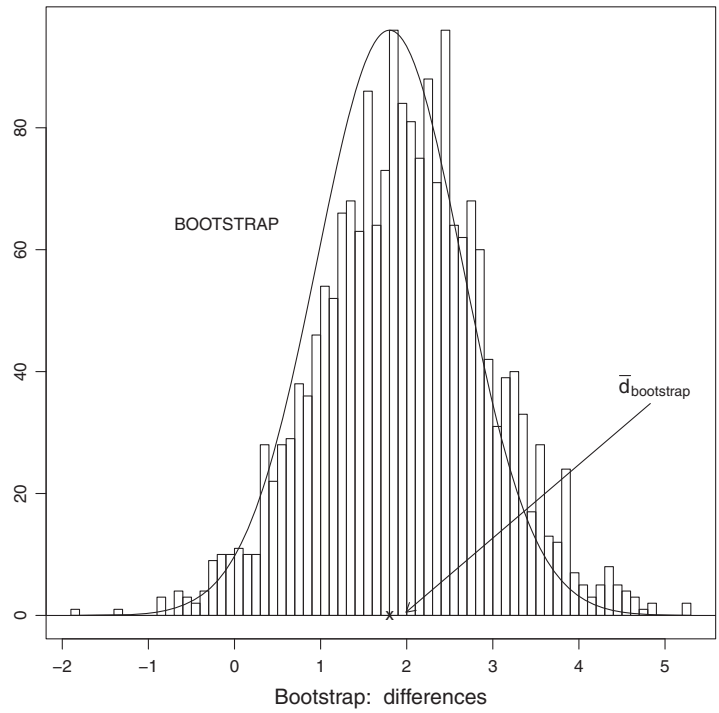


Figure 11.11 Estimated Distribution of Mean Differences in Weight Gain during Pregnancy Created from Bootstrap Replicate Sampling of $n = 187$ Mothers ($n_0 = 141$ and $n_1 = 46 - k = 2000$ Replicate Samples)

The normal distribution–based test statistic

$$z = \frac{0 - \bar{d}_{[.]}}{S_{\bar{d}_{[i]}}} = \frac{0 - 1.985}{0.974} = -2.038$$

and associated p-value of 0.021 again indicate a likely systematic difference in weight gain associated with maternal age. The distribution of replicate estimated mean differences directly indicates the likelihood the difference between observed mean value 1.985 and zero is not likely due to chance alone. Among the 2000 replicate differences, only 47 bootstrap mean values are less than zero, making the empirical and nonparametric p-value $= 47/2000 = 0.024$ (Figure 11.11).

The two kinds of estimated 95% confidence intervals are

$$\text{parametric: } \bar{d}_{[.]} \pm 1.985 S_{\bar{d}_{[i]}} = 1.985 \pm 1.960\,(0.974) \rightarrow (0.076, 3.895)$$

and

$$\text{nonparametric: 2.5th percentile and 97.5th percentile} \rightarrow (0.072, 3.894).$$

Randomization and bootstrap approaches produce similar assessments of the role of maternal age in pregnancy weight gain. The similarity is expected.

Both randomization and bootstrap-estimated normal distributions displayed on the same set of axes illustrate the relationship between the two assessments of an estimated mean

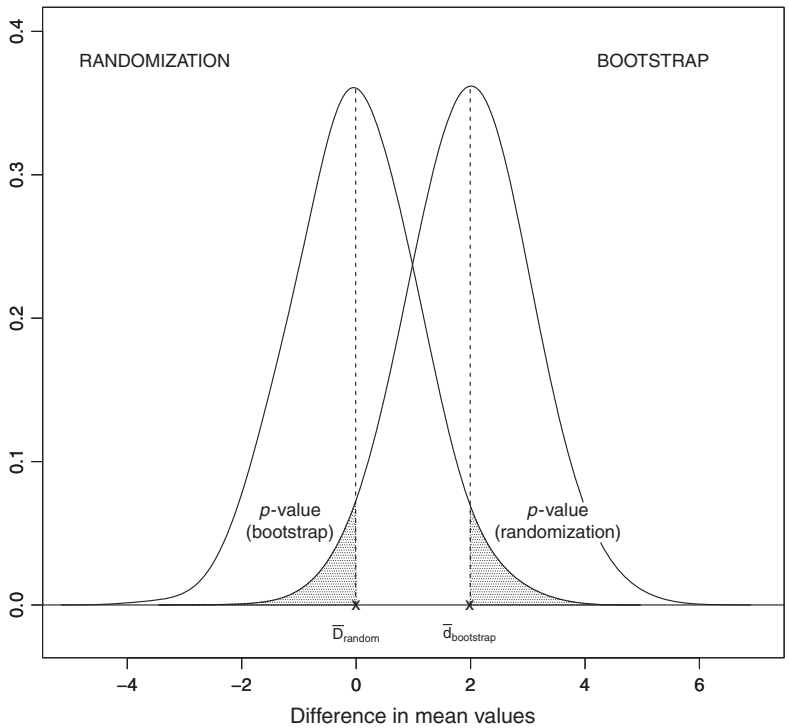


Figure 11.12 Estimated Randomization and Bootstrap Normal Distributions of the Difference between Two Samples from Two Populations

difference (Figures 11.10 and 11.11 create Figure 11.12). First, in the language of hypothesis testing, the distribution estimated from randomization sampling describes the distribution of the mean difference $\bar{D}$ under the hypothesis that only random differences exist between the two sampled populations, called the *null distribution*. The estimated distribution of the mean difference $\overline{d}$ from bootstrap replicate sampling, also in the language of hypothesis testing, is called the *alternative distribution.* It is an estimate of the distribution of the mean difference without a specific hypothesis. Both distributions produce estimates of the same variance because both values are estimated from random resampling of the same sample of data.

Thus, the two methods have similar test statistics and *p*-values. In symbols, these test statistics are

$$\text{randomization test statistic: } z = \frac{\bar{d} - 0}{S_{\bar{D}_{[i]}}}$$

and

$$\text{bootstrap test statistic: } z = \frac{0 - \bar{d}_{[.]}}{S_{\bar{d}_{[i]}}}.$$

The two test statistics differ only because the components of the estimated values z (randomization versus bootstrap) arise from different sampling strategies of the same sample of observations. More specifically, the observed mean value $\bar{d}$ and bootstrap-estimated

mean value $\bar{d}_{[.]}$ estimate the same value. Furthermore, the variances also estimate the same value, namely the variance of the estimated mean difference. Because the two test statistics have approximately the same normal distribution, they also produce similar significance probabilities (Figure 11.12). Specifically, for the weight gain analysis, the test statistic are $z(randomization) = 1.872$ and $z(bootstrap) = -2.038$. Normal distributions are symmetric, therefore, the positive and negative test statistics estimate the same p-value (Figure 11.12). For the example weight gain data, the two significance probabilities are again p-value $(randomization) = 0.031$ and p-value $(bootstrap) = 0.021$.

One last note: when the assumptions underlying parametric based statistical methods are correct or at least approximately correct, the parametric and bootstrap techniques generally yield similar results. Nevertheless, it is worthwhile exploring nonparametric resampling methods because when a parametric approach is not ideal, assumption-free methods usually remain an effective analytic strategy. Comparing parametric and nonparametric analyses is often a useful evaluation of parametric assumptions (Chapter 18).

12

Graphical Analysis

A Confidence Interval and a Boxplot

Confidence intervals and boxplots are popular statistical tools producing similar perspectives on the same issue. Namely, a parsimonious visual description of a data set or properties of an estimated value that almost always enhances interpretation (Figure 12.1).

Central to a typical confidence interval is the normal probability distribution used to estimate the location of the interval bounds (Chapter 2). An alternative that does not depend on an underlying probability distribution and is relatively uninfluenced by extreme values is a *boxplot*. In contrast to the analytic confidence interval, it is basically a visual description.

Properties of the mean and median values are roughly analogous to these two summary descriptions. Mean values indicate location and typically have at least approximate normal distributions leading to their assessment in terms of probabilities. Median values also indicate location but unlike the mean value are insensitive to possible disproportional influences of extreme observations and are not naturally related to probabilities.

A boxplot is constructed from the five data elements:

1. The location indicated by the median value (denoted m)
2. The first quartile (25th percentile, denoted F_{lower})
3. The third quartile (75th percentile, denoted F_{upper})
4. The interquartile range ($F_{upper} - F_{lower}$) and
5. The locations of data values more extreme than $m \pm 1.5\ (F_{upper} - F_{lower})$.

Figure 12.2 displays two boxplots, one created from data sampled from a normal distribution (mean = median = 0 and variance = 1) and one created from data sampled from a *log*-normal distribution (median = 1 and variance = 2.7) (Chapter 14); both plots are based on $n = 100$ random observations.

The width of the box part of the boxplot is proportional to the square root of the sample size with height equal to the interquartile range of ($F_{upper} - F_{lower}$), displaying the location of 50% of the sampled values. The line inside the box identifies the location of the median value. Two lines are extended from the ends of the box, sometimes called *whiskers*. Whiskers are defined in a number of ways. One definition is to calculate the bounds $\hat{m} \pm 1.5 \times (F_{upper} - F_{lower})$, then select the smallest data value larger than the upper bound and the largest data value smaller than the lower bound to establish the whisker end points. Values beyond the whiskers are considered extreme and possibly outlier observations. Parenthetically, the distinction between an extreme and an outlier value is that an extreme value is a member of the population sampled, while an outlier value, for one reason or another, does not belong to the

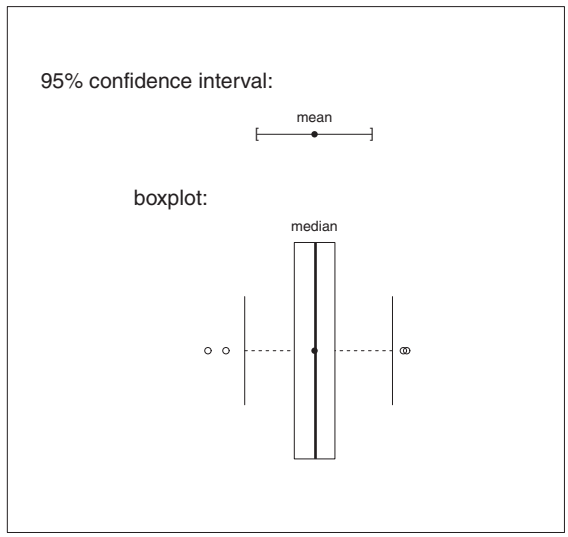


Figure 12.1 95% Confidence Interval and Boxplot Applied to the Same Data

population sampled. Outlier observations that are not extreme exist in sampled data but are rarely identified and, therefore, are rarely eliminated from the analysis, creating a usually harmless and unnoticed bias.

Differences between two distributions of data or estimated values are often easily identified by a visual comparison of boxplots. In Figure 12.2, the normal distribution is clearly close to symmetric, and the *log*-normal distribution is not (Chapter 14). The normal distribution

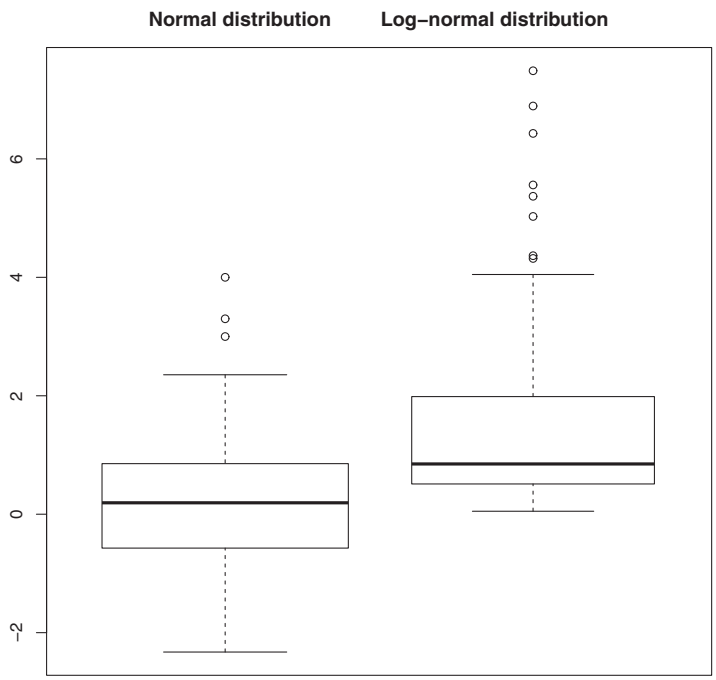


Figure 12.2 Two Distributions Described by Boxplots

Table 12.1 *Summary: Four Summary Values That Define Two Boxplots (Figure 12.2) Based on Data from Normal and Log-Normal Distributions* ($n = 100$)

	Normal		*Log*-normal	
	Observed	Theory	Observed	Theory
Median $= m$	0.025	0.000	1.003	1.000
First quartile $= F_{lower}$	−0.581	−0.674	0.486	0.509
Fourth quartile $= F_{upper}$	0.661	0.674	1.925	1.963
Interquartile range $= F_{upper} - F_{lower}$	1.242	1.349	1.440	1.454

has considerably fewer extreme observations than the *log*-normal distribution, indicating differences in variability. The four summary values calculated from $n = 100$ observations used to construct these two boxplots are presented in Table 12.1.

Multiple Comparisons – A Visual Approach

One of the more difficult statistical issues arises when a number of statistical comparisons are desired within the same data set. Modern genetic data often produce hundreds, even thousands, and occasionally tens of thousands of statistical comparisons. As the number of comparisons increases, the likelihood of errors increases; that is, large but random differences become increasingly likely to be falsely identified as systematic effects. A variety of methods have been created to control the overall error rate incurred when it is necessary to make large numbers of comparisons within the same data set. They range from simply using a significance level divided by the number of statistical comparisons (such as Bonferroni's inequality) to special and sophisticated test statistics that control the overall error rate for any number of comparisons (such as Shaffe's S-method). An often neglected alternative is a graphical approach that allows any number of visual comparisons but foregoes formal statistical significance testing.

Confidence intervals or boxplots displayed side by side are examples of such informal comparisons among estimates or among groups of estimates or among most statistical summary values. The goal is to identify differences among the relationships under investigation. What is missing is a single rigorous quantitative assessment of the likelihood of misleading results caused by random variation called the *family-wise error rate*. Confidence intervals or boxplots do not solve the "multiple testing problem" of increasing likelihood of errors with increasing number of comparisons. Nevertheless, both kinds of statistical descriptions include measures of variability as a guide to assessing the influences of random variation among compared estimates. Two examples are displayed in Figure 12.3. Six confidence intervals are a comparison of the distributions of the frequencies of three kinds of twin pairs (male/male, male/female, and female/female) occurring during two time periods. Six boxplots are a comparison of the distributions of the cholesterol levels for three smoking exposure categories among subjects with and without a coronary heart disease event.

The Cumulative Probability Distribution Function

The cumulative probability distribution function (denoted *cdf*) is a familiar statistical tool introduced in elementary statistics texts where normal, *t*-statistic, and chi-square cumulative

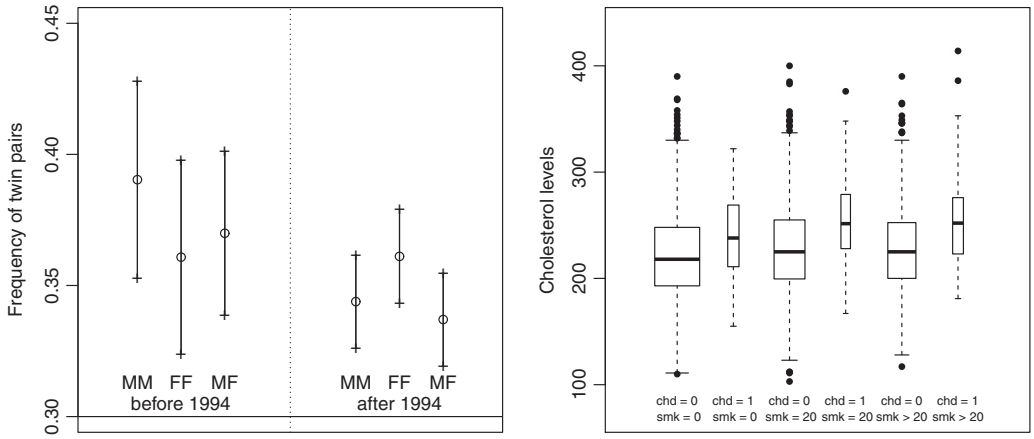


Figure 12.3 Two Examples of Multiple Comparisons Displayed as 95% Confidence Intervals (Left) and Boxplots (Right)

probabilities play central roles (Chapter 1). In addition, cumulative probabilities are often basic to fundamental data analytic techniques (Chapters 16–18 and 22).

A cumulative probability distribution of a variable labeled X is define by the expression

$$\text{cumulative distribution function} = cdf = F(x) = P(X \leq x)$$

creating cumulative probabilities denoted $F(x)$ for values x. For example, a typical cumulative probability from the standard normal distribution, for the value $x = 1.5$, is $F(x) = P(X \leq x) = F(1.5) = P(X \leq 1.5) = 0.933$ (Chapter 1). When an expression for the function $F(x)$ is known, the *cdf*-probabilities are directly calculated and, in other cases, produced with computer software or read from published tables (for example, normal and chi-square distributions) (Chapter 1).

Perhaps the simplest cumulative distribution function is created from a uniform probability distribution (Chapter 1) where all values between 0 and 1 have the same associated probability. The cumulative distribution function is then $F(p) = P(U \leq p) = p$. A slightly more complicated example is the exponential cumulative distribution function $F(x) = 1 - e^{-0.05x}$ and, for $x = 10$, then $F(x) = F(10) = P(X \leq 10) = 1 - e^{-0.05(10)} = 0.393$ (Chapter 16). For any cumulative distribution, the *cdf*-function $F(m) = P(X \leq m) = 0.5$ makes the value m the median value.

When an explicit expression for a cumulative probability distribution $F(x)$ is not known, an estimate can be directly obtained from observed data. A *cdf*-function with a known probability distribution, such as the normal and chi-square distribution, is referred to as *parametric*. When the *cdf*-function is estimated entirely from observed data, the estimate is referred to as *nonparametric*.

The nonparametric estimate of a cumulative probability function is assumption-free and estimated from the expression

$$\text{cumulative distribution function} = \hat{F}(x_i) = \frac{\text{the number of observed values } \leq x_i}{\text{total number of values observed}} = \frac{i}{n}$$

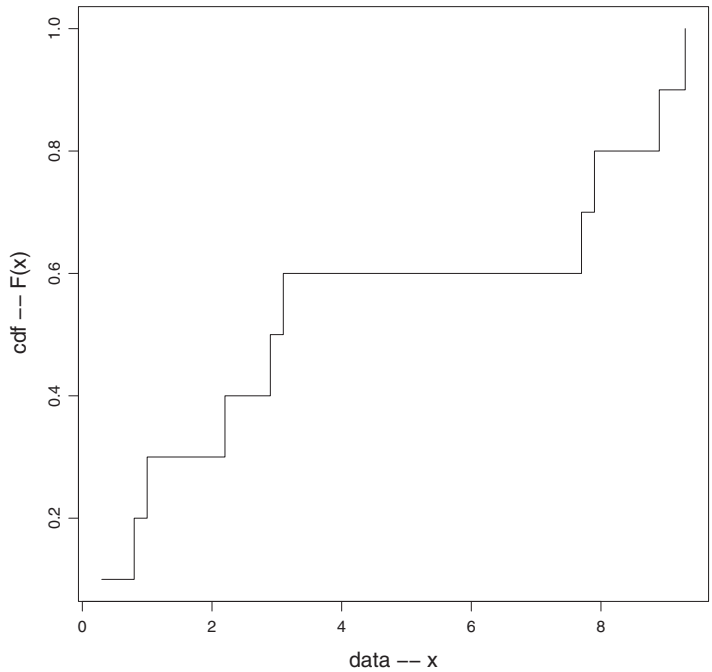


Figure 12.4 Estimated Cumulative Distribution Function from the Example 10 Sample Values

where the symbols $\{x_1, x_2, \ldots, x_n\}$ represent a sample of n observations ordered from smallest to largest value.

For the sample of $n = 10$ observations,

$$x = \{3.1, 1.0, 7.7, 2.2, 9.3, 0.8, 0.3, 2.9, 7.9 \text{ and } 8.9\},$$

when these values are ordered

$$x = \{0.3, 0.8, 1.0, 2.2, 2.9, 3.1, 7.7, 7.9, 8.9 \text{ and } 9.3\}, \text{ then}$$

the estimate of the cumulative distribution function (Figure 12.4) becomes the following:

	1	2	3	4	5	6	7	8	9	10
Ordered data (x_i)	0.3	0.8	1.0	2.2	2.9	3.1	7.7	7.9	8.9	9.3
$\hat{F}(x_i) = P(X \leq x_i) = i/10$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

Although the example applies to 10 values, the estimation from more extensive data is not more complicated. The estimate of a cumulative probability associated with the observation x_i [denoted $\hat{F}(x_i)$] remains simply the proportion of observations less than or equal to the specified value x_i in the ordered sample or again $\hat{F}(x_i) = i/n$. From another point of view, each value $\hat{F}(x_i)$ is the ith quantile or ith percentile of the collected data. From the example 10 observations, the estimated 0.9-quantile level or 90th percentile is $x_9 = 8.9$ because $\hat{F}(x) = P(X \leq 8.9) = 9/10 = 0.9$. A half-dozen versions of nonparametric estimated *cdf*-functions exist that are "fine-tuned" to estimate the cumulative probability

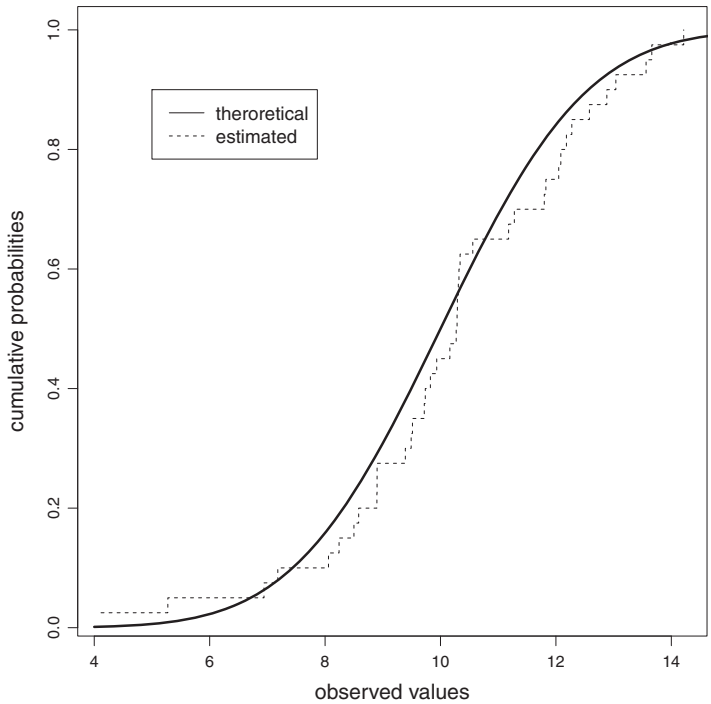


Figure 12.5 Nonparametric Estimated Cumulative Distribution Function (Dotted Line – $n = 40$) and Theoretical Normal Distribution Sampled (Solid Line)

distribution $F(x)$ with slightly improved accuracy. In several cases, the form of these estimates is $\hat{F}(x_i) = (i - a)/(n + 1 - 2a)$ and various values of a are suggested, such as $1/3$.

Figure 12.5 displays a typical nonparametric estimated cumulative distribution function based on a random sample of $n = 40$ normally distributed observations (dashed line). The corresponding parametric *cdf*-function is also displayed (solid line). Both the data and theoretical distributions are created from a normal distribution with a mean value $\mu = 10$ and standard deviation $\sigma = 2$. For example, for $x_{12} = 9.06$, then from the data $\hat{F}(9.06) = 12/40 = 0.300$ and from the normal distribution $F(9.06) = F([9.06 - 10])/2) = F(-0.47) = 0.319$.

Figure 12.6 illustrates four kinds of differences between nonparametric estimated *cdf*-functions. Specifically, for samples of $n = 40$ observations, the plots are graphical comparisons of two samples with the same *cdf*-functions (random differences), two *cdf*-functions that differ by mean values (locations), two *cdf*-functions that differ by variances (variability), and two *cdf*-functions that differ by both mean values and variances. Plots of estimated *cdf*-functions frequently make properties of sampled data obvious.

Inverse Functions for Statistical Analysis

Consider directions for traveling from place A to B.

Start at A:

Go 1 mile, then turn right
Go 1.5 miles, then turn left

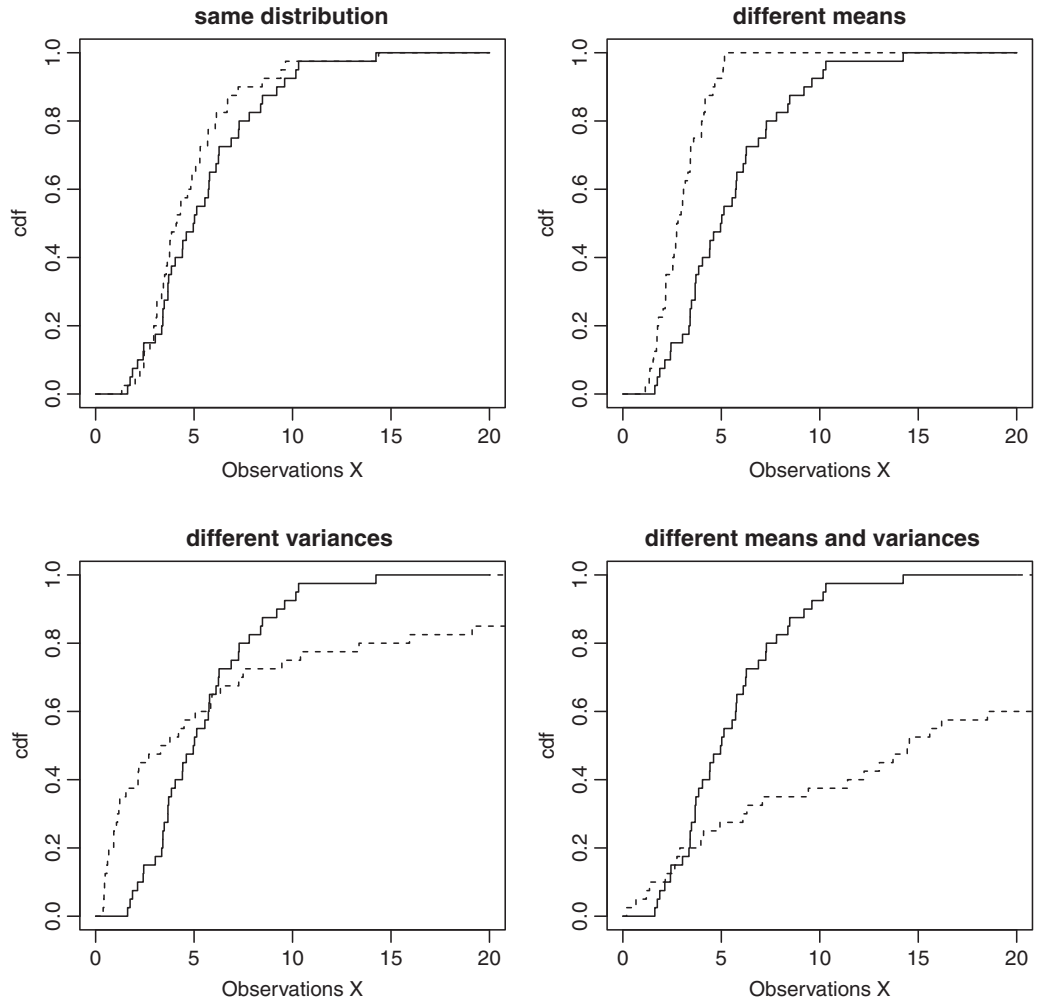


Figure 12.6 Four Plots Illustrating Comparisons of Nonparametric Estimated Cumulative Distribution Functions ($n = 40$)

Go 0.5 miles, then turn left
Go 3 miles, to arrive at B.

Then, for the return trip,
Start at B:

Go 3 miles, then turn right
Go 0.5 miles, then turn right
Go 1.5 miles, then turn left
Go 1 mile, to arrive at A.

In mathematical terms, when the trip from A to B is described by a function denoted $F(x)$, the return trip from B to A is described by the *inverse function*, denoted $F^{-1}(x)$. In more

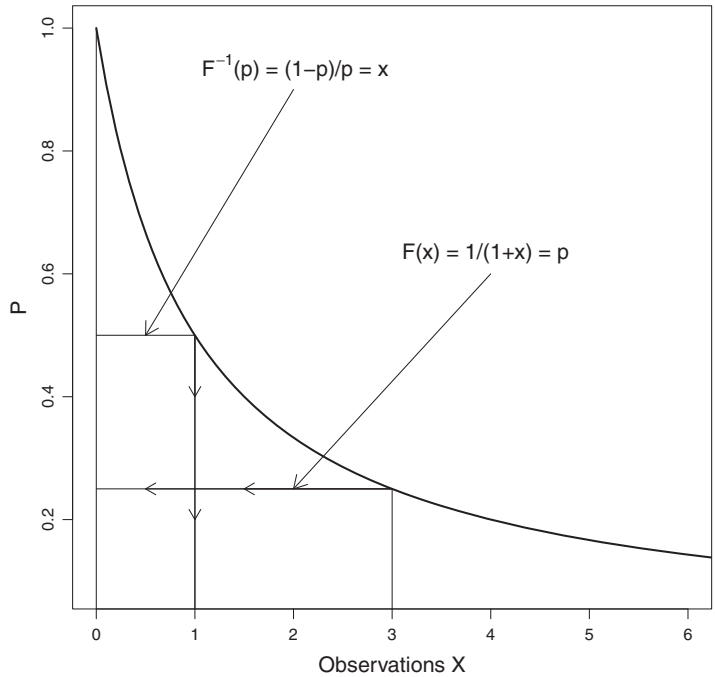


Figure 12.7 Example of Function $F(x) = 1/(1 + x) = p$ for $(x = 3 \rightarrow p = 0.25)$ and Inverse Function $F^{-1}(p) = (1 - p)/p = x$ for $(p = 0.5 \rightarrow x = 1.0)$

detail, the function $F(x) = p$ describes the path from x to p, and the inverse function $F^{-1}(p) = x$ describes the return path from p to x.

For example, the function

$$F(x) = \frac{1}{1+x} = p$$

describes the path from x to p (Figure 12.7). The inverse function

$$F^{-1}(p) = \frac{1-p}{p} = x$$

then describes the return path from p to x (Figure 12.7). A property that characterizes a function and its inverse function is

$$F[F^{-1}(p)] = p \text{ and } F^{-1}[F(x)] = x.$$

In less formal language, the function F applied to its inverse function F^{-1} is the "round trip." That is, the "trip" from x to p, followed by the return "trip" from p to x.

Specifically,

$$\text{when } x = 3, \text{ then } F(x) = \frac{1}{1+x} = \frac{1}{1+3} = \frac{1}{4} = 0.25 = p \quad x = 3 \rightarrow p = 0.25$$

Table 12.2 *List: Seven Functions and Their Inverse Functions Useful for Statistical Applications*

	Functions	Inverse functions
Previous variable	$F(x) = 1/(1+x) = p$	$F^{-1}(p) = (1-p)/p = x$
Maximum two uniform variables	$F(x) = x^2 = p$	$F^{-1}(p) = \sqrt{p} = x$
Uniform variable	$F(x) = p$	$F^{-1}(p) = x$
Linear variable	$F(x) = (x-a)/b = p$	$F^{-1}(p) = bp + a = x$
Exponential distributed variable	$F(x) = 1 - e^{-\lambda x} = p$	$F^{-1}(p) = -log(1-p)/\lambda = x$
Normal distributed variable[a]	$F(x) = pnorm(x) = p$	$F^{-1}(p) = qnorm(p) = x$
Normal distributed variable[b]	$F(x) = \dfrac{1+\sqrt{1-e^{-2x^2}/\pi}}{2} = p$	$F^{-1}(p) = \sqrt{-\dfrac{\pi}{2} log[1-(2p-1)^2]} = x$

[a]Exact normal cumulative distribution probabilities and quantile values must be looked up in tables or calculated with computer software. Nevertheless, when $F(x) = pnorm(1.645) = 0.95$ and $F^{-1}(p) = qnorm(0.95) = 1.645$, then $pnorm[qnorm(p)] = p$ and $qnorm[pnorm(x)] = x$.
[b]Approximate expressions for the cumulative standard normal distribution and its inverse function.

and

$$\text{when } p = 0.25, \text{ then } F^{-1}(p) = \frac{1-p}{p} = \frac{1-0.25}{0.25} = 3 = x \quad p = 0.25 \rightarrow x = 3.$$

Table 12.2 lists seven functions useful for statistical applications and their inverse functions.

Inverse functions are key elements to a number of statistical techniques. Two important applications are the following.

First, for a cumulative probability function $F(x)$ and a random probability p, then $F^{-1}(p) = x$ produces a random value of x from the cumulative probability distribution $F(x)$. This feature of an inverse function is used to create simulated "data" with known statistical properties (to be discussed).

Second, an inverse function applied to a known or hypothetical cumulative probability distribution produces theoretical quantile values that can be directly compared to observed quantile values. Such a comparison effectively contrasts a specific theoretical distribution to a distribution estimated from data. Plots generating the comparison are called *quantile/quantile plots* or *qqplots* (to be discussed).

As already discussed (Chapter 2), inverse functions play a role in the estimation of confidence intervals.

The second function in Table 12.2 is the cumulative distribution function of the random variable X where X represents the maximum value of two independent random values selected from a uniform distribution (Chapter 1). More formally, when u represents a random uniform value between 0 and 1 and X represents the maximum of two of these independent random values, then the cumulative probability distribution function $F(x)$ is

$$\text{cumulative probability function} = P(\text{two values of } u \leq x) = F(x) = P(X \leq x) = x^2 = p.$$

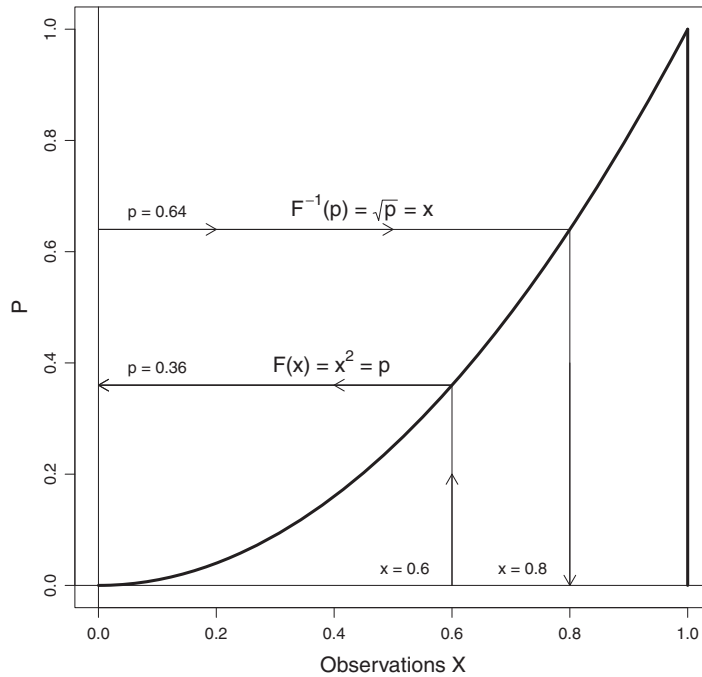


Figure 12.8 Cumulative Distribution Function of $X = \text{maximum}(u_1, u_2)$ When Values u_1 and u_2 Are Independently and Randomly Sampled Values between 0 and 1

That is, the probability that two independent and randomly selected uniformly distributed values are both less than a specified value x is x^2 and is represented by p $(0 \leq x \leq 1)$. Necessarily, the probability the maximum value is less than x is also p. Therefore, the cumulative probability distribution function of the maximum value is $F(x) = x^2 = p$ for random values x between 0 and 1. For example, the cumulative probability that $X = \text{maximum}(u_1, u_2)$ is less than $x = 0.4$ is the $cdf = F(0.4) = P(X \leq 0.4) = 0.4^2 = 0.16$. Figure 12.8 displays the *cdf*-function $F(x) = x^2 = p$ $(x \to p)$. The inverse function is $F^{-1}(p) = \sqrt{p} = x (p \to x)$. Therefore, for the value for $p = 0.16$, the inverse function generates the value $x = \sqrt{0.16} = 0.4$ $(p \to x)$. Like all functions and their inverse function,

$$F[F^{-1}(p)] = F(\sqrt{p}) = p \text{ and } F^{-1}[F(x)] = F^{-1}(x^2) = x.$$

Although this cumulative distribution function has little practical importance, it provides a simple illustration of three ways to display the comparison of two cumulative distribution functions, one calculated from theory and the other estimated from data.

Consider a sample of $n = 100$ values of the random variable X and the question: Are the sampled values accurately described by the cumulative distribution function $F(x) = x^2$? Using the assumption-free and nonparametric *cdf* estimate $\hat{F}(x_i) = i/n$ provides an answer. A simple plot of the two *cdf*-functions [$F(x)$ and $\hat{F}(x_i)$] on the same set of axes displays the differences between theory and data (Figure 12.9).

Alternatively, the same comparison is achieved by plotting n pairs of quantile values from two cumulative distribution functions [$F(x_i)$ and $\hat{F}(x_i)$] on the same set of axes (Figure 12.10).

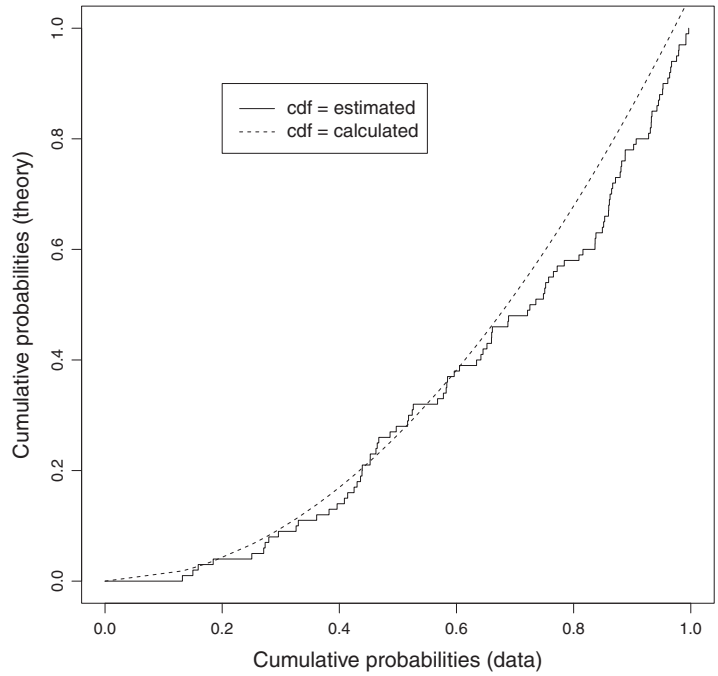


Figure 12.9 Comparison of Two Cumulative Distribution Functions – $\hat{F}(x_i) = i/n$ (Data) versus $F(x) = x^2$ (Theory)

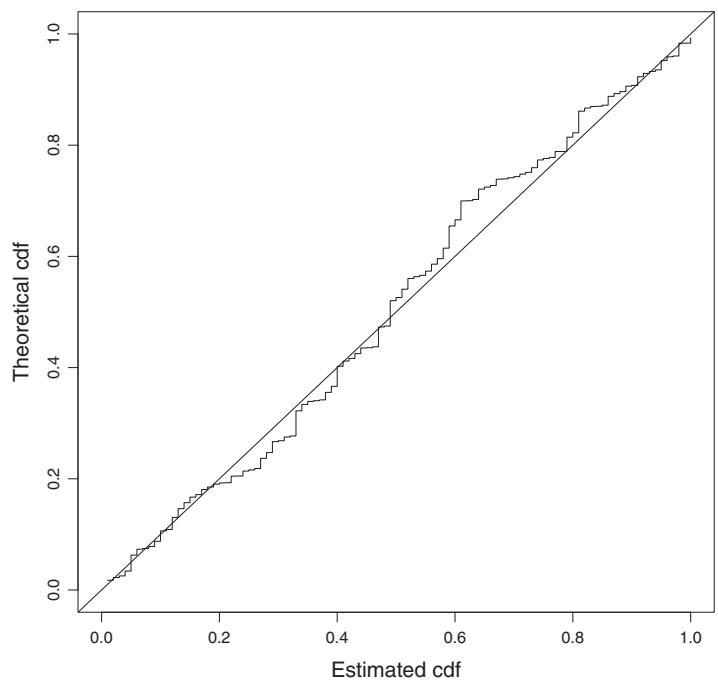


Figure 12.10 Another Version of the Comparison Estimated and Theoretical Cumulative Distribution Functions – (Data) and $F(x_i) = x_i^2$ (Theory)

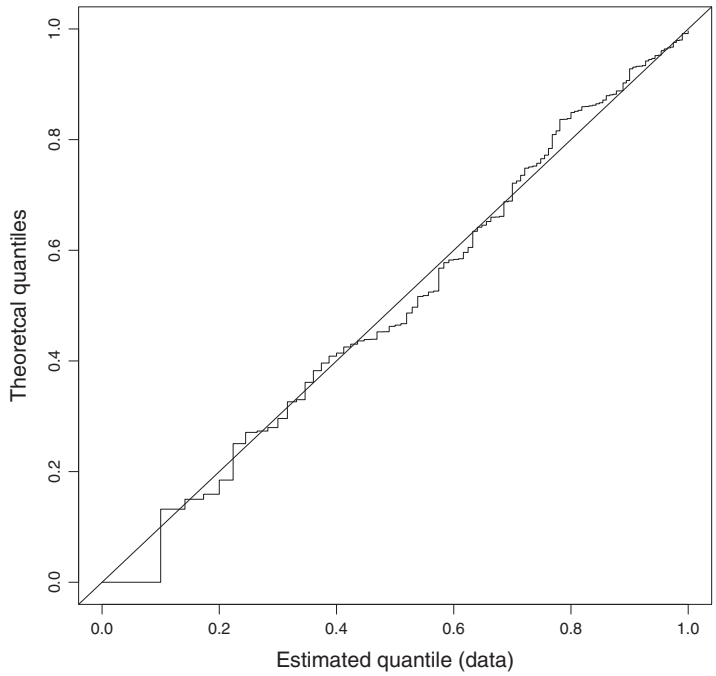


Figure 12.11 Qqplot Comparison of Cumulative Distribution Functions in Terms of Estimated versus Theoretical Quantiles – $\hat{F}(x_i) = i/n$ (Data) and $F(x_i) = x_i^2$ (Theory)

The advantage of this choice is that when compared distributions differ by chance alone, the plotted values randomly deviate from a straight line.

A fundamental strategy of visual comparisons is the following:

> Whenever possible create comparison between straight lines because differences between straight lines are usually obvious, easily interpreted, readily compared, intuitive, and the term parallel takes on concrete meaning.

A third comparison of cumulative distributions uses an inverse function of $F(x)$ to produce theoretical values with a specified distribution (denoted $\tilde{x}_i$). These calculated values, with known statistical properties of the *cdf*-function described by $F(x)$, are then directly compared to the corresponding observed data values. Each ordered data value x_i is an estimate of ith quantile regardless of the sampled distribution. The inverse function $F^{-1}(cdf_i)$ generates the theoretical ith quantile value (denoted $\tilde{x}_i$) from the specified distribution $F(x)$. The n pairs of values (x_i, $\tilde{x}_i$) are then plotted on the same set of axes. When $\tilde{x}_i$ (theory – $F(x_i) = x_i^2$) and x_i (data – $\hat{F}(x_i) = i/n$) values have the same cumulative probability distribution, the plotted pairs again randomly deviate from a straight line with intercept = 0 and slope = 1 (Figure 12.11, solid line). The resulting plot, as noted, is called a *quantile/quantile plot* or *qqplot* because the value $\hat{x}_i$ is a quantile value calculated from a known probability distribution compared to the ordered data value x_i that is an estimate of the same quantile value from the sampled data.

The exponential probability distribution is an important and widely applied statistical distribution (Table 12.2). The cumulative distribution function $cdf = 1 - F(x)$ is frequently

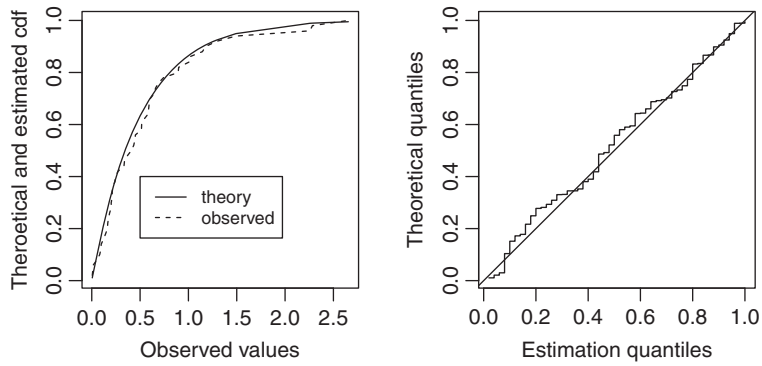


Figure 12.12 Two Kinds of Comparisons of $n = 100$ Sampled Observations to the Theoretical Exponential Cumulative Distribution Function ($\lambda = 0.05$)

basic to the analysis of survival time data (Chapters 17 and 18). For example, it has been used to describe relapse times of childhood leukemia and the time between HIV diagnosis and the onset of AIDS, and it applies to many kinds of "time-to-failure" data. Regardless of the application, the question inevitably arises: Does the theoretical exponential distribution accurately represent the relationships observed within the sample data? The visual display of theoretical and observed cumulative distribution functions is the same in principle as the previous examples.

Specifically, the exponential cumulative distribution function and its inverse function (Table 12.2) are

$$\text{cumulative probability distribution function} = F(x) = 1 - e^{-\lambda x} = p \; (x \rightarrow p)$$

and

$$\text{inverse cumulative probability distribution function} = F^{-1}(p) = -\frac{1}{\lambda} log(1-p) = x \; (p \rightarrow x).$$

The cumulative probability function $F(x) = 1 - e^{-\lambda x}$ produces a cumulative probability p for a specific quantile value x. The inverse function $F^{-1}(p) = -log(1 - p)/\lambda$ produces the quantile value x for a specific value of p. For example, when $\lambda = 0.05$, for $x \rightarrow p$

$$\text{for } x = 20, \text{ then } F(20) = 1 - e^{-0.05(20)} = 0.632 \text{ and}$$
$$\text{for } x = 40, \text{ then } F(40) = 1 - e^{-0.05(40)} = 0.845.$$

Because the inverse function is $F^{-1}(p) = -\,log(1 - p)/0.05$, for $p \rightarrow x$, then

$$\text{for } p = 0.632, F^{-1}(0.632) = -log(1 - 0.632)/0.05 = 20 \text{ and}$$
$$\text{for } p = 0.845, F^{-1}(0.845) = -log(1 - 0.845)/0.05 = 40.$$

The comparison of theoretical to estimated cumulative distribution functions is illustrated (Figure 12.12, top) using a sample of $n = 100$ random and exponentially distributed values ($\lambda = 0.05$). The corresponding qqplot, another version of the same comparison based the estimated quantile values $(\tilde{x}_i, x_i)$, is not substantially different (Figure 12.12, bottom) but produces more obvious visual differences.

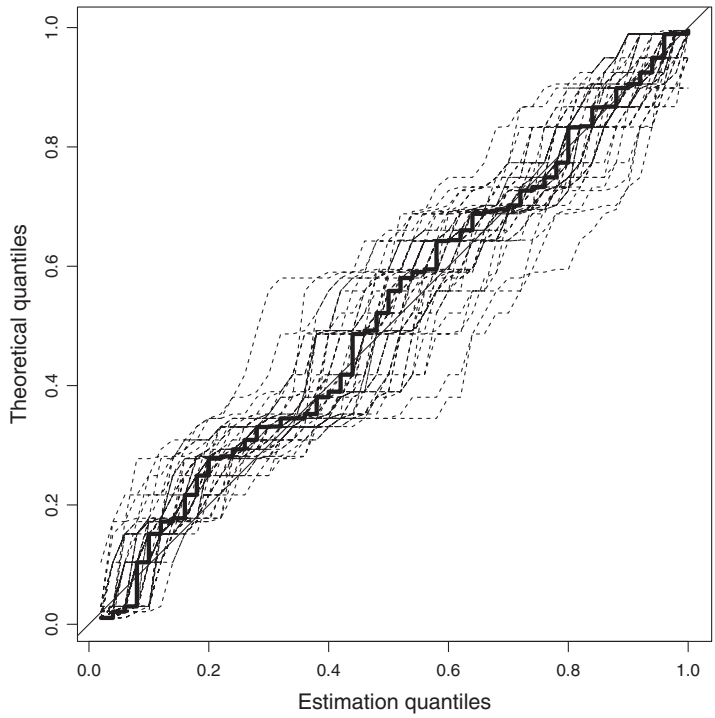


Figure 12.13 Estimated Qqplots from Replicate Random Sampling of Data (Dashed Lines) and Qqplot from Observed Data (Solid Line – Figure 12.12)

The influence of random variation on the compared cumulative distribution functions frequently is an important issue. Sampling variation is almost always present when observed data are plotted. Occasionally it is desirable to indicate the influence of random variation as part of a visual comparison. As noted (Chapter 11), a bootstrap estimate from the sampled data requires only resampling the collected values with replacement. Such a replicate sample generates an estimated cumulative distribution function that differs for the original data estimate by chance alone. A number of these replicate *cdf*-functions added to the original plot graphically display the pattern and extent of influence of sampling variation. Figure 12.13 is an example of 50 replicate plots of bootstrap estimated *cdf*-functions added to a qqplot comparison (Figure 12.12).

Perhaps the simplest comparison of two estimated cumulative distribution functions occurs when data are sampled from two different sources. An often important question is: Are the distributions different? All that is required is to sort the two sets of sampled observations into ascending order and plot the pairs of corresponding values. When the two samples have the same distribution, the within-pair differences are due only to chance. Technically, the ordered sample values produce two sequences of the same estimated quantiles/percentiles. Such a plot, once again when no systematic differences exist, yields an estimate of a straight line with intercept = 0 and slope = 1, sometimes called "*the* $x = y$ *line*" (Figure 12.14). This kind of visual comparison, for example, can be applied as an assumption-free enrichment of the classic two-sample Student's *t*-test analysis of the observed difference between data sampled from two sources.

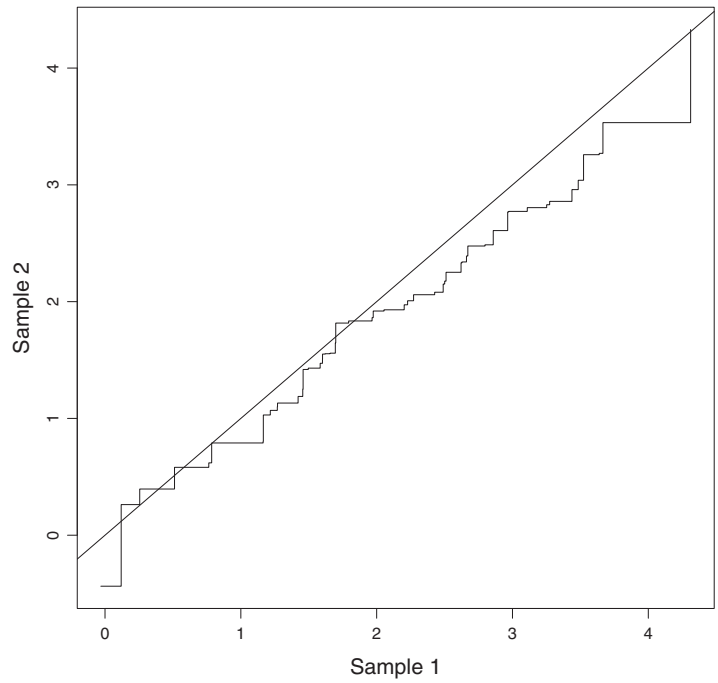


Figure 12.14 Two-Sample Comparison of Estimated *cdf*-Functions Based on $n = 50$ Observations Sampled from the Same Population, $F_1(x) = F_2(x)$

When the compared samples do not consist of equal numbers of observations, the comparison is not as simple. When the sample sizes are large and almost equal, it is likely that randomly removing a few observations will have negligible influences on a visual comparison. Also, rigorous interpolation schemes exist to produce "equal" sample sizes. The necessary interpolation process is tedious and best left to computer software.

Kolmogorov One-Sample Test

A visual comparison of two *cdf*-functions displays a "complete picture." Plots identify the locations, shapes, and magnitudes of the differences across the range of the entire distribution. In addition, statistical tests exist that focus on the influence of random variation and enhance the interpretation of visual impressions. The *Kolmogorov* test is a statistical evaluation of the difference between two *cdf*-functions. This one-sample test is a comparison of a data-estimated *cdf*-function $\hat{F}(x_i)$ based on n sampled values (denoted x_i) to a specified *cdf*-function with known cumulative probability distribution $F(x_i)$. Soviet mathematician Andrey Kolmogorov (b. 1903) derived the distribution of the maximum difference between a data-estimated *cdf*-function and a completely determined theoretical *cdf*-function when only random differences exist. In symbols, the Kolmogorov test statistic is

$$D_0 = maximum(|F(x_i) - \hat{F}(x_i)|) = maximum\left(\left|F(x_i) - \frac{i}{n}\right|\right)$$

where $i = 1, 2, \ldots, n$ = number of ordered x_i-values. The test is nonparametric making the results assumption-free and exact for small sample sizes.

Table 12.3 *Data: Random Sample of* $n = 23$ *Observations from an Exponential Probability Distribution* ($\lambda = 0.10$) *and Estimated Cumulative Distribution*[a]

Data	0.087	0.723	0.885	1.013	1.230	1.813	1.884	2.633
Cdf	0.04	0.09	0.13	0.17	0.22	0.26	0.30	0.35
Data	2.794	3.499	3.807	4.156	5.071	6.041	6.145	8.225
Cdf	0.39	0.43	0.48	0.52	0.57	0.61	0.65	0.70
Data	10.201	14.820	15.703	35.778	42.501	45.016	57.150	
Cdf	0.74	0.78	0.83	0.87	0.91	0.96	1.00	

[a] $cdf = F(x_i) =$ theoretical cumulative distribution function $= 1 - e^{-\lambda x_i} = 1 - e^{-0.10x_i}$.

A sample of $n = 23$ simulated observations with an exponential distribution ($\lambda = 0.10$) illustrates (Table 12.3). The comparison between data ($\hat{F}[x_i]$) and theoretical exponential cumulative distribution functions ($F[x_i]$) clearly shows expected differences (Figure 12.15). The statistical question becomes: Is the maximum difference likely due entirely to random variation? The Kolmogorov test statistic applied to the example data yields the test statistic $D_0 = 0.193$ with associated computer generated p-value $= 0.316$, suggesting a likely absence of systematic influences (Figure 12.15).

Frequently a completely defined cumulative distribution function is unavailable to compare to an estimated distribution. In this case, even for large data sets, using estimated parameters to specify the theoretical cumulative distribution function potentially leads to substantially biased results. For the example, using an estimate from the simulated data of $\hat{\lambda} = 0.085$ rather

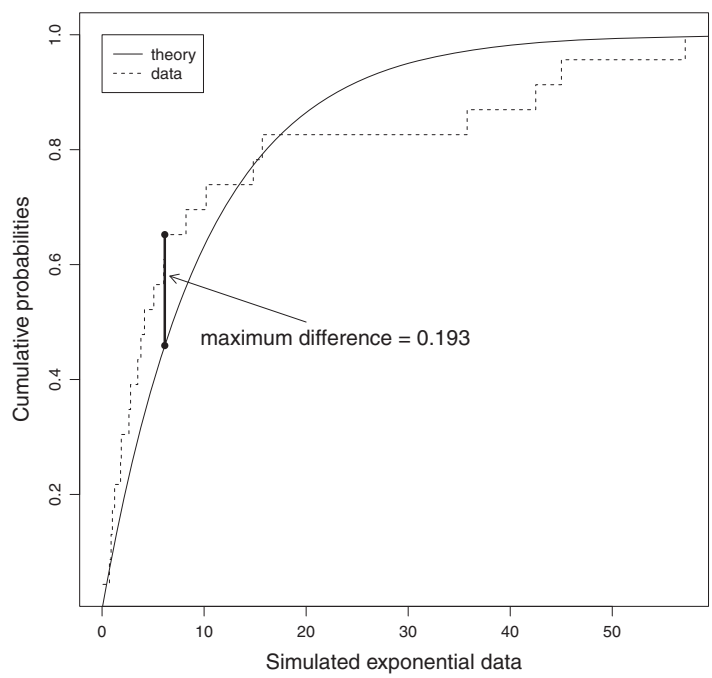


Figure 12.15 Comparison of Data-Generated and Theoretical Cumulative Exponential Probability Distribution Functions

Table 12.4 *Selected Approximate and Exact Critical Values for the Kolmogorov Test Statistic with Significance Level of* 0.05 *(Two-Tail Test)*

n	c(approx)	c(exact)	Difference
5	0.563	0.608	0.045
10	0.409	0.430	0.021
15	0.338	0.351	0.013
20	0.294	0.304	0.010
40	0.210	0.215	0.005
60	0.170	0.176	0.006
80	0.150	0.152	0.002

than the known value $\lambda = 0.10$ produces the considerably different Kolmogorov statistic of $D_0 = 0.246$ with a p-value of 0.104.

The p-values and exact critical values (denoted c) for the Kolmogorov statistical test are found in published tables or computer calculated. In addition, approximate critical values are easily calculated and accurate for sample sizes larger than 20 (error < 0.01). Four examples are

$$\text{Significance level} = P(D_0 \leq c) = 0.15, \text{ then } c = 1.14/\sqrt{n},$$
$$\text{Significance level} = P(D_0 \leq c) = 0.10, \text{ then } c = 1.22/\sqrt{n},$$
$$\text{Significance level} = P(D_0 \leq c) = 0.05, \text{ then } c = 1.36/\sqrt{n} \text{ and}$$
$$\text{Significance level} = P(D_0 \leq c) = 0.01, \text{ then } c = 1.63/\sqrt{n}.$$

A small table comparing approximate and exact significance levels gives a sense of the accuracy of these approximate values for increasing sample sizes n from 5 to 80 (Table 12.4).

Kolmogorov-Smirnov Two-Sample Test

The comparison of two data-estimated cumulative distribution functions follows the same pattern as the one-sample Kolmogorov test. This extended application is a comparison of two *cdf*-functions, both estimated from data. Parallel to the one-sample analysis, the two-sample test statistic is the maximum difference between estimated *cdf*-functions. This nonparametric approach is an alternative to the nonparametric Wilcoxon two-sample rank test or the parametric two-sample t-test (Chapter 8). The Kolmogorov-Smirnov test statistic for the comparison of estimated *cdf*-functions, denoted $\hat{F}_1(x_i)$ and $\hat{F}_2(x_i)$, is

$$D = maximum(|\hat{F}_1(x_i) - \hat{F}_2(x_i)|).$$

The previous data describing the distributions of cholesterol levels for behavior type A and type B individuals once again provide an example (Figure 12.16) (Chapters 8, 10, and 13).

The two-sample Kolomogorv-Smirnov test statistic $D = 0.40$ yields a computer-generated p-value of 0.082. For contrast, the two-sample Wilcoxon rank sum test yields a p-value of 0.012 and the parallel t-test applied to compare mean values yields a p-value of 0.014 (Chapter 8).

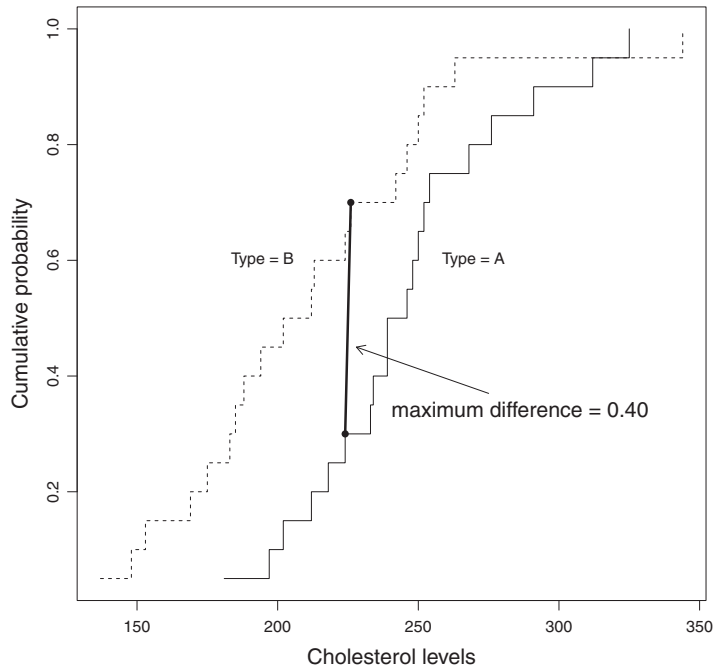


Figure 12.16 Comparison of Cumulative Probability Functions Estimated from $n_1 = 20$ Type A and $n_2 = 20$ Type B Study Subjects Describing the Relationship between Behavior Type and Cholesterol Levels for Men with Weights Exceeding 225 Pounds

Like the one-sample test, the two-sample test has simple approximate critical values (c) given by

$$\text{Significance level} = P(D \le c) = 0.15, \text{ then } c = 1.14\sqrt{\frac{n_1 + n_2}{n_1 n_2}},$$

$$\text{Significance level} = P(D \le c) = 0.10, \text{ then } c = 1.22\sqrt{\frac{n_1 + n_2}{n_1 n_2}},$$

$$\text{Significance level} = P(D \le c) = 0.05, \text{ then } c = 1.36\sqrt{\frac{n_1 + n_2}{n_1 n_2}} \text{ and}$$

$$\text{Significance level} = P(D \le c) = 0.01, \text{ then } c = 1.63\sqrt{\frac{n_1 + n_2}{n_1 n_2}}.$$

Otherwise, exact significance probabilities for a wide range of sample sizes are available from published tables or calculated with computer algorithms.

A *cdf*-plot, a *t*-test, a Wilcoxon rank sum test, and a Kolmogorov-Smirnov test similarly explore differences between two samples of data. Applying these four statistical approaches produces two possibilities. When they essentially agree, the choice of statistical approach

becomes unimportant. Or, when they significantly differ, identifying the source of difference potentially leads to a more complete understanding of the collected data.

Simulation of Random "Data" with a Specific Probability Distribution

In mathematics the value a is said to be congruent to the value b, modulo m if $a - b$ is an integer multiple of m. In less technical terms, one might say congruent means that the difference $a - b$ is contained in the set of all possible multiples of m. For example, the values $a = 16$ and $b = 8$ are congruent modulo $m = 4$ because $a - b = 8 = 2 \times m = 2 \times 4$. Modulo arithmetic plays a key role in statistics as a method to produce *pseudo-random numbers*.

Random, unlike most mathematical concepts, is defined by what it is not. A sequence of values is said to be random if future values cannot be predicted based on knowledge from previous values. The textbook example of random is the toss of a coin because the result of a second toss cannot be predicted from the result observed on the first toss. Modulo arithmetic produces pseudo-random values. A pseudo-random sequence of values mimics the random property of coin tosses, but in reality each value is directly generated from the previously generated value. Nevertheless, a pattern is virtually impossible to detect.

A pseudo-random sequence is generated by the expression

$$x_{i+1} = ax_i \; mod(m).$$

Specific choices of integer values of a and m produce values that appear randomly distributed. Furthermore, sequences produced are essentially indistinguishable for a truly random pattern. These pseudo-random sequences have a specific period made up of nonrepeating and, therefore, equally frequent values created by a process called the *power residue method*. Then dividing each value by the number of values in the sequence produces equally probable values between 0 and 1 forming a sequence with an apparent "random" *uniform probability distribution* (to be discussed). These values are then used to generate "random" values with a specified continuous or discrete probability distribution.

Starting with a value x_0, in this context called a *seed value*, selected values of a and m produce a pseudo-random series of values. The method consists of subtracting m for ax_0 until the difference is less than m. The remaining value is the first value x_1.

Repeating the process, the value ax_1 generates x_2, and, in general, the value ax_i generates x_{i+1}. For example, for the value $a = 5$ and $m = 19$, a sequence generated by $x_{i+1} = 5x_i \; mod(19)$ has a period equal to 9. When the seed value is set at $x_0 = 1$, then

$$\begin{aligned}
x_1 &= 5(1) - 0 = 5 \\
x_2 &= 5(5) - 19 = 6 \\
x_3 &= 5(6) - 19 = 11 \\
x_4 &= 5(11) - 19 = 36 - 19 = 17 \\
x_5 &= 5(17) - 19 = 66 - 19 = 47 - 19 = 28 - 19 = 9 \\
x_6 &= 5(9) - 19 = 26 - 19 = 7 \\
x_7 &= 5(7) - 19 = 16 \\
x_8 &= 5(16) - 19 = 61 - 19 = 42 - 19 = 23 - 19 = 4 \\
x_9 &= 5(4) - 19 = 1.
\end{aligned}$$

The sequence consists of nine different values. The same sequence of nine values then continues to repeat. Example values are

$$\{[5 \quad 6 \quad 11 \quad 17 \quad 9 \quad 7 \quad 16 \quad 4 \quad 1] \quad 5 \quad 6 \quad 11 \quad 17 \quad 9 \quad 7 \quad 16 \quad 4 \quad 1 \cdots\},$$

and, therefore, each value in the complete period is unique and random selections have probability $1/9$.

Another not very different approach to calculating x_{i+1} from x_i requires a value denoted $[x_i/m]$ where the square brackets indicate the integer value of the division. For example, the value $[45/19] = [2.368] = 2$. The pseudo-random sequence is then $x_{i+1} = ax_i - [ax_i/m] \times m$. Again the values $a = 5$ and $m = 19$ generate the same sequence with period equal to 9. Starting with $x_0 = 1$, then

$$\begin{aligned}
x_1 &= 5(1) - [5/19] \times 19 = 5 - 0(19) = 5 \\
x_2 &= 5(5) - [25/19] \times 19 = 25 - 1(19) = 6 \\
x_3 &= 5(6) - [30/19] \times 19 = 30 - 1(19) = 11 \\
x_4 &= 5(11) - [55/19] \times 19 = 55 - 2(19) = 17 \\
x_5 &= 5(17) - [85/19] \times 19 = 85 - 4(19) = 9 \\
x_6 &= 5(9) - [45/19] \times 19 = 45 - 2(19) = 7 \\
x_7 &= 5(7) - [35/19] \times 19 = 35 - 1(19) = 16 \\
x_8 &= 5(16) - [80/19] \times 19 = 80 - 1(19) = 4 \\
x_9 &= 5(4) - [20/19] \times 19 = 20 - 1(19) = 1
\end{aligned}$$

and, as before, the pseudo-random values are

$$\{[5 \quad 6 \quad 11 \quad 17 \quad 9 \quad 7 \quad 16 \quad 4 \quad 1] \quad 5 \quad 6 \quad 11 \quad 17 \quad 9 \quad 7 \quad 16 \quad 4 \quad 1 \cdots\}$$

An effective sequence of pseudo-random values employs a large value of m that produces a substantial period. A variety of suggestions exist for choices of values for a and m. Typical suggestions are as follows:

1. Choose the value m as a power of two or 2^k (k = large integer)
2. Choose an integer value a in the neighborhood of $\sqrt{2^k}$
3. Choose starting value x_0 to be any odd number. The choice of x_0 determines a starting point of the sequence and not its values.

Dividing the values in a complete sequence by the length of the period produces a series of equally likely unique pseudo-random values between 0 and 1.

To illustrate, a sequence for modulo $m = 327{,}619$ and starting value $a = \sqrt{m} + 3 = 575$ creates the generating expression $x_{i+1} = 575x_i \; mod(327{,}619)$. The sequence produced has a period of length 54,603 yielding a nonrepeating sequence of "random" values $p_0, p_1, p_2, \ldots, p_{54{,}603}$ that are equally likely unique pseudo-random probabilities between 0 and 1.

No guarantee exists that a particular choice of a and m produces acceptable "random" values. To find an acceptable sequence of values, a large number of statistical methods are available to evaluate "randomness." A uniform distribution of random values between 0 and

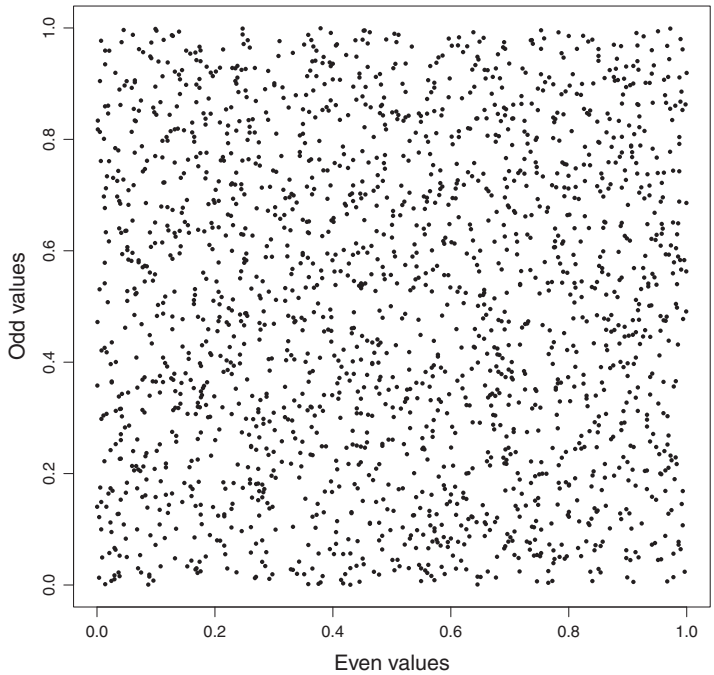


Figure 12.17 Pseudo-random Probabilities Plotted Odd against Even Values

1 has an expected mean value of 0.5 and variance of $1/12$ (Chapter 1). Two simple statistical tests, therefore, to evaluate "randomness" are: a test of the mean value of the sequence ($\bar{p}$)

$$z = \frac{\bar{p} - 0.5}{\sqrt{1/(12n)}}$$

and a test of the variance of the sequence (S_p^2)

$$z = \frac{S_p^2 - 1/12}{\sqrt{1/(180n)}}$$

where $\bar{p}$ and S_p^2 are the usual estimated mean value and variance calculated from the sequence of n "random probabilities." The test statistics z have approximate standard normal distributions when the generated values p_i are a random sample from a uniform probability distribution. For the example, the sequence of 54,603 values yields the mean value of $\bar{p} = 0.5002$ and a test statistic $z = 0.141$ with the associated p-value $= 0.888$. Similarly, the test statistic to assess the variance $S_p^2 = 0.0832$ is $z = 0.834$ with an associated p-value $=$ 0.812.

An effective graphical approach to identifying patterns among pseudo-random values is a plot of the odd values in the sequence against the even values for a substantial number of "random" values. The absence of patterns is certainly required if these values are to be treated as random (Figure 12.17). Many other and more extensive methods exist to identify patterns within sequences of values. Of course, reliable computer-generated random values

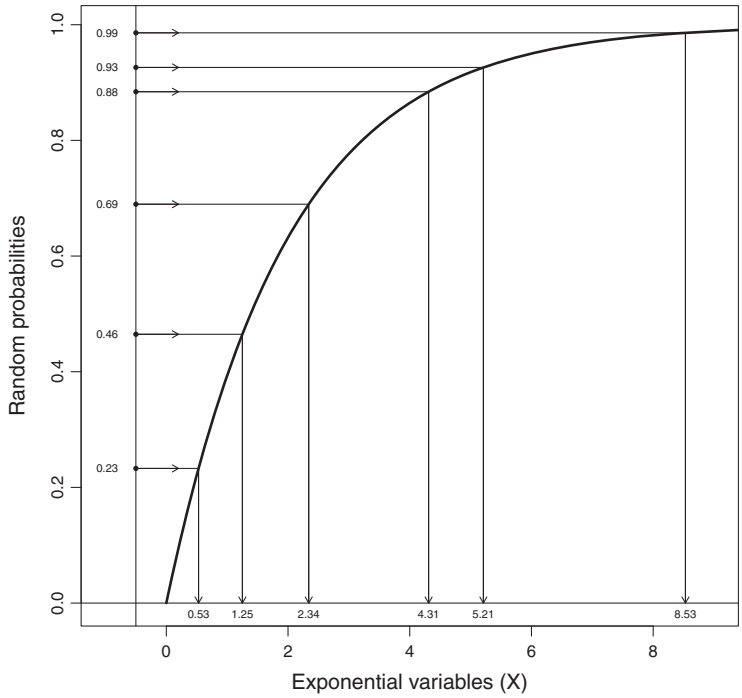


Figure 12.18 Illustration of the Generation of Six "Random" Exponentially Distributed Values (x) from Six "Random" Uniform Distributed Values (p)

are widely available. Some hand-held calculators and even a few wrist watches are capable of generating pseudo-random probabilities.

Simulation of "Data" with a Continuous Probability Distribution

Simulated values with a specified probability distribution are created by essentially the same process used to construct a qqplot. Random uniform probabilities (p_i) and inverse function $F^{-1}(p_i)$ directly create a sequence of random values x_i with probability distribution $F(x)$. The exponential cumulative probability function

$$F(x) = 1 - e^{-\lambda x} \text{ and its inverse function } F^{-1}(p) = -\frac{1}{\lambda} log(1 - p)$$

illustrate (Table 12.2). When the parameter value $\lambda = 0.5$ is selected, ordered pseudo-random probabilities $p_i = \{0.23, 0.46, 0.69, 0.88, 0.93, \text{and } 0.99\}$ produce "random" exponentially distributed values $F^{-1}(p_i) = x_i = \{0.53, 1.25, 2.34, 4.31, 5.21, \text{and } 8.53\}$ with cumulative exponential distribution $F(x) = 1 - e^{-0.5x}$ (Figure 12.18). For example, the "random" value $p_1 = 0.23$ yields the exponentially distributed "random" value $F^{-1}(0.23) = -log(1 - 0.23)/0.5 = 0.53$. The same approach applies to simulating random "data" from any continuous probability distribution defined by its cumulative probability distribution $F(x)$ using its inverse function $F^{-1}(p)$.

Simulation of "Data" with a Discrete Probability Distribution

The computer simulation of values with a specific discrete probability distribution follows the pattern of continuous simulated values. Similar to the role of an inverse function, a classification table links uniform probabilities to discrete values with a specified probability distribution. The first step is to construct a table of cumulative probabilities from the discrete probability distribution to be simulated. For example, Poisson distribution ($\lambda = 2$) cumulative probabilities are the following:

Values (x)	0	1	2	3	4	5	6	7	8	9
Probabilities ($F(x)$)	0.0	0.135	0.406	0.677	0.857	0.947	0.983	0.995	0.999	1.000

The second step is to classify n "random" values sampled from a uniform distribution (pseudo-random probabilities) into this probability table. The categories are essentially a kind of "inverse cumulative function" and produce n "random" integers with the probability distribution used to create the categories. Note that the interval widths of each category equal the discrete probabilities of the distribution to be simulated. For example, the third interval in the Poisson example table is 0.406 to 0.677, and the width 0.270 equals the probability that a randomly chosen value from a uniform distribution will be contained in the third interval creating a Poisson distributed "random" value of 3 with probability 0.270. Thus, pseudo-random uniform probabilities are transformed into random discrete values with a specified probability distribution. For example, the ordered pseudo-random uniform variables $p_i = \{0.04, 0.11, 0.21, 0{,}31, 0.38, 0.51, 0.56, 0.66, 0.72, 0.81, 0.88,$ and $0.96\}$ produce the "random" Poisson distributed counts $x_i = \{0, 0, 1, 1, 1, 2, 2, 2, 3, 3, 4$ and $5\}$ using the table of cumulative Poisson probabilities ($\lambda = 2$). Figure 12.19 displays this process for a Poisson distribution. The same approach applies to simulating random values from most discrete probability distributions.

A Case Study – Poisson Approximation

A case study illustrates a typical use of simulated "data" to provide answers to questions where a theoretical approach would be considerably more difficult. A square root transformation is frequently suggested to simplify the description or evaluation of Poisson distributed values; that is, when variable Y has a Poisson distribution with mean value and variance represented by the parameter λ, then $X = \sqrt{Y}$ has an approximate normal distribution with mean value $= \sqrt{\lambda}$ and variance $= 0.25$ (Chapters 21 and 27). The transformation becomes more accurate as Poisson distribution becomes increasingly more symmetric (larger values of λ). An important practical question becomes: What is the relationship between the parameter λ and accuracy of the approximation?

The simulation of four samples of 10,000 Poisson random values gives a sense of this relationship. Using the Poisson parameter values $\lambda = 2, 3, 6,$ and 12 describes a range of accuracy of a normal distribution as an approximation of the Poisson probability distribution (Table 12.5 and Figure 12.20). The simulated distributions reflect the extent of bias incurred by the use of the transformation $\sqrt{Y}$ as an approximation, providing guidance to its use.

Table 12.5 *Bias: Simulated Mean Values and Variances from Poisson Distribution* (Y) *and the Same Estimates from the Transformed Approximate Normal Distribution* $\sqrt{Y} = X$ *for* $\lambda = \{1, 2, 6,$ *and* $12\}$

	Parameters			
	$\lambda = 1$	$\lambda = 3$	$\lambda = 6$	$\lambda = 12$
Poisson (Y)				
Simulated mean values	1.002	2.993	6.005	12.041
Simulated variances	1.003	2.990	5.939	12.040
Theoretical mean values	1.000	3.000	6.000	12.000
Transformed (X)				
Simulated mean values	0.775	1.628	2.394	3.434
Simulated variances	0.401	0.341	0.271	0.258
Theoretical mean values	1.000	1.732	2.449	3.464
Theoretical variances	0.250	0.250	0.250	0.250

A Graphical Approach to Regression Analysis Diagnostics

A central element in successfully applying a regression analysis is verification of the underlying model assumptions providing evidence of model accuracy, often called *regression diagnostics* or, sometimes, *calibration*. Such an evaluation can be extensive and frequently employs sophisticated and frequently subtle statistical methods. In fact, the process often

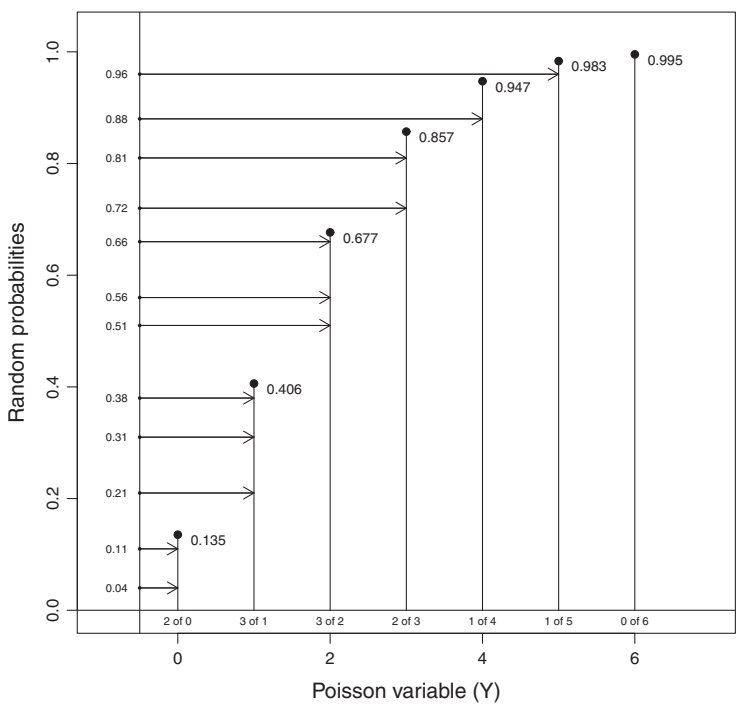


Figure 12.19 Illustration of the Simulation of "Data" with a Poisson Probability Distribution Estimated from "Random" Uniformly Distributed Values (dot = Cumulative Poisson Probability – $\lambda = 2$)

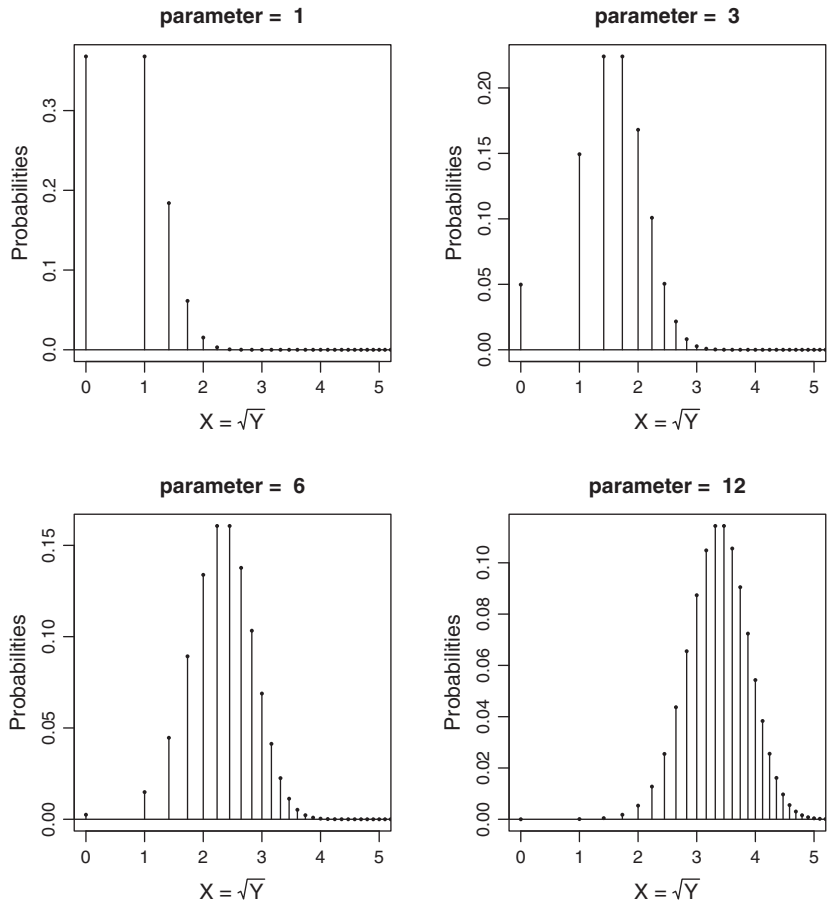


Figure 12.20 Four Example Distributions of Poisson Transformed Variables ($X = \sqrt{Y}$)

borders on an art form. The following discussion describes a few graphical tools to explore the effectiveness of regression models. These methods apply to some extent to most regression analyses but the present focus is on the multivariable linear regression model to illustrate the general pattern of regression diagnostics based on visual assessment.

A natural place to start a "diagnosis" is a description of the correspondence between data and model estimates. A direct comparison is achieved by plotting pairs of values $(y_i, \hat{y}_i)$ where y_i represents an observation and $\hat{y}_i$ represents the same observation estimated from the regression model. The comparison of model values to data values is fundamental to all regression analysis.

To be concrete, consider the following additive linear regression model and analysis relating infant birth weight (y_i) to maternal weight gain by trimesters. The regression model is

$$y_i = b_0 + b_1 gain1_i + b_2 gain2_i + b_3 gain3_i + b_4 gest_i + b_5 wt_i + b_6 ht_i + b_7 age_i + b_8 parity_i.$$

Eight independent variables and estimated model coefficients ($\hat{b}_i$), based on $n = 101$ mother-infants pairs, are summarized in Table 12.6.

Table 12.6 *Regression Model*[a] *Results: Infant Birth Weight*[b] *and Maternal Weight Gain by Trimesters (Kilograms – n = 101)*

	Estimates	*s. e.*	*p*-values
intercept	−5.005	–	–
*gain*1	−7.397	0.024	0.756
*gain*2	1.483	0.064	0.576
*gain*3	−1.179	0.111	0.999
gest	19.500	3.419	<0.001
wt	−10.960	1.289	0.397
ht	0.605	0.086	0.485
age	−3.836	0.066	0.674
parity	12.970	6.226	0.040

[a] *gain*1 = maternal weight gain in first trimester, *gain*2 = maternal weight gain in second trimester, *gain*3 = maternal weight gain in third trimester, *gest* = infant gestational age, *wt* = maternal prepregnancy weight, *ht* = maternal height, *age* = maternal age, and *parity* = infant parity.
[b] Birth weight measured in kilograms.

A first task is to display the correspondence between each of the $n = 101$ observed infant birth weights (y_i) and the regression model estimates for the same infants ($\hat{y}_i$). The comparison is displayed by plotting the model-estimated values on vertical axes and the corresponding observed values on the horizontal axes (Figure 12.21). When all model/observed differences

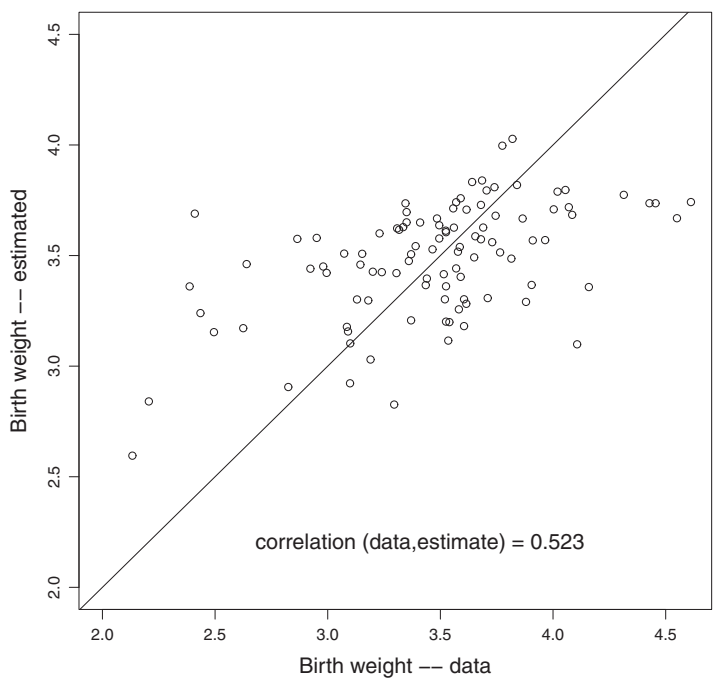


Figure 12.21 Model: Estimated Infant Birth Weights ($\hat{y}_i$) Plotted against Observed Infant Birth Weights (y_i)

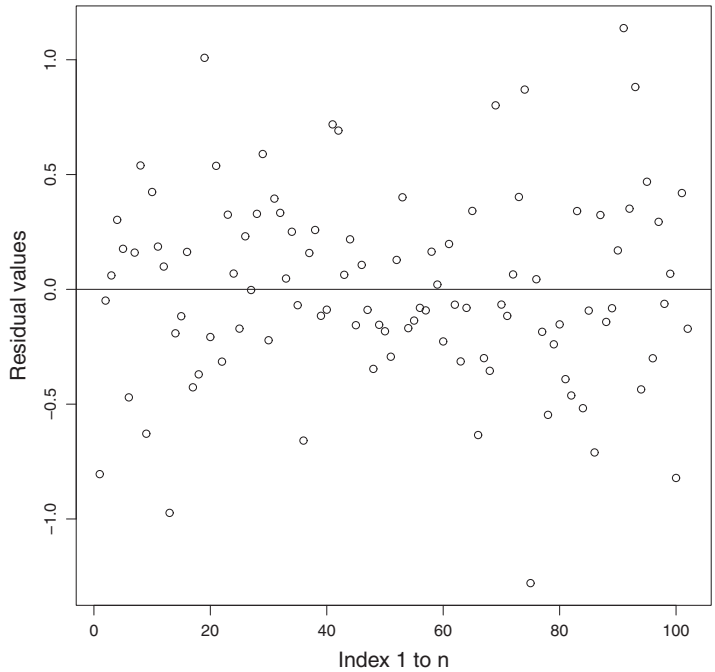


Figure 12.22 Residual Values Plotted against the Index Values 1 to $n = 101$ Observations

($y_i - \hat{y}_i$), the *residual values*, are exactly zero, the model estimates perfectly duplicate the observations, and the plotted pairs create exactly a straight line (Figure 12.21, solid line). Otherwise, the extent of deviations of these plotted pairs from this straight line (intercept = 0 and slope = 1.0) produce a description of the accuracy of the model to represent the relationships within the collected data, a process frequently called "goodness-of-fit." A product-moment correlation coefficient provides a statistical summary indicating the degree a straight line represents the plotted pairs of values y_i and $\hat{y}_i$ (Chapters 5 and 7).

An accurate statistical model produces residual values that are not only relatively small but randomly distributed. Conversely, any pattern among the differences between model and observed values indicates a failure of the model to adequately represent some aspect of the relationships within the sampled data. For example, the model might not include other important variables, causing a pattern among the residual values, or linear terms $b_i x_i$ may not accurately represent a nonlinear influence of specific independent variables x_i. The appearance of patterns among residual values is a strong indication that more sophisticated representations of one or more independent variables are needed.

Plotting the residual values on the vertical axes generated by an index on the horizontal axes potentially identifies nonrandom patterns. Typically and in Figure 12.22, the index chosen is the integers 1 to n = sample size. More relevant index values are sometimes available based on subject matter considerations.

Furthermore, for an accurate model, no pattern exists when the estimated residual values are plotted against each independent variable included the regression model. Figure 12.23 illustrates four plots (gestational age, parity, maternal weight, and maternal height) that

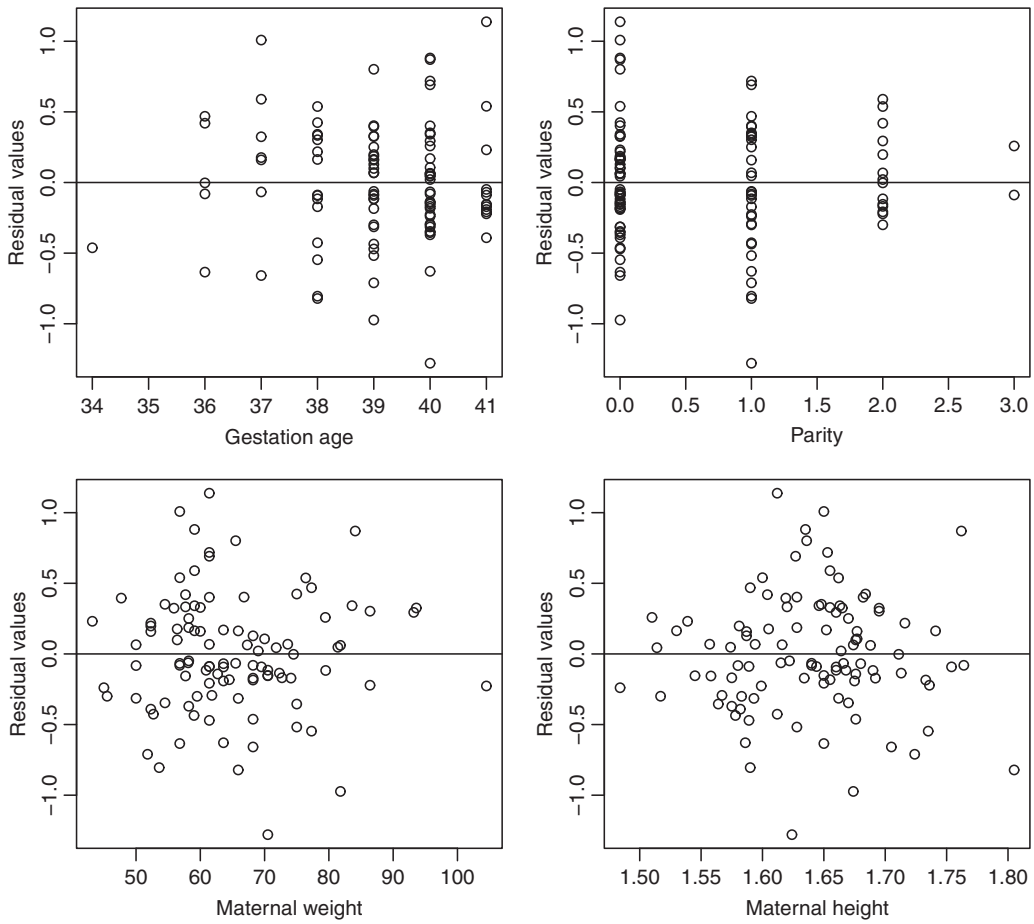


Figure 12.23 Residual Values Plotted against Values from Four Independent Variables Used in Regression Analysis – Gestational Age, Parity, Maternal Weight, and Height

potentially identify patterns among the residual values associated with specific independent variables. As noted, observed patterns indicate specific model failures but can also suggest ways to improve the performance to the model.

Regression models typically include a theoretical stochastic influence described by a specified probability distribution. The choice of this distribution is almost always a normal distribution for the multivariable linear model. Consequently, the residual values should not only be relatively small and random but should have at least the approximate probability distribution of the stochastic contribution to the model. A visual assessment of this requirement is simply accomplished by comparing two cumulative distribution functions. For the example data, a *cdf*-function for a normal distribution with the same mean and variance as the residual values is computer generated. The mean value of the residual values is always exactly zero. Their variance can be either estimated directly from estimated residual values or extracted from summary statistics that are typically produced as a part of the regression

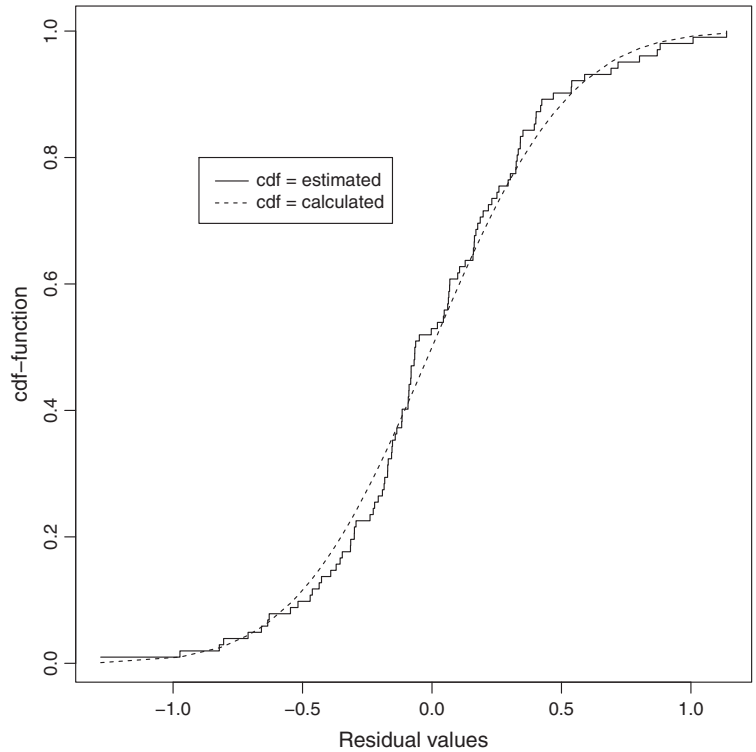


Figure 12.24 Assessment of the Assumption of Normality: Residual Values from Birth-Weight Regression Model ($n = 101$ – Solid Line) Compared to Computer-Generated Normal Distribution Values (Dashed Line) with the Same Mean and Variance

analysis. Figure 12.24 displays a theoretical cumulative distribution of residual values as if they have exactly a normal distribution (dashed line).

Also included is the plot of the nonparametric estimated cumulative distribution function estimated from the $n = 101$ data-generated residual values (solid line). The graphic comparison leaves little doubt that a normal probability distribution accurately represents the stochastic contribution to the maternal weight gain regression model (Figure 12.24).

It is always a good idea to investigate observed values that are least accurately represented by the statistical model. A variety of techniques exist. A few of statistical techniques are designed to identify different kinds of particularly influential data values. Cook's distance generated is typical. Cook's distance measures the influence of each observation on the estimated regression coefficients. Thus, various kinds of aberrant or extreme observations are potentially identified. The expression used to estimate this statistical measure is complicated and is only efficiently calculated with computer software.

A plot of $n = 101$ Cook's distances and the corresponding residual values is displayed in Figure 12.25 for the birth-weight regression model (Table 12.6). A note of warning: An extreme value of Cook's distance does not automatically identify an outlier observation. Sample data typically contain extreme values. Any conclusion that an extreme value is indeed

Table 12.7 *A Comparison: Values of a Specific Mother-Infant Pair (Observation 74) and Mean Values of the Other Mother-Infant Variables* ($n = 100$)

	Model variables								
	Bwt	*gain*1	*gain*2	*gain*3	*gest*	*wt*	*ht*	*age*	*parity*
Mean values	3.49	3.31	6.88	8.34	39.11	64.84	1.64	26.56	–
Observation 74	4.61	9.43	12.89	20.29	40.00	84.10	1.76	21.00	0

an outlier requires further investigation. A final determination primarily rests on nonstatistical subject matter considerations supported by statistical evidence. The final decision becomes an application of common sense. Mark Twain noted that the trouble with common sense is that it is not common. Marked on the plot are the 10 data values that have the greatest influence on maternal weight generated by the regression analysis (Figure 12.25, dots). Observation number 74 is the most outstanding (upper right).

To follow up, the observed values of the variables associated with this specific mother-infant pair (record 74, top right) are compared to the mean values from the remaining $n = 100$ subjects in the regression analysis (Table 12.7).

The substantial pre-pregnancy weight of the mother and a large weight gain, particularly for a first pregnancy (*parity* = 0), are likely explanations from the notable influence of this subject.

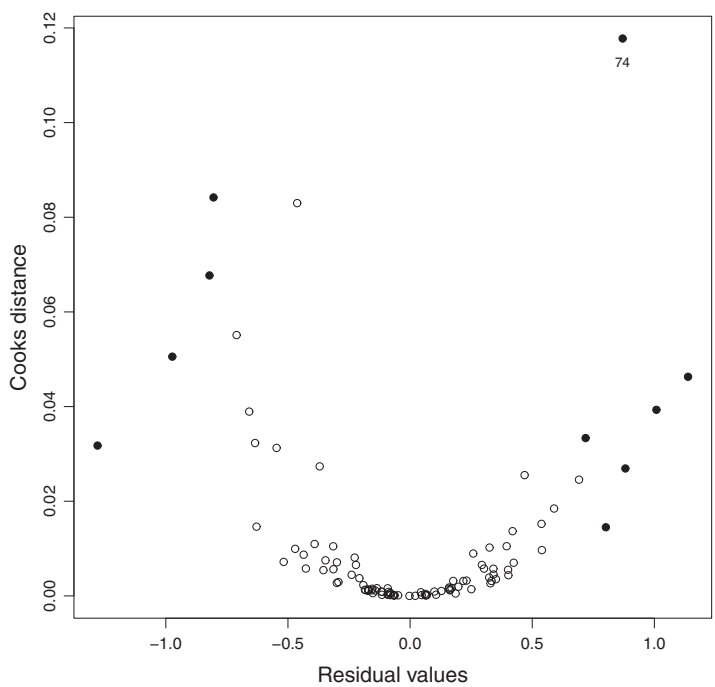


Figure 12.25 Cook's Distances Plotted against $n = 101$ Residual Values from Birth Weight Regression Analysis

Table 12.8 *Regression Model*[a] *Results: Infant Birth Weight and Maternal Pregnancy Weight Gain* ($n = 651$)

$log\text{-}odds(bwt_i) = a + b_1 gest_i + b_2 gain_i + b_3 bmi_i + b_4 age_i + b_5 parity_i$

	Estimates	*s. e.*	*p*-values
intercept	−28.53	–	–
gest	0.677	0.072	<0.001
gain	0.052	0.017	0.017
bmi	0.048	0.019	0.013
age	−0.004	0.018	0.838
parity	0.447	0.123	<0.001

[a] *gest* = infant gestational age, *gain* = total maternal pregnancy weight gain (kilograms), *bmi* = maternal body mass index, *age* = maternal age, and *parity* = infant parity.

A Graphical Goodness-of-Fit for the Logistic Regression Model

Assessment of accuracy of a logistic model is complicated by the fact that model-estimated values are either *log*-odds values or probabilities and the observed outcome is a binary variable. Thus, model-estimated values differ from the observed values regardless of the accuracy of the proposed model. Nevertheless, a graphical comparison provides a valuable evaluation of the theoretical logistic model as a summary of the relationships within sampled data.

Again, data on birth weight of newborn infants and maternal weight gain are used to illustrate ($n = 651$ mothers). The outcome measure is a binary variable identifying male infants with birth weights less than 3500 grams ($bwt_i = 1$) and male infants with birth weights of 3500 grams or greater ($bwt_i = 0$), that is, infants below and above the mean birth weight. The model and results of logistic regression analysis are described in Table 12.8.

A goodness-of-fit plot starts by creating a table based on estimated logistic model probabilities. Using these estimates, categories are created with increasing likelihood of the occurrence of the outcome under study. For example, 10 percentiles (deciles) calculated from the logistic model-estimated probabilities. The study subjects are classified into these 10 categories, and 10 probabilities describing the likelihood of the outcome variable are directly estimated, one estimate for each category. Each probability is estimated from approximately 10% of the sample data. The logistic model is then used to also estimate the mean probabilities for the same 10 categories. A plot of 10 data-estimated probabilities and 10 model-estimated probabilities produces a direct goodness-of-fit comparison.

The logistic regression model assessment of the influence of maternal weight gain on the risk of a below-mean-birth-weight infant (<3500 grams) produces an estimated logistic curve. This curve is then used to produce 10 increasing levels of risk (deciles). Using these deciles, the $n = 651$ corresponding subjects are classified into the same 10 categories. Ten data-estimated probabilities of a below-mean-birth-weight infant are then calculated directly from the observed data (circles) within each category and plotted (Table 12.9 and Figure 12.26, dashed line). For example, for the 65 infants contained in the

Table 12.9 *Goodness-of-Fit:* 10 *Model and Data-Estimated Probabilities of a Below-Mean-Birth-Weight Infant (<*3500 *Grams) for* 10 *Percentiles of Risk Created from the Logistic Model–Estimated Probabilities*

	Decile interval limits									
Lower bounds	0.00	0.18	0.30	0.39	0.46	0.47	0.53	0.64	0.76	0.94
Upper bounds	0.18	0.30	0.39	0.46	0.47	0.53	0.64	0.76	0.94	1.00
Deciles	1	2	3	4	5	6	7	8	9	10
Data probabilities	0.138	0.262	0.231	0.477	0.538	0.508	0.662	0.646	0.738	0.818
Model estimates	0.119	0.241	0.348	0.429	0.492	0.553	0.610	0.666	0.725	0.836

model-estimated fifth percentile (0.46 to 0.47), the observed probability of below-mean-birth-weight infant is $35/65 = 0.538$ (<3500 grams). The corresponding logistic model-estimated probability is 0.492. Adding the 10 model-estimated probabilities (dots) from the estimated logistic curve for the same categories to the plot creates a visual evaluation of the correspondence between model and data (Table 12.9 and Figure 12.26, solid versus dashed line).

The number of categories created from the logistic model-estimated probabilities can be increased producing additional corresponding data-estimated probabilities at a cost of considerably increased variation due to the reduced sample size within each category. Statistically smoothing these estimated values into a single curve produces a simple and sensitive visual comparison between model and data.

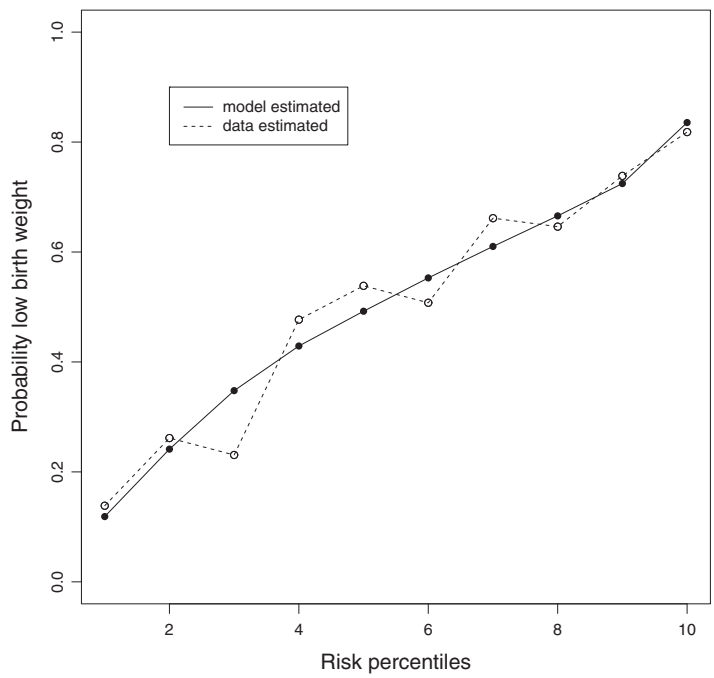


Figure 12.26 Model-Generated (Dots) and Data-Estimated Probabilities (Circles) – Goodness-of-Fit Plot for the Logistic Regression Analysis

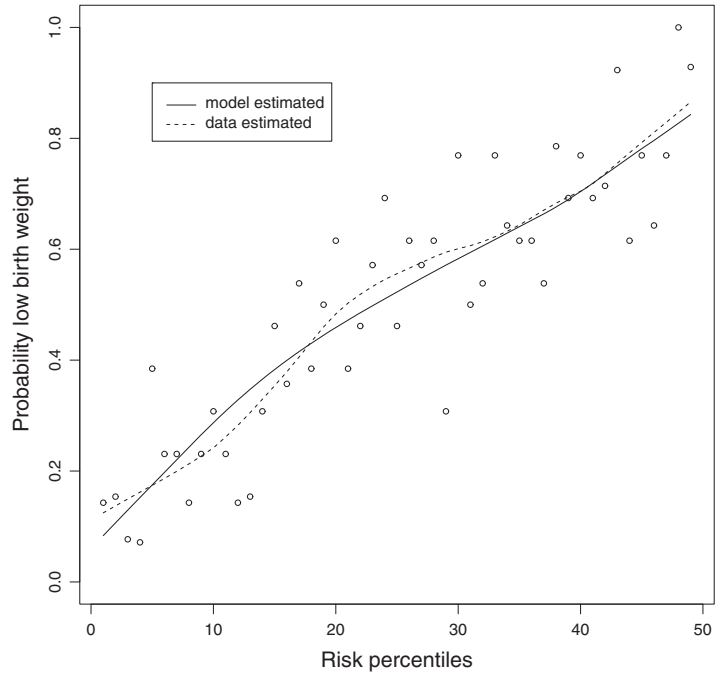


Figure 12.27 Logistic Regression Model (Solid Line) and Data-Estimated Probabilities (Dashed Line) Forming a Diagnostic Plot (50 Data-Estimated Probabilities – Circles)

For the birth-weight example, employing 50 categories and a weighted average smoothing process (Chapter 3) produces a data-generated line (dashed) and a model-generated line (solid). The comparison again clearly shows little difference between data and theoretical logistic model-estimated probabilities of an infant birth weight of less than 3500 grams (Figure 12.27), again indicating an exceptionally accurate model.

13

The Variance

Mean values dominate the statistical world primarily because of their role in statistical testing and estimation. Furthermore, mean values typically have at least approximate normal distributions making many applications simple and direct. Beginning courses in statistics are predominately about mean values. This popularity is entirely justified. A mean value, however, would not be nearly so valuable without its constant companion, its variance. Indeed, a mean value can be assessed using a normal probability distribution only when its variance plays a role. Variance also has other important and sometimes underappreciated functions. When several mean values or other summary statistics are compared, issues arise as to equality among the estimated values. Evaluation of the equality among a series of estimates becomes an evaluation of variability. In addition, answers to specific statistical questions depend on direct comparisons of sample variability.

The normal probability distribution is key to assessment of mean values and their estimation. Equally, the chi-square probability distribution is key to assessment of variability among estimates. A theorem defining the chi-square probability distribution (Chapter 1) is the following:

> If $Z_1, Z_2, \ldots, Z_m$ represent m independent and normally distributed variables, each with mean $= 0$ and variance $= 1$, then the sum of m squared z-values,
>
> $$X^2 = Z_1^2 + Z_2^2 + \cdots + Z_m^2,$$
>
> has a chi-square probability distribution with m degrees of freedom.

In much the same way the normal distribution is essential to evaluation of estimated mean values, the chi-square distribution is essential to evaluations of estimated variability. Thus, under specific conditions, probabilities associated with the variability observed in sampled data or among estimates or summary values are described by a chi-square distribution.

A Brief Note on Estimation of the Variance

The usual expression to estimate the variance represented by σ_X^2 from a sample of n sampled observations $\{x_1, x_2, \ldots, x_n\}$ is

$$\textit{sample variance of } x = S_X^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2.$$

Although this estimate is basic to many statistical analyses, the reason the sum of the squared deviations is dividing by $n - 1$ and not n is rarely explained. One might argue that the choice of being divided by n yields an estimate directly interpreted as the mean of the

squared deviations. However, the divisor n is rarely used. The reason lies in the fact that the mean squared deviation produces a slightly biased estimate of the variance of the population sampled.

Note that, for a population with mean value denoted μ, the sum of squared deviations based on $\bar{x}$ is

$$\begin{aligned}\text{sum of squares} &= \sum (x_i - \bar{x})^2 \\ &= \sum ([x_i - \mu] - [\bar{x} - \mu])^2 \\ &= \sum (x_i - \mu)^2 - 2(\bar{x} - \mu) \sum (x_i - \mu) + n(\bar{x} - \mu)^2 \\ &= \sum (x_i - \mu)^2 - n(\bar{x} - \mu)^2 \qquad i = 1, 2, \ldots, n.\end{aligned}$$

Therefore, the sum of deviations from an estimated mean value other than μ produces a smaller sum of squared deviations that underestimates the variability of the value x. In symbols, then $\Sigma(x_i - \bar{x})^2 \leq \Sigma(x_i - \mu)^2$. The sum of squares based on the estimated mean value $\bar{x}$ is an estimate of $n\sigma_X^2 - \sigma_X^2 = (n-1)\sigma_X^2$. Therefore, dividing the sum of squared deviations by $n - 1$ compensates for the reduction caused by the usually necessary use of $\bar{x}$ as an estimate of the mean value μ producing a more accurate estimate (unbiased) of the variance σ_X^2. Clearly, the distinction between $n - 1$ and n is an issue only when the sample size is small. Parenthetically, the sum of squared deviations divided by n is the slightly biased maximum likelihood estimate of the population variance σ_X^2 from a normal distribution (Chapter 27).

A Confidence Interval

Although the normally distributed values represented by z_i in the previous chi-square theorem have a number of different forms, the most elementary is

$$z_i^2 = \left[\frac{x_i - \mu}{\sigma x}\right]^2$$

where the value x_i has a normal distribution with mean value represented by μ and variance represented by σ_X^2. Then, for a sample of n independent observations x_i, the value of the expression

$$X^2 = \sum z_i^2 = \sum \left[\frac{x_i - \mu}{\sigma x}\right]^2 \approx \sum \left[\frac{x_i - \bar{x}}{\sigma x}\right]^2 = \frac{(n-1)S_X^2}{\sigma_X^2} \quad i = 1, 2, \ldots, n$$

has a chi-square distribution because, as the theorem requires, each value z_i has a normal distribution with mean value $\mu = 0$ and variance $\sigma_Z^2 = 1.0$. Therefore, a chi-square statistic can be viewed as comparison of a sample estimated variance S_X^2 based on n data values to a theoretical known variance σ_X^2. Typically, the mean value μ is not known and is estimated by the sample mean value $\bar{x}$; that is, the estimated variance is $S_X^2 = \sum (x_i - \bar{x})^2/(n-1)$. The value X^2 continues to have a chi-square distribution but with $n - 1$ degrees of freedom. Note that when S_X^2 exactly equals σ_X^2, the chi-square statistic becomes $X^2 = n - 1$ and, as might be expected, the value $n - 1$ is the mean value of the chi-square distribution (degrees of freedom) (Chapter 1).

Consider a small example of $n = 10$ observations sampled from a normal distribution with mean value $\mu = 10$ and variance $\sigma_X^2 = 2$:

$$\{10.92, 9.20, 11.02, 9.20, 13.48, 11.24, 10.41, 12.22, 9.59 \text{ and } 9.24\}.$$

The sample mean value is $\bar{x} = 10.652$ and the estimated variance is $S_X^2 = 2.045$.

Like all estimates, the estimate S_X^2 is subject to the influence of sampling variation (Chapter 2). As usual, a confidence interval describes the accuracy and precision of the estimated value. A direct application of the chi-square probability distribution produces a 95% confidence interval based on the estimated variance S_X^2.

To start, from a chi-square probability distribution with $n - 1$ degrees of freedom,

$$P\left[X^2_{n-1,0.025} \le X^2 \le X^2_{n-1,0.975}\right] = P\left[X^2_{n-1,0.025} \le \frac{(n-1)S_X^2}{\sigma_X^2} \le X^2_{n-1,0.975}\right] = 0.95$$

where $X^2_{n-1,q}$ represents a q-level quantile value. In symbols, then $P(X^2 \le X^2_{n-1,q}) = q$.

Rearrangement of this probability statement yields the 95% confidence interval $(\hat{A}, \hat{B})$ for the variance σ_X^2

$$95\% \text{ confidence interval} = P\left[\frac{(n-1)S_X^2}{X^2_{n-1,0.975}} \le \sigma_X^2 \le \frac{(n-1)S_X^2}{X^2_{n-1,0.025}}\right] = P(\hat{A} \le \sigma_X^2 \le \hat{B}) = 0.95$$

based on its estimate S_X^2. Thus, the estimated confidence interval $(\hat{A}, \hat{B})$ has a probability of 0.95 of containing the value of the underlying variance σ_X^2 associated with the population sampled.

For the example data ($n = 10$), the quantile values for probabilities $q = 0.025$ and $1 - q = 0.975$ from a chi-square distribution with $n - 1 = 9$ degrees of freedom are

$$X^2_{9,0.025} = 2.700 \quad \text{and} \quad X^2_{9,0.975} = 19.023.$$

The 95% confidence interval bounds become the following: lower bound $= \hat{A} = 9(2.045)/19.023 = 0.968$ and upper bound $= \hat{B} = 9(2.045)/2.700 = 6.816$ based on the estimated variance $S_X^2 = 2.045$. Therefore, the value estimated by S_X^2 (namely, σ_X^2) has a 0.95 probability of being contained within the confidence interval $(\hat{A}, \hat{B}) = (0.968, 6.816)$. Unlike the confidence interval constructed from a mean value ($\bar{x}$) (Chapter 2), the confidence interval constructed from an estimated variance (S_X^2) is not symmetric and is sensitive to the requirement that the data sampled be normally distributed.

Test of Variance

Occasionally, it is desired to compare a variance estimated from a sample of data to a variance of a specific postulated probability distribution. A chi-square distribution applied to make this comparison is called a *test of variance*. A test of variance is what the name suggests. The variance estimated from sampled data is directly compared to the variance that would be expected from a specified distribution. For example, if n sampled values arise from a binomial distribution, then the variability among the observed values would be that of a

Table 13.1 *Data: Observations with Potential Poisson Distribution – Count per Minute of Radioactive Emissions* ($n = 60$)

1	3	1	0	1	0	0	1	0	2	2	0	0	0	1	1	0	1	2	2	0	1	2	1	1	0	2	2	1	0
2	1	1	1	1	0	0	1	1	3	2	0	0	4	3	1	0	1	2	2	1	2	1	1	0	0	1	1	2	2

Table 13.1 *(Continued) Summary: Frequency of the Occurrence of Each Count per Minute*

Summary counts							
Occurrences	0	1	2	3	4	≥5	Total
Counts	18	24	14	3	1	0	60

binomial distribution, $variance = np(1 - p)$ (Chapter 4). A chi-square test statistic to assess this comparison of estimated and postulated variances is

$$X^2 = \frac{\sum (x_i - n\hat{p})^2}{n\hat{p}(1 - \hat{p})} = \frac{(n - 1)S_X^2}{n\hat{p}(1 - \hat{p})} \quad i = 1, 2, \ldots, n$$

and X^2 has an approximate chi-square distribution when in fact the n observed values x_i are sampled from a binomial distribution. Because the binomial probability p is estimated from the sampled values (denoted $\hat{p}$) when it is unknown, the degrees of freedom become $n - 1$ (Chapter 27). Thus, the chi-square statistic is a comparison of the data-estimated variance S_X^2 to the postulated binomial distribution variance $\sigma_X^2 = variance\,(\hat{p}) = np(1 - p)$.

Similarly, an assessment of the conjecture that sampled data have a Poisson probability distribution follows the same pattern. The chi-square test statistic is

$$X^2 = \frac{\sum (x_i - \hat{\lambda})^2}{\hat{\lambda}} = \frac{(n - 1)S_X^2}{\hat{\lambda}} \qquad i = 1, 2, \ldots, n.$$

The mean value and the variance of a Poisson distribution are identical and denoted λ (Chapter 4). Again, the chi-square test of variance is a comparison of the data-estimated variance S_X^2 to the postulated Poisson probability distribution variance of $\sigma_X^2 = variance\,(x) = \lambda$, estimated by $\hat{\lambda} = \bar{x}$ (Chapter 4).

Consider a sample of counts of emissions of radio-active particles from one of the experiments of the early physicist Ernest Rutherford (b. 1871). The number of alpha-particle emissions, detected by a counter, observed within each minute were recorded for one hour. The 60 values from a single experiment are listed in Table 13.1.

The estimated assumption-free sample variance is $S_X^2 = 0.891$ (Table 13.1). For a Poisson probability distribution, the sample mean value $\bar{x} = \hat{\lambda}$ estimates both the Poisson mean value and variance. The Poisson distribution generated estimate of the variance from the emissions data is $\bar{x} = \hat{\lambda} = 65/60 = 1.083$. The chi-square test statistic, $S_X^2 = 0.891$ compared to $\hat{\lambda} = 1.083$, becomes

$$X^2 = \frac{(n - 1)S_X^2}{\hat{\sigma}_X^2} = \frac{59(0.891)}{1.083} = 48.538$$

or, in more detail,

$$X^2 = \frac{\sum (x_i - \bar{x})^2}{\bar{x}} = \frac{(1-1.083)^2 + (3-1.083)^2 + \cdots + (2-1.083)^2}{1.083} = \frac{52.583}{1.083} = 48.538.$$

The test of variance value $X^2 = 48.538$ has a chi-square distribution with $n - 1 = 59$ degrees of freedom when the data are a random sample from a Poisson probability distribution. The resulting p-value is $P(X^2 \geq 48.538|\textit{Poisson distribution}) = 0.833$. A chi-square distribution based on $n - 2$ degrees of freedom is slightly more accurate (p-value = 0.807).

Homogeneity/Heterogeneity

The chi-square distribution is central to assessing differences among a series of estimated values. For example, probabilities a low birth weight infant estimated from several different ethnic groups are certainly not identical. A basic question is: Are the observed differences due to characteristics of the groups (heterogeneous), or are these differences due to sampling variation alone (homogeneous)? The statistical question becomes: Are the underlying probabilities from k groups equal? or, in symbols, does $p_1 = p_2 = \cdots = p_k = p$? Because the observed differences among the estimated probabilities would be expected to vary, a chi-square statistic provides an evaluation of the likelihood the observed variation is strictly random. Thus, variability reflects heterogeneity among a series of estimates, and a chi-square statistic provides a formal statistical evaluation of the observed heterogeneity.

Consider summary statistics denoted $\hat{g}_j$ estimated from k groups, denoted $\{\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_k\}$, with at least an approximate normal distribution with variance $\sigma^2_{\hat{g}}$. When these k estimates randomly differ from a single theoretical value represented by g_0, the observed variation among the summary values is described by a chi-square distribution. Specifically, for k estimated values $\hat{g}_j$,

$$X^2 = \sum Z_j^2 = \sum \left[\frac{\hat{g}_j - g_0}{\sigma_{\hat{g}}}\right]^2 = \frac{\sum (\hat{g}_j - g_0)^2}{\sigma^2_{\hat{g}}} = \frac{(n-1)S^2_{\hat{g}}}{\sigma^2_{\hat{g}}} \quad j = 1, 2, \ldots, k$$

has a chi-square distribution with k degrees of freedom when each of the k estimates $\hat{g}_j$ differ from the specified value g_0 by chance alone. To repeat, the chi-square test statistic is a comparison of sample variance $S^2_{\hat{g}}$ to theoretical variance $\sigma^2_{\hat{g}}$.

Three examples of frequently encountered chi-square comparisons are the following:

1. Test of homogeneity of k at least approximate normal normally distributed estimated mean values each calculated from n_j observations: $\hat{g}_j = \bar{x}_j$ and $g_0 = \mu$, then

$$X^2 = \frac{\sum n_j(\bar{x}_j - \mu)^2}{\sigma^2_X} \qquad j = 1, 2, \ldots, k = \text{number of estimates } \bar{x}_j,$$

2. Test of homogeneity of k at least approximate normal normally distributed estimated probabilities each calculated from n_j observations: $\hat{g}_j = \hat{p}_j$ and $g_0 = P$, then

$$X^2 = \frac{\sum n_j(\hat{p}_j - P)^2}{P(1-P)} \qquad j = 1, 2, \ldots, k = \text{number of estimates } \hat{p}_j,$$

and

3. Test of homogeneity of k at least approximate normal normally distributed estimated log-rates each calculated from n_j observations: $\hat{g}_j = \hat{r}_j$ and $g_0 = R$, then

$$X^2 = \Sigma d_j[log(\hat{r}_j) - log(R)]^2 \qquad j = 1, 2, \ldots, k = \text{number of estimates } \hat{r}_j,$$

because $variance[log(\hat{r}_j)] = 1/d_j$ where d_j represents the number of events used to estimate the rate $r_j = d_j/n_j$ from the jth group (Chapters 16 and 27).

All three examples address the fundamental chi-square question: Is the observed variation among the estimated values random? If the answer is no, the chi-square test statistic likely reflects the nonrandom variation yielding evidence of heterogeneity (small p-value). That is, a large chi-square statistic is likely due to additional systematic variation among the groups sampled. If the answer is yes, the chi-square test statistic likely reflects only random variation (homogeneity) among the groups sampled (large p-value). In practice, the specific value of g_0 is frequently not known and is replaced by an estimate $\hat{g}_0$ that causes the degrees of freedom to be reduced.

Analysis of Variance – Mean Values

Undoubtedly, the most important test of homogeneity is the comparison of k estimated mean values sampled from k populations. A sample of independent observations (denoted x_{ij} – ith value sampled from the jth population) allows the assessment of differences among a series of k mean values, denoted $\bar{x}_j$, each calculated from k groups of n_j sampled observations.

The total variability among the sampled x_{ij}-values consists of into two components. Specifically, the components summarize the differences (variability) among observed values within each group and differences (variability) between group mean values from each group or, in symbols,

$$\text{total} = \text{within} + \text{between or } x_{ij} - \bar{x} = (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x}).$$

Then, squaring these deviations and adding the squared values of the observations produces three summary sums of squared deviations. The summary values are

$$\Sigma\Sigma(x_{ij} - \bar{x})^2 = \Sigma\Sigma(x_{ij} - \bar{x}_j)^2 + \Sigma n_j(\bar{x}_j - \bar{x})^2$$

where again $i = 1, 2, \ldots, n_j$ = number of observations in each group and $j = 1, 2, \ldots, k$ = number of groups (Chapter 27). A more succinct notation for these three sums of squared deviations is $SS_t = SS_w + SS_b$ where SS_t (t for total) measures the variability observed among all sampled values without regard to group membership, SS_w (w for within groups) combines measures of the variability from within each of the k sampled groups into a single estimate and SS_b (b for between groups) measures the variability among the k mean values $\bar{x}_j$. The degrees of freedom associated with these three summary sums of squared values are $N - 1$ (total), $N - k$ (within) and $k - 1$ (between). The symbol N represents the total number of observations or $N = \Sigma n_j$. Like the sums of squares, the total degrees of freedom separates into two corresponding components, namely, $N - 1 = (N - k) + (k - 1)$.

This partition of the total variability leads to two fundamental estimated variances. First, when all observations are sampled from k groups with the same variance (σ_X^2), then an estimate of the common within group variance is

$$S_w^2 = \frac{1}{N-k} SS_w = \frac{1}{N-k} \sum (n_j - 1) S_j^2 \quad j = 1, 2, \ldots, k$$

where S_j^2 represents the variance estimated from each of the k sampled groups based on n_j observations that makeup the jth group (degrees of freedom $= N - k$). In symbols, these k estimated variances are $S_j^2 = \Sigma(x_{ij} - \bar{x}_j)^2/(n_j - 1)$. Thus, the estimate $S_w^2 = \Sigma w_j S_j^2$ is a weighted average of the k estimated within group variances S_j^2 sometimes called *a pooled variance.* The weights are $w = (n_j - 1)/(N - k)$.

From the mean values calculated from each of the k groups (denoted $\bar{x}_j$), the between group variance is estimate by

$$S_b^2 = \frac{1}{k-1} SS_b = \frac{1}{k-1} \sum n_j (\bar{x}_j - \bar{x})^2 \quad j = 1, 2, \ldots, k.$$

The extent of heterogeneity among the k sample mean values is directly measured by the variance estimate S_b^2 (degrees of freedom $= k - 1$). Furthermore, when these k groups consist of independently sampled observations from populations with the same mean value (homogeneous) and the same variance, the two variances S_w^2 and S_b^2 estimate the same quantity, namely, the variance of the observations x_{ij}, denoted σ_X^2. Thus, the ratio of these two estimated variances will likely be close to one when the k population mean values are identical, particularly when the sample size in all groups is large. Therefore, the ratio of the estimated between group variance to the estimated within group variance (S_b^2/S_w^2) reflects the homogeneity/heterogeneity of the observed differences among sample mean values.

As might be suspected, the situation is not quite that simple. More generally, the distribution of the ratio of these two estimates of variability is described by an *f*-distribution when the data are independently sampled from k normal distributions with the same mean value and variance (Chapter 1). An *f*-distribution is defined by two parameters, namely, degrees of freedom of the estimated value in the numerator ($k - 1$) and the degrees of freedom of the estimated value in the denominator ($N - k$). The F-ratio comparison of the between group to within group estimated variances is

$$F = \frac{SS_b/(k-1)}{SS_w/(N-k)} = \frac{S_b^2}{S_w^2}$$

and has an *f*-distribution when the observed x_{ij} values are independently sampled from k populations with the same normal distribution (Chapter 1).

The *f*-distribution produces probabilities reflecting the likelihood of homogeneity among k sample mean values. When the sample mean values randomly differ, the F-ratio statistic certainly differs from a ratio of 1.0, but when they systematically differ, the F-ratio likely substantially differs from a ratio of 1.0. A rigorous statistical evaluation of the equality of between group and within group estimated variances consists of calculating an F-ratio statistic and applying the *f*-distribution to assess the likelihood the observed deviation from ratio of 1.0 occurred by chance alone. This assessment of homogeneity/heterogeneity among k estimated mean values is called a *one-way analysis of variance.*

Table 13.2 *Data: Birth Weights (Kilograms) and Birth Order of* $n = 40$ *Newborn Infants to Illustrate Application of One-Way Analysis of Variance (Small Sample from the University of California, San Francisco Perinatal Database)*

Birth weight	2.48	2.92	3.73	4.16	3.80	3.42	3.62	3.82	3.92	3.34	3.26
Birth order	0	1	3	1	1	2	0	2	2	1	0
Birth weight	3.87	2.92	3.20	4.10	4.06	3.35	3.40	4.48	4.00	3.42	2.76
Birth order	1	1	0	1	0	0	1	0	1	1	0
Birth weight	2.98	4.58	4.23	4.26	3.66	4.13	2.80	3.62	3.40	3.70	3.10
Birth order	0	2	3	0	0	3	0	1	1	0	0
Birth weight	3.56	4.08	2.38	3.74	3.56	3.62	3.52				
Birth order	0	0	1	0	0	0	1				

Table 13.2 *Data (Continued) Summary: Infant Birth Weight Mean Values (Kilograms) Classified into One-Way Table by Birth Order* ($n = 40$)

	Birth order				
Group (k)	1	2	3	4	Summary
Birth order	0	1	2	≥3	–
n_j	19	14	4	3	$N = 40$

A sample of infant birth weights classified into a one-way table by birth order serves to illustrate ($n = 40$ – Table 13.2). To assess the conjecture that birth weight is unrelated to birth order (no difference in mean birth weights – homogeneity), an evaluation of the observed differences among four sample mean values is based on the difference of between and within estimated variances (Tables 13.2 and 13.3). The F-ratio statistic is

$$F = \frac{SS_b/(k-1)}{SS_w/(N-k)} = \frac{S_b^2}{S_w^2} = \frac{1.386/3}{9.123/36} = \frac{0.462}{0.253} = 1.824$$

where $k = 4$ groups and $N = 40$ observations. From an f-distribution with 3 and 36 degrees of freedom, the p-value is $P(F \geq 1.830|\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu) = 0.160$. Thus, the significance probability indicates that the observed variability among the four mean birth weights is not unlikely to have occurred by chance alone. The analysis and, therefore the data, produce no strong evidence that infant birth weights differ among birth order groups (p-value = 0.160).

Table 13.3 *Analysis of Variance: Summary Results from the Analysis of One-Way Classification of Infant Birth Weight by Birth Order (Table 13.2)*

	Analysis of variance table		
	Degrees of freedom	Sums of squares	Estimated variances
Birth order (*between*)	$k - 1 = 3$	$SS_b = 1.386$	$S_b^2 = 0.462$
Variability (*within*)	$N - k = 36$	$SS_w = 9.123$	$S_w^2 = 0.253$
Total	$N - 1 = 39$	$SS_t = 10.509$	–

Table 13.4 *Data: Artificial Probabilities to Illustrate Interpretation of Three Sources of Variability*

Groups (j)	x_j	n_j	$\hat{p}_j$	$n_j\hat{p}_j(1-\hat{p}_j)$
1	1	10	0.1	0.9
2	5	25	0.2	4.0
3	14	35	0.4	8.4
4	24	40	0.6	9.6
5	72	90	0.8	14.4
Total	116	200	0.58	48.7

A chi-square test of variance produces a similar result. For the birth weight example, the approximate chi-square test statistic is

$$X^2 = \frac{\sum n_j(\bar{x}_j - \bar{x})^2}{S_w^2} = \frac{(k-1)S_b^2}{S_w^2} = \frac{3(0.462)}{0.253} = 5.471$$

yielding a p-value of 0.140 (degrees of freedom = 3). Again the statistical assessment of the differences between the k mean values is a comparisons of variances: That is, when estimated variances S_w^2 and S_b^2 are close to equal, the chi-square statistic is close to its mean value (degrees of freedom). In general, the one-way analysis of variance F-ratio statistic and the parallel chi-square test of variance produce similar results, particularly when all sample sizes n_j are large (Chapter 1).

Another Partitioning of Variability

The partition that creates a one-way analysis of variance has a parallel version describing estimated probabilities calculated from k groups. Again, the expression for the total variability separates into within group and between group components. Specifically, for samples consisting of n_j values used to estimate probabilities $\hat{p}_j$ from each of k groups. The partition of the total variation (*total* = *within* + *between*) is

$$N\hat{p}(1-\hat{p}) = \Sigma n_j\hat{p}_j(1-\hat{p}_j) + \Sigma n_j(\hat{p}_j - \hat{p})^2 \quad \text{again, } N = \Sigma n_j \text{ and } j = 1, 2, \ldots, k.$$

The value $\hat{p}$ represents the estimated probability ignoring the classification of the observations into groups and is estimated directly the total from N observations.

To illustrate, consider artificial probabilities estimated from five groups (Table 13.4).

The total variability $N\hat{p}(1-\hat{p}) = 48.72$ divides into within group variability $\sum n_j\hat{p}_j(1-\hat{p}_j) = 37.30$ and between group variability $\Sigma n_j(\hat{p}_j - \hat{p})^2 = 11.42$ calculated from the $N = \Sigma n_j = 200$ observations. This description of variability among estimated probabilities does not differ in principle from the previous analysis of variance partition applied to sample mean values.

These example data illustrate an important insight into the interpretation of summary values. The summary value $\hat{p}$ ignoring group membership is $\hat{p} = \Sigma x_j/N = 116/200 = 0.58$. A weighted average calculation of the probability $\hat{p}$ accounting for group membership

yields the identical value $\hat{p} = \Sigma n_j \hat{p}_j / N = 0.58$. The total variation of $\hat{p}$ ignoring group membership is 48.72 and within group variation accounting for the group membership is 37.30. Thus, ignoring differences among the groups produces a larger value than the value calculated accounting for differences among the groups. The partition of the total variation shows that the difference is due to the between group variation included in the estimate of the total variation. Specifically, the difference between the two estimates of variation (*total versus within*) is due to differences among the group estimated values $\hat{p}_j$. For the example probabilities, this measure of variation is $\Sigma n_j(\hat{p}_j - \hat{p})^2 = 11.42$.

The difference between the total variation and within group variation has important implications because an estimate of a single probability from a sample of data implicitly assumes a homogeneous population with respect to the probability estimated. The estimate is calculated and interpreted as if no between group heterogeneity exists. In reality, only in rather special situations are probabilities among all members of a sampled population random deviations from a single value. For example, a probability estimated from a sex-, ethnicity-, age-specific population creates an estimate that more closely represents the situation where a single value within the sampled population exists but nonrandom heterogeneity undoubtedly remains. The influence of this usually unmeasured, and, therefore, unavoidable heterogeneity increases the estimated variability associated with a single estimate causing lengths of confidence intervals to increase and values of test statistics to decrease making analytic results conservative. In other words, the usually necessary assumption that the value estimated is a single constant value, when in fact it is not, produces understated statistical results. That is, the bias incurred by not completely accounting for heterogeneity among groups within the values sampled increases the estimated variance, which reduces the likelihood that nonrandom differences will be detected.

Two-Sample Test of Variance

A number of statistical methods exist to compare specific properties of two samples of data. The most well known is Student's *t*-test (Chapter 1) assessment of the observed difference between two estimated mean values. In addition, the log-rank test (Chapter 17) is a two-sample assessment of survival data, and the Wilcoxon (Mann-Whitney) is a nonparametric rank test (Chapter 8) designed to compare data sampled from two populations. The two-sample comparisons of estimated variance follows in the same vain.

A variety of techniques are designed to compare two estimated sample variances (denoted S_1^2 and S_2^2) to evaluate possible differences between two population variances (denoted σ_1^2 and σ_2^2). Four of these methods are the following:

1. Ratio of variances – *F*-ratio test
2. Differences in the logarithms of variances – Bartlett's test
3. Analysis of mean variability – Levene's test and
4. A nonparametric rank comparison – Siegel-Tukey rank test.

Data from a coronary heart disease study consisting of $n = 40$ high-risk men who weight more than 225 pounds classified by type-A and type-B behavior types and their cholesterol levels illustrate (Table 13.5).

Table 13.5 *Data: Cholesterol Levels of* 40 *Men at High Risk of Coronary Heart Disease Who Weight More than* 225 *Pounds Classified by A/B-Behavior Type and Ranked to Compare Variability (Chapter 8)*

Cholesterol	137	148	153	169	175	181	183	185	188	194	197
Behavior	B	B	B	B	B	A	B	B	B	B	A
Ranks	1	4	5	8	9	12	13	16	17	20	21
Cholesterol	202	202	212	212	213	218	224	224	226	233	234
Behavior	A	B	B	A	B	A	A	B	B	A	A
Ranks	24	25	28	29	32	33	36	37	40	39	38
Cholesterol	239	239	242	246	246	248	250	250	252	252	254
Behavior	A	A	B	B	A	A	B	A	A	B	A
Ranks	35	34	31	30	27	26	23	22	19	18	15
Cholesterol	263	268	276	291	312	325	344				
Behavior	B	A	A	A	A	A	B				
Ranks	14	11	10	7	6	3	2				

Table 13.5 *(Continued) Summary: Cholesterol Data for Two Behavior Types (B-Type and A-Type)*

	Behavior types		
	A-type	B-type	Summaries
Sample size	$n_1 = 20$	$n_2 = 20$	$n = 40$
Mean value	$\bar{x}_1 = 245.1$	$\bar{x}_2 = 210.3$	$\bar{x} = 227.7$
Variance	$S_1^2 = 1342.4$	$S_2^2 = 2336.7$	$S^2 = 2102.0$

F-Ratio Test of Variance

Like the one-way analysis of variance, a ratio of two estimated variances addresses the question of the equality of two variances, denoted σ_1^2 and σ_2^2. Formally, the test statistic is

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}.$$

As before, the F-ratio test statistic has an f-distribution when both samples of observations have normal distributions with same variance ($\sigma_1^2 = \sigma_2^2 = \sigma^2$). Applied to the behavior type data (Table 13.5), the F-ratio test statistic is

$$F = \frac{S_2^2}{S_1^2} = \frac{2336.7}{1342.4} = 1.741.$$

Under the conjecture the variances are equal and the cholesterol measures have normal distributions, the test statistic F has an f-distribution with degrees for freedom $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$. Thus, for the cholesterol data, the p-value is $P(F \geq 1.741 | \sigma_1^2 = \sigma_2^2) = 0.118$ (two-sided: p-value $= 0.236$) based on degrees of freedom $df_1 = 19$ and $df_2 = 19$ providing little persuasive evidence of a systematic difference in variability of cholesterol levels between behavior type groups.

For normally distributed data, the fact that the *F*-ratio has an *f*-distribution allows a direct calculation of a 95% confidence interval $(\hat{A}, \hat{B})$ based on the ratio of two estimated variances. From an *f*-distribution,

$$P\left[F_{df_1,df_2,0.025} \leq F \leq F_{df_1,df_2,0.975}\right] = P\left[F_{df_1,df_2,0.025} \leq \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \leq F_{df_1,df_1,0.975}\right] = 0.95$$

where $F_{df_1,df_2,q}$ represents a q-level quantile value from an *f*-distribution with degrees of freedom df_1 and df_2. After a straight forward rearrangement, then

$$95\% \text{ confidence interval} = P\left[\frac{1}{F_{df_1,df_2,0.975}}\frac{S_1^2}{S_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{1}{F_{df_1,df_2,0.025}}\frac{S_1^2}{S_2^2}\right] = 0.95$$

produces a 95% confidence interval for the ratio of two variances (σ_1^2/σ_2^2) based on the ratio of estimates S_1^2 and S_2^2 (S_1^2/S_2^2).

For the behavior type data, $q = 0.025$ and $q = 0.975$ quantile values from the *f*-distribution are $F_{19,19,0.025} = 0.396$ and $F_{19,19,0.975} = 2.526$.

The confidence interval bounds $\hat{A} = 1.741/2.526 = 0.689$ and $\hat{B} = 1.741/0.396 = 4.398$ create a 95% confidence interval of $(\hat{A}, \hat{B}) = (0.689, 4.398)$, based on the estimated *F*-ratio = 1.741.

The statistical term *robust* applies. In a statistical context robust means that estimates or analytic techniques are resilient to failure of the underlying statistical assumptions (Chapter 15). The *t*-test comparison of two mean values is an example of a robust statistical technique. Thus, analytic results remain relevant and useful when the data have approximate normal distributions and the variances are not extremely unequal. An *f*-ratio comparison of two sample variances is not robust. When the data are not normally distributed, the *f*-ratio test is not likely credible. Statistician R. Miller states that the failure could be "catastrophic."

When the normality of the sampled data is suspect, the distribution of an *F*-ratio–like statistic can be estimated with bootstrap sampling, again $F = S_1^2/S_2^2$. For the cholesterol data, using 5000 replicate bootstrap samples produces an estimate of the distribution of the ratio of two estimated variances regardless of the underlying distributions of the sampled data (Figure 13.1).

From the bootstrap-estimated distribution, the estimated *F*-ratio–like statistic contrasting the sample variances from the cholesterol data is $F = 1.766$ producing a data-estimated and assumption-free one-sided *p*-value of $P(F \geq 1.766|\sigma_1^2 = \sigma_2^2) = 850/5000 = 0.170$ (two-sided = *p*-value = 0.340).

As usual, a bootstrap-estimated distribution produces a 95% confidence interval directly from the 2.5th percentile and the 97.5th percentile values (Figure 13.1). For the cholesterol data, the nonparametric 95% confidence interval is (0.608, 4.772).

Bartlett's Test of Variance

Bartlett's test of variance is theoretically complicated but simply understood and applied. The essence of the approach is again a comparison of two estimated variances. The conjecture that two sample variances come from populations with the same variance $(\sigma_1^2 = \sigma_2^2)$ is compared

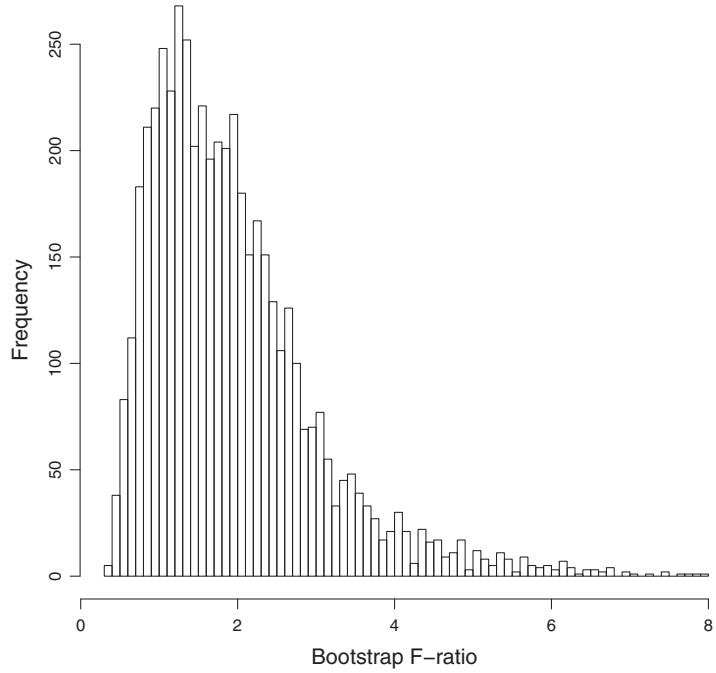


Figure 13.1 Bootstrap-Estimated Distribution for the F-Ratio–Like Test Statistic from Cholesterol/Behavior Data

to the conjecture that sampled populations have different variances ($\sigma_1^2 \neq \sigma_2^2$). Bartlett's approximate chi-square distributed test statistic to asses these two possibilities is

$$X^2 = (N-2)log\left(S_w^2\right) - \left[(n_1-1)log\left(S_1^2\right) + (n_2-1)log\left(S_2^2\right)\right]$$

for $N = n_1 + n_2$ total observations. The estimate of a single common variance from both samples (said to be pooled) is again a weighted average of within group estimated variances [weights $= (n_j - 1)/(N-2)$], specifically

$$S_w^2 = \frac{1}{N-2}\left[(n_1-1)S_1^2 + (n_2-1)S_2^2\right].$$

The estimated variances S_1^2 and S_2^2 are the usual estimates calculated separately from each group. Notice that when these estimated variances are identical ($S_1^2 = S_2^2 = S_w^2$) Bartlett's test statistic is $X^2 = 0$. To the extent that S_1^2 and S_2^2 differ, the test statistic X^2 measures the difference and has a chi-square distribution with one of freedom when the underlying population variances are equal ($\sigma_1^2 = \sigma_2^2$). Therefore, a chi-square statistic produces the probability that the estimated variances differ by chance alone. As before, the data from both groups are required to have at least an approximate normal distributions. Like the F-test, the accuracy of Bartlett's test is sensitive to this requirement.

For the cholesterol data, the separate estimated variances are $S_1^2 = 1342.4$, $S_2^2 = 2336.7$ and the single pooled estimated variance is $S_w^2 = 1839.6$. Bartlett's chi-square test statistic is $X^2 = 1.441$ with one degree of freedom producing the p-value $P(X^2 \geq 1.441 | \sigma_1^2 = \sigma_2^2) = 0.230$.

Table 13.6 *Analysis of Variance Results: Levene's Test of Variance Applied to Compare Variability of Cholesterol Levels between Type-A or Type-B Individuals ($n = 40$ Observations and $k = 2$ Groups)*

Analysis of variance table			
	Degrees of freedom	Sums of squares	Estimated variances
Between	1	$SS_b = 950.62$	$S_b^2 = 950.62$
Within	$N - 2 = 38$	$SS_w = 27928.50$	$S_w^2 = 734.96$
Total	$N - 1 = 39$	$SS_t = 28879.12$	–

Levene's Test of Variance

Levene's test of variance is theoretically simple and routinely applied. A one-way analysis of variance is typically used to assess variability among mean values calculated from k samples. Levene suggests a comparison among mean values that reflects differences in variability using a one-way analysis of variance. Specifically, a one-way analysis of variance is applied to the transformed observations $y_{ij} = \mid x_{ij} - \bar{x}_j \mid$ constructed from each observation and used to evaluate observed differences in mean variability (denoted $\bar{y}_1$ and $\bar{y}_2$), one mean value estimated from each group.

For the example behavior type data, when the y_{ij}-values are the deviations constructed from the 40 observations (Table 13.5), an analysis of variance applied to these "data" produces a statistical evaluation of the difference in variability between the two behavior types (type-A and type-B – Table 13.6). The two mean values are $\bar{y}_A = 36.90$ and $\bar{y}_B = 27.15$. Thus, the value $\bar{y}_A - \bar{y}_B$ reflects the mean difference in variability. The analysis of variance F-ratio comparison estimated from two measures of variability based on comparison of the two mean values of the transformed y_{ij}-values is

$$F = \frac{S_b^2}{S_w^2} = \frac{950.62}{734.96} = 1.293.$$

The value of the test statistic F has an approximate f-distribution with degrees of freedom one and $n_1 + n_2 - 2 = 38$ when the variability of the cholesterol levels between behavior groups differs by chance alone. The p-value is $P(F \geq 1.293 | \sigma_1^2 = \sigma_2^2) = 0.263$. In less formal language, no evidence again exists of a systematic difference in variability of cholesterol levels between type-A and type-B individuals as measured by the difference in mean values $\bar{y}_A$ and $\bar{y}_B$. A two-sample t-test produces the identical result ($t = 1.137$ and $F = 1.293 = (1.137)^2$) (Chapter 1).

A special feature of Levene's approach is that the statistical requirements underlying the classic analysis of variance do not hold. The constructed measures of variability (y_{ij}) are not normally distributed and are not independent. Investigations by a number of statisticians concluded that failure to fulfill the formal requirements of a one-way analysis of variance is usually not a substantial influence and the results from Levene's test of variance based on comparing mean variability typically remain useful. In short, Levene's test statistic is remarkably robust.

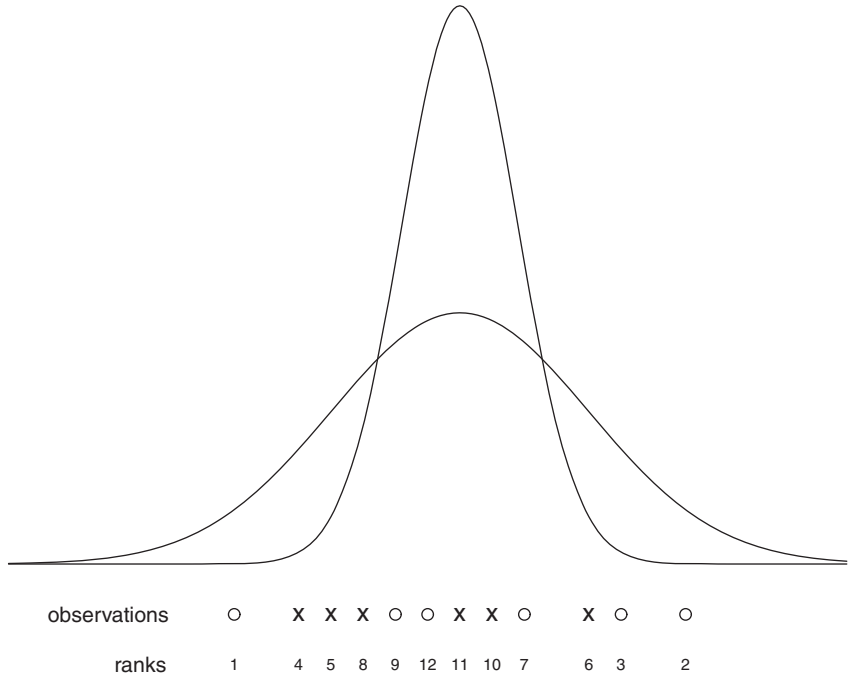


Figure 13.2 Illustration of Siegel-Tukey Pattern of Ranking Observations for Comparison of Variances between Two Groups (x = Group 1 and ○ = Group 2)

Siegel-Tukey Two-Sample Test of Variance

The Siegel-Tukey comparison of two sample variances is a nonparametric (rank) procedure. It differs from the Wilcoxon (Mann-Whitney) two-sample rank sum test primarily in the pattern of assigning ranks. The Wilcoxon pattern is designed to detect differences in mean values between two compared groups (Chapter 8). The Siegel-Tukey pattern is designed to detect differences in variability between two compared groups. Otherwise, the mechanics of calculating and evaluating the test summary statistic are the same.

The Siegel-Tukey pattern of assigning ranks starts with combining n_1 and n_2 sampled observations into a single sample and ranking the observations without regard to group membership. The smallest observation is ranked 1, the largest and next largest observations are ranked 2 and 3, then the second and third smallest observations are ranked 4 and 5, and so on for all sampled observations. This "back and forth" pattern is displayed in Figure 13.2. The specific ranking of the behavior/cholesterol data is included in Table 13.5. When the two distributions compared have the same mean value but different variances, the smaller ranks are likely associated with the distribution with the larger variance (Figure 13.2).

Because ranks are randomly distributed between two groups when the observations are themselves randomly distributed between the two groups (no difference in variability), the mechanics of Siegel-Tukey nonparametric evaluation is not different from the Wilcoxon two-sample rank test. The sum of the ranks (denoted W), therefore, calculated from one group randomly deviates from the sum of the ranks expected when no difference in variability exists between the groups (denoted W). Specifically, for the group labeled 1, when the

samples compared have the same mean value and variance, then the expected value of W is

$$W = \frac{1}{2}n_1(N+1) \quad \text{with} \quad variance\,(W) = \frac{1}{12}[n_1 n_2(N+1)]$$

where again n_1 and n_2 represent the number of observations in each of the compared samples and $N = n_1 + n_2$ represents the total sample size (Chapter 8).

For the cholesterol data (Table 13.5), the observed sum of the ranks is $W_1 = 447$ from the $n_1 = 20$ type-A individuals with associated $variance\,(W) = 20(20)(41)/12 = 1366.67$. The expected sum of the ranks when no difference in variability exists between behavior groups is $W = 20(41)/2 = 410$ (Chapter 8). The comparison of these two values produces the approximate chi-square distributed test statistic with one degree of freedom

$$X^2 = \left[\frac{W_1 - W}{\sqrt{variance(W)}}\right]^2 = \left[\frac{447 - 410}{\sqrt{1366.67}}\right]^2 = 1.002$$

that yields a p-value of $P(W \geq 447|\sigma_1^2 = \sigma_2^2) = P(X^2 \geq 1.002|\sigma_1^2 = \sigma_2^2) = 0.317$. The result is potentially biased by the possibility that the mean values of the two groups compared are not equal.

It should be noted that the Siegel-Tukey test statistic is nonparametric, but it is not assumption-free. The test requires the assumption or knowledge that the mean values of the two compared distributions are the same. When the mean values differ, a comparison can be made between samples with identical mean values by subtracting the respective estimated mean values of the compared samples from each observation making the mean values in both groups zero. However, the test is then no longer distribution-free.

Comparison of Several Variances

Bartlett's and Levene's tests of variance are directly extended to a comparison of more than two estimated variances. The expression for Bartlett's test statistic becomes

$$X^2 = (N-k)\,log\left(S_w^2\right) - \Sigma(n_j - 1)log\left(S_j^2\right) \quad j = 1, 2, \ldots, k$$

where again n_j represents the number of observations in the jth group and $N = \Sigma n_j$ represents the total number of normally distributed observations. As before, the within estimated variance (pooled) is the weighted average

$$S_w^2 = \frac{1}{N-k}\sum (n_j - 1)S_j^2 \quad j = 1, 2, \ldots, k$$

and the estimated sample variances from each of the k groups are, as usual,

$$S_j^2 = \frac{1}{n_i - 1}\sum (x_{ij} - \bar{x}_i)^2 \quad i = 1, 2, \ldots, n_j.$$

In addition, the accuracy of Bartlett's test statistic X^2 is improved by an adjustment labeled C where

$$C = 1 + \frac{1}{3(k-1)}\left[\sum \frac{1}{n_j - 1} - \frac{1}{N-k}\right] \quad j = 1, 2, \ldots, k.$$

Table 13.7 *Summary* ($N = 750$ *and* $k = 5$)*: Maternal Weight (Kilograms) Retained Six Months after Delivery Classified into Five Weight Categories* (*Quintiles* – $n_j = 150$)

	Weight gain categories					
Quintiles	0–16.7	16.7–24.4	24.4–30.6	30.6–34.2	>34.2	Total
n_j	150	150	150	150	150	750
$\bar{x}_j$	3.556	3.435	3.538	3.222	3.477	3.506
S_j^2	0.234	0.294	0.239	0.236	0.263	0.254

Adjustment of the chi-square test statistic $X_c^2 = X^2/C$ is important only when one or more of the sample sizes n_j are small.

Maternal weight retained six months after delivery classified into five categories of weight gained during pregnancy illustrates Bartlett's test of variance applied to compare variability among more than two groups (Table 13.7).

The resulting Bartlett's chi-square test statistic $X^2 = 2.873$ (degrees of freedom $= k - 1 = 4$) produces a p-value of $P(X^2 \geq 2.873|\textit{no difference}) = 0.579$ indicating little evidence of unequal variability among the five maternal weight gain categories. The adjusted chi-square test statistic is $X_c^2 = 2.875$, and, as expected, adjustment has only a slight influence when samples sizes n_j are large for all groups.

Like Bartlett's test, Levene's test of variance simply extends to any number of groups. A one-way analysis of variance is directly applied to measure variability by again creating the transformed variables $y_{ij} = |x_{ij} - \bar{x}_j|$ for all observations. The k mean values $\bar{y}_j$, one from each group, as before, reflect differences in variability and are assessed with the usual one-way analysis of variance (homogeneity/heterogeneity). This approach is illustrated again with the weight gain data (Tables 13.7 and 13.8). The mean values $\bar{y}_j$ are

$$\bar{y}_1 = 0.369, \bar{y}_2 = 0.412, \bar{y}_3 = 0.388, \bar{y}_4 = 0.378 \text{ and } \bar{y}_5 = 0.396.$$

summarizing the variability among the five groups.

The one-way analysis of variance F-ratio

$$F = \frac{S_b^2}{S_w^2} = \frac{0.042}{0.101} = 0.414$$

Table 13.8 *Analysis of Variance: Variability of Weight Retained during Pregnancy among Five Maternal Weight Gained Categories* ($N = 750$ *and* $k = 5$)

	Analysis of variance		
	Degrees of freedom	Sums of squares	Estimated variances
Weight gain	$k - 1 = 4$	$SS_b = 0.168$	$S_b^2 = 0.042$
Variability	$N - k = 745$	$SS_w = 74.096$	$S_w^2 = 0.101$
Total	$N - 1 = 749$	$SS_t = 74.264$	

has an approximate *f*-distribution when the five weight gain categories have the same variance. The *p*-value from an *f*-distribution with 4 and 745 degrees of freedom is $P(F \geq 0.414|\sigma_1^2 = \cdots = \sigma_5^2) = 0.799$. Levene's test of variance also produces little evidence of an influence from maternal weight gain on the variability of weight retention; that is, the mean values measuring variability among the five groups show no evidence of systematic differences (Table 13.7).

14

The Log-Normal Distribution

The log-normal distribution has been referred to as the "Cinderella" distribution because its older "sister" is the dominant overbearing normal probability distribution. More technically, the logarithm of a log-normally distributed value has a normal distribution. A normal distribution is symmetric, and its mean value and variance are unrelated. Changes in the mean value shift the location but do not influence its shape (Chapter 1). Similarly, changes in the variance change the shape but do not shift its location (mean value). Figure 14.1 displays four normal distributions that differ in mean values and variances.

When the variable X has a normal distribution, the value $Y = e^X$ has a log-normal distribution. The resulting distribution of Y is no longer symmetric, and the mean value and variance are no longer unrelated. Figure 14.2 displays the four log-normal distributions corresponding to the normal distributions in Figure 14.1. Unlike the normal distribution, changes in the mean value produce changes in both location and shape. Similarly, changes in the variance also do not have simple consequences. Of primary importance is the basic fact that natural and everyday properties of the normal distribution dictate the less intuitive and more complex properties of the log-normal distribution.

Table 14.1 contains details of the relationship between the parameters of these two distributions. The relationships originate from the fact that for variable Y with a log-normal distribution, the variable $X = log(Y)$ has a normal distribution with mean value denoted by μ_x and variance by σ_x^2. For example, the median value of a log-normal distribution is e^{μ_x} and follows directly from a normal distribution because

$$\text{median value} = P(X \geq \mu_x) = 0.5, \quad \text{then } P(e^X \geq e^{\mu_x}) = P(Y \geq e^{\mu_x}) = 0.5$$

where again μ_x represents the mean of a normal distribution. The justifications of the three other relationships are not as simple (Table 14.1). Nevertheless, they depend entirely on the normal distribution parameters μ_x and σ_x^2.

The most fundamental relationship between these two probability distributions involves cumulative log-normal probabilities $P(Y)$ and their corresponding cumulative normal distribution probabilities $P(Z)$. Specifically, the relationship is

$$\begin{aligned} P(Y) &= P(Y \leq c) = P(log[Y] \leq log[c]) = P(X \leq log[c]) \\ &= P\left(Z \leq \frac{log[c] - \mu_x}{\sigma_x}\right) = P(Z), \end{aligned}$$

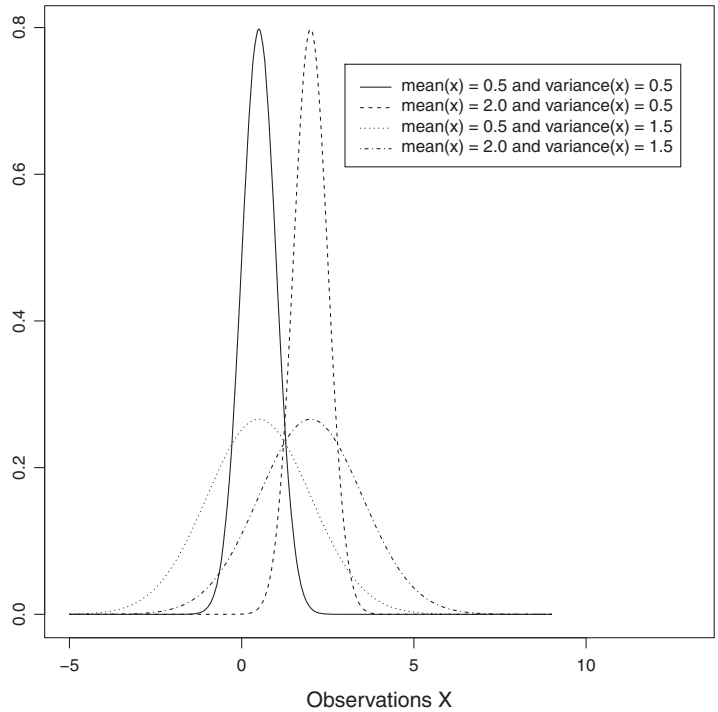


Figure 14.1 Four Normal Distributions

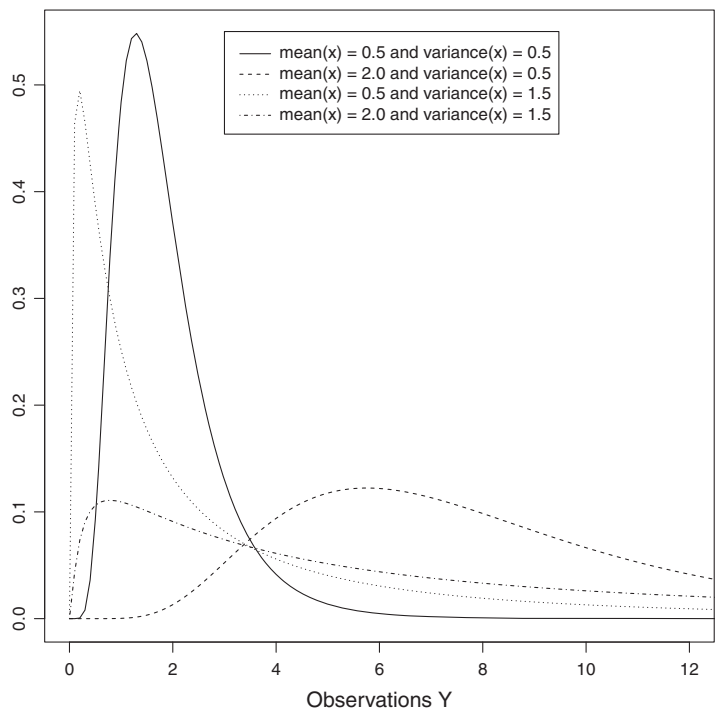


Figure 14.2 Four Log-Normal Distributions

Table 14.1 *Summary Values: Relationships between Median Values, Mean Values, Variances, and Modes of Normal and Log-Normal Distributions*

	Normal	Log-normal
Median	μ_x	e^{μ_x}
Mean	μ_x	$e^{\mu_x+\frac{1}{2}\sigma_x^2}$
Variance	σ_x^2	$(e^{2\mu_x+\sigma_x^2})(e^{\sigma_x^2}-1)$
Mode	μ_x	$e^{\mu_x-\sigma_x}$

where Z has a standard normal distribution ($\mu = 0$ and $\sigma = 1$). This relationship is displayed in Figure 14.3 for $\mu_x = 1.0,\ \sigma_x^2 = 0.7$, and $c = 4.48$. Then

$$P(Y \geq 4.48) = P(log\,(Y) \geq log\,[4.48]) = P(X \geq 1.5) = P(Z \geq 0.598) = 0.275,$$

where $z = (log[c] - \mu_x)/\sigma_x = (1.5 - 1.0)/\sqrt{0.7} = 0.598$, is calculated from a standard normal distribution.

For the log-normal distribution, the mean value is greater than the median value, and the median value is greater than the mode. Because the variances σ_x^2 is always positive, then necessarily mean > median > mode (Table 14.1).

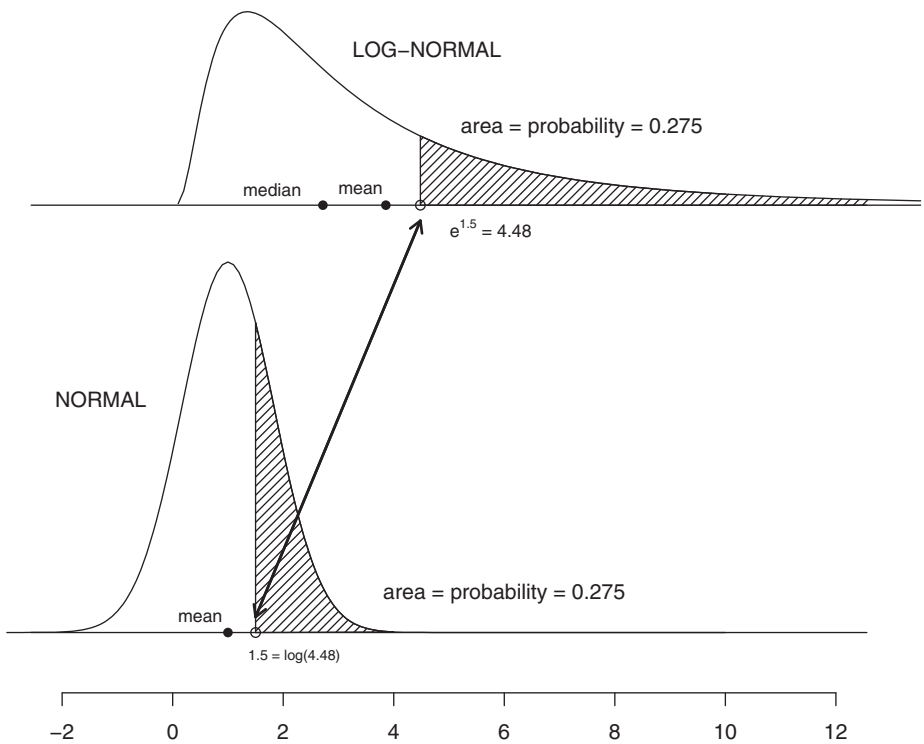


Figure 14.3 Relationship between Log-Normal Distribution and Normal Distribution Probabilities

Table 14.2 *Data: Randomly Generated Observations from a Log-Normal Distribution (Parameters: Mean* $= \mu_x = 0.5$*, Variance* $= \sigma_x^2 = 1.0$ *and* $n = 30$*)*

	Log-normal distributed data (Y)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Y	2.69	1.11	2.75	1.11	8.49	3.07	2.02	5.00	1.34	1.13	1.22	1.74	0.25	1.59	16.63
X^a	0.99	0.10	1.01	0.10	2.14	1.12	0.70	1.61	0.29	0.12	0.20	0.55	−1.38	0.47	2.81
	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Y	4.36	4.33	0.96	3.23	2.72	0.22	2.07	0.75	5.89	6.94	2.53	0.29	1.61	0.37	0.96
X^a	1.47	1.46	−0.04	1.17	1.00	−1.53	0.73	−0.28	1.77	1.94	0.93	−1.24	0.47	−0.99	−0.04

[a] $X = log(Y)$.

A fundamental feature of the relationship between log-normal and normal distributions is that estimates of the parameters and probabilities of a log-normal distribution are functions of the usual estimates of the mean value and variance from a normal distribution. For example, the mean value and variance of the log-normal variable Y are estimated from the mean value and variance estimated from the normally distributed value $X = log(Y)$. Therefore, to characterize the log-normal distribution all that is needed are two maximum likelihood estimated values from the normal distribution, namely,

$$\text{estimated mean value of } X = \hat{\mu}_x = \frac{1}{n}\sum x_i = \frac{1}{n}\sum log(y_i)$$

and

$$\text{estimated variance of } X = \hat{\sigma}_x^2 = \frac{1}{n}\sum (x_i - \hat{\mu}_x)^2 = \frac{1}{n}\sum (log(y_i) - \hat{\mu}_x)^2,$$

where $i = 1, 2, \ldots, n$ = number of sampled observations. Functions of these estimates are also maximum likelihood estimates, making the estimated log-normal summary values maximum likelihood estimates (Chapter 27). As is sometimes the case with maximum likelihood estimates, these estimated summary values can be biased, sometimes substantially biased, particularly when the variance σ_x^2 is large. Rather complicated bias corrections exist and, alternatively, other kinds of unbiased estimates are available.

Example: Lognormal Distributed Data

Consider artificial data created specifically to illustrate estimates of the parameters and describe the properties of a log-normal distribution (Table 14.2). A simulated sample of $n = 30$ random values from a log-normal distribution with mean value of $\mu_x = 0.5$ and variance of $\sigma_x^2 = 1.0$ illustrates. To simulate "data" with a log-normal distribution, when z is a random value from a standard normal distribution, then $y = e^{\mu_x + z\sigma_x}$ is a random value with a log-normal distribution with parameters μ_x and σ_x^2 (Table 14.2 – $\mu_x = 0.5$ and $\sigma_x^2 = 1.0$).

As noted, the estimates of the mean value and variance from a normal distribution ($\hat{\mu}_x$ and $\hat{\sigma}_x^2$) allow estimation of the log-normal distribution and its properties as illustrated (Figure 14.4, solid line) (Chapter 12).

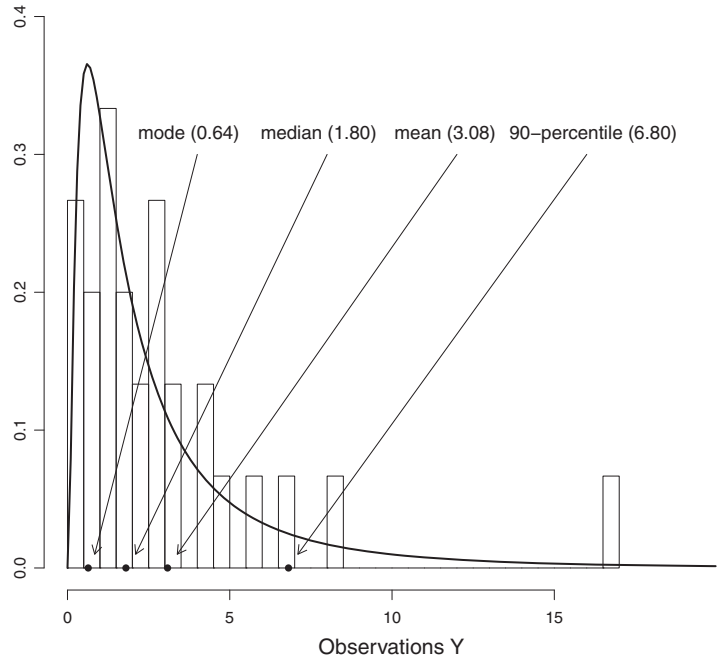


Figure 14.4 Log-Normal Distribution Estimated (Solid Line) from Simulated Data (Table 14.2 – $n = 30$)

Estimated parameters					
Normal distribution			Log-normal distribution		
Mean	Variance	Mean	Variance	Median	Mode
$\hat{\mu}_x$	$\hat{\sigma}_x^2$	$e^{\hat{\mu}_x + \frac{1}{2}\hat{\sigma}_x^2}$	$e^{2\hat{\mu}_x + \hat{\sigma}_x^2} \times (e^{\hat{\sigma}_x^2} - 1)$	$e^{\hat{\mu}_x}$	$e^{\hat{\mu}_x - \hat{\sigma}_x}$
0.589	1.073	3.083	18.286	1.803	0.640

In addition, these same data produce estimates of a variety of other descriptive summary statistics of a log-normal distribution, such as quantiles/percentiles. For the example data, an estimate of the 90th percentile is

$$\hat{P}_{90} = e^{\hat{\mu}_x + 1.282\hat{\sigma}_x} = e^{0.589 + 1.282(1.036)} = e^{1.917} = 6.799$$

where 1.282 is the 90th percentile of the standard normal distribution. Confidence intervals follow the same pattern where, again from the data in Table 14.2, the 95% the confidence interval for the median value of a log-normal distribution is

$$\text{lower bound} = e^{\hat{\mu}_x - 1.960\hat{\sigma}_x} = e^{0.589 - 1.960(0.189)} = e^{0.219} = 1.244 \text{ and}$$
$$\text{upper bound} = e^{\hat{\mu}_x + 1.960\hat{\sigma}_x} = e^{0.589 + 1.960(0.189)} = e^{0.960} = 2.612$$

yielding a 95% confidence interval of (1.244, 2.612) based on the estimated median value $e^{\hat{\mu}_x} = e^{0.589} = 1.803$.

The properties of a ratio of two log-normally distributed values, similar to the properties of a log-normal distribution, directly relate to the properties of two normal distributions. If Y_1 represents a log-normally distributed value with parameters μ_1 and σ_1^2 and Y_2 represents an independent log-normally distributed value with parameters μ_2 and σ_2^2, then the distribution of the estimated ratio $\hat{R} = Y_1/Y_2$ has a log-normal distribution with median value $e^{\mu_1-\mu_2}$. For the estimated ratio $\hat{R} = Y_1/Y_2$, then $log(\hat{R}) = log(Y_1) - log(Y_2) = X_1 - X_2$ has a normal distribution with mean value $\mu_1 - \mu_2$ and $variance(log[\hat{R}]) = \sigma_1^2 + \sigma_2^2$.

The median of a log-normal distribution is also called the geometric mean. In symbols, for a sample $\{y_1, y_2, \ldots, yn\}$,

$$\text{median} = \text{estimated geometric mean} = e^{\hat{\mu}_x} = e^{\frac{1}{n}\sum x_i} = e^{\frac{1}{n}\sum log(y_i)} = \left[\prod y_i\right]^{\frac{1}{n}}.$$

A Left-Censored Log-Normal Distribution

An issue occasionally arises in sampling values from log-normal distributions when, for one reason or another, values are not measured. For example, they could be expensive to collect or they could be relatively unimportant or both. Such data occur in the analysis of toxic materials when the focus is primarily on high levels of a potentially harmful substance. Observations below a specified value, referred to as the limit of detection (denoted *ld*), are sometimes not measured, causing the data values to be *left-censored* (Chapters 16 and 17). That is, the number of occurrences of the unmeasured values is known, but their actual values are not measured. Estimation of log-normal distribution parameters are then biased unless the estimation accounts for the presence of censored values (Chapter 16). A number of statistical strategies exist. Sometimes unmeasured values that lie below the limit of detection are simply replaced by constant values such as $ld/2$ or $ld/\sqrt{2}$. In other situations, specialized maximum likelihood estimates are used or logistic regression techniques applied.

To explore the issues that arise from a left-censored log-normal distribution, a direct and natural approach is presented that is easily implemented and provides an intuitive estimate of the mean value and variance not biased by unmeasured values that lie below the limit of detection.

Figure 14.5 (upper left) displays a log-normal distribution with unmeasured values below the limit of detection, left-censored (small values). Naturally the corresponding normal distribution is similarly left-censored (upper right). Thus, unmeasured log-normally distributed values below the limit of detection *ld* are also unmeasured normally distributed values below the logarithm of the limit of detection, specifically the value $log(ld)$. Normal distributions are symmetric and, therefore, provide an opportunity to use values from the upper tail of the sampled distribution to create "measured" values below the limit of detection to estimate the "missing" lower tail. All observed values above $2\hat{m} - log(ld)$, where $\hat{m}$ represents the estimated median value of the normal distribution, are used to create corresponding values in the lower tail below the limit of detection. Specifically, each observation x_i greater than $2\hat{m} - log(ld)$ produces a theoretical value $x_i' = 2\hat{m} - log(y_i) = 2\hat{m} - x_i$ that symmetrically corresponds to a value below $log(ld)$ (Figure 14.5, lower left). The "completed" data then consist of n measured observations and n_0 left-tail "observations" that are exactly symmetric copies of the upper tail observations (mirror images). These "data" then allow the estimation of the parameters μ_x and σ_x^2 as if the censored values were measured (Figure 14.5, lower

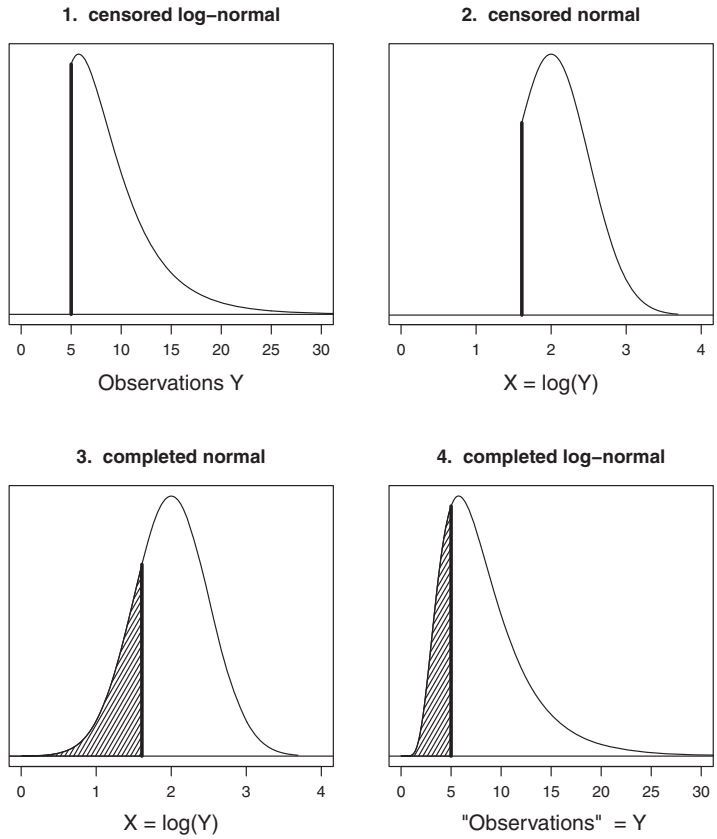


Figure 14.5 Illustration of Estimation from a Left-Censored Log-Normal Distribution

right). Using the "complete" normal distribution, the estimated mean value $\hat{\mu}_x$ and variance $\hat{\sigma}_x^2$ are calculated in the usual manner. The process, therefore, accounts for unmeasured values below the limit of detection and produces estimates from the collected data as if all values were measured. Estimation is no longer biased by the left-censored observations. The estimates, however, are not fully efficient (smallest possible variance), and the limit of detection must be below the median value. Nevertheless, the estimates $\hat{\mu}_x$ and $\hat{\sigma}_x^2$ estimated from the "complete" data allow a simple and unbiased description of the sampled log-normal distribution parallel to the estimation when the data are complete.

An Applied Example

In a case/control study of childhood leukemia interest focused on a herbicide called Dacthal as a possible contributor to disease risk. From $n = 230$ children participating in a study of the Dacthal exposure, the collected data contained 76 low-level values reported but not measured, left-censored values. A mean value based directly on $n = 154$ measured levels of this herbicide is biased (too large) caused by the absent unmeasured small values.

Assuming that the distribution of the herbicide exposure levels is accurately described by a log-normal distribution, the logarithms of the 154 observed values, denoted $x_i = log(y_i)$, have

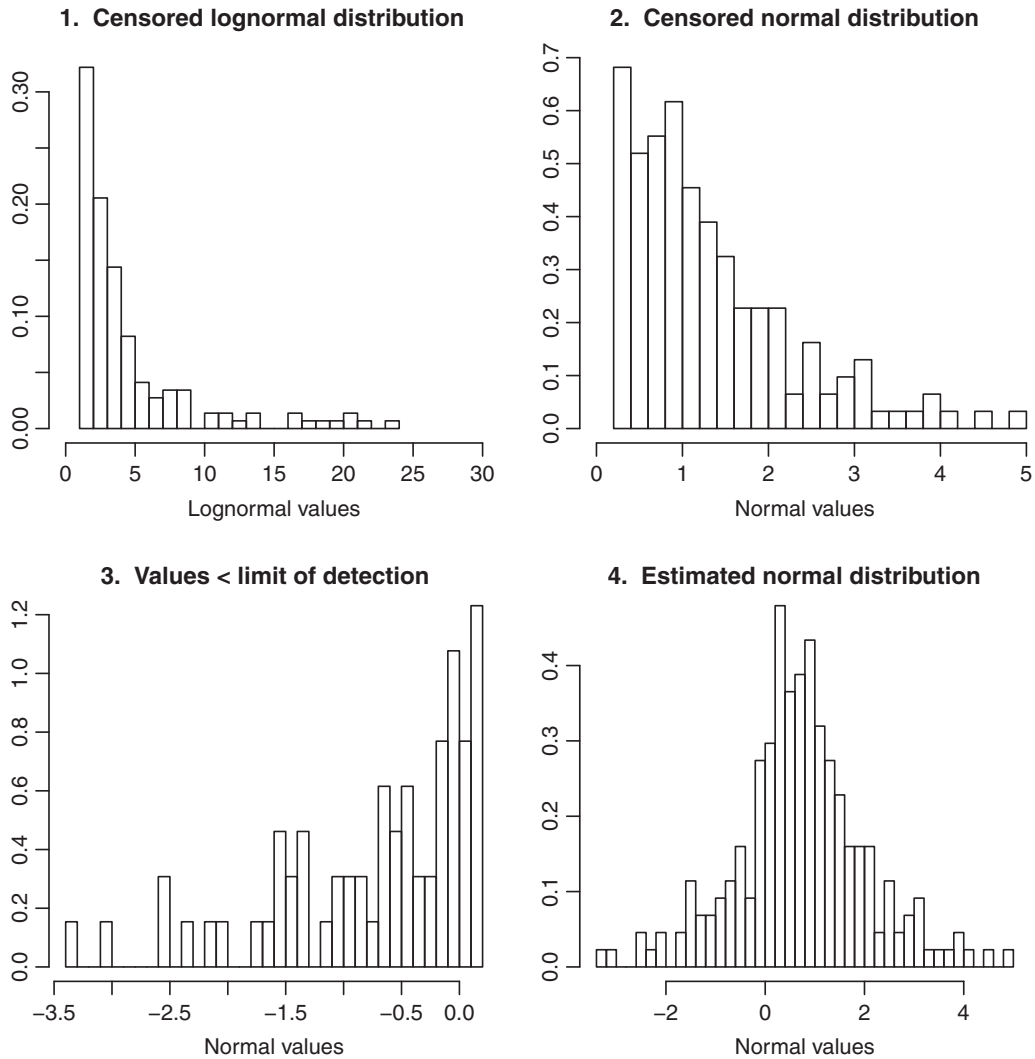


Figure 14.6 The Four Steps in the Estimation of the Log-Normal Distribution of the Herbicide Dacthal

a normal distribution. Then, using the upper tail of this normal distribution to estimate values for the censored lower tail values produces a "complete" normal distribution of herbicide measurements. That is, the $n_0 = 80$ estimates of lower tail "observations" based on 80 upper tail values plus the $n = 154$ original measured observations create a total of $N = n + n_0 = 234$ values; that is, the n_0 values in the upper tail produce perfectly symmetric values that provide an estimate of the "missing" lower tail. The mean value and variance are then estimated in the usual fashion using the N values of "complete" data. These unbiased estimates are the usual,

$$\text{estimated mean value} = \textit{mean}\,(x_i) = \hat{\mu}_x = \frac{1}{N}\sum x_i \text{ and}$$

$$\text{estimated variance} = \textit{variance}\,(x_i) = \hat{\sigma}_x^2 = \frac{1}{N-1}\sum (x_i - \hat{\mu}_x)^2.$$

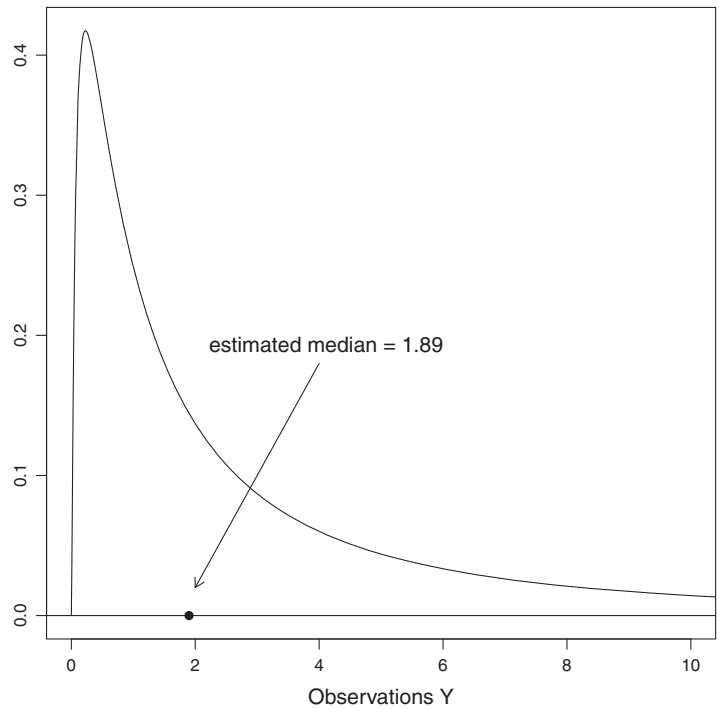


Figure 14.7 Log-Normal Distribution Estimated from Left-Censored Dacthal Exposure Data (Sample "Size" $N = 234$)

Figure 14.6 displays the four steps in the estimation of the log-normal distribution from the Dacthal pesticide data. The estimated mean value is $\hat{\mu}_x = 0.638$, and the estimated variance is $\hat{\sigma}_x^2 = 2.117$ from the "complete" normal distribution. The estimated mean value and variance then yield estimates of the parameters of the log-normal distribution. Specifically, for the log-normal distribution, the estimated mean value and the variance are

$$\text{mean of distribution of } Y = e^{\hat{\mu}_x + \frac{1}{2}\hat{\sigma}_x^2} = e^{0.638 + \frac{1}{2}(2.117)} = 5.455$$

and

$$\begin{aligned}\text{variance of distribution of } Y &= e^{2\hat{\mu}_x + \hat{\sigma}_x^2}(e^{\hat{\sigma}_x^2} - 1) = \left(e^{[2(0.638)+2.117]}\right)(e^{2.117} - 1) \\ &= 217.396.\end{aligned}$$

The estimated median (geometric mean) value is

$$\text{median of a log-normal distribution of } Y = e^{\hat{\mu}x} = e^{0.638} = 1.893.$$

Figure 14.6 graphically displays the estimated log-normal distribution of Dacthal levels based on the assumption that the sampled data have a log-normal distribution (Figure 14.7). Of course, computer software is available to calculate the parallel estimates with more sophisticated and optimum maximum likelihood methods. Nevertheless, the less efficient estimation leads to a simple and intuitive understanding of the issues of accounting for the influences from censored observations.

15

Nonparametric Analysis

Consider two estimated mean values (denoted $\bar{x}_0$ and $\bar{x}_1$) and the mean difference $\bar{x}_0 - \bar{x}_1$ that becomes the focus of analysis. An often effective strategy to explore the differences between two mean values is called a *matched pairs design*. For each individual sampled from one population an individual is purposely chosen from second population who matches a variable or variables of the member sampled from the first population. The collection of these pairs constitute a *matched pairs* data set. For example, the classic "nature/nurture" matched pairs are identical twins (matched for genetics) raised in different families and compared for specific traits. Measures on the same subjects before and after a treatment are another example of matched pairs of observations (Chapter 1). It is said that these subjects serve as their own "controls." More usually, individuals sampled from two different sources such as cases with a specific property and controls without the specific property are purposely matched for specific variables such as sex, ethnicity, and age.

Two fundamental properties of the matched pair strategy are the following:

> Each pair is identical for a variable or a set of variables, and any observed difference within a matched pair, therefore, cannot be attributed to the *matched variables*. They are the same for both members of pair. Thus, the analysis is not influenced by these intentionally matched variables, sometimes called confounder or nuisance variables. The data collection strategy guarantees that the mean values of the matched variables are identical in the two compared samples and, therefore, have no influence on the mean difference $\bar{x}_0 - \bar{x}_1$.

Differences are easier to identify when observations compared are similar. In statistical terms, similar means variability associated with a comparison is reduced. Reduced variability leads to an increased likelihood of systematic differences becoming apparent.

Reduction in variability due to matching is illustrated by the expression for variance of the difference $\bar{x}_0 - \bar{x}_1$ from two groups with the same variance (denoted σ^2) and a correlation between members within each pair denoted ρ (Chapter 5). Specifically, an expression for the variance of the mean difference between n matched pairs is (Chapter 27)

$$variance(\bar{x}_0 - \bar{x}_1) = \frac{\sigma_0^2}{n} + \frac{\sigma_1^2}{n} - \frac{2\sigma_{01}}{n} = \frac{\sigma_0^2}{n} + \frac{\sigma_1^2}{n} - \frac{2\rho\sigma_0\sigma_1}{n},$$

and when $\sigma_0^2 = \sigma_1^2 = \sigma^2$ making the $covariance(x_0, x_1) = \sigma_{01} = \sigma\rho^2$, then

$$variance(\bar{x}_0 - \bar{x}_1) = \frac{2\sigma^2(1-\rho)}{n} = \frac{2\sigma^2}{\left[\frac{n}{1-\rho}\right]}.$$

Table 15.1 *Data (Chapter 11): Pairs of before and after Measurements of Thiocyanate Levels from Extremely Heavy Cigarette Smokers Participating in Intensive Intervention Trial (Subset of $n = 20$ Subjects)*

	Thiocyanate levels									
	1	2	3	4	5	6	7	8	9	10
Before	131	154	45	89	39	103	127	129	148	181
After	111	13	16	35	23	115	87	120	122	228
Differences	20	141	29	54	14	−11	40	9	26	−47
Signs	+	+	+	+	+	−	+	+	+	−
Ranks	8	20	11	18	7	6	15	4	10	17
Signed ranks	8	20	11	18	7	−6	15	4	10	−17
	11	12	13	14	15	16	17	18	19	20
Before	122	120	66	122	115	113	163	67	156	191
After	118	121	20	153	139	103	124	37	163	246
Differences	4	−1	46	−31	−24	10	39	30	−7	−55
Signs	+	−	+	−	−	+	+	+	−	−
Ranks	2	1	16	13	9	5	14	12	3	19
Signed ranks	2	−1	16	−13	−9	5	14	12	−3	−19

The expression $n/(1 - \rho)$ determines the *effective sample size*. For correlation $\rho = 0$, the effective size is n, when $\rho = 0.5$, the effective sample size increases to $2n$, and when $\rho = 0.8$, the effective sample size becomes $5n$. Thus, a matching strategy acts as if additional "data" were collected, and the increase depends on the within-pair correlation.

Costs are associated with a matched pair strategy. Briefly, a few are: the influences of the matched variables are not available for study, difficulties arise in finding matches for exceptional observations, and subjects chosen to match are not likely be representative of any specific population. In addition, as the association between a matching variable and the variable under study increases, the amount of usable data decreases, producing a phenomenon called *overmatching*. That is, the sample size decreases because sampled pairs become more likely "matched" for the variable under study. An extreme example occurs in a study of air pollution as a possible contributor to the risk of asthma. If an affected child (case) is matched to a neighborhood child (control), air quality, the variable of interest, would likely be the same for both case and control.

An often used analytic strategy is to transform sampled data into a more tractable form for analysis. One of the simplest transformations is the differences within matched pairs. Matched pair differences then produce a direct and simple statistical analysis. The transformed values become the differences $d_i = x_{0i} - x_{1i}$ created from each pair and are summarized by the single mean value

$$\bar{d} = \frac{1}{n}\sum d_i = \frac{1}{n}\sum (x_{0i} - x_{1i}) = \bar{x}_0 - \bar{x}_1$$

for $i = 1, 2, \ldots, n$ = number of matched pairs ($2n$ observations). For the thiocyanate measurements from the intervention trial data (Table 15.1), the mean value of $n = 20$

within-pair differences is

$$\bar{d} = 119.050 - 104.700 = 14.350.$$

Note that the estimated correlation between before and after thiocyanate measurements $r = 0.784$ produces an effective sample size of $n/(1 - r) = 20/(1 - 0.784) = 92$ from the 20 original pairs.

A matched pair analysis then becomes a comparison of the estimated mean value $\bar{d}$ to zero using Student's t-test. For the smoking intervention trial data, the t-test statistic is

$$T = \frac{\bar{d} - 0}{S_{\bar{D}}} = \frac{14.350 - 0}{\sqrt{89.728}} = 1.515.$$

The variance of the mean value is directly estimated from the n differences d_i in the usual way and is, for the example data, $S^2_{\bar{D}} = S^2_D/n = 1794.6/20 = 89.728$. Furthermore, variables such as the subject's age, sex, age began smoking, and socioeconomic status are a few of many matched variables that do not influence the before and after comparisons among the trial participants. The test statistic T then has a t-distribution with $n - 1 = 19$ degrees of freedom when only random differences exist within pairs. The p-value is $P(T \geq 1.515 \mid \textit{no difference within pairs}) = 0.073$. A t-test requires the values observed from each member of the x_0/x_1-pair to be sampled from normal distributions with the same variance. For small samples ($n < 30$), these two properties are virtually impossible to verify statistically.

When requirements and assumptions are fulfilled for a parametric analysis, a parallel nonparametric analysis addressing the same issues rarely produces substantially different results. When the results differ, the parametric assumptions become suspect. Thus, applying both kinds of analytic approaches to the same data is frequently a constructive strategy. Statistician Elizabeth Scott frequently said, "Compared to the difficulty and expense of collecting data, analysis is easy and inexpensive."

The Sign Test

The *sign test* is a nonparametric alternative to the paired t-test. Again the differences within pairs are used to construct a test statistic. As the name indicates, the sign test is an assessment of the pattern of the signs observed among the matched pair differences. The magnitude of the within-pair difference is ignored. Therefore, counts of the positive or negative differences indicate the plausibility that only random variation causes the observed differences within pairs. Key to assessment is the binomial distribution with parameters $p = 0.5$ and sample size n. A sign test analysis consists of simply calculating a p-value directly from the data using a binomial distribution (Chapter 4).

The smoking intervention trial data yield 13 positive differences or a proportion of $\hat{p} = 13/20 = 0.65$ from the before and after measurements among $n = 20$ intervention trial participants (Table 15.1). The statistical question is: What is the probability that the observed proportion of positive differences differs from 0.5 by chance alone? This question translates into comparing the data estimate $\hat{p} = 0.65$ to the theoretical probability $p_0 = 0.5$, the value expected when only random differences occur within pairs. Relevant probabilities from the binomial distribution with parameters $p_0 = 0.5$ and $n = 20$ are the following:

	Binomial probabilities[a]							
Counts	10	11	12	13	14	15	16	≥17
$p = \text{count}/n$	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85
Binomial probabilities (p)	0.176	0.160	0.120	0.075	0.037	0.015	0.001	0.000
$P(\hat{p} \geq p \mid \textit{random})$	0.588	0.412	0.252	0.132	0.058	0.021	0.006	0.001

[a]Right-tail cumulative probabilities of the binomial distribution (Chapter 4).

The sign test yields the exact p-value of $P(\hat{p} \geq 0.65 \mid p_0 = 0.5) = 0.132$ calculated directly from the binomial distribution ($n = 20$ and $p_0 = 0.5$). Like most nonparametric techniques, the sign test yields exact p-values for small sample sizes.

For sample sizes greater than 20 or so, a normal distribution–derived approximation to the binomial distribution is an often used alternative. The test statistic

$$z = \frac{\hat{p} - p_0 - \frac{1}{2}(1/n)}{\sqrt{p_0(1-p_0)/n)}} = \frac{\hat{p} - 0.5 - \frac{1}{2}(1/n)}{0.5\sqrt{1/n}}$$

has an approximate standard normal distribution when no systematic differences exist within the matched pairs ($p_0 = 0.5$). The test statistic from the smoking intervention trial data becomes

$$z = \frac{0.65 - 0.5 - \frac{1}{2}(0.05)}{0.5\sqrt{0.05}} = 1.118.$$

The approximate p-value becomes $P(\hat{p} \geq 0.65 \mid \textit{no systematic difference}) = P(Z \geq 1.118) = 0.132$ (Chapter 6).

A sometimes important question is: How likely is a specific nonrandom difference detected? For the sign test, the normal distribution approximation provides an answer. A formal accept/reject statistical test starts with establishing a critical point c such as $P(\hat{p} \geq c \mid p_0 = 0.5) = 0.05$. Based on the value c, a value $\hat{p}$ less that c leads to the inference that no persuasive evidence exists of a systematic influence (likely random). For the sign test, the value of c is

$$c = p_0 + z_{0.95}\sqrt{p_0(1-p_0)/n} = 0.5 + 1.645(0.5)\sqrt{1/n}.$$

Thus, an observed value of $\hat{p}$ greater than c leads to the inference that systematic influences likely cause at least some within-pairs differences (likely not random). For the smoking intervention trial data, the value c is 0.684 because $P(\hat{p} \geq 0.684 \mid p_0 = 0.5) = 0.05$ ($n = 20$). An inference based on c that a nonrandom difference exists when, in fact, no systematic influences are present will be wrong with probability 0.05 (Table 15.2).

Then, for a specified probability denoted p and a sign test constructed from n observations, the question of the likelihood that the sampled data produce an estimate $\hat{p}$ greater than c is answered by the expression

$$P(\hat{p} \geq c|p) = P(Z \geq z) \quad \text{where } z = \frac{c - p}{0.5\sqrt{1/n}}.$$

Table 15.2 *Computations: The Approximate Probability That a Sign Test Identifies a Nonrandom within Pair Difference Based on the Estimate* $\hat{p}$ *When* $p = \{0.5, 0.6, 0.7, 0.8, \text{and } 0.9\}$ *for Sample Sizes* $n = 20, 30, 40,$ *and* 50

	Values of p				
c	0.5	0.6	0.7	0.8	0.9
$P(\hat{p} \geq c \mid n = 20)$	0.684	0.05	0.23	0.56	0.85	0.97
$P(\hat{p} \geq c \mid n = 30)$	0.650	0.05	0.29	0.71	0.95	1.00
$P(\hat{p} \geq c \mid n = 40)$	0.630	0.05	0.35	0.81	0.98	1.00
$P(\hat{p} \geq c \mid n = 50)$	0.616	0.05	0.41	0.88	1.00	1.00

Note: $P(\hat{p} \geq c \mid p_0 = 0.5) = 0.05$ for all sample sizes.

For the smoking intervention trial data, for example, when $p = 0.6$, then again

$$c = 0.5 + 1.645(0.5)\sqrt{1/20} = 0.684 \quad \text{and} \quad z = \frac{0.684 - 0.6}{0.5\sqrt{1/20}} = 0.751$$

yields the probability $P(\hat{p} \geq 0.684 \mid p = 0.6) = P(Z \geq 0.751) = 0.23$. Thus, the probability is approximately 0.23 that an observed estimate $\hat{p}$ produces a test statistic that leads to correctly identifying a nonrandom difference based on $n = 20$ observations when in fact $p = 0.6$. Estimates for other values of p and sample sizes n follow the same pattern. The probability $P(\hat{p} \geq c \mid p)$ is typically called *the power of the sign test* (Table 15.2). It is the probability that the analysis identifies a nonrandom difference when a nonrandom difference exists.

A historical note: John Graunt (1662) in his early study of human data was the first to note an excess of male over female births based on 139,728 male and 130,782 female records of christenings in London during the period 1629–1661. In addition, a colleague, John Arbuthnott, observed that each year male outnumbered female christenings for 82 consecutive years. He calculated the probability of this occurrence by chance alone as 0.5^{82} and concluded that likelihood of only positive differences for 82 years was so improbable that chance could not be the explanation. This was likely the first formal use of a sign test. Also remarkable is the use of a statistical test/inference 250 years before this fundamental statistical concept was generally recognized.

The Wilcoxon Signed Rank Tests

The matched pairs *t*-test is designed for analysis of data as recorded, and the sign test uses only the sign of within-pair differences. An intermediate strategy is the *Wilcoxon signed rank test*. The absolute values of within-pair differences are ranked from 1 to n (the number of observed pairs). The largest differences are replaced with the largest ranks and the smallest differences with the smallest ranks. These ranked values are then given the sign of the original difference. The *signed ranks* reflect the magnitude of each within-pair difference. The test statistic (denoted W^+) is the sum of the signed ranks with a positive sign. The sum of the negative signed ranks could be equally used because the sum of the positive and negative

Table 15.3 *Example: Illustration of the Wilcoxon Signed Rank Test – Four Matched Pairs (Differences, Ranks of Absolute Values, Signs, and Signed Ranks)*

	Pairs			Statistics		
	x_0	x_1	Differences	Ranks	Signs	Signed ranks
1	17	10	7	4	+	4
2	3	8	−5	3	−	−3
3	4	6	−2	1	−	−1
4	5	1	4	2	+	2

ranks is $n(n+1)/2$ (Chapters 5, 8, and 27). The ranks become the "data" analyzed making the technique nonparametric.

A small example illustrates the details with artificial data consisting of four matched pairs (Table 15.3).

When the differences within pairs are due only to random variation, the 2^n possible sums of the signed ranks occur each with equal probability. For the example, the $2^4 = 16$ outcomes of the test statistic W^+ each occurs with the same probability of $1/16$ when no systematic differences exist within pairs (Table 15.4). From another point of view, each signed rank is included in the sum of positive ranks with probability p or is not included with probability $1 - p$. If only random differences exist within pairs, then $p = p_0 = 0.5$. The complete probability distribution associated with the test statistic W^+ for $n = 4$ pairs directly follows from 16 equally likely outcomes (Table 15.5 – $p = p_0 = 0.5$).

Table 15.4 *Example: The* $2^4 = 16$ *Equally Likely Sums of Signed Ranks of the Wilcoxon Signed Rank Test Based on* $n = 4$ *Matched Pairs (Table 15.2)*

	Sums of signed ranks					
	1	2	3	4	W^+	W^-
1	0	0	0	0	0	10
2	0	0	0	1	4	6
3	0	0	1	0	3	7
4	0	0	1	1	7	3
5	0	1	0	0	2	8
6	0	1	0	1	6	4
7	0	1	1	0	5	5
8	0	1	1	1	9	1
9	1	0	0	0	1	9
10	1	0	0	1	5	5
11	1	0	1	0	4	6
12	1	0	1	1	8	2
13	1	1	0	0	3	7
14	1	1	0	1	7	3
15	1	1	1	0	6	4
16	1	1	1	1	10	0

Table 15.5 *Example (Continued): Complete Distribution of the Wilcoxon Signed Rank Test for n = 4 When Only Random Differences Occur within Pairs – Probabilities and Cumulative[a] Probabilities*

Sum of ranks W^+											
0	1	2	3	4	5	6	7	8	9	10	Total
0.0625	0.0625	0.0625	0.1250	0.1250	0.1250	0.1250	0.1250	0.0625	0.0625	0.0625	1.00
1.0000	0.9375	0.8750	0.8125	0.6875	0.5625	0.4375	0.3125	0.1875	0.1250	0.0625	–

[a]Right-tail exact cumulative probability $= P(W \geq W^+ \mid p_0 = 0.5)$.

For the example, when the sum of the positive ranks is $W^+ = 6$, then $P(W^+ \geq 6 \mid p_0 = 0.5) = 7/16 = 0.4375$ (Table 15.5). Note that this test is exact (no approximations) for a sample size of four.

As with many nonparametric rank-based tests, a normal distribution provides a useful approximation of the exact probability distribution for moderately large sample sizes. The signed rank test follows the pattern of the sign test with both an exact and normal distribution approximation–derived *p*-values. For the matched pairs Wilcoxon signed rank test, the expected value and variance of the test statistic W^+ are

$$E\,W^+ = \frac{n(n+1)}{4} \quad \text{and} \quad variance(W^+) = \frac{n(n+1)(2n+1)}{24}$$

for n randomly differing matched pairs and are the elements that create an approximate normal distribution test statistic (Chapter 27).

The corresponding exact and normal distribution approximate probabilities from the distribution of W^+ are calculated for $n = 10$ matched pairs for select values (denoted $\tilde{P}$ – Table 15.6). The approximate probabilities from the normal distribution are calculated in

Table 15.6 *Approximation: Exact (P) and Normal Approximation ($\tilde{P}$) Probabilities[a] from Distribution of the Wilcoxon Signed Rank Test for n = 10 Matched Pairs*

W^+	Exact P	Approximate $\tilde{P}$
12	0.947	0.954
16	0.884	0.899
20	0.784	0.807
24	0.652	0.677
28	0.500	0.520
32	0.348	0.361
36	0.216	0.222
40	0.116	0.121
44	0.053	0.057
48	0.019	0.023
52	0.005	0.008

[a]P and $\tilde{P}$ are right-tail cumulative probabilities and $W^+ =$ sum of positive signed ranks.

routine fashion. These probabilities are $\tilde{P} = P(W \geq W^+ \mid \textit{random differences}) = P(Z \geq z)$ where

$$z = \frac{W^+ - n(n+1)/4}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

has an approximate standard normal distribution when no systematic differences occur within pairs. The exact probabilities (denoted P – Table 15.6) associated with the sum of positive ranks are found in specialized tables or computer calculated.

The smoking intervention trial provides an example of the signed rank test applied to actual data (Table 15.1). The signed ranks are displayed in the last line of Table 15.1. The sum of the positive signed ranks is

$$W^+ = 8 + 20 + 11 + 18 + 7 + 15 + 4 + 10 + 2 + 16 + 5 + 14 + 12 = 142$$

and the sum of the negative signed ranks is

$$W^- = 6 + 17 + 1 + 13 + 9 + 3 + 19 = 68.$$

As required, the sum $W^+ + W^- = n(n+1)/2 = 20(21)/2 = 210$ for the $n = 20$ pairs of before and after differences. The exact computer calculated p-value is $P(W^+ \geq 142 \mid \textit{random difference only}) = 0.088$.

A large sample normal distribution approximation allows an accurate estimated p-value, as noted. For the before and after smoking trial data ($n = 20$), the signed rank test statistic expected value is

$$E\,W^+ = \frac{n(n+1)}{4} = \frac{20(21)}{4} = 105,$$

and the variance of the distribution of test statistic W^+ is

$$\textit{variance}(W^+) = \frac{n(n+1)(2n+1)}{24} = \frac{20(21)41}{24} = 717.5.$$

When all within-pair differences are due only to random variation, then

$$z = \frac{W^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} = \frac{142 - 105}{\sqrt{717.5}} = 1.381$$

has an approximate standard normal distribution yielding an approximate p-value $= P(W \geq 142 \mid \textit{random differences only}) = P(Z \geq 1.381) = 0.0836$ (computer exact $= 0.884$).

Kruskal-Wallis Nonparametric Comparison of *k* Sample Mean Values

A one-way analysis of variance is a parametric statistical technique that provides an evaluation of the differences (variation) observed among k mean values (Chapter 13). When observations denoted y_{ij} classified into k groups of size n_j have normal distributions with the same variance, a one-way analysis of variance yields an f-distributed test statistic that reflects the plausibility that the underlying k population mean values are equal (homogeneous). The fundamental feature of the analytic approach is a partitioning of the total variation yielding a comparison of two meaningful components, one measuring the within and the other measuring between group variation (Chapters 13 and 27).

Table 15.7 *Data: Weight Gained (Kilograms) During Pregnancy for Four Ethnic Groups (Small Sample from a Large Database – N = 44)*

White														
Gain	2.24	1.59	0.82	9.82	8.55	7.04	5.40	3.73	5.44	6.58	7.43	8.14	8.76	2.88
Ranks	12	8	6	40	35	28	22	16	23	26	30	33	37	14
African American														
Gain	4.51	3.05	1.69	0.57	0.00	7.95	8.13	8.27						
Ranks	20	15	9	4	1	31	32	34						
Asians														
Gain	9.62	8.55	7.14	6.23	5.04	3.84	2.76	4.47						
Ranks	38	36	29	25	21	17	13	19						
Hispanics														
Gain	11.91	11.52	11.08	10.50	9.75	2.08	1.09	0.30	0.01	0.58	2.10	3.99	5.62	6.79
Ranks	44	43	42	41	39	10	7	3	2	5	11	18	24	27

Weight gained during pregnancy from four ethnic groups illustrates this partitioning of variability to explore the likelihood that the estimated mean values differ only randomly among four compared groups. These data are a small subset of $N = \Sigma\, n_j = 44$ mothers from the University of California, San Francisco Perinatal Database (Table 15.7). The mean weight gained summarized for white, African American, Asian, and Hispanic mothers ranges from 4.27 to 5.99 kilograms (Table 15.8).

The results of a one-way analysis of variance applied to these 44 mothers from four groups are presented in Table 15.9 (Chapter 13). The *f*-test statistic is $F = 0.357$ calculated under the parametric conjecture that the data are normally distributed in all four groups with same mean value and variance (Table 15.8). The resulting *p*-value of 0.785 provides essentially no evidence of systematic differences among the four mean values.

The Kruskal-Wallis nonparametric approach is also an analysis of a one-way classification of *k* mean values and is mechanically the same as the parametric analysis

$$\sum\sum(y_{ij} - \bar{y})^2 = \sum\sum(y_{ij} - \bar{y}_j)^2 + \sum n_j(\bar{y}_j - \bar{y})^2$$
$$\text{for } i = 1, 2, \ldots, n_j \text{ and } j = 1, 2, \ldots, k.$$

Table 15.8 *Results: Summary Means and Variances from Observed Weight Gain (Kilograms) and Ethnicity Data*

	White	African American	Asian	Hispanic	Summaries
Number of observation	14	8	8	14	44
Estimated mean value[a]	5.601	4.271	5.990	5.523	5.41
Estimated variance	8.501	12.081	5.725	21.505	12.29

[a]Mean values $\bar{y}_j$ each based on n_j observed values ($j = 1, 2, 3$, and 4).

Table 15.9 *Results: Analysis Variance Table for Ethnic/Gain Data (Parametric Analysis)*

	Sum of squares	*df*	Mean squares	*F*	*p*-value
Between	$B = 13.8$	3	4.59	0.356	0.785
Within	$W = 514.7$	40	12.87	–	–
Total	$T = 528.5$	43	12.29	–	–

Note: total sum of squares (T) = *within sum of squares* (W) + *between sum of squares* (B).

The observations are replaced with their corresponding ranks. In symbols, data values denoted y_{ij} become ranks or $r_{ij} = rank(y_{ij})$. The observations ranked from 1 to sample size N without regard to the group membership are treated as "data." The ranks for the maternal weight gain data are included in Table 15.6. The identical analysis of variance calculations measure the within group and between group variation but based on ranked values (Tables 15.10 and 15.11).

Because the "data" are ranks, the following simplifications occur:

$$\bar{r} = (N+1)/2 = 22.5,\ T = \sum\sum (r_{ij} - \bar{r})^2 = N(N+1)(N-1)/12 = 7095$$

and

$$variance(\bar{r}) = (N+1)/12 = 3.75.$$

The between sum of squares (denoted B – Table 15.11) measures the differences among the k mean values of ranked observations from the four ethnic groups ($\bar{r}_j$); that is, the statistic

$$B = \sum n_j(\bar{r}_j - \bar{r})^2 \quad j = 1, 2, \ldots, k = \text{number of groups}$$

effectively measures the variation among the k mean values of the ranked observations calculated from each of k groups (homogeneity/heterogeneity – Table 15.10). The Kruskal-Wallis test statistic becomes $X^2 = (N-1)B/T$ and has a chi-square distribution (degrees of freedom $= k - 1$) when the mean ranked values differ only because of random variation among the k groups. Like all rank procedures, the properties of distributions sampled are not relevant to the results of the analysis.

Table 15.10 *Results: Summary Mean Values and Variances of Ranked Weight Gain and Ethnicity Data* ($N = 44$)

	White	African American	Asians	Hispanics	Summaries
Number of observation	14	8	8	14	44
Mean value[a]	23.61	18.25	24.69	22.57	22.5
Variance	121.93	171.36	79.21	273.49	165

[a]Mean ranks $\bar{r}_j$ each based on n_j ranked values ($j = 1, 2, 3$, and 4).

Table 15.11 *Results: Analysis of Variance Table for Ethnic/Gain Ranks (Nonparametric Analysis)*

	Sum of squares	*df*	Mean squares
Between	$B = 200.01$	3	67.67
Within	$W = 6894.99$	40	172.36
Total	$T = 7095$	43	165

Note: Again, $\sum\sum(r_{ij} - \bar{r})^2 = \sum\sum(r_{ij} - \bar{r}_j)^2 + \sum n_j(\bar{r}_j - \bar{r})^2$ or $T = W + B$.

Specifically, for the weight gain data, the nonparametric test statistic based on $N = 44$ ranked observations yields between sum of squares $B = 200.013$ and total sum of squares of $T = 7095$ (Table 15.11). The test statistic becomes $X^2 = 43(200.013)/7095 = 1.212$. The value X^2 has an approximate chi-square distribution (degrees of freedom $= k - 1 = 3$) when only random differences occur among the four mean values $\bar{r}_j$. The associated p-value is $P(X^2 \geq 1.212 \mid \textit{random differences}) = 0.750$. Like the parametric analysis, the nonparametric analysis produces no evidence of systematic differences in mean weight gain among the four compared ethnic groups.

One last note: The Wilcoxon two-sample nonparametric analysis (Chapter 8) is identical to the Kruskal-Wallis analysis of variance when samples from two groups are compared ($k = 2$).

Three-Group Regression Analysis

An extremely simple nonparametric approach to an analysis based on the regression model $y = a + bx$ provides an estimated summary straight line resistant to possible disruptive influences caused by extreme or outlier values. Resistant in this regression context means extreme or nonlinear observations have minimal influence on the estimation of the summary regression line, particularly the estimated slope (denoted $\hat{b}$). The method is said to be *robust* (Chapter 13). The cost is a decrease in precision of the estimated slope and intercept relative to the corresponding ordinary least squares estimates (Chapters 3 and 5). Occasionally situations arise where it is useful to trade increased accuracy (reduced bias) at the cost of decreased precision.

The three-group estimate of a straight line calculated to summarize a sample of x/y-pairs starts with dividing the ordered observations into three groups made as close as possible to equal in size. Based on the x-values, three median values are calculated, and based on the y-values, three median values are calculated. These median values form three pairs denoted (X_L, Y_L), (X_M, Y_M), and (X_R, Y_R) and are displayed in Figure 15.1 (dots) using a sample of blood pressure and body weight data ($n = 21$ – Table 15.12). The left-hand median value (L) and the right-hand median value (R) become the basis for the estimated slope of the summary line. Using median values provides considerable resistance to the influences of extreme values, which is particularly important when the sample size is small. Specifically, a nonparametric expression for this robust estimate of the slope is

$$\hat{B} = \frac{Y_R - Y_L}{X_R - X_L}.$$

Table 15.12 *Data: Systolic Blood Pressure Measurements and Body Weights of Subjects at High Risk for Coronary Heart Disease with Cholesterol Levels That Exceed 350 ($n = 21$ x/y Pairs)*

Subjects	1	2	3	4	5	6	7	8	9	10	11	12	13
Weight (x)	133	196	175	205	165	145	198	120	151	180	155	210	191
Blood pressure (y)	108	120	128	130	120	104	136	134	110	137	120	154	125
Subjects	14	15	16	17	18	19	20	21					
Weight (x)	146	164	170	157	162	160	135	148					
Blood pressure (y)	146	130	148	126	136	127	120	128					

Two points are required to establish a straight line. An estimate of the intercept a is usually chosen as the second point. The slope b of any straight line is

$$b = \frac{\text{change in } y}{\text{change in } x} = \frac{y_i - y_j}{x_i - x_j}.$$

For the special case where $y_i = \bar{y}$, $y_j = a$, $x_i = \bar{x}$, and $x_j = 0$, then, for an intercept represented by a, the expression

$$b = \frac{\bar{y} - a}{\bar{x} - 0} \text{ yields the value } a = \bar{y} - b\bar{x}.$$

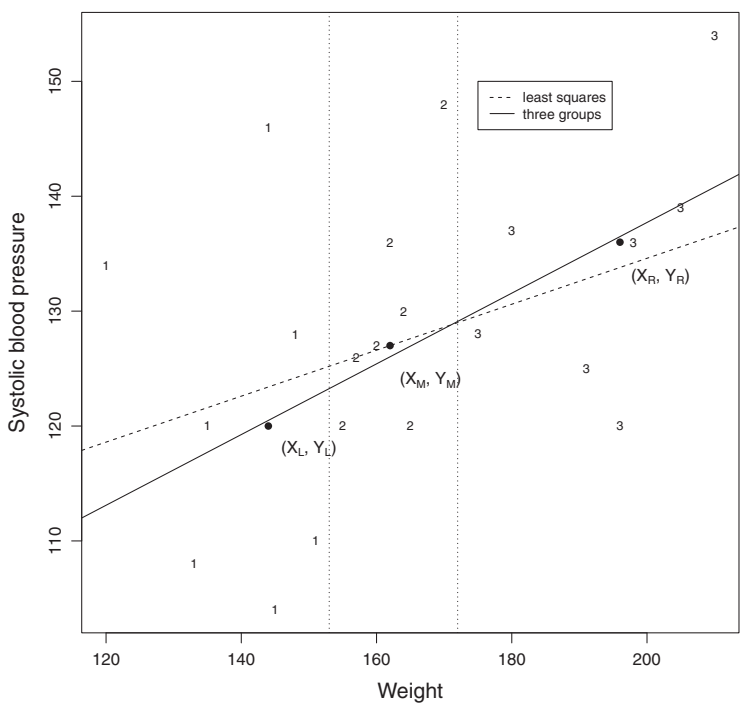


Figure 15.1 Three Groups of Blood Pressure Measurements and Body Weights of $n = 21$ Individuals with Extremely High Cholesterol Levels (>350)

Table 15.13 *Model Results: Least Squares Estimated Parameters from Linear Model Analysis of Weight and Its Relationship to Blood Pressure (Table 15.12)*[a]

Parameters	Estimates	*s. e.*	*p*-value
$\hat{a}$	93.864	18.093	–
$\hat{b}$	0.207	0.109	0.072

[a]Estimated line: $\hat{y}_i = 93.864 + 0.207x_i$.

For the three-group regression analysis, the estimated intercept is then

$$\hat{A} = \frac{Y_L + Y_M + Y_R}{3} - \hat{B}\left[\frac{X_L + X_M + X_R}{3}\right] = \bar{Y} - \hat{B}\bar{X}.$$

The estimated line becomes $\hat{Y}_i = \hat{A} + \hat{B}X_i$.

From the blood pressure and body weight data (Table 15.12), the three X/Y-median values are the following:

x-Values	$X_L = 145$	$X_M = 162$	$X_R = 196$
y-values	$Y_L = 120$	$Y_M = 128$	$Y_R = 137$

Therefore, the estimated slope is

$$\hat{B} = \frac{137 - 120}{196 - 145} = 0.333$$

with estimated intercept

$$\hat{A} = \frac{1}{3}(120 + 128 + 137) - 0.333\left[\frac{1}{3}(145 + 162 + 196)\right] = 72.444.$$

The robust estimated line is $\hat{Y}_i = 72.\,444 + 0.333x_i$ (Figure 15.1, solid line).

Ordinary least squares estimation applied to the same $n = 21$ subjects with extremely high cholesterol levels is summarized in Table 15.13 (Figure 15.1, dashed line).

The two estimated lines are similar, but the precision of the estimates dramatically differs. The bootstrap estimated standard error (2000 replicate samples) of the estimated slope $\hat{B} = 0.333$ from the nonparametric three-group regression analysis has *standard error* $(\hat{B}) = 0.446$. The corresponding parametric estimated value $\hat{b} = 0.207$ from the least squares analysis has *standard error* $(\hat{b}) = 0.109$. Similarly, the standard errors of the estimated intercepts are *standard error* $(\hat{A}) = 81.476$ and *standard error* $(\hat{a}) = 18.093$. A least squares approach produces estimates with standard errors considerably smaller than the median-derived robust and nonparametric estimates. Nevertheless, when extreme or nonlinear observations are likely present in the data, unbiased estimates, even with high degree of variability, are likely to be superior to badly biased estimates with high precision. Statistician J. W. Tukey expressed the same thought as "an approximate answer to the right question is worth a great deal more than a precise answer to the wrong question."

Tukey's Quick Test

In response to a challenge made at the 1956 Royal Statistical Society meeting, J. W. Tukey (b. 1915) developed a simple, easily implemented, and effective statistical test to assess differences between two samples of observations. He called the statistical technique a *quick test*, or it is sometimes referred to as a "pocket test." Tukey's nonparametric quick test is compact, self-contained, and easily applied without calculations or tables and provides a rigorous nonparametric comparison of two independent samples of data.

The term "quick" is an understatement. Directly for the original paper (Technometrics, 1959), Tukey completely describes his test:

> Given two groups of measurements, taken under conditions (treatments etc.) *A* and *B*, we feel more confident of our identification of the direction of difference the less the groups overlap one another. *If one groups contains the highest value and the other groups the lowest value*, then we may choose (*i*) to count the number of values in one group exceeding all value in the other, (*ii*) to count the number of values in the other group falling below all those on the one, and (*iii*) to sum these two counts (we *require* that neither count be zero). If the two groups are roughly the same size, then the critical values of the total count are *roughly*, 7, 10, 13, i.e., 7 for a two-sided 5% level, 10 for a two sided 1% level and 13 for a two sided 0.1% level.

The compactness of the procedure is made clear by the description, complete with critical values, in this short paragraph.

To repeat, in a different form: Denote the hypothesis that two samples differ by chance alone (*null hypothesis*) with the symbol H_0. When H_0 is true, the two-sided error probabilities are the following:

$T \leq 7$, then accept H_0
$7 < T \leq 10$, then reject H_0; $\alpha = 0.05$
$10 < T \leq 13$, then reject H_0; $\alpha = 0.01$ and
$13 \leq T$, then reject H_0; $\alpha = 0.001$

where the test statistic T represents the total count of the number of the nonoverlapping observations. The symbol α represents the probability of the error of rejecting the null hypothesis when it is true or, in symbols, $\alpha = P(\textit{reject } H_0 \mid H_0 \textit{ is true})$. That is, extreme random variation leads to declaring a difference exists when it does not, sometimes called a *type I error*.

A quick test is most accurate when the sample sizes are equal. Tukey recommends that the ratio of the sample sizes should "be roughly equal." Others have derived strategies that are less influenced by differing sample sizes. Tied values are generally not a problem unless they are "boundary" values used to determine the count T. When ties occur at a boundary, suggested strategies exist to adjust the counts that make up the test statistic T. The power of a statistical test, as noted, is defined as the probability the analysis identifies a nonrandom difference when a nonrandom difference exists. The power of Tukey's quick test has been described under different conditions and compared to other approaches. On the other hand, dealing with these issues makes the quick test less "quick." Tukey quotes mathematician Churchill Eisenhart as defining the "practical power" of a test as statistical power multiplied by the probability the test is used.

Table 15.14 *Data: Artificial Data to Illustrate Tukey's Quick Test to Evaluate Differences from Two Sources (Labeled A and B – n = 10)*

	Observed values									
	1	2	3	4	5	6	7	8	9	10
A	[0.2	1.5	2.1	3.0]	4.0	4.2	4.4	4.8	5.3	6.7
B	3.1	3.4	3.7	4.1	5.3	5.8	6.3	6.6	[7.8	7.9]

Note: Bracketed values are the nonoverlapping observations that makeup the test statistic count *T*.

For an example, consider the artificial data in Table 15.14. The same data displayed from smallest to largest value (bold = sample *A*) readily identify the nonoverlapping values (square brackets):

[**0.2**, **1.5**, **2.1**, **3.0**], 3.1, 3.4 3.7, **4.0**, 4.1, **4.2**, **4.4**, **4.8**, **5.3**, 5.3, 5.8, 6.3, 6.5, **6.7**, [7.8, 7.9].

The quick test statistic requires the following:

1. The sample sizes be close to equal
2. One sample contains the largest value and the other sample the smallest value and
3. A null hypothesis that states the two independent samples differ by chance alone.

Therefore, for the artificial data, the values in sample *A* less than the smallest value in *B* (3.1) are 0.2, 1.5, 2.1, and 3.0. The values in sample *B* greater than the largest value in sample *A* (6.7) are 7.8 and 7.9. Thus, the count of nonoverlapping values is $T = 4 + 2 = 6$. Compared to the 5%-critical value $T \leq 7$, the observed value $T = 6$ indicates marginal evidence that sample *A* systematically differs from sample *B*. Formally, the hypothesis H_0 of no systematic difference is accepted.

The same count of nonoverlapping values can be made from ordinal data. The previous example data (Table 15.14) represented as ordinal values are the following:

ordinal data (ordered)

$[A_1, A_2, A_3, A_4], B_1, B_2, B_3, A_5, B_4, A_6, A_7, A_8, A_9, B_5, B_6, B_7, B_8, A_10, [B_9, B_{10}].$

The ordered values of ordinal observations, represented by A_i and B_i, produce the previous count *T*. That is, four values from sample *A* are smaller than the smallest value from sample *B* (B_1) and two values in sample *B* are larger than the largest value in sample *A* (A_{10}) making the test statistic *T* again $T = 2 + 4 = 6$. The quick test count *T* only requires a meaningful ordering of the observed values.

Cholesterol levels of participants in a study of coronary heart disease classified by behavior type-A and type-B illustrate a quick test applied to $n = 34$ subjects randomly selected from a study of 3153 men at high risk (Table 15.15) (Chapter 6).

The cholesterol levels of type-B individuals less than the smallest value among type-A individuals (176) are {132, 140, 149, 155, and 172}. The cholesterol levels of type-A individuals greater than the largest value among the type-B individuals (255) are {271, 274, and 325}. Thus, the total number of nonoverlapping values $T = 5 + 3 = 8$ indicates a

Table 15.15 *Data: Study Subjects Ordered by Cholesterol Levels for Type-A and Type-B Behaviors Illustrate Tukey's Quick Test (Subset of Large Coronary Heart Disease Study – $n = 34$)*

	Cholesterol levels																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Type-B	[**132**	**140**	**149**	**155**	**172**]	179	181	182	185	189	220	220	225	247	254	255		
Type-A	176	177	177	180	192	202	209	210	211	223	227	237	238	239	242	[**271**	**274**	**325**]

Note: The eight bold values between brackets are the nonoverlapping observations that makeup the test statistic count *T*.

likely nonrandom difference between the two distributions of cholesterol levels for type-A and type-B study participants. The test statistic T exceeds the α-level 5% critical value of 7. Formally, the hypothesis that the observed difference occurred by chance alone (H_0) is rejected ($\alpha = 0.05$). For contrast, applied to the same data, the classic Student's t-test estimated p-value is 0.035, and the Wilcoxon ranked sum test estimated p-value is 0.076 (Chapter 8). The example demonstrates the quickness of Tukey's quick test.

Friedman Rank Test

The Friedman rank test is a nonparametric comparison of mean values not very different in principle from the Kruskal-Wallis rank test. Both test statistics are a comparison of differences in mean values of ranks calculated to reflect differences among two or more groups. The Friedman analysis, however, employs a strategy of assigning ranks designed to explore data classified into a $r \times c$ table (Chapters 8 and 10).

To assess the degree of association between two categorical variables used to construct an $r \times c$ table, the values in each row are ranked from 1 to c. The mean of the rank values from each of the c columns becomes nonparametric measures of the differences among the table columns. The row and column categories are essentially unrestricted. The row or columns, for example, can be strata or blocks of data or categories or levels of ordinal variables. The ranked values also can be a wide variety of possibilities such as counts or rates or proportions or mean values.

Consider the small example consisting of a 3×3 table of artificial data that illustrates the mechanics of the Friedman assignment and assessment of the ranks:

	Data			Ranks		
	A_1	A_2	A_3	A_1	A_2	A_3
B_1	2.3	1.6	3.5	2	1	3
B_2	4.3	1.9	1.2	3	2	1
B_3	5.3	6.9	0.2	2	3	1

The values in the three rows B_1, B_2, and B_3 are ranked from 1 to 3, and the column sums of these ranked values are $R_1 = 7$, $R_2 = 6$, and $R_3 = 5$ creating the mean rank values $\bar{r}_j = R_j/r$. Specifically, the mean rank values are: $\bar{r}_1 = 2.33$, $\bar{r}_1 = 2.00$ and $\bar{r}_3 = 1.67$. Because the

ranks simplify the "data," the mean value and the total sum of squares of the $r \times c$ ranked values are

$$\text{mean value} = \bar{r} = \frac{c+1}{2}$$

and

$$\text{total sum of squares} = T = \sum\sum \left(r_{ij} - \frac{c+1}{2}\right)^2 = r(c-1)c(c+1)/12.$$

The Friedman test statistic, similar to the Kruskal-Wallis test statistic, is a comparison of the variation of the $\bar{r}_j$ mean ranks among the c columns relative to the total variation. In symbols, the summary measure of the variation among the mean rank values (denoted B) is

$$B = r \sum \left(\bar{r}_j - \frac{c+1}{2}\right)^2 \quad j = 1, 2, \ldots, c,$$

and the Friedman test statistic becomes

$$S = \frac{r(c-1)B}{T}.$$

Like the Kruskal-Wallis test, the test statistic is created from the between sum of squares from an analysis of variance applied to ranks that replace the original observed values (not shown). For small sample sizes, published tables and computer software provide the exact probability the test statistic S results from strictly random variation among the c mean ranks of the column values. In addition, for moderate sample sizes, the test statistic S has an approximate chi-square distribution with $c - 1$ degrees of freedom.

From the example 3×3 table of ranks:

$$\bar{r} = \frac{4}{2} = 2.0, \quad T = \frac{3(2)3(4)}{12} = 6$$

and

$$B = 3[(2.33 - 2.0)^2 + (2.0 - 2.0)^2 + (1.67 - 2.0)^2] = 0.667,$$

making the Friedman test statistic $S = r(c-1)B/T = 3(2)0.677/6 = 0.667$. A usually noted "short cut" expression is

$$S = \frac{12}{rc(c+1)} \sum R_j^2 - 3r(c+1)$$

and again from the ranked data

$$S = \frac{12}{3(3)4}[7^2 + 6^2 + 5^2] - 3(3)(4) = 0.667.$$

The exact distribution of the Friedman test statistic S consists of $(c!)^r$ equally likely values when only random variation causes the observed differences among the mean rank values. For the 3×3 table, there are $(3!)^3 = 216$ equally likely 3×3 tables but only six different values of the Friedman test statistic S (Table 15.16, row 1). When no association exists, of the 216 possible tables, then 204 produce a value of S equal or greater than the observed value $S = 0.667$. The exact cumulative probability associated with the test statistic S is

Table 15.16 *Summary: Distribution of Friedman's Test Statistic and Computer-Generated Exact Cumulative Probabilities for a* 3×3 *Table*

Values of the friedman test statistic (S)	0	0.667	2.00	2.667	4.667	6.00
Counts of occurrence of values of S	12	90	36	36	36	6
Cumulative counts of the values of S	216	204	114	78	42	6
Cumulative probabilities[a]	1.00	0.944	0.528	0.361	0.194	0.028

[a]Right-hand tail values or $P(S \geq s)$.

p-value $= P(S \geq 0.667 \mid$ random differences only$) = 204/216 = 0.944$. The approximate chi-square value (degrees of freedom $c - 1 = 2$), which would not be expected to be accurate, yields the p-value $= 0.717$. As expected, the approximation improves as the size of the table increases. For a 5×5 table, where $(5!)^5 = 2.5 \times 10^{10}$ possible tables exist, a few chi-square approximate and exact probabilities give a sense of accuracy (Table 15.17).

Infant mean birth weights classified into a 6×6 table by maternal age and parity illustrate an application of Friedman's rank test (Tables 15.18 and 15.19). The table of ranks itself is helpful. Two trends within the table become apparent. For parity values 0 and 1, the mean birth weights tend to decrease with increasing maternal age. For parity values 3, 4 and ≥ 5, the mean birth weights consistently increase with increasing maternal age.

The Friedman test statistic is

$$B = 41 \text{ and } T = 105, \text{ therefore, } S = \frac{r(c-1)B}{T} = \frac{6(5)41}{105} = 11.714.$$

The value $S = 11.714$ has an approximate chi-square distribution (degrees of freedom $= c - 1 = 5$) when maternal age and parity are unrelated. The Friedman test statistic S produces an approximate p-value of $P(S \geq 11.714 \mid$ random column differences$) = 0.039$.

A special case of the Friedman statistic occurs for a table with two columns ($c = 2$). An example 10×2 table illustrates (Table 15.20). From Table 15.20, the Friedman analysis yields

$$B = 10[(1.2 - 1.5)^2 + (1.8 - 1.5)^2] = 1.8 \quad \text{and } T = 10(1)(2)3/12 = 5,$$

making the test statistic $S = r(c-1)B/T = 10(1)(1.8)/5 = 3.6$. For the special case of a $r \times 2$ table, the Friedman test is simply a different version of the sign test; that is, the test

Table 15.17 *Approximation: For a* 5×5 *Table Exact and Approximate Friedman Significance Probabilities Close to Selected Values* {0.20, 0.10, 0.05, 0.02*, and* 0.01}

S	Selected	Chi-square ($df = 4$)	Exact
5.92	0.20	0.205	0.218
7.52	0.10	0.111	0.107
8.96	0.05	0.062	0.050
10.56	0.02	0.032	0.019
11.52	0.01	0.021	0.010

Table 15.18 *Data (Chapter 10): Mean Birth Weights (Kilograms) of Newborn Infants Classified by Parity and Maternal Age*

	Maternal age					
Parity	<20	20–25	25–30	30–35	35–40	≥40
0	3.281	3.278	3.280	3.220	3.202	3.333
1	3.305	3.354	3.360	3.345	3.332	3.315
2	3.291	3.371	3.397	3.405	3.399	3.398
3	3.258	3.356	3.390	3.422	3.434	3.443
4	3.225	3.322	3.374	3.412	3.435	3.467
≥5	3.202	3.303	3.335	3.392	3.417	3.482

Table 15.19 *Data: Ranked Values within Rows of Mean Birth Weights (Kilograms) of Newborn Infants Classified by Parity and Maternal Age*

	Maternal age					
Parity	<20	20–25	25–30	30–35	35–40	≥40
0	5	3	4	2	1	6
1	1	5	6	4	3	2
2	1	2	3	6	5	4
3	1	2	3	4	5	6
4	1	2	3	4	5	6
≥5	1	2	3	4	5	6
Sum of ranks R_j	10	16	22	24	24	30
Mean value ($\bar{r}_j$)	1.67	2.67	3.67	4.00	4.00	5.00

Table 15.20 *Data: The Friedman Test Applied to a* 10×2 *Table That Is Identical to a Sign Test*

	Columns	
Rows	1	2
1	1	2
2	1	2
3	1	2
4	2	1
5	1	2
6	1	2
7	2	1
8	1	2
9	1	2
10	1	2
R_j	12	18
Mean values $\bar{r}_j$	1.2	1.8

statistic for the sign test of the estimated proportion ($\hat{p}$) and Friedman rank test statistic (S) are identical. From the first column of Table 15.20, the proportion of values of rank two is $\hat{p} = 2/10 = 0.2$ ($rows = r = 10$). The second column has the same distribution and provides no new information. Then the chi-square test statistic X^2 from the sign test statistic ($\hat{p}$) and the Friedman test statistic S are

$$X^2 = \left[\frac{\hat{p} - 0.5}{1/\sqrt{4r}}\right]^2 = \left[\frac{0.2 - 0.5}{1/\sqrt{4(10)}}\right]^2 = 3.6 \quad \text{and} \quad S = \frac{r(c-1)B}{T} = \frac{10(1)(1.8)}{5} = 3.6.$$

The value $X^2 = S = 3.6$ has an approximate chi-square distribution (degrees of freedom $= c - 1 = 1$) when the within-pair differences reflect only random variation (p-value $= 0.058$).

Survival

16

Rates

Particularly in the context of epidemiology and medical research, rates are used and typically understood at a practical level as a measure of risk. A complete and rigorous description of a rate and its properties provides a deeper understanding of this important statistical tool. The following explanations and illustrations apply to the properties of many kinds of rates, but mortality rates are described to simplify terminology and provide concrete examples.

To enhance application and interpretation of an estimated rate, such as the examples in Table 16.1, detailed descriptions of the origins and properties answer six basic questions:

1. What are the defining elements of a rate?
2. What assumptions are necessary to accurately estimate a rate?
3. What conditions allow the estimation of a rate when the exact time of death is not known?
4. What is the relationship between a rate and time of survival?
5. What is the fundamental difference between a rate and probability?
6. What assumptions are necessary to statistically evaluate a rate?

An Average Mortality Rate

The sample mean value is certainly one of the most used and basic of statistical summaries. A less used summary statistic is the ratio of two sample mean values. On Wall Street, the mean stock price and mean earnings are reported as a ratio. In medicine, the level HDL cholesterol to LDL cholesterol are sometimes expressed as a ratio of mean values. In a number of fields, the mean signal-to-mean noise ratio measures of performance. A mortality rate is a ratio of two mean values.

The numerator of a mortality rate is the mean number of deaths, more frequently called the proportion of deaths. In symbols, denoted $\bar{d}$, the mean value is

$$\text{mean number of deaths} = \text{proportion} = \frac{0+1+0+0+\cdots+1}{n} = \frac{\sum d_i}{n} = \frac{d}{n} = \bar{d},$$

where d_i represents a binary variable that takes on values zero (alive = 0) or one (death = 1) among n individuals and d represents the sum of these binary values. That is, the sum represented by d is the total number of deaths among n at risk individuals. In the language of baseball, this kind of mean value is called a batting average, but elsewhere it is usually referred to as a proportion or an estimated probability.

The denominator of a mortality rate is the mean person-years of survival relevant to the observed deaths. The mean survival time, calculated the same way as most mean values, is

Table 16.1 *Data: Selected Female Breast Cancer Ethnic- and Age-Specific Mortality Rates per* 100,000 *Persons at Risk and Approximate 95% Confidence Intervals for U.S. Women (1999–2005)*

Ethnicity	Mortality ages	Populations	Deaths	Lower[a]	Rates	Upper[a]
African American	25–34	11,139,194	234	1.85	2.10	2.39
White	25–34	63,818,574	549	0.79	0.86	0.94
African American	65–74	3,446,912	2012	55.88	58.37	60.98
White	65–74	32,426,185	14,725	44.68	45.41	46.15

[a]Approximate 95% confidence interval bounds.

the total accumulated time at risk divided by the number of individuals (again denoted n) who survived or died during a specific time period. In symbols, the estimated mean survival time is

$$\text{mean survival time} = \frac{\text{total time at-risk}}{\text{total number individuals at-risk}} = \frac{\sum t_i}{n} = \bar{t},$$

where t_i represents an observed time alive of each of n at-risk individuals ($i = 1, 2, \ldots, n$). The expression for an average mortality rate (denoted $\hat{R}$) is the ratio of these two mean values or

$$\text{estimated average mortality rate} = \hat{R} = \frac{\text{mean number of deaths}}{\text{mean survival time}} = \frac{\bar{d}}{\bar{t}}.$$

The units of the numerator $\bar{d}$ are deaths and the units of the denominator $\bar{t}$ are person-years, making the units of the estimated rate $\hat{R}$ deaths per person-years. Any units of time can be used. Person-years are a frequent choice for mortality data.

Equivalently, but less intuitively, an average mortality rate is often defined as the total number of deaths divided by the total observed time-at-risk or

$$\text{estimated average mortality rate} = \hat{R} = \frac{\bar{d}}{\bar{t}} = \frac{d/n}{\sum t_i/n} = \frac{d}{\sum t_i}.$$

The estimated rate $\hat{R}$ requires some fine-tuning to be exactly described (details follow).

One important reason to express mortality risk as a rate is to create commensurate comparisons. To illustrate, for two populations:

Population I: among 1000 people at risk, 10 died during a six-month period and
Population II: among 500 people at risk, 40 died during a two-year period.

It is not immediately obvious which population sustained the greater risk. The rate from population I is 20 deaths per 1000 person-years compared to the rate from population II of 40 deaths per 1000 person-years clearly identifies a twofold difference.

For a cohort of n persons observed until all individuals die (denoted d), the average mortality rate is

$$\text{average mortality rate} = \hat{R} = \frac{\bar{d}}{\bar{t}} = \frac{d}{\sum t_i} = \frac{n}{\sum t_i} = \frac{1}{\bar{t}},$$

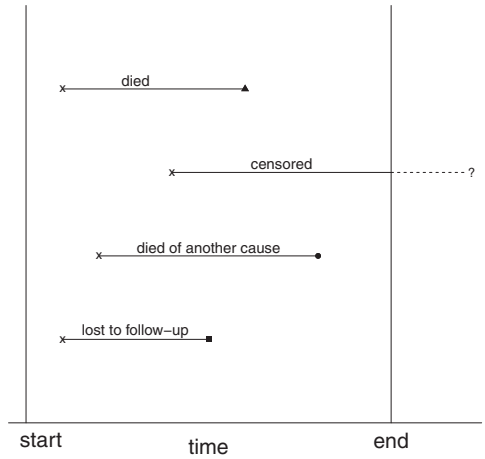


Figure 16.1 Schematic Representation of Four Kinds of Mortality Data Observed within an Interval

where Σt_i again is the total person-years lived by all $n = d$ individuals. In addition, the mean time lived by these n individuals $\sum t_i / n = \bar{t}$ is $1/\hat{R}$. It is not uncommon that the reciprocal of a rate equals the mean time or reciprocal of the mean time equals a rate. For example, when a car travels at a rate of 60 miles per hour or 1 mile per minute, the mean travel time is $1/rate$ or 1 hour per 60 miles or 1 minute per mile.

Mortality data from a cohort observed until every person dies ($n = d$) are rare. It is far more common that observed data contain individuals who died and individuals who did not die during the period of observation ($d < n$). Individuals who survived beyond the observation period have unknown survival times (Figure 16.1). Furthermore, their survival times are greater than their observed times, and, therefore, these individuals would add more survival time to the total accumulated if their times of death were known. Such survival times are said to be *censored or incomplete* (Chapter 17).

Data collected to describe survival experience over a specified time period can include survival times of individuals who were lost from observation and, perhaps, others who died of causes other than the cause under study (Figure 16.1). These individuals do not bias the estimated mortality rate if their absence from observation is completely unrelated to the issues under study. In other words, when the "missing" individuals are not systematically different from the remaining individuals, no bias occurs. They are treated as randomly censored observations. The loss of these individuals decreases the precision of an estimated rate, but the estimate remains unbiased. Otherwise, the survival times of these individuals bias, sometimes considerably bias, the estimation of summary statistics. For example, clinical trial participants could be lost because their symptoms disappeared and they withdrew from the trial, or their illness could have become more severe, causing them to drop out of the trial. Typically and in the following discussion, individuals who died from other causes or were lost from observation causing their survival times to be censored are assumed to randomly differ from individuals with complete survival times, called *noninformative* censored observations.

The total survival time of n individuals who either died during the study period (again denoted d) or who were randomly censored is, as before, the sum of all observed times alive

regardless of outcome status (Σt_i). As noted, the $n - d$ censored individuals would contribute additional survival time to the total had the period of data collection been longer. Thus, the total observed survival time is an underestimate of the total time survived by all subjects. Despite this underestimate, an accurate estimated mortality rate is possible.

The assumption that during the period under consideration mortality risk is constant allows creation of an estimate of a mortality rate that accounts for incomplete survival times. When risk of death is constant, expected survival time is the same for all observed individuals at any time. That is, all sampled individuals always have the same expected survival time, denoted μ. Therefore, the unobserved survival times, on average, remain μ for each individual with an incomplete survival time, and the number of these censored individuals is $n - d$. An estimate of the total unobserved survival time these censored individuals would have added to the total observed is $(n - d)\ \mu$ person-years as if the survival times of these individuals were complete. Adding this estimate, based on the assumption of constant mortality risk and noninformative censoring, to the total observed survival time, an estimate of the mean survival time is

$$\mu = \frac{\text{known survival time} + \text{estimated unobserved survival time}}{n}$$
$$= \frac{\text{estimated total survival time}}{n}.$$

The estimated mean survival times becomes

$$\hat{\mu} = \frac{\sum t_i + (n-d)\hat{\mu}}{n} \quad \text{or} \quad n\hat{\mu} = \sum t_i + n\hat{\mu} - d\hat{\mu} \quad \text{and} \quad \hat{\mu} = \frac{\sum t_i}{d}.$$

The denominator is the number of deaths d and not the number of subjects n, compensating for the bias incurred from the underestimated total survival time. Thus, when censoring is noninformative and risk is constant, the estimated average mortality rate $\hat{R}$ is identical to the previous rate defined as the mean number of deaths divided by the mean survival time ($\hat{R} = \bar{d}/\bar{t} = d/\sum t_i$). The average mortality rate is, as before, the reciprocal of the mean survival time or

$$\text{average mortality rate} = \hat{R} = \frac{1}{\hat{\mu}} = \frac{d}{\sum t_i}.$$

Mortality rates are not constant over the entire life time. They typically vary from close to zero (children and teenagers) to around 200 deaths per 1000 person-years in older women (individuals aged 90 years). Thus, an overall mortality rate, such as the U.S. white, male rate for the year 2010 of 13.7 deaths per 1000 person-years, does not effectively reflect a specific individual's risk. A single estimate insufficiently summarizes anything but a single and constant value. Therefore, for a single estimated rate to be useful, the rate estimated must be constant or at least approximately constant for the time period considered. For one-, five-, or even 10-year periods, a single rate frequently provides a reasonably accurate estimate of an underlying risk of human disease and death. For example, the mortality rate for U.S. white, 50-year-old males (549 deaths per 100,000 person-years) and for 55-year-old males (697 deaths per 100,000 person-years) are approximately constant over the age interval 50 to 55 years. Rates for the very young (under one year old), the very old (over 80 years old), and broad age intervals (greater than 10 years) require more sophisticated estimation.

The number of deaths per persons-at-risk, frequently called a rate, is in fact a proportion or an estimated probability. Examples are illustrated in Table 16.1 for breast cancer mortality. As with most estimates, the interpretation of these rates is improved by an accompanying confidence interval. Typically, deaths are rare events occurring independently with approximately constant and small probability among large numbers of at-risk individuals (n). In this case, the variability observed in the number of deaths is likely described accurately by a Poisson probability distribution, which then provides a statistical basis for constructing a confidence interval (Chapters 4, 9, and 27). Specifically, the estimated variance of a mortality rate $\hat{r} = \frac{d}{n}$ based on a Poisson probability distribution is

$$variance(\hat{r}) = \frac{\text{rate}}{\text{number of persons at-risk}} = \frac{\hat{r}}{n},$$

and, for the logarithm of an estimated rate $\hat{r}$, the estimated variance becomes

$$variance[log(\hat{r})] = \frac{1}{d},$$

where d again represents the total number of observed deaths (Chapter 27). Logarithms of rates tend to have more symmetric distributions than the rates themselves, improving the use of a normal distribution as an approximation. Note that the precision of $log(\hat{r})$ depends only on the number of deaths.

An approximate 95% confidence interval based on the normal distribution and the estimate $log(\hat{r})$ is $log(\hat{r}) \pm 1.960\sqrt{1/d}$, and the confidence interval for the rate $\hat{r}$ becomes

$$\text{lower bound} = \hat{A} = base \times e^{[log(\hat{r}) - 1.960\sqrt{1/d}]}$$

and

$$\text{upper bound} = \hat{B} = base \times e^{[log(\hat{r}) + 1.960\sqrt{1/d}]},$$

where the *base* is an arbitrary constant chosen to make the estimated rate a value usually greater than one for purely aesthetic reasons.

The breast cancer mortality rate among African American women ages 25 to 34 years is $\hat{r} = 234/11{,}139{,}194 = 0.00002101$ or 2.101 deaths per 100,000 persons at risk (Table 16.1). Therefore, the associated 95% confidence interval for the underlying rate r estimated by the rate $\hat{r} = 2.101$ deaths per 100,000 persons-at-risk becomes

$$\text{lower bound} = 100{,}000 \times e^{log(234/11,139,194) - 1.960\sqrt{1/234}} = 100{,}000 \times e^{-10.899} = 1.848$$

and

$$\text{upper bound} = 100{,}000 \times e^{log(234/11,139,194) + 1.960\sqrt{1/234}} = 100{,}000 \times e^{-10.643} = 2.388,$$

where the estimated variance is based on the assumption that the distribution of the observed number of deaths is at least approximately described by a Poisson probability distribution (Chapter 9).

A popular choice for a comparison of two rates is a ratio. A rate ratio is no more than a rate estimated from one source (denoted $\hat{r}_1$) divided by a rate estimated from another independent source (denoted $\hat{r}_2$). In symbols, the estimate of a rate ratio (denoted $\hat{rr}$) is then

$$\text{estimated rate ratio} = \hat{rr} = \frac{\hat{r}_1}{\hat{r}_2}.$$

Again from Table 16.1, the breast cancer mortality rate among African American women ages 25 to 34 is $\hat{r}_1 = 2.101$ per 100,000 persons-at-risk, and the corresponding rate among white women is $\hat{r}_2 = 0.860$ per 100,000 persons-at-risk producing, an estimated rate ratio of $\hat{rr} = 2.101/0.860 = 2.442$.

A confidence interval constructed from the logarithm of a rate ratio again improves the accuracy of a normal approximation. The logarithm of an estimated rate ratio $\hat{rr} = 2.442$ is $log\,(\hat{rr}) = log(2.442) = 0.893$. The variance of the logarithm of the estimated rate ratio is

$$\begin{aligned} variance(log[\hat{rr}]) &= variance(log[\hat{r}_1/\hat{r}_2]) \\ &= variance(log[\hat{r}_1]) + variance(log[\hat{r}_2]) = \frac{1}{d_1} + \frac{1}{d_2}, \end{aligned}$$

where d_1 represents the number of deaths that produced the estimated rate $\hat{r}_1$ and d_2 represents the number of deaths that produced the estimated rate $\hat{r}_2$.

Because the two compared breast cancer rates (ages 25–34 – Table 16.1) consist of $d_1 = 234$ (African American women) and $d_2 = 549$ (white women) deaths, the estimated variance associated with the distribution of the logarithm of the estimated rate ratio $log(\hat{rr}) = 0.893$ is

$$variance(log[\hat{rr}]) = \frac{1}{234} + \frac{1}{549} = 0.0061,$$

yielding an approximate normal distribution based 95% confidence interval for $log(rr)$ of (0.740, 1.046). A 95% confidence interval for the rate ratio rr becomes $(e^{0.740}, e^{1.046}) =$ (2.095, 2.846) based on the estimate $\hat{rr} = e^{0.893} = 2.442$. Again only the total numbers of deaths is relevant to the precision of the estimate.

An Approximate Average Rate

The vast majority of disease and mortality data are available in the form of tables made up of age or time intervals (for example, Table 16.1). These data are readily obtained from public sources (for example, National Center for Health Statistics or the National Cancer Institute websites: http://www.cdc.gov/nchs/ or http://www.nci.nih.gov). From these public sources, the exact survival time associated with each observed case of disease or death is almost never reported. Reported mortality data typically consist of the number of individuals at-risk at the beginning of an interval (denoted l) and the number of individuals who died during the interval (again, denoted d). To estimate a specific mortality rate from interval data it is necessary to make an assumption about the pattern of deaths within the interval considered.

An almost always used and frequently realistic assumption is that the probability of death, particularly over a short period of time, has a uniform probability distribution (Chapter 1). Necessarily for an interval from t_1 to t_2, individuals at one end of the time interval then have at least approximately the same probability of death as individuals at the other end of the

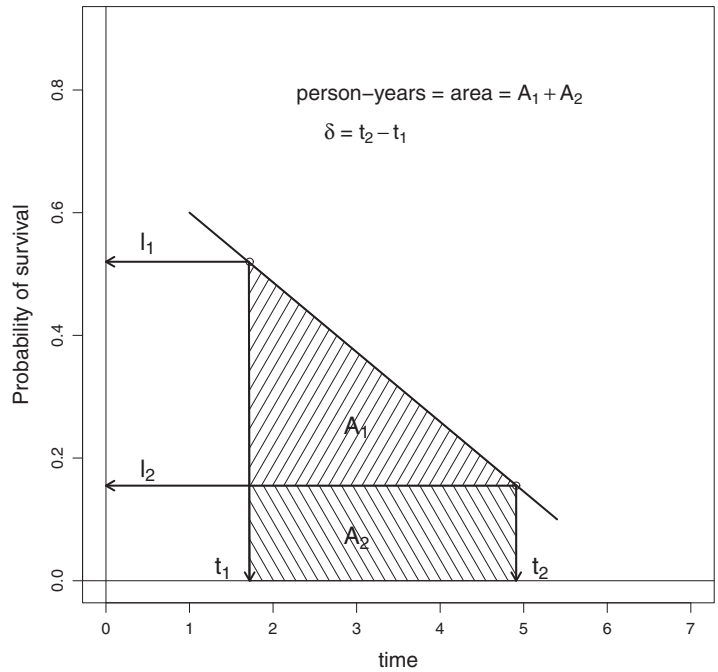


Figure 16.2 Illustration of Approximate Total Survival Time of Those Individuals Who Died (Triangle – A_1) and Total Survival Time of Those Individuals Who Survived the Entire Interval (Rectangle – A_2)

interval. When the probability of death is the same for all individuals, the mean time survived by those individuals who died becomes one-half the length of the interval. In symbols, the mean time survived by those who died is $\frac{1}{2}\delta$ where $\delta = t_2 - t_1$ represents the length of the interval. This assumption is sometimes called the *actuarial assumption.*

The approximate total survival time accumulated within an interval (length $= \delta$) is often accurately approximated by a simple geometric area made up of a triangle and a rectangle (labeled A_1 and A_2 in Figure 16.2). Specifically, when l_2 individuals survive the entire interval, their survival times add an estimated total of $\delta \times l_2$ person-years to the total person-years lived (a rectangle – A_2). The individuals who died during the interval add a total of $\frac{1}{2}\delta(l_1 - l_2) = \frac{1}{2}\delta \times d$ person-years to the total survival time (a triangle – A_1). The approximate total person-years at-risk for l_1 individuals alive at the beginning of the interval become

$$\text{total person-years} = \text{rectangle} + \text{triangle} = \delta l_2 + \tfrac{1}{2}\delta d \quad \text{or} \quad \delta l_1 - \tfrac{1}{2}\delta d$$

because $l_2 = l_1 - d$. Therefore, an expression for the estimated average mortality rate for l individuals alive at the start of an interval of length δ is

$$\text{approximate average rate} = \hat{R} = \frac{\text{deaths}}{\text{person-years}} = \frac{d}{\delta\left(l - \frac{1}{2}d\right)}.$$

This average rate accurately reflects risk when the pattern of survival is at least approximately uniform within a specific time period, which, as noted, is frequently the case for

disease or mortality in human populations. Again, an exception is newborn infants, where the probability of death is extremely high at the beginning the first year of life; that is, the probability of death is not constant. The mean time lived by those infants who died before one year of age is about 0.1 years and not 0.5 years. For large intervals (>10 years) and older individuals (>80 years old), the probability of death also is not likely constant. Otherwise, the assumption of at least an approximate uniform probability of death frequently produces useful approximate average rates.

The probability of death (denoted q) is estimated by

$$\text{probability of death} = \hat{q} = \frac{d}{l}.$$

The estimate $\hat{q}$ is directly the count of those who died d divided by the number of individuals who could have died, denoted l. Time is implicit in the calculation but does influence the calculated probability. For example, during an eight-year study of $l = 3154$ men at high risk of coronary heart disease, a total of $d = 279$ study participants experienced a coronary event, yielding an estimated probability of $\hat{q} = d/l = 279/3154 = 0.088$. The fact that these individuals were observed over an eight-year period is not directly evident from the estimated probability.

An average rate is usually reported in units of person-years and a probability is unitless. An average rate can be any positive number, and a probability is always between 0 and 1. An average rate directly incorporates time into the estimate, and a probability does not. A lifetime mortality rate in many populations is about 15 deaths per 1000 person-years, and the lifetime probability of death is 1.0. Nevertheless, for short time intervals, these two apparently different quantities are often closely related.

Specifically,

$$\text{approximate average rate} = \hat{R} = \frac{d}{\delta\left(l - \frac{1}{2}d\right)},$$

$$\frac{\frac{d}{l}}{\delta\left(\frac{l}{l} - \frac{1}{2}\frac{d}{l}\right)} = \frac{q}{\delta\left(1 - \frac{1}{2}q\right)} \approx \frac{q}{\delta}.$$

Thus, an estimated average mortality rate $\hat{R}$ approximately equals the estimated probability of death q divided by the length of the time interval δ associated with the probability. For most human diseases and causes of death, the probability q is small, making $1 - \frac{1}{2}q \approx 1$. For the previous coronary heart disease example, the approximate average rate of a coronary event is then $\hat{R} \approx q/\delta = 0.088/8 = 0.011$ or 11 *chd*-events per 1000 person-years.

Analogously, when the same time period is considered, a ratio of two rates is approximately equal to the ratio of two probabilities. In symbols, the ratio of rates is

$$\text{rate ratio} = \frac{\hat{R}_1}{\hat{R}_2} \approx \frac{q_1/\delta}{q_2/\delta} = \frac{q_1}{q_2}.$$

Two seventeenth-century scientists, John Graunt and Edmund Halley, independently suggested an ingenious method to characterize mortality patterns. Their approach used current mortality data to create the mortality experience of a hypothetical cohort of individuals as

Table 16.2 *Life Table Probabilities: Numbers Surviving Individuals and Numbers Deaths in Two Communities – 17th-Century London and 21st-Century Berkeley, CA (2005)*

	London			Berkeley		
Age	Probability[a]	Survived	Died	Probability[a]	Survived	Died
0	1.00	1000	0	1.0000	100,000	0
0–10	0.54	540	460	0.9985	99,848	152
10–20	0.34	340	200	0.9952	99,518	329
20–30	0.21	210	130	0.9871	98,708	811
30–40	0.14	140	70	0.9725	97,245	1462
40–50	0.08	80	60	0.9451	94,509	2737
50–60	0.05	50	30	0.8872	88,720	5789
60–70	0.02	20	30	0.7365	73,647	15,072
>70	0.01	10	10	0.4047	40,473	33,175

[a]Probability of surviving beyond interval limit l_i.

if they were observed from birth of the first individual to death of last individual. This description of mortality risk frequently is called a *life table*.

A life table consists of a series of current age-specific survival probabilities applied to describe the mortality experience of a hypothetical cohort of individuals as if they were observed over time. For example, John Graunt's original life table was based on age-specific probabilities of death estimated from then-current London mortality data, collected primarily from church records. Specially, these seventeenth-century survival probabilities are the following:

	Age Intervals (10 Years)							
Years	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Probability	0.54	0.34	0.21	0.14	0.08	0.05	0.02	0.01

The consecutive application of these survival probabilities to a hypothetical cohort of 1000 individuals reproduces Graunt's life table "cohort" (Table 16.2).

The same process applied to 100,000 members of a hypothetical cohort produces a life table created from modern mortality data again using current survival probabilities estimated from Berkeley, CA (2005). These two patterns of survival are displayed in Figure 16.3.

An estimated total years lived by the 1000 life-table "Londoners" (Table 16.2) is

$$\begin{aligned}\text{total survival time} = \bar{t} &= \delta \sum \left(l_i - \tfrac{1}{2}d_i\right) \\ &= 10\left[\left(1000 - \tfrac{1}{2}460\right) + \left(540 - \tfrac{1}{2}200\right) + \cdots + \left(10 - \tfrac{1}{2}10\right) + 10\right] \\ &= 770 + 440 + \cdots + 10 = 18{,}950 \text{ person-years}\end{aligned}$$

calculated from intervals with lengths $= \delta = 10$ years.

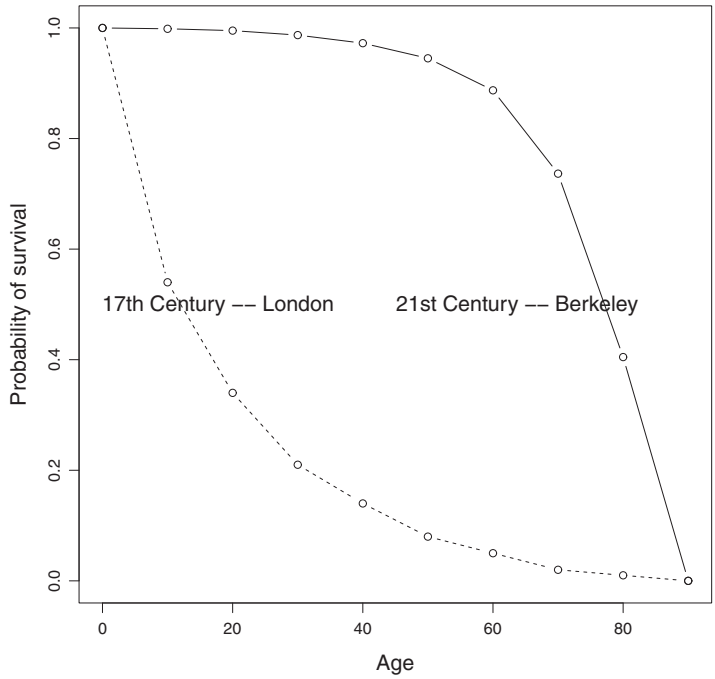


Figure 16.3 Survival Probabilities for Two Communities –17th-Century London and 21st-Century Berkeley, CA

The mean survival time estimated from the London life table cohort becomes $\bar{t} = 18{,}950/1000 = 18.95$ years. Similarly calculated, the mean survival time estimated from the Berkeley cohort is $\bar{t} = 7\ 629{,}045/100{,}000 = 76.29$ years. Both calculations depend on the actuarial assumption that deaths are uniformly distributed within all 10-year age intervals. This assumption is least accurate for the older age intervals and improved estimates are available.

A life table constructed from current mortality rates creates a description of a hypothetical cohort of individuals as if they were observed until all members of the cohort died. For a life table cohort, a newborn infant is assumed to have exactly the same mortality risk in 60 years experienced by a person 60 years old during the year the infant was born. Clearly, mortality risk changes over time. Nevertheless, changes in human mortality are generally small, making life table summaries frequently useful for short-term predictions and excellent for comparing current mortality patterns among different groups and populations. Table 16.3 contains a few representative estimates of mean survival times calculated from recent life tables for a variety of counties.

A Rate Estimated from Continuous Survival Times

A well-known mathematical expression leads to Euler's value, symbolized by e. One form of this expression is

$$\left[1 - \frac{x}{n}\right]^n$$

Table 16.3 *Example: Representative World-Wide 21st-Century Mean Years of Survival (WHO, 2009)*

Countries	Means	Countries	Means	Countries	Means
Afghanistan	48	Finland	80	Switzerland	82
Australia	82	India	65	United Kingdom	80
Canada	81	Japan	83	United States	79
Congo	55	Mexico	76	Vietnam	72
China	74	Pakistan	63	Zambia	48

and becomes approximately e^{-x} when the value n is large and becomes equal to e^{-x} when n is infinitely large (Chapter 27). For example, when $x = 0.5$, then

$$\left[1 - \frac{0.5}{20}\right]^{20} = 0.603 \text{ or } \left[1 - \frac{0.5}{50}\right]^{50} = 0.605 \text{ and ultimately, } e^{-0.5} = 0.607.$$

Euler's expression provides a link between discrete and continuous survival probabilities.

A fundamental measure of continuous survival time is a *survival probability*, denoted $S(t)$, and is the probability of surviving beyond a specific time t, described by the expression

$$S(t) = P(\text{surviving from time 0 to } t) = P(\text{surviving beyond time } t) = P(T \geq t).$$

Consider a time interval of length t divided into n segments of length δ such as



where the survival probability at any time t (denoted p) is identical within each subinterval of time (length $= \delta$) for the interval length $n\delta = t$. Under these conditions, the survival probability

$$S(t) = P(T \geq t) = \prod_{i=1}^{n} p_i = p^n = (1-q)^n = \left(1 - \frac{nq}{n}\right)^n \approx e^{-qn} = e^{-\lambda\delta n}$$
$$= e^{-\lambda t} \text{ (Chapter 27).}$$

The mortality rate represented by λ is, as before, the probability of death (q) divided by the length of the interval ($\lambda = q/\delta$ or $q = \lambda\delta$) and is constant within all subintervals. The survival probability $S(t)$ becomes exactly $e^{-\lambda t}$ when n is infinitely large ($\delta \to 0$), which is another way of saying the survival probability becomes $S(t) = P(T \geq t) = e^{-\lambda t}$ when time t is a continuous variable ($\delta = 0$) and risk of death is constant. More technically, as the interval length δ converges to zero, the survival time distribution converges to an *exponential survival probability*, described by the expression $S(t) = e^{-\lambda t}$.

For another point of view, the probability of surviving from time $= 0$ to time $= t_2$ divides into two intervals, 0 to t_1 and t_1 to t_2. Thus, the probability of surviving beyond time t_2 is

$$\text{survival probability} = S(t_2) = P(T \geq t_2) = P(T \geq t_2 | T \geq t_1) P(T \geq t_1).$$

For exponentially distributed survival times, this survival probability becomes

$$S(t_2) = e^{-\lambda t_2} = P(T \geq t_2 | T \geq t_1) e^{-\lambda t_1},$$

and, for interval from t_1 to t_2, the survival probability is then

$$P(T \geq t_2 | T \geq t_1) = e^{-\lambda(t_2 - t_1)}.$$

This straightforward relationship indicates the central property of exponentially distributed survival times. The probability of surviving from time t_1 to t_2 depends only on the difference $t_2 - t_1$ and not the actual times t_1 and t_2. For example, when $t_2 - t_1 = 20$ years and λ is 0.04, the survival probability is $P(T \geq t_2 / \mathrm{T} \geq t_1) = e^{-0.04(20)} = 0.449$ for $t_2 = 100$ and $t_1 = 80$ and for $t_2 = 25$ and $t_1 = 5$. The probability of surviving 20 years is the same for an 80 year old as for a five year old. The survival probability is 0.449 for any 20-year period. In general, exponential survival describes the risk for individuals who do not age or objects that do not wear out. In terms of a mortality rate, represented by λ, risk is constant. In symbols, the mortality rate is expressed as $\lambda(t) = \lambda$.

The expression to estimate the exponential survival time rate (λ) from continuous time survival data is the same as the previously described average mortality rate $\hat{R}$. The data estimated rate described by an exponential survival probability distribution is again

$$\text{estimated rate} = \hat{\lambda} = \frac{d}{\sum t_i},$$

where t_i represents both complete and noninformative censored survival times. In addition, the estimated variance of the logarithm of the estimate $\hat{\lambda}$ is again $variance[log(\hat{\lambda})] = 1/d$. As always, to accurately estimate a rate with a single value, it is necessary that the underlying rate estimated is a single constant value, which is exactly the property that produces an exponential survival distribution.

Consider the survival times of $n = 23$ African American participants in the San Francisco Men's Health Study (1990). Their survival times are either weeks from diagnosis of AIDS to death (complete) or weeks from diagnosis to the end of participation in study (censored). The survival times of the study participants classified by cigarette smokers and nonsmokers are the following:

Nonsmokers: 2^+, 42^+, 27^+, 22, 26^+, 16, 31, 37, 15, 30, 12^+, 5, 80, 29, 13, 1, and 14

Smokers: 21^+, 13, 17, 8, 23, and 18,

where "+" indicates a noninformative censored survival time. These data contain 17 nonsmokers with five censored survival times and six smokers with one censored survival time.

For the nonsmokers, the estimated mortality rate based on $n_1 = 17$ complete (12) and censored (5) survival times is $\hat{\lambda}_1 = 12/402 = 0.030$. The estimate $log(\hat{\lambda}_1)$ is -3.512 with $variance(log[\hat{\lambda}_1]) = 1/d_1 = 1/12 = 0.083$. The 95% confidence interval based on the estimated log-rate bounds $(-4.077, -2.946)$ yields an approximate 95% confidence interval for the rate λ_1 of $(e^{-4.077}, e^{-2.946}) = (0.017, 0.053)$. An estimate of the mean survival time for nonsmokers is $\bar{t} = 1/\hat{\lambda}_1 = 1/0.030 = 33.5$ weeks.

Similarity, for the smokers, the estimated mortality rate based on $n_2 = 6$ complete (5) and censored (1) survival times is $\hat{\lambda}_2 = 5/100 = 0.05$. The estimate $log(\hat{\lambda}_2)$ is -2.996 with $variance(log[\hat{\lambda}_2]) = 1/d_2 = 1/5 = 0.200$. The 95% confidence interval based the estimated

log-rate bounds $(-3.872, -2.119)$ yields an approximate 95% confidence interval for the rate λ_2 of $(e^{-3.872}, e^{-2.119}) = (0.021, 0.120)$. An estimate of the mean survival time for smokers is $\bar{t} = 1/\hat{\lambda}_2 = 1/0.050 = 20.0$ weeks.

A 95% confidence interval for the mean survival time can be calculated directly from the confidence interval bounds of an estimated rate. For a specific estimated survival time $\bar{t}$, the lower bound is the reciprocal of the upper bound of the 95% confidence interval for the rate λ. Similarly, the upper bound is the reciprocal of the corresponding corresponding lower bound of the rate (Chapter 2).

Specifically, from the AIDS data for nonsmokers based on an estimated survival time of $\bar{t} = 1/0.030 = 33.5$ weeks, then

$$\text{lower bound} = \frac{1}{0.053} = 19.025 \quad \text{and} \quad \text{upper bound} = \frac{1}{0.017} = 58.989$$

is an approximate 95% confidence interval. For smokers, based on an estimated survival time of $\bar{t} = 1/0.05 = 20$ weeks, then again

$$\text{lower bound} = \frac{1}{0.120} = 8.324 \quad \text{and} \quad \text{upper bound} = \frac{1}{0.021} = 48.051$$

is an approximate 95% confidence interval.

The median value is a frequent choice for a single estimate to characterize a "typical" value from an asymmetric distribution. Because survival times are never less than zero and the nature of risk rarely produces symmetric influences, survival time distributions are often asymmetric.

For exponential survival times with a rate represented as λ, then $S(t) = e^{-\lambda t}$ makes the median survival time (denoted m) the value such that $S(m) = 0.5$. The estimated median and its estimated variance are then

$$\text{median} = \hat{m} = \frac{log(2)}{\hat{\lambda}} \quad \text{with} \quad variance(\hat{m}) = \frac{\hat{\lambda}^2}{d},$$

where d again represents a number of deaths. To improve accuracy, an approximate 95% confidence interval is once again estimated from the logarithm of the estimated median value. Furthermore, the variance of the logarithm of the estimated median is $variance(log[\hat{m}]) = 1/d$. Thus, a normal-based approximate 95% confidence interval becomes

$$95\% \text{ confidence interval} = e^{[log(\hat{m}) \pm 1.960\sqrt{1/d}]}$$

from the logarithm of the estimated median value $\hat{m}$. Note that (Chapter 27),

$$\text{variance}[log(\hat{m})] = variance(log[log(2)] - log(\hat{\lambda})) = variance[log(\hat{\lambda})] = \frac{1}{d}.$$

Consider an example from the AIDS data. For nonsmokers, the estimated mortality rate is $\hat{\lambda}_1 = 0.030$, making the estimated median survival time $\hat{m} = log(2)/0.030 = 23.220$ weeks. The approximate 95% confidence interval based on the estimate $log(\hat{m})$ is $log(\hat{m}) \pm 1.960\sqrt{1/d_1} \rightarrow (2.579, 3.711)$ and from the estimate $\hat{m} = 23.220$, then $(e^{2.579}, e^{3.711}) = (13.187, 40.888)$. Because the confidence interval bounds constructed from the estimate mortality rate $\hat{\lambda}_1 = 0.030$ are the previous bounds $(\hat{A}, \hat{B}) = (0.017, 0.053)$, the same

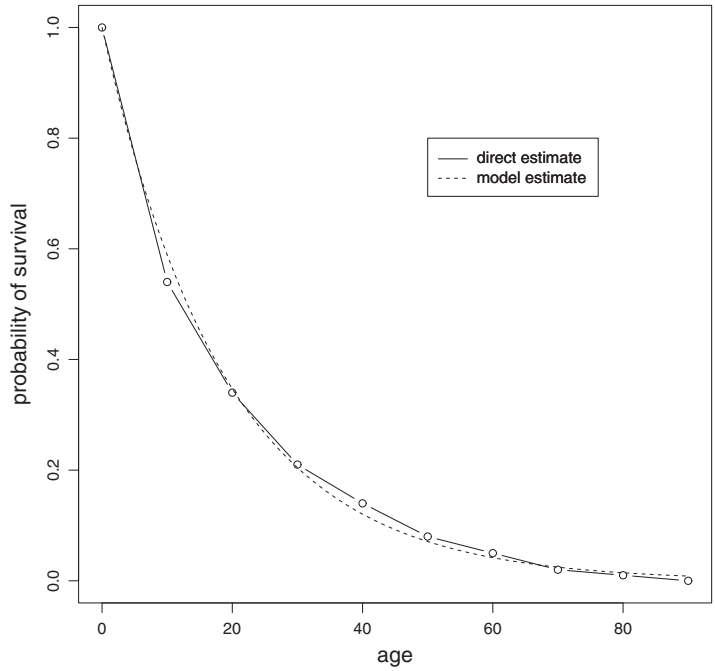


Figure 16.4 Comparison of Theoretical Continuous Survival Probabilities [$S(t)$] and Observed Life Table Values from 17th-Century London

confidence bounds based directly on the estimated $\hat{m} = log(2)/0.030 = 23.220$ weeks are

$$\text{lower} = log(2)/\hat{B} = log(2)/0.053 = 13.187 \quad \text{and}$$
$$\text{upper} = log(2)/\hat{A} = log(2)/0.017 = 40.888.$$

For smokers, similarly, the estimated mortality rate is $\lambda_2 = 0.050$, making the estimated median survival time $\hat{m} = log(2)/0.50 = 13.863$ weeks. The approximate 95% confidence interval based on $log(\hat{m})$ is

$$log(\hat{m}) \pm 1.960\sqrt{1/d_2} \rightarrow (1.753, 3.506)$$

and from the estimate $\hat{m} = 13.863$, then $(e^{1.753}, e^{3.506}) = (5.770, 33.307)$. Again from the previous confidence bounds for $\hat{\lambda}_2$, namely, $(\hat{A},\ \hat{B}) = (0.021, 0.120)$, the confidence bounds based on the estimated $\hat{m} = log(2)/0.050 = 13.863$ weeks are

$$\text{lower} = log(2)/\hat{B} = log(2)/0.120 = 5.770 \quad \text{and}$$
$$\text{upper} = log(2)/\hat{A} = log(2)/0.021 = 33.307.$$

A historic application of the conjecture that survival times have an exponential distribution is the mortality pattern from 17th-century London. From John Graunt's life table (Table 16.2), the mean life time is $\bar{t} = 18.950$, making the estimated constant mortality rate $\hat{\lambda} = 1/\bar{t} = 1/18.950 = 0.053$. The assumption of constant mortality risk for all ages then yields the estimated exponential survival probability function of $\hat{S}(t) = e^{-\hat{\lambda}t} = e^{-0.053t}$ for any age t.

For example, the probability of surviving any $t = 20$-year period in 17th-century London is estimated as $\hat{S}(t) = e^{-0.053(20)} = 0.346$. This estimate is a plausible description of the pattern of 17th-century mortality only if the mean survival time is the same for all ages (constant mortality risk). The estimated exponential survival curve $\hat{S}(t)$ (Figure 16.4, dashed line) and the survival pattern calculated from the London life table (circles – solid line), consisting of 10 survival probabilities estimated without an assumption of a constant lifetime mortality risk, are strikingly similar (Table 16.2).

17

Nonparametric Survival Analysis

The term "survival" applies to a variety of kinds of data. Applied to human populations, survival typically refers to time to death or time to a specific disease outcome. In other contexts, survival data are called *time to failure* data, which refers to such things as lifetimes of consumer products or times to equipment failures. In the following, the term survival refers to the observed time survived until death again to simplify terminology and provide concrete examples, but the methods discussed generally apply. The analysis of survival data is wide ranging in terms of approaches and complexity. Nevertheless, the principal methods are primarily specific applications of basic statistical techniques. These methods are parametric and depend on the properties of the population sampled (Chapter 18) or nonparametric and do not require knowledge or assumptions about the population sampled. The latter is the topic to be discussed.

Like many statistical methods, probabilities are key to analysis and description of survival time data. Particularly important is the *survival probability* (Chapter 16), which describes the probability of surviving beyond a specific time (denoted t) and, in symbols, is defined by

$$\begin{aligned}\text{survival probability} &= S(t) = P(T \geq t) \\ &= \frac{\text{number individuals who survived beyond time } t}{\text{total number who could have survived beyond time } t}.\end{aligned}$$

A nonparametric estimate of the probability distribution associated with survival times is typically referred to as a *Kaplan-Meier* or a *product-limit* estimate. Nevertheless, when all survival outcomes have known times of occurrence, the estimation is not basically different from the nonparametric estimate of a cumulative probability distribution (Chapter 12). The estimated survival probability (denoted $\hat{S}[t]$) is then the proportion

$$\begin{aligned}\hat{S}(t) &= \frac{\text{number surviving beyond time } t}{n} \\ &= \frac{\text{number surviving from time 0 to time } t}{n} \quad \text{time} = t \geq 0,\end{aligned}$$

where n represents the number of individuals who could have died, referred to as *at-risk* individuals. The corresponding estimated cumulative probability function is

$$\hat{F}(t) = \hat{P}(T \leq t) = 1 - \hat{S}(t).$$

To illustrate estimation of a survival probability distribution, a sample of $n = 10$ patients with advanced symptoms associated with AIDS was selected from participants in a large

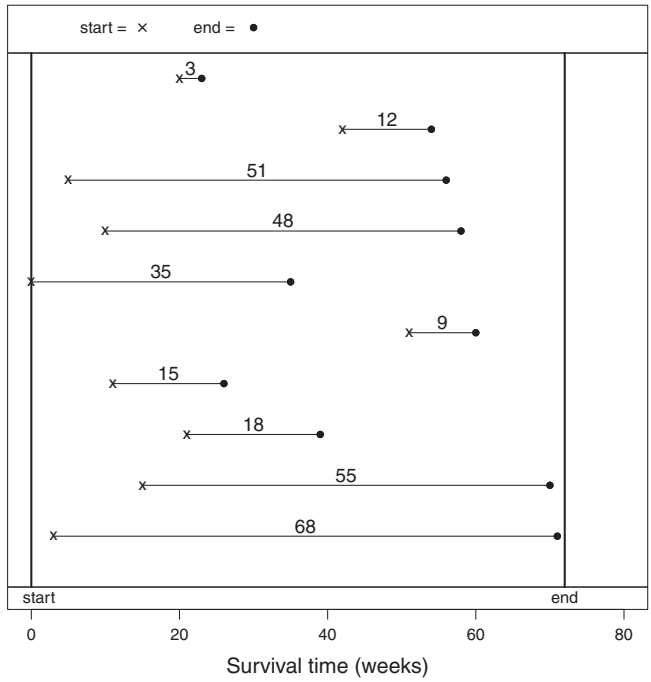


Figure 17.1 Survival Times: Display of Observed Survival Times (Weeks) of $n = 10$ AIDS Patients (SFMHS – 1983)

study of the disease. The data are part of the San Francisco Men's Health Study (SFMHS), and the 10 survival times are displayed in Figure 17.1. The specific survival times are

$$t_i = \{3, 9, 12, 15, 18, 35, 48, 51, 55 \text{ and } 68 \text{ weeks}\}.$$

To estimate the distribution of survival times, the observed values (denoted t_i) are classified into a table created so that exactly one death occurs in each of n specially constructed time intervals (Table 17.1). Then it is straightforward to estimate the conditional probability of death for each interval. This conditional probability (denoted q_i) is estimated by

$$\hat{q}_i = P(\textit{death in interval } i \mid \textit{alive at the beginning of interval } i) = \frac{1}{n_i},$$

where n_i represents the number individuals at risk at the beginning of the ith interval. The complementary conditional probability of surviving the ith interval is

$$\begin{aligned}\hat{p}_i &= P(\textit{alive at the beginning interval } i+1 \mid \textit{alive at the beginning of interval } i) \\ &= 1 - \hat{q}_i = 1 - \frac{1}{n_i} = \frac{n_i - 1}{n_i}.\end{aligned}$$

The 10 conditional survival probabilities ($\hat{p}_i$) estimated from the AIDS data are shown in Table 17.1. For example, the estimated conditional probability of death for specific time interval 15 to 18 weeks (interval $i = 5$) is $\hat{q}_5 = 1/6 = 0.167$, making the conditional probability of survival $\hat{p}_5 = 1 - \hat{q}_5 = 5/6 = 0.833$.

Table 17.1 *Estimation: Kaplan-Meier (Product-Limit) Estimated Survival Probabilities* (P) *from* $n = 10$ *AIDS Patients (SFMHS – 1983)*

	Data			Estimates		
i	$t_{i-1} - t_i$	deaths	n_i	$\hat{p}_i$	$\hat{P}_i$	$s.e.^a$
1	0–3	1	10	0.900	0.900	0.095
2	3–9	1	9	0.888	0.800	0.127
3	9–12	1	8	0.875	0.700	0.145
4	12–15	1	7	0.857	0.600	0.155
5	15–18	1	6	0.833	0.500	0.158
6	18–35	1	5	0.800	0.400	0.155
7	35–48	1	4	0.750	0.300	0.145
8	48–51	1	3	0.667	0.200	0.127
9	51–55	1	2	0.500	0.100	0.095
10	55–68	1	1	0.000	0.000	–

[a]Estimated standard error of the distribution of the estimate $\hat{P}_i$.

The estimated unconditional survival probability for the kth interval is

$$\begin{aligned}\hat{P}_k &= P(\textit{survival from time } t_0 = 0 \textit{ to time } t_k) = P(\textit{survival beyond time } t_k)\\ &= \frac{\text{number surviving beyond time } t_k}{\text{number originally at-risk at time } t_0}.\end{aligned}$$

For the example data, the estimated unconditional probability of surviving beyond 18 weeks (again interval $i = 5$) is $\hat{P}_5 = 5/10 = 0.5$; that is, among the $n = 10$ observed AIDS patients, five survived beyond 18 weeks. The estimate of the unconditional survival probability for interval k is $\hat{P}_k = (n - k)/n$. The estimated variance of this estimated probability, from a binomial probability distribution (Chapter 4), is

$$\textit{variance}(\hat{P}_k) = \frac{\hat{P}_k(1 - \hat{P}_k)}{n}.$$

Therefore, the estimated variance associated with the estimate (Table 17.1)

$$\begin{aligned}\hat{P}_5 = 0.5 \text{ is } \textit{variance}(\hat{P}_5) &= 0.5(0.5)/10\\ &= 0.025 \ (\textit{standard error} = \hat{se} = \sqrt{0.025} = 0.158).\end{aligned}$$

The Kaplan-Meier estimate of a survival probability distribution is also called the *product-limit estimate.* The name comes from a more general approach based on the product of estimated conditional survival probabilities $\hat{p}_i$. Specifically, the estimate of the unconditional survival probability is the product of conditional interval-specific survival probabilities where

$$\begin{aligned}\hat{P}_k = P(\textit{survive } t_0 \textit{ to } t_1 \mid \textit{alive } t_0) &\times P(\textit{survive } t_1 \textit{ to } t_2 \mid \textit{alive } t_1)\\ &\times \cdots \times P(\textit{survive } t_{k-1} \textit{ to } t_k \mid \textit{alive } t_{k-1})\end{aligned}$$

or

$$\hat{P}_k = \hat{p}_1 \times \hat{p}_2 \times \hat{p}_3 \ \times \cdots \times \ \hat{p}_k$$

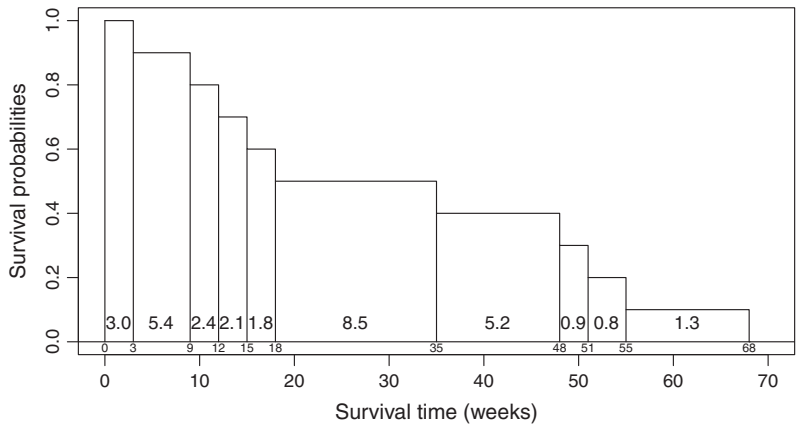


Figure 17.2 Nonparametric Estimated Survival Probability Distribution ($n = 10$) and the Area Enclosed by Each Interval

again for the *k*th interval. A product-limit estimated survival probability is identical to the previous Kaplan-Meier estimate because

$$\hat{p}_i = \frac{n-i}{n-i+1}, \text{ then } \hat{P}_k = \frac{n-1}{n} \times \frac{n-2}{n-1} \times \frac{n-3}{n-2} \times \cdots \times \frac{n-k}{n-k+1} = \frac{n-k}{n}.$$

From the AIDS example, the estimated survival probability for the fifth interval (Table 17.1) is, again,

$$\hat{P}_5 = \frac{9}{10} \times \frac{8}{9} \times \frac{7}{8} \times \frac{6}{7} \times \frac{5}{6} = \frac{5}{10} = 0.5.$$

A plot of the 10 estimated unconditional probabilities displays the entire estimated survival probability distribution based on the AIDS survival times (Table 17.1 and Figure 17.2).

The mean survival time calculated directly from the 10 observed survival times t_i is

$$\text{mean survival time} = \bar{t} = \frac{1}{n}\sum t_i = \frac{1}{10}[3 + 9 + 12 + \cdots + 68] = 31.4 \text{ weeks}.$$

Alternatively, the same mean survival time equals the area enclosed by the estimated survival probability distribution and is

$$\text{mean survival time} = \bar{t} = \sum \hat{P}_{i-1}(t_i - t_{i\ -1}) \quad i = 1, 2, \ldots, n$$

where $P_0 = 1.0$ when $t_0 = 0$. For the example AIDS data ($n = 10$ rectangles – Figure 17.2),

$$\begin{aligned}\text{mean survival time} = \bar{t} &= \frac{\text{total person-weeks lived}}{n} = \frac{\text{total area}}{n} \\ &= 1.0(3-0) + 0.9(9-3) + 0.8(12-9) + \cdots + 0.1(68-55) \\ &= 1.0(3) + 0.9(6) + 0.8(3) + \cdots + 0.1(13) \\ &= 3.0 + 5.4 + 2.4 + \cdots + 1.3 = 31.4 \text{ weeks}.\end{aligned}$$

To repeat, this mean value is the sum of the areas of ten rectangles (weeks lived), each with area $a_i = \hat{P}_{i-1}(t_i - t_{i-1})$, that make up the total area enclosed by the estimated survival distribution (Figure 17.2) ($\bar{t} = \sum a_i = 34.4$).

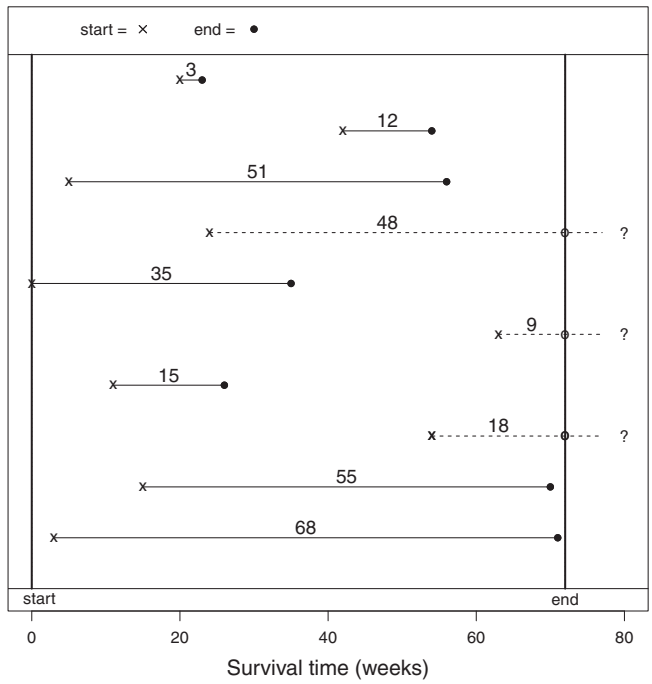


Figure 17.3 Survival Data from $n = 10$ AIDS Patients with $n - d = 3$ Censored Individuals at End of the Study (Version 1)

The expression to estimate of the variance of the survival times t_i, as usual, is

$$variance(t) = S_T^2 = \frac{1}{n-1}\sum(t_i - \bar{t})^2 \quad i = 1, 2, \ldots, n.$$

The variance of the mean survival time is then estimated by $variance\,(\bar{t}) = S_T^2/n$. For the $n = 10$ example survival times (Table 17.1), the estimated variance of the mean value $\bar{t}$ is $variance(\bar{t}) = 522.49/10 = 52.249$.

For most samples of survival data not all survival times end in an observed outcome such as death. Survival times are often incomplete because the data collection ended before the outcome occurred. These observations are said to be *right censored* (Chapter 16). Thus, survival data frequently contain a number of complete (denoted d) and incomplete (denoted $n - d$) observations. Figures 17.3 and 17.4 display two perspectives on the same data containing right censored observations (Table 17.2). The figures are constructed from the previous AIDS data but with $d = 7$ complete and $n - d = 3$ censored survival times, denoted 9^+, 18^+, and 48^+; that is, the data are

$$t_i = \{3, 9^+, 12, 15, 18^+, 35, 48^+, 51, 55 \text{ and } 68 \text{ weeks}\}.$$

To calculate accurate estimates from data containing censored survival times, the estimation strategy is designed to account for unobserved survival time to produce unbiased estimates.

The first step in calculating an unbiased estimate of a survival time distribution is to construct a table similar to the table of complete survival times (Table 17.1). This table is again constructed so that each interval contains a single death. Thus, the d complete

Table 17.2 *Kaplan-Meier Estimated Survival Probabilities* (P) *from* $d = 7$ *Complete Survival Times and* $n - d = 3$ *Censored Observations (SFMHS – 1983)*

	Data			Estimates		
i	$t_{i-1} - t_i$	Deaths	n_i	$\hat{p}_i$	$\hat{P}_i$	$s.e.^a$
1	0–3	1	10	0.900	0.900	0.095
2	3–12	1	8	0.875	0.787	0.134
3	12–15	1	7	0.857	0.675	0.155
4	15–35	1	5	0.800	0.540	0.173
5	35–51	1	3	0.667	0.360	0.187
6	51–55	1	2	0.500	0.180	0.158
7	55–68	1	1	0.000	0.000	–

[a]Estimated standard error of the distribution of the estimate $\hat{P}_i$.

values produce a table with d intervals (d rows) summarizing both complete and censored observations (Table 17.2).

The usual assumption underlying estimation of a survival probability distribution from data containing censored survival times is that only random differences exist between individuals with censored and individuals with complete survival times; that is, the only systematic difference between these two kinds of observations is that the time of the end point is not known for randomly censored individuals. When only random differences exist, the censoring

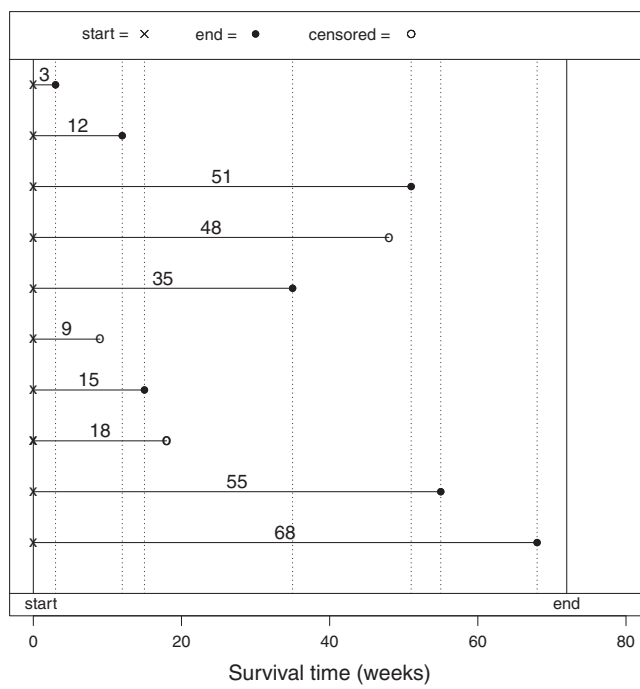


Figure 17.4 Survival Data from $n = 10$ AIDS Patients with $n - d = 3$ Censored Individuals at End of the Study (Version 2)

is, as before, *noninformative* (Chapter 16). The estimation then proceeds in the same way as described for complete data. The primary distinction is that no pattern exists among conditional survival probabilities p_i because no pattern exists among the occurrences of noninformative censored observations. For the example, the estimated conditional survival probability from the interval 35 to 51 weeks (the fifth interval – Table 17.2) is $q_5 = 1/n_5$ with $n_5 = 3$ at-risk individuals at the beginning of the interval, making the estimate of the conditional probability of death $\hat{q}_5 = 1/3 = 0.333$ and, therefore, the estimated conditional probability of surviving the interval is $\hat{p}_5 = 2/3 = 0.667$; that is, the value n_i is the number of individuals who survived or died in the ith interval. Regardless of the absence of a regular pattern, the conditional survival probabilities remains the components of the product-limit estimated unconditional survival probability P_i. The estimated value is again the product of estimated conditional survival probabilities. An unbiased estimated value $\hat{P}_5$ is, for example,

$$\begin{aligned}\hat{P}_5 &= P(\textit{survival beyond interval } t_5) = P(\textit{survival beyond } 51 \textit{ weeks})\\ &= \hat{p}_1 \times \hat{p}_2 \times \hat{p}_3 \times \hat{p}_4 \times \hat{p}_5\\ &= \frac{9}{10} \times \frac{7}{8} \times \frac{6}{7} \times \frac{4}{5} \times \frac{2}{3} = 0.360.\end{aligned}$$

The reason the estimated survival probability P_k is not biased by the presence of censored observations is displayed in Figure 17.4. The number of individuals in each interval (the n_i-counts) are the individual who died and $n_i - 1$ others who survived the entire interval. For example, the first interval (0 to 3 weeks) includes all 10 individuals ($n_1 = 10$). The second interval (3 to 12 weeks) contains eight individuals, one who died and seven who completed the interval ($n_2 = 8$). Only the eight complete observations are used to estimate the conditional probability $\hat{p}_2 = 7/8 = 0.875$. The censored survival time (9^+ weeks) is not included in the second interval calculation or in subsequent calculations. Thus, the count n_i represents the number of individuals who died or completed the interval (Table 17.2). For each interval, the conditional estimated probability $\hat{p}_i$ is based on only deaths and complete observations and, therefore, is not influenced by the incomplete nature of censored data. The product of interval-by-interval unbiased estimates produces a product-limit estimated unconditional probability $\hat{P}_k$ that is also unbiased. A plot of the estimated survival probability distribution based on estimation from $d = 7$ complete and $n - d = 3$ censored AIDS survival times (Table 17.2) is displayed in Figure 17.5.

The variance of the distribution of survival probabilities ($\hat{P}_i$) estimated from data that include censored observations ($d < n$) is not as simple as the binomial case (Table 17.1) where all survival times are complete ($d = n$). An expression for the estimation of the variance, due to early statistician Major Greenwood (b. 1880), is

$$\textit{variance}(\hat{P}_k) = \hat{P}_k^2 \sum \frac{\hat{q}_i}{n_i \hat{p}_i} \quad i = 1, 2, \ldots, k,$$

sometimes called *Greenwood's variance* (Chapter 27). For example, the estimated variance associated with the estimated survival probability $\hat{P}_5 = 0.360$ (Table 17.2) is

$$\begin{aligned}\textit{variance}(\hat{P}_5) &= (0.360)^2 \left[\frac{0.100}{10(0.900)} + \frac{0.125}{8(0.875)} + \frac{0.143}{7(0.857)} + \frac{0.200}{5(0.800)} + \frac{0.333}{3(0.667)}\right]\\ &= 0.035.\end{aligned}$$

The estimated standard error becomes $\hat{se} = \sqrt{0.035} = 0.187$ (Table 17.2).

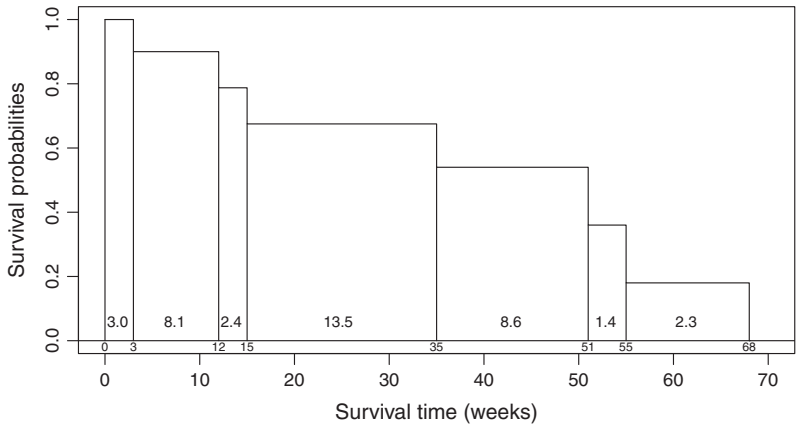


Figure 17.5 Nonparametric Estimated Survival Probability Distribution from Complete ($d = 7$) and Censored ($n - d = 3$) Observations and Areas Enclosed by Each Interval (Total Survival Time)

When no censored values occur ($d = n$), the Greenwood expression is identical to the binomial variance $variance(\hat{P}_k) = \hat{P}_k(1 - \hat{P}_k)/n$. For example, from the complete survival data (Table 17.1), the estimated survival probability is $\hat{P}(surviving\ beyond\ 18\ weeks) = \hat{P}_5 = 0.500$ with a Greenwood estimated variance of

$$variance(\hat{P}_5) = (0.\ 500)^2 \left[\frac{0.100}{10(0.900)} + \frac{0.111}{9(0.889)} + \frac{0.125}{8(0.875)} + \frac{0.143}{7(0.857)} + \frac{0.167}{6(0.833)} \right]$$
$$= 0.0.025.$$

or, as before,

$$variance(\hat{P}_5) = \frac{\hat{P}_5(1 - \hat{P}_5)}{n} = \frac{0.5(0.5)}{10} = 0.025.$$

The estimated standard error is again $\hat{s}e = \sqrt{0.025} = 0.158$ (Table 17.1).

The mean survival time estimated from the AIDS data ($n = 10$) is, as before, the area enclosed by the estimated survival probability distribution ($d = 7$ rectangles – Figure 17.5 and Table 17.2):

$$mean\ survival\ time = \bar{t} = \sum a_i = \sum \hat{P}_{i-1}(t_i - t_{i-1})$$
$$= 1.000(3 - 0) + 0.900(12 - 3) + \cdots + 0.180(68 - 55)$$
$$= 3.0 + 8.1 + \cdots + 2.3 = 39.383 \text{ weeks}.$$

The mean value calculated directly from the data is biased (likely too small) because the censored survival times t_i are understated.

The estimated variance of the distribution of the estimated mean value $\bar{t}$ requires a sum of calculated values denoted A_k. These values are sums of the rectangular areas (denoted a_i) that make up the product-limit estimated survival probability distribution (Tables 17.1 and 17.2). The sum represented by A_k is defined by the expression

$$A_k = \sum \hat{P}_{i-1}(t_i - t_{i\ -1}) = \sum a_i \quad i = k + 1, \ldots, d,$$

where again a_i represents the area of the ith rectangle (Figure 17.5).

Table 17.3 *Calculation: Worked Example for Estimate of Mean Survival Time and Its Variance from Incomplete Data (Table 17.2)*

	Intervals						
	1	2	3	4	5	6	7
$\hat{P}_{i-1}$	1.00	0.900	0.788	0.675	0.540	0.360	0.180
$t_i - t_{i-1}$	3	9	3	20	16	4	13
a_i	3.00	8.100	2.362	13.500	8.640	1.440	2.340
n_i	10	8	7	5	3	2	1
$n_i\,(n_i - 1)$	–	90	56	42	20	6	2
A_i	39.383	36.383	28.283	25.920	12.420	3.780	2.340
$\sum A_i^2/\,[n_i\,(n_i - 1)]$	–	14.708	14.284	15.996	7.713	2.381	2.738

Note: $a_i = \hat{P}_{i-1}(t_i - t_{i-1})$ and $\bar{t} = \sum a_i = 39.383$

Geometrically, the value A_k is the area under the survival curve to the right starting with the kth interval. For example, when $k = 5$, then $a_5 + a_6 + a_7 = 8.640 + 1.440 + 2.340 = 12.420$ (Figure 17.5). Note that when $k = 1$, then $A_1 = \sum a_i = \bar{t}$ and, as before, is the estimated mean survival time. The estimated variance of the distribution of the estimated mean survival time $\bar{t}$ becomes

$$variance(\bar{t}) = \frac{d}{d-1}\sum \frac{A_i^2}{n_i(n_i-1)} \quad i = 2, 3, \ldots, d,$$

where d again represents the number of intervals in the product-limit table (the number of complete observations if no identical survival times occur). For the example data from the $n = 10$ AIDS study participants (Table 17.2), the estimated mean survival time is $\bar{t} = \sum a_i = 39.383$ weeks. The estimated variance associated with the estimate is $variance(\bar{t}) = 67.457$. The details of the calculations are displayed in Table 17.3. As usual, the normal distribution provides an approximate 95% confidence interval estimated by $\bar{t} \pm 1.960\sqrt{variance(\bar{t})}$. Specifically for the AIDS data (Table 17.2), the 95% confidence interval based on the estimated mean survival time of $\bar{t} = \sum a_i = 39.383$ is

$$39.383 \pm 1.960(8.219) \rightarrow (23.285, 55.480).$$

The computational details are displayed for the complete AIDS data (Table 17.1) in Table 17.4 *($n = d$)*.

The estimated mean survival time is again $\bar{t} = \sum a_i = 31.4$ weeks and the variance of $\bar{t}$ is, as before,

$$variance\,(\bar{t}) = \frac{d}{d-1}\sum \frac{A_i^2}{n_i\,(n_i-1)} = \frac{10}{9}\,[8.96 + 7.35 + \cdots + 0.85] = 52.249.$$

Clearly, a direct calculation easily achieves the identical estimates but applying the general approach, illustrated by this simplest case (no censoring), provides additional appreciation of the logic and mechanics of the estimation process.

Table 17.4 *Calculation: Worked Example for Estimate of Mean Survival Time and Its Variance from Complete Data (Table 17.1)*

	Intervals									
	1	2	3	4	5	6	7	8	9	10
$\hat{P}_{i-1}$	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
$t_i - t_{i-1}$	3	6	3	3	3	17	13	3	4	13
a_i	3.0	5.4	2.4	2.1	1.8	8.5	5.2	0.9	0.8	1.3
n_i	10	9	8	7	6	5	4	3	2	1
$n_i\,(n_i - 1)$	–	90	72	56	42	30	20	12	6	2
A_i	31.4	28.4	23.0	20.6	18.5	16.7	8.2	3.0	2.1	1.3
$\sum A_i^2/n_i\,(n_i - 1)$	–	8.96	7.35	7.58	8.15	9.30	3.36	0.75	0.74	0.85

Cumulative Hazard Function

A sometimes alternative to the product-limit description of survival data is the *cumulative hazard function*. Parallel to the product-limit estimated survival probability distribution, the cumulative hazard function is a summary value derived from cumulative estimates based again on interval specific estimated conditional survival probabilities ($\hat{p}_i$). This summary provides an additional prospective on survival experience, but its role in the description of survival data is not very different from a product-limit estimated survival distribution.

The estimated cumulative hazard function is defined by the expression

$$\hat{H}(t_k) = -log(\hat{P}_k) = -log\left(\prod \hat{p}_i\right) = -\sum log(\hat{p}_i) \quad i = 1, 2, \ldots, k,$$

where, once again, $\hat{P}_k$ is the product-limit estimated survival probability constructed from the estimated conditional survival probabilities $\hat{p}_i$.

Because $log\,(p) = log\,(1 - q) \approx -q$, an approximate cumulative hazard function can be expressed as

$$\hat{H}(t_k) \approx \sum \hat{q}_i \qquad i = 1, 2, \ldots, k \quad \text{and} \quad k = 1, 2, \ldots, d - 1$$

and is an accurate approximation when q is small. The interval specific values of $\hat{H}(t_i)$ are

$$\begin{aligned}
\hat{H}(t_1) &\approx \hat{q}_1 \\
\hat{H}(t_2) &\approx \hat{q}_1 + \hat{q}_2 \\
\hat{H}(t_3) &\approx \hat{q}_1 + \hat{q}_2 + \hat{q}_3 \\
&\text{- - -} \\
&\text{- - -} \\
&\text{- - -} \\
\hat{H}(t_k) &\approx \hat{q}_1 + \hat{q}_2 + \cdots + \hat{q}_k
\end{aligned}$$

indicating the reason $\hat{H}(t_k)$ measures cumulative hazard (Chapter 18). For the 10 AIDS patients (Table 17.2), the estimated values $\hat{H}(t_k)$ are presented in Table 17.5 and Figure 17.6.

Table 17.5 *Estimated Cumulative Hazard Function* $\hat{H}(t_i)$ *from 10 AIDS Survival Times (Table 17.2)*

	Data			Estimates	
i	$t_{i-1} - t_i$	d_i	n_i	$\hat{p}_i$	$\hat{H}(t_i)$
1	0–3	1	10	0.900	0.105
2	3–12	1	8	0.875	0.239
3	12–15	1	7	0.675	0.393
4	15–35	1	5	0.540	0.616
5	35–51	1	3	0.360	1.022
6	51–55	1	2	0.180	1.715
7	55–68	1	1	0.000	–

Description of the Median Survival Time

An estimated mean survival time is frequently not the best choice to characterize survival data. Survival time distributions are typically asymmetric (skewed to the left or long right "tail") making a mean value less representative of a "typical" observation. In general, a median value is a more natural summary when the distribution that generated data is not symmetric.

The estimation of the median survival time (denoted $\hat{m}$) is the same for complete or censored survival data. The median value is the time such that the unconditional survival

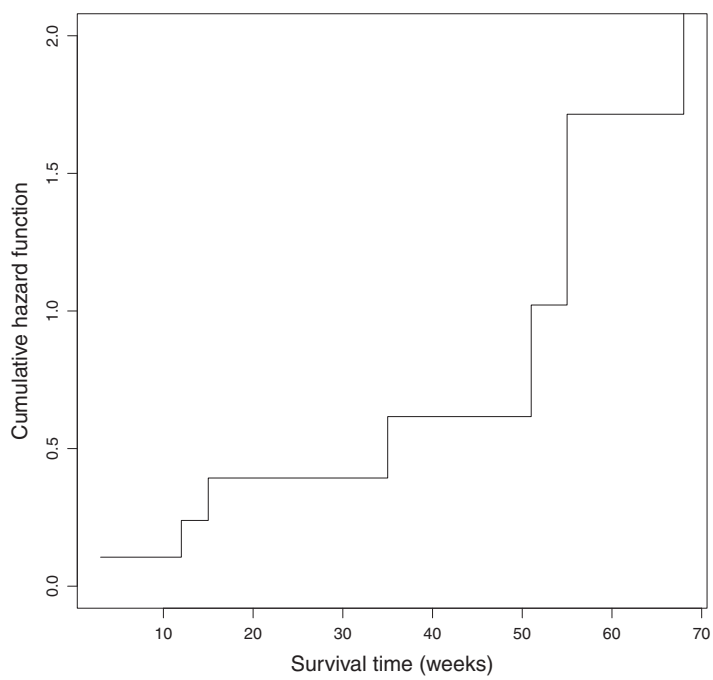


Figure 17.6 Estimated Cumulative Hazard Function $\hat{H}(t_k)$ from the 10 AIDS Survival Probabilities ($\hat{p}_i$ – Table 17.5)

probability is $P_m = 0.50$. The exact determination is rarely possible from a product-limit estimated survival function, but it is a simple task to locate the interval containing the median survival time. For the censored AIDS survival data (Table 17.2), the survival probability $\hat{P}_m = 0.50$ is greater than $\hat{P}_5 = 0.360$ and less than $\hat{P}_4 = 0.540$, and, therefore, the median survival time lies between 35 and 51 weeks (interval = 5th). Several possibilities are available to refine this interval estimate because no theoretically optimum choice exists. Candidates are lower bound = 35, upper bound = 51, mean bound = $(35 + 51)/2 = 43$, and the linear interpolated value = $35 + 16(0.540 - 0.50)/(0.540 - 0.360) = 38.556$. The commonly chosen estimate for the median value is the upper bound. For the AIDS data, this upper bound estimated median value is $\hat{m} = 51$ weeks.

Comparison of Two Survival Probability Distributions – The Log-Rank Test

Statistical analysis is frequently about comparisons. A nonparametric comparison of two estimated survival probability distributions is achieved with an analytic technique called the *log-rank test*.

Comparison of survival experiences between two groups employs a strategy similar to the product-limit estimation of a survival probability distribution. A sequence of 2×2 tables, each containing only deaths and subjects with complete survival times, is constructed so strata-specific estimates are again not influenced by censored observations. Parallel to the construction of product-limit estimate time intervals, each table contains only deaths and complete survival times. Values calculated from these tables are, therefore, not biased by the incomplete nature of censored observations. As before, a summary value calculated by combining strata-specific unbiased estimates is unbiased.

To illustrate, two sets of survival data (again subsets of AIDS patients from the SFMHS data) are the following:

For smokers, $n_1 = 8$ observations with $d_1 = 7$ (complete) and $n_1 - d_1 = 1$ (censored -24^+) survival times:

$$\{2, 8, 13, 24^+, 25, 38, 43 \text{ and } 48 \text{ weeks}\}$$

For nonsmokers, $n_0 = 17$ observations with $d_0 = 14$ (complete) and $n_0 - d_0 = 3$ (censored -30^+, 50^+, and 66^+) survival times:

$$\{12, 18, 21, 34, 40, 46, 30^+, 33, 39, 42, 44, 50^+ 56, 58, 61, 66^+ \text{ and } 77 \text{ weeks}\}.$$

A plot of the product-limit estimated survival probability distributions for smokers and nonsmokers displays the comparison to be analyzed (Figure 17.7).

To compare the survival experiences between smokers and nonsmokers, a sequence of 2×2 tables is constructed from the $d = d_1 + d_0 = 7 + 14 = 21$ complete survival times. The general notation for the strata-specific tables at time t_i (ith strata or ith time interval) follows the previous pattern (Table 17.6) where t_i represents a complete survival time (Chapters 6 and 9). To repeat, censored survival times are included in the calculations only when they span the entire interval.

Table 17.7 displays the first 12 strata-specific tables from the smoking/survival data.

Table 17.6 *Notation: Observed* 2×2 *Table for the* i*th-Strata/Interval of the Log-Rank Analysis*

Strata $= t_i$	Death	Alive	Total
Smokers	a_i	b_i	$a_i + b_i$
Nonsmokers	c_i	d_i	$c_i + d_i$
Total	$a_i + c_i$	$b_i + d_i$	n_i

Each of the $d = 21$ data-generated tables is compared to a corresponding table created so that exactly no smoking-survival association exists. The first of these 2×2 tables illustrates the pattern of construction of these 21 "no association" tables (strata $t_1 = 2$ – Table 17.7 and 17.8).

A number of ways exist to create a table so that survival and smoking status are exactly unrelated. One approach requires that the probability of death among smokers to be identical to the probability of death among nonsmokers. In symbols, these two conditional probabilities $P(death \mid smoker)$ and $P(death \mid nonsmoker)$ are illustrated in Table 17.8 where $P(death \mid smoker) = 0.32/8 = P(death \mid nonsmoker) = 0.68/17 = P(death) = 1/25 = 0.04$ (Chapter 6).

An expression to calculate the expected number of smokers who would theoretically die in each strata if no association exists between smoking and survival is

$$A_i = n_i \left[\frac{a_i + b_i}{n_i}\right]\left[\frac{a_i + b_i}{n_i}\right] \text{ or more usually } A_i = \left[\frac{a_i + b_i}{n_i}\right] \text{ when } a_i + c_i = 1.0,$$

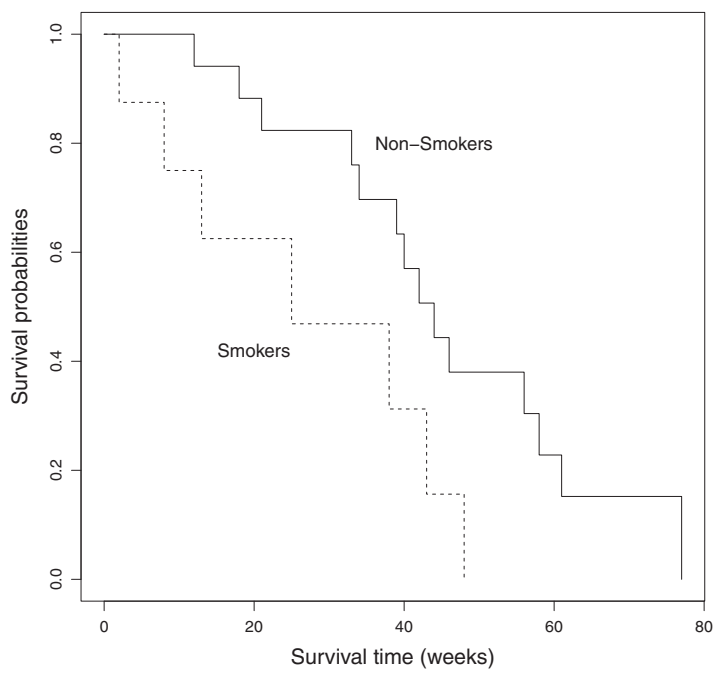


Figure 17.7 The Product-Limit Estimated Survival Probability Distributions – Smokers (Dashed Line) and Nonsmokers (Solid Line) from the SFMHS HIV/AIDS Data

Table 17.7 *Examples: Observed Log-Rank* 2×2 *Tables for First 12 Strata from Smoking/AIDS Data* ($n = 25$, $d = 21$, *and* $n - d = 4$)

	Death	Alive	Total		Death	Alive	Total
$t_1 = 2$				$t_2 = 8$			
Smokers	1	7	8	Smokers	1	6	7
Nonsmokers	0	17	17	Nonsmokers	0	17	17
Total	1	24	25	Total	1	23	24
$t_3 = 12$				$t_4 = 13$			
Smokers	0	6	6	Smokers	1	5	6
Nonsmokers	1	16	17	Nonsmokers	0	16	16
Total	1	22	23	Total	1	21	22
$t_5 = 18$				$t_6 = 21$			
Smokers	0	5	5	Smokers	0	5	5
Nonsmokers	1	15	16	Nonsmokers	1	14	15
Total	1	20	21	Total	1	19	20
$t_7 = 25$				$t_8 = 33$			
Smokers	1	3	4	Smokers	0	3	3
Nonsmokers	0	14	14	Nonsmokers	1	12	13
Total	1	17	18	Total	1	15	16
$t_9 = 34$				$t_{10} = 38$			
Smokers	0	3	3	Smokers	1	2	3
Nonsmokers	1	11	12	Nonsmokers	0	11	11
Total	1	14	15	Total	1	13	14
$t_{11} = 39$				$t_{12} = 40$			
Smokers	0	2	2	Smokers	0	2	2
Nonsmokers	1	10	11	Nonsmokers	1	9	10
Total	1	12	13	Total	1	11	12

calculated from the marginal frequencies of the ith table. For the example ($t_1 = 2$ weeks – Table 17.8), the theoretical number of deaths among smokers in the first interval is $A_1 = 8/25 = 0.320$. The observed value is $a_1 = 1$. All $d = 21$ values of a_i (observed) and A_i (calculated) are presented in Table 17.9.

Table 17.8 *Example: Calculated Log-Rank* 2×2 *Table with Survival Time Exactly Independent of Smoking Status for the First Strata* $t_1 = 2$ *(Table 17.7)*

$t_1 = 2$	Death	Alive	Total
Data			
Smokers	1	7	8
Nonsmokers	0	17	17
Total	1	24	25
Calculated (exactly no association)			
Smokers	0.32	7.68	8
Nonsmokers	0.68	16.32	17
Total	1	24	25

Table 17.9 *Summary Table: Number of Observed Deaths among Smokers* (a_i)*, Variance* ($\hat{v}_i$)*, and Theoretical Number of Deaths among Smokers* (A_i) *Calculated as if Survival Time* (t_i) *and Smoking Status Are Unrelated (No Association)*

	Data						Estimates	
i	t_i	a_i	b_i	c_i	d_i	n_i	A_i	$\hat{v}_i$
1	2	1	7	0	17	25	0.320	0.218
2	8	1	6	0	17	24	0.292	0.207
3	12	0	6	1	16	23	0.261	0.193
4	13	1	5	0	16	22	0.273	0.198
5	18	0	5	1	15	21	0.238	0.181
6	21	0	5	1	14	20	0.250	0.188
7	25	1	3	0	14	18	0.222	0.173
8	33	0	3	1	12	16	0.188	0.152
9	34	0	3	1	11	15	0.200	0.160
10	38	1	2	0	11	14	0.214	0.168
11	39	0	2	1	10	13	0.154	0.130
12	40	0	2	1	9	12	0.167	0.139
13	42	0	2	1	8	11	0.182	0.149
14	43	1	1	0	8	10	0.200	0.160
15	44	0	1	1	7	9	0.111	0.099
16	46	0	1	1	6	8	0.125	0.109
17	48	1	0	0	6	7	0.143	0.122
18	56	0	0	1	4	5	0.000	0.000
19	58	0	0	1	3	4	0.000	0.000
20	61	0	0	1	2	3	0.000	0.000
21	77	0	0	1	1	1	0.000	0.000

The summary statistics from a log-rank analysis are the total number of observed deaths among individuals who smoke ($\sum a_i$) and the corresponding total number of theoretical deaths among individuals who smoke calculated as if no smoking/survival association exists ($\sum A_i$). These two summary values from the smoking data example are $\sum a_i = 7$ deaths observed among smokers and $\sum A_i = 3.539$ calculated deaths among smokers as if no association exists (Table 17.9). The difference yields an increase of $(7 - 3.539) = 3.461$ deaths among smokers. The statistical question becomes: Is the observed increase likely to have occurred by chance? As usual, the answer requires an estimated variance and, in the log-rank case, an estimate of the variance of $\sum a_i$. The variance of the total number of deaths ($\sum a_i$) is estimated by $\hat{V} = variance(\sum a_i) = \sum variance(a_i)$, where

$$variance\,(a_i) = \hat{v}_i = \frac{(a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)}{n_i^2(n_i - 1)}.$$

The variance $\hat{v}_i$ is estimated from each 2×2 table based on the hypergeometric probability distribution (Chapter 6). For example, the estimated variance from the first table is $\hat{v}_1 = [8(17)(1)(24)]/[25^2(24)] = 0.218$ ($t_i = 2$ – first row in Table 17.9).

Three summary values, $\sum a_i = 7$, $\sum A_i = 3.539$ and $\hat{V} = \sum \hat{v}_i = 2.746$, produce an approximate chi-square distributed test statistic that sheds light on the difference between

observed and theoretical numbers of smokers who died. Specifically, the chi-square statistic (degrees of freedom = 1)

$$X^2 = \frac{(\sum a_i - \sum A_i)^2}{\hat{V}} = \frac{(7 - 3.539)^2}{2.746} = 4.363$$

yields the p-value $= P(X^2 \geq 4.363 \mid \textit{no association}) = 0.037$ as if a smoking/survival association does not exist. Thus, the increased number of observed deaths among smokers is not likely due entirely to chance alone.

This comparison of survival experiences is called the log-rank test. Of note, the analysis and chi-square statistic do not involve either logarithms or ranks. The name comes from an entirely different description of an alternative calculation that is not as intuitive as viewing the analysis as a comparison of summary values estimated from of a sequence of strata consisting of 2×2 tables. Also, the variance of a sum equals the sum of the variances only when the values that make up the sum are uncorrelated (Chapter 27). The estimated log-rank variance $\hat{V}$ consists of the sum of estimated values $\hat{v}_i$ calculated from related tables. The fact that the same individuals appear in consecutive tables introduces a dependence among the estimated variances $\hat{v}_i$. Statistical investigations indicate that this sometimes critical requirement of independence does not have an important influence on log-rank test results.

Table 17.10 summarizes the log-rank comparison of survival experience of smokers and nonsmokers, particularly displaying the parallel relationship between the log-rank table ($d = 7 + 14 = 21$ strata) and the compared product-limit estimated survival probability distributions ($d = 21$ complete survival times – Figure 17.7). A last reminder: The log-rank test requires, like product-limit estimation, any censoring of observations to be noninformative.

Proportional Hazards Rates

A popular and frequently effective description of survival data is achieved by comparing hazard functions as a *hazards ratio* (denoted *hr*). In its simplest form, a hazard function describes survival time in terms of a specific rate. For example, survival times t, 0, 3, 5, 9, and 10, and corresponding rates, 0.05, 0.06, 0.08, 0.3, and 1.0, form a discrete hazard function consisting of four values (Figure 17.8, upper left). The same function is displayed for nine survival times (upper right) and with yet more survival times (lower left). The fourth plot (lower right) displays the function based on huge number of survival times producing virtually a continuous relationship between survival time t and a hazard rate, denoted $h(t)$.

An exact hazard rate is not an intuitive statistical measure or simple technically. The formal definition is

$$\text{hazard rate at time } t = h(t) = \frac{-\frac{d}{dt}S(t)}{S(t)}$$

where, as before, $S(t)$ represents a survival probability function.

An approximate hazards ratio is illustrated by the comparison of the U.S. male-female age-specific average mortality rates. The U.S. life table mortality rate for males age 60 to 65 is 1238.0 deaths per year (77,820 – 71,630 = 6190 deaths), and for U.S. females of the same age, the mortality rate is 783.2 deaths per year (87,034 – 83,118 = 3916 deaths). The rate ratio is $1238.0/783.2 = 1.581$. However, during the five-year period approximately 10%

Table 17.10 *Summary Table: Link between Log-Rank Test and Product-Limit (Kaplan-Meier) Estimated Survival Distributions*

	Log-rank statistics						Product-limit probabilities			
	t_i	a_i	b_i	c_i	d_i	n_i	$\hat{p}_1$	$\hat{P}_1$	$\hat{p}_2$	$\hat{P}_2$
1	2	1	7	0	17	25	$7/8 = 0.875$	0.875	–	–
2	8	1	6	0	17	24	$6/7 = 0.857$	0.750	–	–
3	12	0	6	1	16	23	–	–	$16/17 = 0.941$	0.941
4	13	1	5	0	15	22	$5/6 = 0.833$	0.625	–	–
5	18	0	5	1	15	22	–	–	$15/16 = 0.938$	0.882
6	21	0	5	1	14	21	–	–	$14/15 = 0.933$	0.824
7	25	1	3	0	13	18	$3/4 = 0.750$	0.469	–	–
8	33	0	2	1	12	16	–	–	$12/13 = 0.923$	0.760
9	34	0	3	1	11	15	–	–	$11/12 = 0.917$	0.697
10	38	1	2	0	11	14	$2/3 = 0.667$	0.312	–	–
11	39	0	2	1	10	13	–	–	$10/11 = 0.909$	0.633
12	40	0	2	1	9	12	–	–	$9/10 = 0.900$	0.570
13	42	0	2	1	8	11	–	–	$8/9 = 0.889$	0.507
14	43	1	1	0	8	10	$1/2 = 0.500$	0.156	–	–
15	44	0	1	1	7	9	–	–	$7/8 = 0.875$	0.443
16	46	0	1	1	6	8	–	–	$6/7 = 0.857$	0.380
17	48	1	0	0	6	6	$0/1 = 0.000$	0.000	–	–
18	56	0	1	1	4	4	–	–	$4/5 = 0.800$	0.304
19	56	0	1	1	3	3	–	–	$3/4 = 0.750$	0.228
20	61	0	1	1	2	2	–	–	$2/3 = 0.667$	0.152
21	77	0	1	1	0	0	–	–	$0/1 = 0.000$	0.000

Note: $\hat{p}_1$ and $\hat{p}_2$ represent the estimated conditional and $\hat{P}_1$ and $\hat{P}_2$ represent the estimated unconditional survival probabilities.

more females than males who could have died (at-risk) making the comparison misleading (87,034 versus 77,820).

An approximate hazards ratio is

$$\hat{hr} \approx \frac{\text{number of male deaths per approximate person-years}}{\text{number of female deaths per approximate person-years}} = \frac{6190/\frac{1}{2}5[77{,}820 + 71{,}630]}{3916/\frac{1}{2}5\,[87{,}034 + 83{,}118]} = \frac{0.0166}{0.0092} = 1.800.$$

A hazard rate is an instantaneous and theoretical measure of risk that is approximated by an observed average rate of deaths per person-time at risk. Technically, an average approximate mortality rate for a time interval of length δ is

$$\begin{aligned}\text{average approximate mortality rate} = R(t) &= \frac{\text{number of deaths}}{\text{person-years}} \\ &= \frac{\text{proportion of deaths}}{\text{mean person-time-at-risk}} \\ &= \frac{S(t+\delta) - S(t)}{\frac{1}{2}\delta[S(t) + S(t+\delta)]}\end{aligned}$$

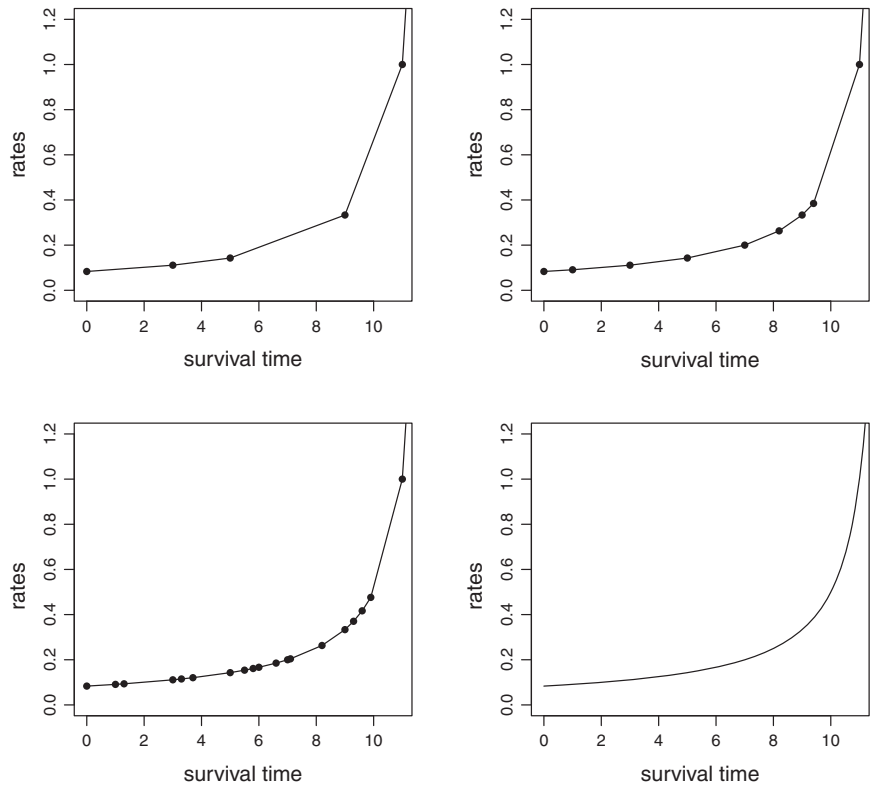


Figure 17.8 Examples of Four Hazard Functions

where $S(t) = P(T \geq t)$ represents again a survival probability function (Chapter 16). The approximate mean person-time-at-risk is the average survival probability $\frac{1}{2}[S(t) + S(t+\delta)]$ for the interval t to $t+\delta$ multiplied by δ.

An average mortality rate converges to the hazard rate as δ becomes small or, in symbols,

$$R(t) = \frac{[S(t+\delta) - S(t)]/\delta}{\frac{1}{2}[S(t) + S(t+\delta)]} \rightarrow \frac{-\frac{d}{dt}S(t)}{S(t)} = h(t).$$

Note: The mathematical definition of the derivative of the function $S(t)$, denoted $\frac{d}{dt}S(t)$, is

$$\frac{d}{dt}S(t) = \frac{[S(t) - S(t+\delta)]}{\delta}$$

as δ becomes small ($\delta \rightarrow 0$). In addition, the expression $\frac{1}{2}[S(t) + S(t+\delta)]$ converges to $S(t)$ as δ becomes small. To summarize, in symbols, as $\delta \rightarrow 0$, then $R(t) \rightarrow h(t)$.

Hazard rates do not exclusively apply to disease and mortality data (Chapter 8). A variety of fields have a "hazard rate." In physics it is called relative velocity, in chemistry it is called a relative rate, in quality control it is called failure rate, in statistics is called an instantaneous rate, and in the early 19th century it was called the force of mortality.

An Example of Survival Analysis

Uniform probabilities of death produce a linear survival function (Chapter 1). Consider a description of uniform mortality risk of women between the ages of 90 and 100. Specifically, the survival probability function is

$$S(t) = 10 - 0.1t \quad 90 \leq t \leq 100.$$

For this linear survival function $S(90) = 1.0$ and $S(100) = 0.0$. The average approximate mortality rate is

$$\begin{aligned} \textit{Rate}(t \textit{ to } t+\delta) &= \frac{S(t) - S(t+\delta)}{\frac{1}{2}\delta[S(t) + S(t+\delta)]} \quad \text{and where } \delta = t_2 - t_1, \text{ then,} \\ &= \frac{0.1(t_2 - t_1)}{\frac{1}{2}(t_2 - t_1)[20 - 0.1(t_1 + t_2)]} = \frac{0.1}{10 - \frac{1}{2}0.1(t_1 + t_2)}. \\ &= \frac{1}{100 - \frac{1}{2}(t_1 + t_2)}. \end{aligned}$$

Two examples of rates are the following:

$$\textit{Rate}(\text{age 96 to 100}) = \frac{1}{100 - \frac{1}{2}(96 + 100)} = 0.50 \quad \text{(age-specific rate) and}$$

$$\textit{Rate}(\text{age 96 to 100}) = \frac{1}{100 - \frac{1}{2}(90 + 100)} = 0.20 \quad \text{(crude rate)}$$

and

$$\text{mean survival time} = \frac{1}{\textit{Rate}\ (\text{age 90 to 100})} = \frac{1}{0.20} = 5 \text{ years}.$$

An average rate increasingly resembles the hazard function as the interval used to estimate the rate becomes smaller. Ultimately, when the difference becomes zero (an instantaneous rate – $\delta = 0$), the rate $R(t)$ becomes the hazard rate $h(t)$. For $\textit{Rate}(t)$, then

$$\textit{Rate}(t) = h\,(t) = \frac{1}{100 - t}$$

becomes the hazard function because $t_2 - t_1 \to 0$ or $t_1 = t_2 = t$.

Note: The conditional probability of death (denoted q) expressed as the ratio of survival functions is

$$q = \frac{S(t) - S(t+\delta)}{S(t)} = \frac{(10 - 0.1t_1) - (10 - 0.1t_2)}{10 - 0.1t_1} = \frac{t_2 - t_1}{100 - t_1}.$$

The relationship between the probability of death q and the hazard function is

$$q = \frac{1}{100 - t_1} \times (t_2 - t_1) = h(t_1) \times (t_2 - t_1)$$

and

$$h(t_1) = \frac{q}{t_2 - t_1} = \frac{q}{\delta}.$$

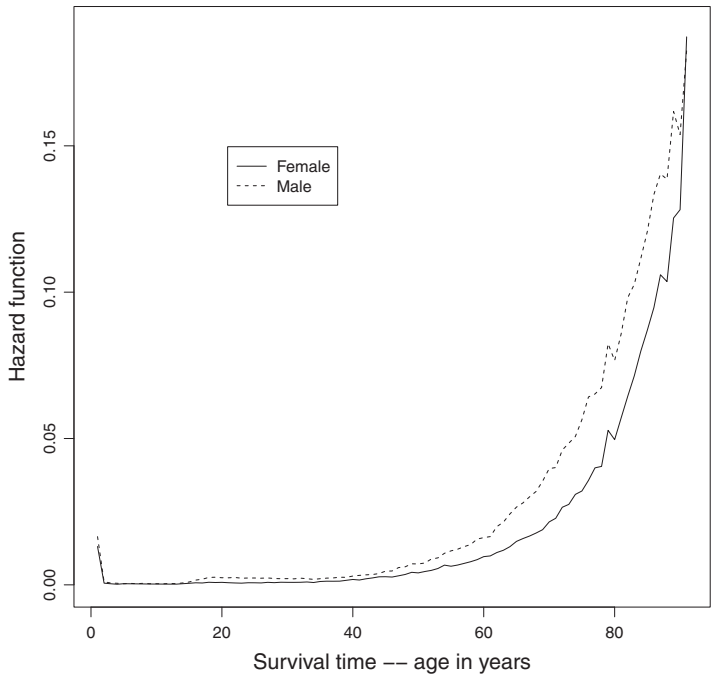


Figure 17.9 Hazard Functions – U.S. Male (Solid Line) and Female (Dashed Line)

Frequently a ratio of two hazard functions is chosen to be the focus of an analysis of survival data. Specifically, for hazard functions denoted $h_1(t)$ and $h_0(t)$, where $h_0(t)$ is sometimes referred to as the "baseline" hazard function, a hazards ratio summary is defined as

$$\text{hazards ratio} = hr = \frac{h_1(t)}{h_0(t)} = c.$$

In this specific case the hazard rates are proportional; that is, the ratio is a constant (represented by c) for all survival times (represented by t). Furthermore, the value of c can be constructed to reflect influences from one or more risk factors. For example, constant c can be made to reflect influences such as ethnicity, age, and sex.

Consider lifetime hazard functions estimated for men and women from U.S. vital records data (Figures 17.9 and 17.10 – year 2000). The male and female approximate hazard functions clearly differ, but it is difficult to get a sense of proportionality from a direct visual comparison. A comparison of the logarithms of two proportional hazard functions yields two lines separated by a constant distance; that is

$$log[hr] = log\left[\frac{h_1(t)}{h_0(t)}\right] = log(h_1[t]) - log(h_0[t]) = log(c),$$

making the distance between plotted lines the constant value $log(c)$ regardless of the survival time. The log-plot comparison of U.S. male and female approximate hazard rates (Figure 17.10) indicates that for age less than 40 years, the hazard functions are certainly not proportional but, after age 40, the male-female hazards ratio appears close to proportional; that is, the plotted log-hazard rates visually appear to be accurately represented by two

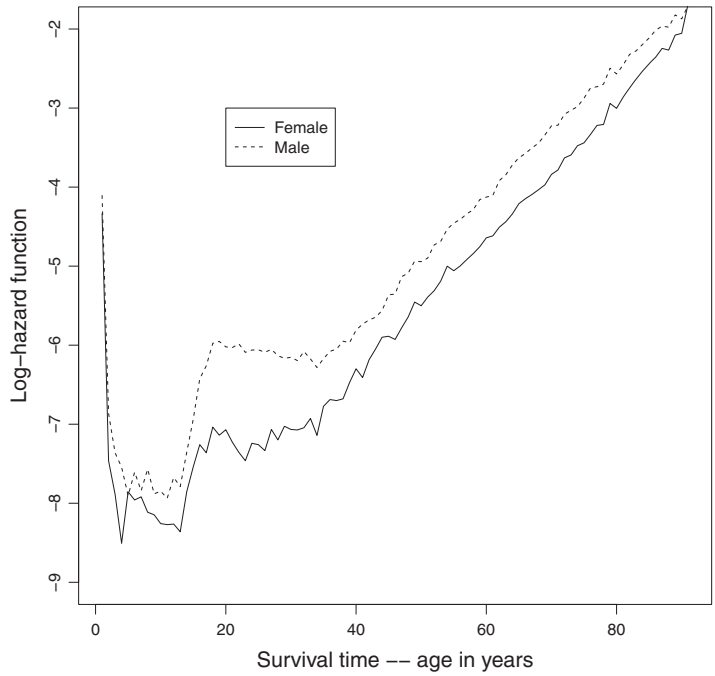


Figure 17.10 Logarithm of Functions – U.S. Male (Solid Line) and Female (Dashed Line)

straight and parallel lines. Thus, the mortality risk among males appears greater than among females for all ages and constant beyond the age 40. Figure 17.10 also indicates a sometimes important property of plotting a hazard function. The details of a hazard function typically become visually more prominent when a logarithmic scale is used.

It is important to emphasize that when hazard functions are proportional, time of survival does not influence the hazards ratio comparison. A constant hazards ratio then becomes a single, direct, and parsimonious summary comparison of risk between sources of survival data. Otherwise, the comparison requires more complicated analytic techniques to produce a useful summary description. In addition, the comparison of rates is traditionally expressed as ratios. From example, a rate of 15 deaths per of 100,000 person-years compared to a rate of 10 deaths per 100,000 person-years is usually reported as 1.5 times larger and not as 5 additional deaths.

A proportional hazards model is designed to describe the influences of specified variables on a hazards ratio. To evaluate the extent of potential influences, a constant of proportionality value c is constructed to be a function of these variables. For example, the proportionality constant could be a linear function of an influence of a variable x such as $c = a + bx$, or, a bit more sophisticated, the value c could be represented as $c = e^{b_0 + b_1 x}$ to evaluate the influence of the variable x on the ratio of two proportional hazard functions in terms of the coefficients b_0 and b_1. A more extensive and often used proportional hazards multivariable model is

$$\text{proportional hazards} = hr = \frac{h_1(t)}{h_0(t)} = c = e^{b_0 + b_1 x_{1j} + b_2 x_{2j} + \cdots + b_k x_{kj}}$$

where the value $log(c)$ is created to be an additive and linear function of k variables each denoted x_{ij} (ith variable and jth observation). Thus, the estimated coefficients b_i succinctly measure the separate influence of each corresponding variable x_i on the multivariate hazards rate ratio.

From another point of view, the logarithm of a proportional hazards ratio [$log(hr)$] yields the more familiar additive and linear function of the k risk variables. Specifically, a log-hazard ratio model becomes

$$\text{log-hazard model} = log[hr] = log\left[\frac{h_1(t)}{h_0(t)}\right] = log(c) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k.$$

Interpretation and inferences from this additive model differ little in principle from most additive models. Typical of regression models, the relationship between the explanatory variables (x_i) and the outcome measure ($log[hr]$ or hr) is created to produce useful interpretations of the estimated coefficients (b_i). This additive proportional hazards model is designed to measure the influence of specific, separate, and multiplicative contributions from each of the components of a multivariate observation in terms of coefficients b_i. Because the model hazards rates are proportional, the estimated components are not influenced by survival time.

Exponentiating the proportionality constant value (e^c) has two primary features. First, hazard ratios are never negative and the value e^c is never negative. Second, the specific influence of each risk variable x_i is measured in terms of a ratio; that is, the overall hazards ratio becomes the product of separate and multiplicative influences. Thus,

$$\textit{proportional hazards ratio} = hr = \frac{h_1(t)}{h_0(t)} = e^c = e^{b_1x_1} \times e^{b_2x_2} \times \cdots \times e^{b_kx_k}.$$

Each additive risk variable increases or decreases the hazards ratio depending only on the value of coefficient b_i and variable x_i regardless of the values of the other variables in the regression model. The model hazards ratio remains proportional as long as the model components continue to be unrelated to survival time. The individual measures of influence are sometimes called *relative hazards ratios*. In symbols, a relative hazards ratio reflecting only the influence of a specific risk variable x_i is $rh_i = e^{b_ix_i}$. The overall hazards ratio is the product $hr = \prod rh_i$. Like all additive models, a multivariate observation is separated into individual measures of the influence associated with each of its components.

The Cox Analysis of Proportional Hazards Models

As might be expected, both parametric and nonparametric statistical approaches are available to analyze survival data based on a regression model. The following describes a nonparametric approach to a *proportional hazards analysis*. This statistical technique is designed to evaluate the influences of a set of risk variables on a hazards ratio based on the ranks of the survival times. The observed survival times are replaced by their ranks, and, as usual, these ranked values produce the analysis. The properties of the population distribution sampled, like all rank strategies, are no longer relevant to the analysis. In more concrete terms, the application to survival data does not require knowledge or assumptions about the properties of the underlying hazard functions, particularly the baseline hazard function $h_0(t)$. Therefore, the results of an analysis based on ranks cannot be used to estimate parameters of

the population sampled. Otherwise, the interpretation of the estimated coefficients from a proportional hazards model analysis is not different from most regression models, but the method to estimate the model coefficients is a unique feature. Using an approach called *conditional likelihood estimation*, statistician David Cox derived a method to estimate model coefficients and their variances without parametric assumptions. Almost unbelievably, these estimates have essentially the same statistical properties as the likelihood estimates used in the analysis of parametric models (Chapters 9, 10, and 27). Nevertheless, critically important requirements remain. The hazard functions must be known or, at least, assumed to be proportional and censoring of survival times must be noninformative.

The simplest proportional hazards model is a two-sample model represented by

$$hr = \frac{h_1(t)}{h_0(t)} = c_i = e^{bFi} \quad \text{or} \quad log[hr] = log\left[\frac{h_1(t)}{h_0(t)}\right] = log(c_i) = bF_i$$

designed to statistically compare two samples of survival time data ($F_1 = 1$ indicates members of one group and $F_0 = 0$ indicates members of another).

Using the previous SFMHS smoking data ($n_1 = 8$ smokers and $n_0 = 17$ nonsmokers – Table 17.9), direct application of the two-sample Cox proportional model yields an estimated regression coefficient $\hat{b} = -1.005$ and its estimated variance $S^2_{\hat{b}} = 0.2315$. Thus, the estimated hazards ratio is $\hat{hr}_i = e^{-1.005F_i}$. A chi-square test statistic to evaluate the influence of random variation on the estimated coefficient $\hat{b} = -1.005$ relative to $b = 0$ (no difference – $hr = 1$) is

$$X^2 = \left[\frac{\hat{b} - 0}{S_{\hat{b}}}\right]^2 = \left[\frac{-1.005 - 0}{\sqrt{0.2315}}\right]^2 = (-2.089)^2 = 4.363$$

and yields the p-value $= 0.037$.

The identical chi-square test statistic results from a log-rank test applied to this same smoking/survival data because these two methods are algebraically identical when no "tied" survival times are present. Although this property of the Cox hazards model is not very useful, the relationship indicates that conditional likelihood estimation of the model coefficients can be viewed as stratifying the ranked data based on the complete survival times to calculate conditional estimates. These conditional estimates, parallel to the log-rank procedure, are combined into a single summary estimate of model coefficients $\hat{b}$ unaffected by noninformative censored survival times.

Proportional Hazards Model – A Case Study

The San Francisco Men's Health Study (SFMHS) includes a cohort of 72 homosexual/bisexual men, ages less than 36 years old, who were diagnosed as seropositive (HIV-positive subjects) and entered into a special study (July 1984 to December 1987). The "survival time" for these study participants is time from entry into the study until diagnosis of AIDS or the end of the study period (weeks). The end point is diagnosis of AIDS not death, but, nevertheless, the length of this AIDS-free interval is referred to as survival time for consistency of terminology. This SFMHS cohort produced 43 complete (AIDS cases) and 29 censored observations (AIDS-free) at the end of the 42-month study period ($n = 72$). Three factors potentially relating to the prognosis of AIDS are explored: *cd*4-lymphocyte

Table 17.11 *Proportional Hazards Model: Survival Times Influenced by cd*4 *Counts from SFMHS Cohort* ($n = 72$)

	Estimate	*s. e.*	*p*-value
*cd*4	−0.003	0.00082	<0.001
Conditional log-likelihood = −154.673			

count (*cd*4), serum β-microglobulin levels (β), and the subject's age (*age*). The variables *cd*4 counts and β measurements reflect immune response to infection.

A good place to start the search for both an accurate and parsimonious proportional hazards model is an analysis based on only *cd*4 counts. The log-hazards ratio version of the Cox proportional analysis is then

$$log[hr_i(t)] = b_1 cd4_i,$$

directly using the $n = 72$ observed *cd*4 counts as reported. The conditional likelihood estimate of the model coefficient $\hat{b}_1$ is −0.003 (Table 17.11).

The model estimated hazards ratio per *cd*4 count is then $\hat{hr} = e^{-0.003} = 0.997$ or, for a difference of 100 in *cd*4 counts, $\hat{hr} = 0.997^{100} = 0.773$ $(1/0.733 = 1.364)$. Notice that the negative coefficient indicates increasing risk associated with decreasing *cd*4 counts. The price of applying a model without a parametric structure is that the Cox conditional estimate of the hazards ratio does not produce estimates of the hazard functions. From a mechanical point of view, the Cox hazards model cannot be used to estimate a constant baseline risk that is possible for parametric regression models (Chapter 18). That is, the usual regression model constant term denoted previously b_0 cannot be estimated (Chapter 7). Nevertheless, evaluation of the influence of random variation on the estimated *cd*4-coefficient parallels the usual approach. The approximate normal probability distribution based test statistic

$$z = \frac{\hat{b}_1 - 0}{S_{\hat{b}1}} = \frac{-0.0031 - 0}{0.00082} = -3.782$$

produces the associated *p*-value of $P(Z \leq -3.782 \mid b = 0) < 0.001$.

Three important statistical questions are relevant. First, can a better statistical representation of the *cd*4 counts be found? Second, are the SFMHS data adequately described by a model that requires the hazards ratio be proportional? Third, what are the additional influences from the variables β-microglobulin and age?

The estimated *cd*4 model requires the log-hazards ratio to linearly increase as the *cd*4 count decreases ($b_1 < 0$). Stratifying the HIV/AIDS data into three categories based *cd*4 counts allows a direct comparison of three estimated coefficients that possibly identify other patterns of risk. For example, a nonlinear pattern would likely produce unequal coefficients for one or more categories. Dividing the study subjects into three strata based on *cd*4 levels (<600, 600 to 900, and >900 *cd*4 counts) and again using the hazards model $log[hr_i(t)] = b_1 cd4_i$ produces the three separate analyses (Table 17.12).

The three estimated coefficients, despite wide 95% confidence intervals, indicate the accuracy of the hazard function model is likely improved by a nonlinear representation of the influence of the *cd*4 counts. Keeping in mind that separate analyses substantially reduce

Table 17.12 *Proportional Hazards Model: Survival Times Influenced by cd4 Counts* ($n = 72$) *from Analysis of SFMHS Data – cd4 Counts Stratified into* intervals <600, 600 *to* 900, *and* >900

	n	Estimates	*s. e.*	p-values	Lower[a]	Upper[a]
cd4: <600	31	-0.006	0.002	<0.001	-0.009	-0.003
cd4: 600 to 900	26	-0.003	0.004	0.440	-0.011	0.005
cd4: >900	15	-0.001	0.003	0.720	-0.007	0.005

[a]Approximate 95% confidence interval bounds.

the sample size and, consequently, increase variability, nevertheless, the estimates compared among strata can suggest additional features of the survival data.

To further explore nonlinear alternative representations of the influence of the *cd*4 variable, a sequence of polynomial models provides an evaluation of different patterns:

1. linear: $log[hr_i(t)] = b_1 cd4_i$
2. *quadratic*: $log[hr_i(t)] = b_1 cd4_i + b_2(cd4_i)^2$ and
3. *cubic*: $log[hr_i(t)] = b_1 cd4_i + b_2(cd4_i)^2 + b_3(cd4_i)^3$

Comparisons of conditional log-likelihood values generated from each polynomial representation of the influence of the *cd*4 counts indicates the potential utility of a nonlinear relationship. Contrasting nested conditional log-likelihood values from the HIV/AIDS survival data (Table 17.13) shows a substantially larger log-likelihood value relative to the linear model when the influence of the *cd*4 counts on the hazards ratio is described by a quadratic relationship (model 2).

Although the log-likelihood comparisons clearly show a strong influence from including a quadratic term, a direct analysis displays the details. Specifically, for the quadratic model

$$log[hr_i(t)] = b_1 cd4_i + b_2(cd4_i)^2,$$

the estimated coefficients and their variances are given in Table 17.14.

The fact that a quadratic model is an improvement over a linear model does not address the always present question of proportionality of the ratio of hazard functions and the requirement of proportionality is crucial. To evaluate this issue, a simple strategy is to create a survival model that contains additional terms that allow the possibility of a relationship between risk variables and survival time, a nonproportional hazards model. The evaluation of proportionality then becomes an assessment of the influence of these time-dependent terms. When the included terms have negligible influence, their addition to the model produces little evidence that the influence of the *cd*4 counts depends on survival time. In other words,

Table 17.13 *Proportional Hazards Model: Survival Times Influenced by cd4 Counts* ($n = 72$) *from SFMHS Data – Conditional Log-Likelihood Values from Nested Models (Linear, Quadratic, and Cubic Polynomials)*

	Log-likelihoods	Chi-square statistics	p-value
Linear	-154.67	–	–
Quadratic	-150.84	7.671	0.022
Cubic	-150.16	1.350	0.245

Table 17.14 *Proportional Hazards Model: Quadratic Influence of* cd*4 Counts from SFMHS Data* ($n = 72$)

	Estimates	*s. e.*	*p*-values
*cd*4	-0.0096	0.0022	–
$(cd4)^2$	4.60×10^{-6}	1.43×10^{-6}	0.001
Conditional log-likelihood $= -150.837$			

the requirement of proportionality does not appear unreasonable. The opposite is true. When these specialized interaction terms are important components of the survival model, it is likely that the model without these terms is not an adequate representation of a proportional *cd*4/risk relationship. Table 17.15 contains the results of an evaluation of the assumption of proportionality of the quadratic *cd*4 model (Table 17.14). The specific model containing the time/*cd*4 count interaction terms is

$$log[hr_i(t)] = b_1 cd4_i + b_2(cd4_i)^2 + b_3[log(t) \times cd4_i] + b_4[log(t) \times (cd4_i)^2].$$

The conditional log-likelihood ratio comparison of the quadratic *cd*4 model with (Table 17.14) and without (Table 17.15) survival time terms included is (Chapter 10)

$$X^2 = 2[L_{interaction} - L_{no\ interaction}] = 2\,[-150.602 - (-150.837)] = 0.476.$$

The test statistic X^2 has an approximate chi-square distribution (degrees of freedom $= 2$) when $b_3 = b_4 = 0$, producing a p-value of $P(X^2 \geq 0.476 \,|\, proportional) = 0.788$. The degrees of freedom, as usual, are the difference between the number of parameters used to define the compared models. The conditional log-likelihood contrast provides no persuasive evidence of nonproportionality. It is important to note that special but readily available computer software is necessary to estimate the model coefficients and conditional log-likelihood values for this extended Cox analysis of the survival time data.

The Cox analysis with two additional risk variables, β-microglobulin and age, follows the pattern of multivariable regression models in general. A three variable additive model is

$$log[hr_i(t)] = b_1 cd4_i + b_2 \beta_i + b_3 age_i.$$

The analytic results are presented in Table 17.16. The three-variable model postulates that the influences of *cd*4 counts, β-microglobulin, and *age* are additive, and, as always for regression models, the assumption of additivity of the independent variables is a key issue.

Table 17.15 *Proportional Hazards Model: Survival Times Influenced by* cd*4 Counts from SFMHS Data –Evaluation of Proportionality* ($n = 72$)

	Estimates	*s. e.*	*p*-values
*cd*4	-0.0115	0.0058	–
$(cd4)^2$	3.97×10^{-6}	5.42×10^{-6}	–
$log(t) \times cd4$	0.0011	0.0017	0.531
$log(t) \times (cd4)^2$	-1.63×10^{-7}	-1.41×10^{-6}	0.908
Conditional log-likelihood $= -150.602$			

Table 17.16 *Proportional Hazards Model: Survival Times Influenced by* cd*4 Counts, β-Microglobulin, and Age from SFMHS Data* ($n = 72$)

	Estimates	*s. e.*	*p*-values
*cd*4	−0.003	0.001	0.003
β	0.366	0.162	0.022
age	−0.016	0.037	0.665
Conditional log-likelihood = −151.895			

The question of additivity is effectively evaluated by comparing the conditional log-likelihood value from the additive model to the conditional log-likelihood value from the model with all pairwise interaction terms included (Chapter 10). The specific interaction proportional hazards model is

$$log[hr_i(t)] = b_1 cd4_i + b_2\beta_i + b_3 age_i + b_4[cd4_i \times \beta_i] + b_5[cd4_i \times age_i] + b_6[\beta_i \times age_i].$$

The details of the analysis of this nonadditive multivariable model are contained in Table 17.17.

Applying a log-likelihood ratio comparison (Tables 17.16 and 17.17), the test statistic

$$X^2 = 2[L_{\text{interaction}} - L_{\text{no interaction}}] = 2[-150.753 - (-151.895)] = 2.285$$

has an approximate chi-square distribution (degrees of freedom = 3), yielding a p-value of $P(X^2 \geq 2.285 \mid b_4 = b_5 = b_6 = 0) = 0.515$. Thus, the conditional log-likelihood comparison of additive versus interaction models produces no statistical evidence that the additive model is not an accurate representation of the relationships within the HIV/AIDS survival data.

The question of proportionality is explored, as before, by comparing the additive model with *time*/*risk*-interaction terms included (Table 17.18) to the model without these terms (Table 17.16). The specific survival time model with the *time*/*risk*-interaction terms included is

$$log[hr_i(t)] = b_1 cd4_i + b_2\beta_i + b_3 age_i + b_4[log(t) \times cd4_i] + b_5[log(t) \times \beta_i] + b_6[log(t) \times age_i].$$

This extended model produces the conditional likelihood estimated coefficients associated with proportionality (b_4, b_5, and b_6 – Table 17.18).

Table 17.17 *Proportional Hazards Model: Survival Times Influenced by* cd*4 Counts, β-Microglobulin, and Age from the SFMHS Data* ($n = 72$)

	Estimates	*s. e.*	*p*-values
cd_4	−0.0047	0.0073	–
β	0.6088	1.5739	–
age	−0.1426	0.2263	–
$cd_4 \times \beta$	−0.0009	0.0009	0.284
$cd_4 \times age$	0.0001	0.0002	0.473
$\beta \times age$	0.0112	0.0477	0.801
Conditional log-likelihood = −150.753			

Table 17.18 *Proportional Hazards Model: Survival Times Influenced by cd*4 *Counts, β-Microglobulin, and Age from SFMHS Data – Assessment of Proportionality* ($n = 72$)

	Estimates	*s. e.*	*p*-values
*cd*4	−0.0073	0.0029	–
β	0.4202	0.6620	–
age	−0.1451	0.1428	–
$log(t) \times cd_4$	0.0013	0.0007	0.084
$log(t) \times \beta$	−0.0240	0.1812	0.894
$log(t) \times age$	0.0387	0.0040	0.332
Conditional log-likelihood = −157.197			

Again, comparison of conditional log-likelihood values provides a useful assessment of the influence of the survival time–dependent interaction terms relative to an additive proportional hazards model. Specifically, the Cox conditional log-likelihood ratio test statistic

$$X^2 = 2[L_{interaction} - L_{no\ interaction}] = 2[-157.197 - (-151.895)] = 5.302$$

has an approximate chi-square distribution (degrees of freedom = 3) and associated *p*-value if $P(X^2 \geq 2.285 \mid b_4 = b_5 = b_6 = 0) = 0.151$.

Therefore, supported by evidence of a serviceable additive and proportional hazards multivariable model, *cd*4 counts appear to be a strong predictor of a diagnosis of AIDS as well as β-microglobulin, to a lesser extent, while age likely does not play an important prognostic role (Table 17.16).

Notice that the variables β-microglobulin and age do not confound the *cd*4 influence. The *cd*4-coefficient $\hat{b}_1 = -0.003$ is essentially the same for the model excluding (Table 17.11) and including (Table 17.16) these two variables. Although influences of *cd*4 counts, β-microglobulin, and age are well established today, when these data were collected and analyzed (circa 1990), details of their response to HIV infection was an important issue.

18

The Weibull Survival Function

Student's *t*-test is among the foremost achievements of early statisticians. It is described in most introductory statistical textbooks. An equally important contribution frequently ignored, Student opened the door to the statistical analysis of small data sets when specific properties describe the population sampled. His *t*-test allowed a statistical analysis based on a single mean value regardless of the sample size. The technique of condensing a sample of observations into a few summary values and focusing on these values to evaluate and describe the properties of collected data today is so routine that it is not given a second thought.

The Weibull parametric survival model is a continuation of the use of a few statistical summary values to analyze and describe collected data based on known or theoretical properties of the population sampled. The model, popularized by Waloddi Weibull (b. 1887), is a two-parameter representation of the relationship between independent variables and the time to an occurrence of a specific outcome. This parametric survival time model has been applied to investigate a wide range of time-to-failure data, such as the lifetime of light bulbs, fatigue failures of metals, time to relapse of a disease, reliability of engine parts, and, as will be described in detail, time to death or to occurrence of disease.

Somewhat analogous to normal distribution parameters mean value μ and variance σ^2, the Weibull distribution is defined by a location parameter (denoted λ) and a shape parameter (denoted γ). Three expressions for this two-parameter model of survival time (denoted t) are

1. The survival probability $= S(t) = P(T \geq t) = e^{-\lambda t^{\gamma}}$,
2. The hazard function $= h(t) = \lambda\gamma t^{\gamma-1}$ and
3. The log-cumulative hazard function $= log[H(t)] = log(\lambda) + \gamma log(t)$.

Any one of the three expressions dictates the other two. For example, analogous to the nonparametric version (Chapter 17), the cumulative hazard function is defined as $H(t) = -log[S(t)]$, and, for the Weibull model,

$$log[H(t)] = log(-log[S(t)]) = log(\lambda t^{\gamma}) = log(\lambda) + \gamma\, log(t)$$

is a linear function of $log(t)$ (intercept $= log(\lambda)$ and slope $= \gamma$).

Examples of the Weibull survival and hazard functions are displayed in Figures 18.1 and 18.2. The parametric Weibull hazard function is a direct application of the general definition

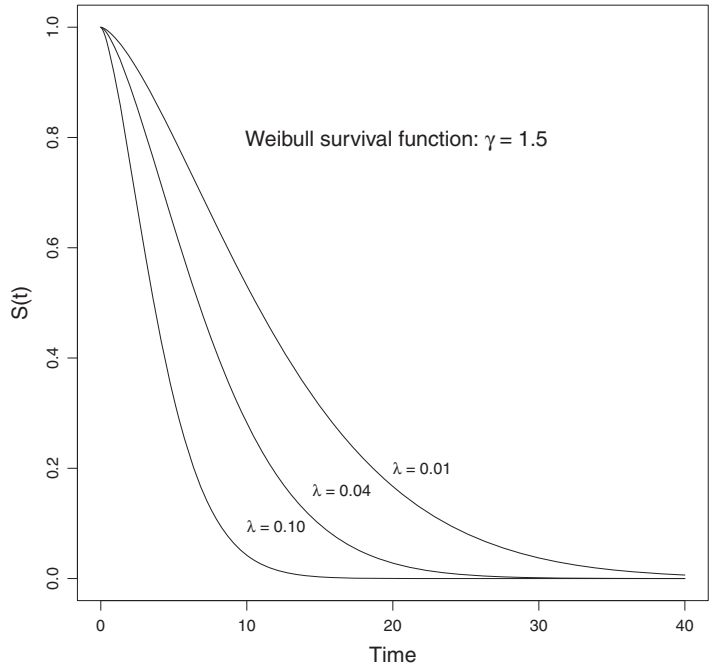


Figure 18.1 Three Weibull Survival Functions $S(t)$ with Location Parameters $\lambda = \{0.01, 0.04 \text{ and } 0.10\}$ for a Shape Parameter $\gamma = 1.5$

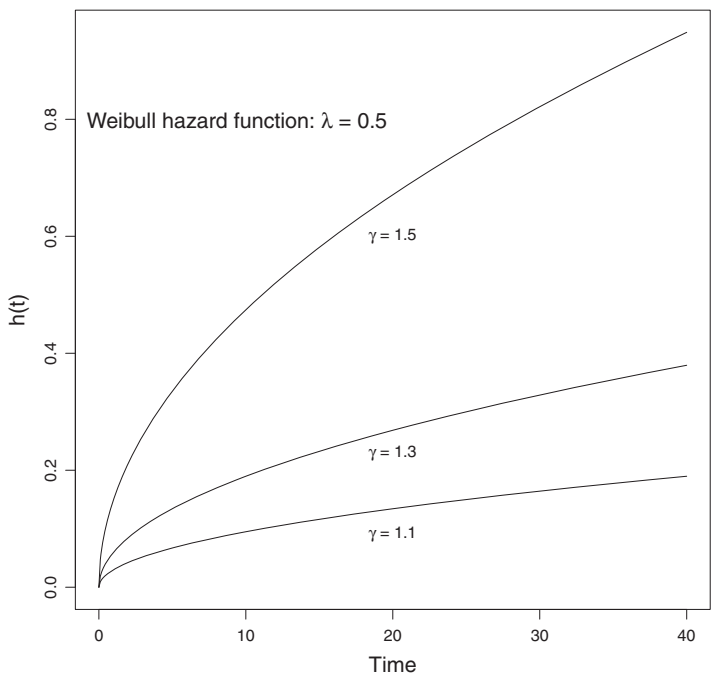


Figure 18.2 Three Weibull Hazard Functions $h(t)$ with Shape Parameters $\lambda = \{1.1, 1.3 \text{ and } 1.5\}$ for a Location Parameter $\lambda = 0.05$

Table 18.1 *Data: Simulated Data to Illustrate Properties of a Weibull Survival Model* ($n = 50 - \lambda = 0.01$, *and* $\gamma = 1.5$)

12.90	2.42	30.35	24.43	13.48	19.49	15.84	8.56	27.15	2.01
35.80	49.89	8.66	17.35	3.24	7.84	9.41	3.91	16.59	37.48
24.06	8.08	14.91	6.93	9.46	37.71	20.43	8.06	23.39	4.68
18.11	2.45	14.78	25.88	10.47	9.38	9.62	24.17	20.85	6.26
22.09	32.77	39.17	43.04	6.69	4.92	8.54	10.83	5.56	6.43

(Chapter 17) and is the derivative of the survival function $S(t)$ with respect to t divided by $S(t)$ or, in symbols,

$$h(t) = \frac{-\frac{d}{dt}\, S(t)}{S(t)} = \frac{-\frac{d}{dt}\, e^{-\lambda t^{\gamma}}}{e^{-\lambda t^{\gamma}}} = \lambda\gamma t^{\gamma-1}.$$

The shape of the Weibull model hazard function is determined by the parameter γ. For $\gamma > 1$, the hazard function is an increasing function of survival time (Figure 18.2). For $\gamma < 1$, the hazard function is a decreasing function of survival time. When $\gamma = 1$, the hazard function is constant ($h(t) = \lambda$), making the exponential survival distribution a special case of the Weibull survival distribution (Chapter 16). In symbols, for $\gamma = 1$, then

$$S(t) = e^{-\lambda t^{\gamma}} = e^{-\lambda t} \quad \text{and} \quad h(t) = \lambda\gamma t^{\gamma-1} = \lambda.$$

To describe the role of the two Weibull model parameters in concrete terms, a sample of 50 simulated random survival times from a Weibull distribution with location parameter $= \lambda = 0.01$ and shape parameter $= \gamma = 1.5$ illustrate. The inverse function of the Weibull survival time distribution $S(t)$ is

$$t = S^{-1}(p) = \left[\frac{-log(p)}{\lambda}\right]^{1/\gamma} = \left[\frac{-log(p)}{0.01}\right]^{0.667}.$$

The inverse function and $n = 50$ random uniform probabilities p generate $n = 50$ Weibull distributed random survival times based on parameter values $\lambda = 0.01$ and $\gamma = 1.5$ (Table 18.1) (Chapter 12).

To get a sense of these survival time data, a few summary values are given in Table 18.2 and the distribution of the data is displayed in Figure 18.3.

The maximum likelihood estimates of the parameters λ and γ are the solutions to a nonlinear equation that is only efficiently solved with a computer algorithm. The computer estimates based on the $n = 50$ random Weibull survival times are $\hat{\lambda} = 0.055$ and $\hat{\gamma} = 1.453$.

Table 18.2 *Data: Summary Statistics from* $n = 50$ *Randomly Generated Survival Times with a Weibull Distribution* ($\lambda = 0.01$ *and* $\gamma = 1.5$)

Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
2.01	7.90	13.19	16.50	23.89	49.90

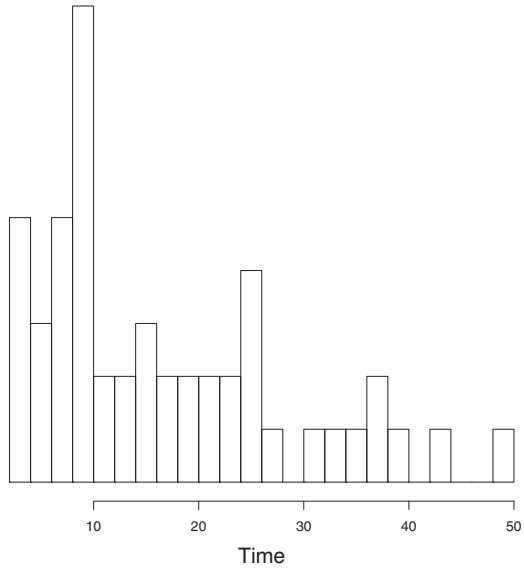


Figure 18.3 Distribution of Simulated Weibull Data ($n = 50$)

Therefore, the specific estimates of the three previous expressions that characterize survival time summarized by a Weibull probability model are the following:

$$\hat{S}(t) = e^{-0.055t^{1.453}},$$

$$\hat{h}(t) = (0.055)1.453t^{0.453}$$

and

$$log[\hat{H}(t)] = log(0.055) + 1.453\, log(t).$$

As part of maximum likelihood estimation, the estimate $log(\hat{\gamma}) = log(1.\,453) = 0.373$ and its standard error of $S_{log(\gamma)} = 0.110$ are typically calculated. These estimates allow an assessment of the utility of the shape parameter. From the simulated data, the test statistic

$$X^2 = \left[\frac{log(\hat{\gamma}) - log(1)}{\sqrt{variance(log[\hat{\gamma}])}}\right]^2 = \left[\frac{0.373 - 0}{0.110}\right]^2 = (3.401)^2 = 11.564$$

has an approximate chi-square distribution (degrees of freedom = 1) when $\gamma = 1$ and yields a p-value of less than 0.001. Thus, the much simpler case of a constant hazard function is not helpful in light of substantial evidence that the shape parameter indicates an increasing hazard function ($\gamma \neq 0$).

Another approach to the same issue is the comparison of exponential ($\gamma = 1$) and Weibull ($\gamma \neq 1$) survival time distributions with likelihood statistics. The estimation of the parameter

λ from an exponential survival model using the simulated data is $\hat{\lambda} = 0.060$ with the associated log-likelihood value of -190.260 (Chapter 16). Note that $\bar{t} = 16.530$ making $\hat{\lambda} = 1/16.530 = 0.060$ (Table 18.1). The estimates $\hat{\lambda} = 0.055$ and $\hat{\gamma} = 1.453$ from the Weibull survival model using the same data produce a log-likelihood value of -185.415. The likelihood ratio test statistic (degrees of freedom = 1)

$$X^2 = 2[-185.415 - (-190.260)] = 9.690$$

yields the p-value of 0.002. Both test statistics are approximate, but the likelihood ratio test is generally the more accurate estimate of a chi-square distributed variable, particularly in the case of small p-values.

The log-cumulative hazard function presents an opportunity to create a simple visual comparison between observed survival data and values from the estimated Weibull model. The first step is an estimate of the cumulative probability distribution $F(t)$ from the sampled survival times. The distribution-free estimate is $\hat{F}(t_i) = i/n$ where the survival times t_i are ordered from smallest to largest (Chapter 12). The next step is to use $\hat{F}(t)$ to estimate the survival function or set $\hat{S}(t) = 1 - \hat{F}(t)$. The final step is to note that the expression of the log-cumulative Weibull hazard function is a simple linear equation of the form $y = a + bx$; that is, the Weibull model log-cumulative hazards function $log[H(t)] = log(-log[S(t)]) = log(\lambda) + \gamma log(t)$ is a straight line function of $log(t)$ with intercept $a = log(\lambda)$ and slope $b = \gamma$.

Applying ordinary least squares estimation to the x/y-pairs of values, $x = log(t_i)$ and $y = log(-log[\hat{S}(t_i)])$ yields a straight line based exclusively on data-generated survival times. Using the simulated example data, the estimated intercept is $\hat{a} = -2.859$ and the estimated slope is $\hat{b} = 1.504$. The survival data estimated straight line $log[\hat{H}(t)]$ is $-2.859 + 1.504log(t)$ (Figure 18.4, dashed line). The Weibull model log-cumulative hazard function is the estimated straight line

$$log[\hat{H}(t_i)] = log(\hat{\lambda}) + \hat{\gamma} log(t_i) = log(0.055) + 1.453log(t_i) = -2.908 + 1.453log(t_i)$$

using the estimated parameters from the simulated data (Figure 18.4, solid line) (Chapter 12).

The data-generated least squares estimates have larger variances than estimated maximum likelihood model values but provide a model-free estimate of a log-cumulative hazard function. Comparison of two straight lines, one based only on data (dashed line) and the other based on a parametric Weibull log-cumulative hazard function (solid line), provides a visual assessment of the adequacy of the Weibull model to represent sampled survival data (Figure 18.4). The lack of difference generated by the artificially generated Weibull survival time "data" is, of course, expected.

Figures 18.5 and 18.6 contrast the survival and hazard functions based on the parametric maximum likelihood estimates ($\hat{\lambda} = 0.055$ and $\hat{\gamma} = 1.453$), the data-generated linear regression estimates ($\hat{\lambda} = e^{\hat{a}} = e^{-2.859} = 0.057$ and $\hat{\gamma} = \hat{b} = 1.504$), and the model values ($\lambda = 0.01$ and $\gamma = 1.5$).

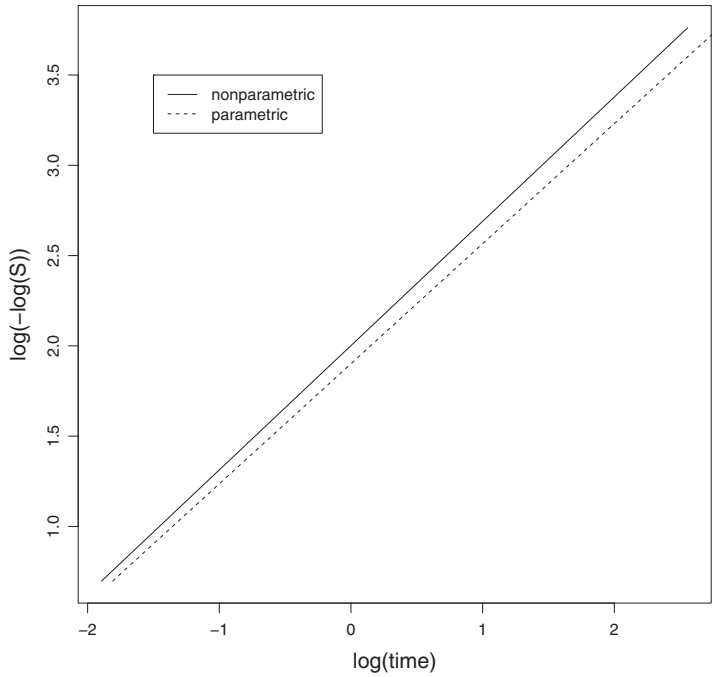


Figure 18.4 Goodness-of-Fit Plot: Model (Solid Line) and Data (Dashed Line) Estimates from the Simulated Survival Time Data (Table 18.1 – $n = 50$)

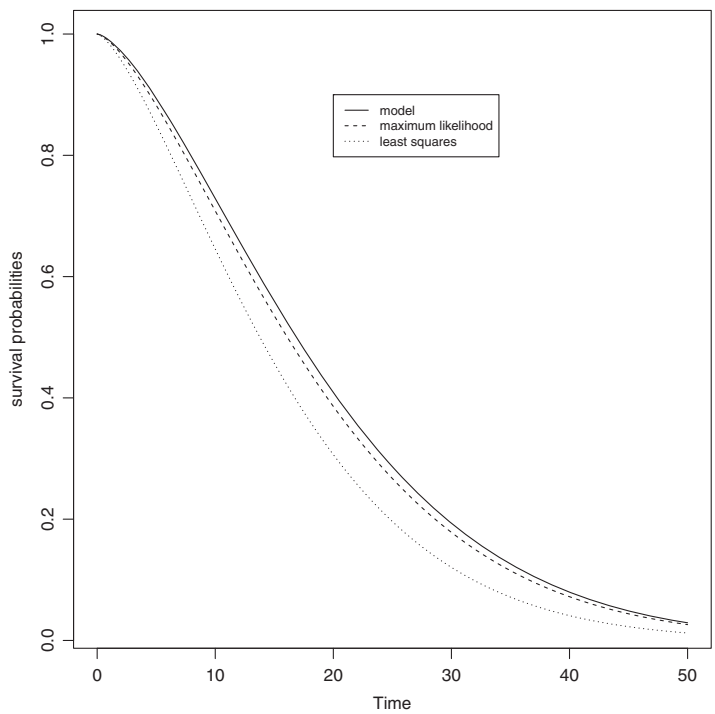


Figure 18.5 Estimation: Survival Functions Based on the Parameter Values, the Maximum Likelihood Estimates and the Linear Regression Estimates (Table 18.1 – $n = 50$)

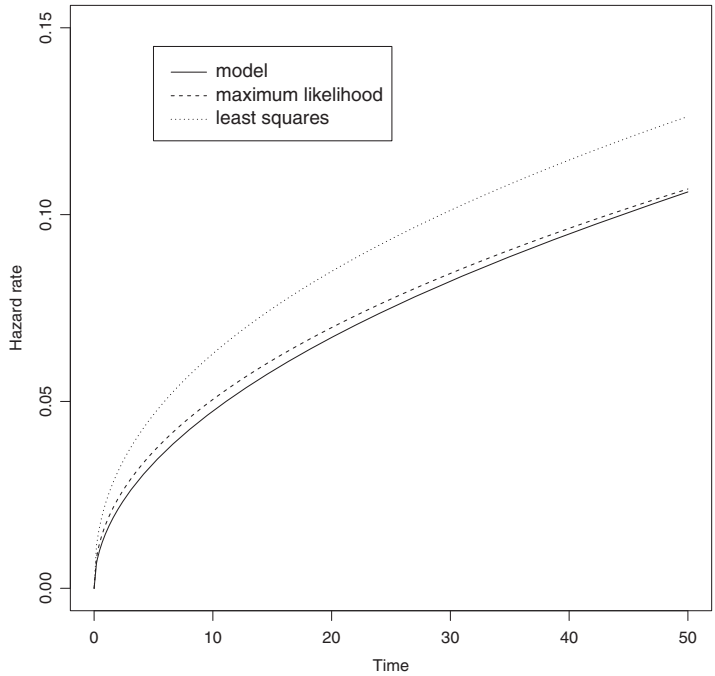


Figure 18.6 Estimation: Hazard Functions Based on the Parameter Values, the Maximum Likelihood Estimates, and the Linear Regression Estimates (Table 18.1 – $n = 50$)

Two-Sample Comparison – The Weibull Model

The comparison of survival experience between two samples of data is frequently a comparison between two proportional hazard functions (Chapter 17); that is, from a sample denoted zero and a sample denoted one, such a comparison is

$$\text{hazards ratio} = hr = \frac{h_1(t)}{h_0(t)} = c \quad \text{or} \quad h_1(t) = h_0(t) \times c.$$

For the Weibull two-sample model, the compared hazard functions are represented by $h(t|F) = \lambda_F \gamma t^{\gamma - 1}$ where $F = 0$ for one sample and $F = 1$ for the other sample. The scale parameter $\lambda_F$is defined as $\lambda_F = e^{[b_0 + b_1 F]\gamma}$. The scale parameter is explicitly constructed to create a Weibull model representation of the hazards ratio with three properties:

1. The scale parameter λ_F produces a hazards ratio that does not depend on survival time. That is, the ratio of Weibull hazard survival functions is proportional ($hr = c$).
2. The scale parameter λ_F is always positive, and, therefore, the model estimated hazards ratio is always positive.
3. The influences measured by coefficients b_0 and b_1 produce multiplicative assessments of two separate components of the hazard function designed to identify differences in survival experience between groups. In symbols,

$$\text{total risk} = \text{(baseline risk)} \times \text{(additional risk)} = e^{[b_0 + b_1]\gamma} = e^{b_0 \gamma} \times e^{b_1 \gamma}.$$

Table 18.3 *Analysis: Weibull Model Estimated Parameters from* $n = 174$ *AIDS Subjects to Compare Survival Times between Smokers and Nonsmokers*

Parameters	Estimates	*s. e.*	*p*-values
b_0	−3.240	–	–
b_1	0.135	0.139	0.331
$log(\gamma)$	0.149	0.062	0.017
log-likelihood (L) = −638.605			

The consequence of these three properties of scale parameter λ_F is a simple description of the Weibull model hazards ratio given by the expressions

$$F = 0\text{: } h(t|F = 0) = e^{b_0\gamma} \times \gamma\, t^{\gamma-1},$$
$$F = 1\text{: } h(t|F = 1) = e^{[b_0+b_1]\gamma} \times \gamma\, t^{\gamma-1}$$

and, therefore,

$$\text{hazards ratio} = hr = \frac{h(t|F = 1)}{h(t|F = 0)} = \frac{e^{[b_0+b_1]\gamma} \times \gamma\, t^{\gamma-1}}{e^{b_0\gamma} \times \gamma\, t^{\gamma-1}} = e^{b_1\gamma}.$$

Data from the San Francisco Men's Health Study again provide survival times (weeks) for $n = 174$ HIV/AIDS subjects divided into groups of 80 men who do not smoke with 71 deaths associated with AIDS and 9 censored observations and 94 men who do smoke with 84 deaths associated with AIDS and 10 censored observations. Summarized by the Weibull two-sample model, the survival experiences of smokers and nonsmokers are succinctly compared. The maximum likelihood computer-generated estimates of the Weibull model coefficients b_0, b_1, and γ produce an evaluation and description of the difference in survival experience (Table 18.3). The influence of the censored survival times is accounted for by computer estimation software when the censoring is noninformative.

Perhaps a good place to start is an evaluation of evidence that the shape parameter substantially contributes to the two-parameter Weibull model ($\gamma \neq 1$) (Chapter 16). Using the AIDS data, an exponential survival model produces an estimate of the same model but with $\gamma = 1$ (Table 18.4). Key, as usual, to contrasting these two models is their accompanying log-likelihood values (Table 18.3 and Table 18.4).

Table 18.4 *Analysis: Exponential Model Estimated Parameters from* $n = 174$ *AIDS Subjects to Compare Survival Times between Smokers and Nonsmokers*

Parameters	Estimates	*s. e.*	*p*-value
b_0	−3.210	–	–
b_1	0.135	0.161	0.404
log-likelihood (L) = −641.216			

The likelihood comparison ($\gamma \neq 1$ versus $\gamma = 1$) yields the approximate chi-square distributed likelihood ratio test statistic

$$X^2 = 2[-638.605 - (-641.216)] = 5.222$$

and a p-value of 0.022 (degrees of freedom = 1). It appears unlikely that the value of the estimated Weibull shape parameter results only from capitalizing on random variation and usefully increases the accuracy of the Weibull model comparison of the estimated survival times between smoker and nonsmoker. Technically, the difference in the log-likelihood values is not likely due to chance alone.

The estimated shape parameter is, $\hat{\gamma} = e^{log(\hat{\gamma})} = e^{0.149}\, 1.160$ and the estimated scale parameter becomes

$$\hat{\lambda}_F = e^{[\hat{b}_0 + \hat{b}_1 F]\hat{\gamma}} = e^{[-3.240 + 0.315F]1.160}.$$

Values of the scale parameters are, therefore, for nonsmokers ($F = 0$), $\hat{\lambda}_0 = e^{-3.240(1.160)} = 0.0233$, and, for smokers ($F = 1$), $\hat{\lambda}_1 = e^{[-3.240 + 0.135](1.160)} = 0.0273$. The model estimated hazards ratio is

$$\hat{hr} = \frac{h(t|F=1)}{h(t|F=0)} = \frac{\hat{\lambda}_1}{\hat{\lambda}_0} = \frac{0.0273}{0.0233} = 1.170 \quad \text{or} \quad \hat{hr} = e^{\hat{b}_1 \hat{\gamma}} = e^{-0.135(1.160)} = 1.170.$$

The median survival time provides another view of respective mortality risks between subjects who smoke ($F = 1$) and do not smoke ($F = 0$). The Weibull survival function produces an estimated median value (denoted m_F) where

$$S(m_F) = 0.5 = e^{-\lambda_F t^{\gamma}},$$

making the estimated median survival time

$$\text{median survival time} = \hat{m}_F = \left[\frac{log(2)}{\hat{\lambda}_F}\right]^{1/\hat{\gamma}}.$$

The model estimated parameters $\hat{\gamma} = 1.160$, $\hat{\lambda}_0 = 0.0233$, $\hat{\lambda}_1 = 0.0273$ yield estimated median values of $\hat{m}_0 = 18.62$ weeks (nonsmokers, $F = 0$) and $\hat{m}_1 = 16.27$ weeks (smokers, $F = 1$).

The hazards ratio, as the name suggests, is a comparison of risk measured on a ratio scale. The difference between estimated median values is a comparison of risk measured on a time scale (weeks), sometimes referred to as a measure of *acceleration*; that is, the risk from smoking accelerates the median survival time by $\hat{m}_0 - \hat{m}_1 = 18.62 - 16.26 = 2.36$ weeks. A Weibull analysis provides both perspectives on the smoker/nonsmoker survival experience.

The Shape Parameter

Student's two-sample t-test requires compared normal distributions to have the same variance. The t-test is, nevertheless, resilient to deviations from this requirement, said to be robust. A similar issue arises when Weibull survival model is used to compare two samples. The estimated Weibull hazards ratio $\hat{hr}$ is only constant with respect to the survival time t when

Table 18.5 *Analytic Results: Three Weibull Models to Address the Question of Equality of the Shape Parameters (HIV/AIDS Data, San Francisco Men's Health Study – n = 174)*

	$\hat{b}_0$	$\hat{\lambda}_0$	$\hat{\lambda}_1$	$\hat{\gamma}$	$log(L)$
smokers	−3.107	0.026	–	1.168	−340.805
nonsmokers	−3.238	–	0.027	1.151	−297.794
combined	−3.240	0.027	0.023	1.160	−638.605

the shape parameters are the same for compared groups. The difference is required to be zero or, in symbols, $\gamma_0 - \gamma_1 = 0$. For example, the difference between two log-cumulative hazard functions

$$\begin{aligned} log[H_0(t)] - log[H_1(t)] &= log(-log[S_0(t)]) - log(-log[S_1(t)]) \\ &= [log(\lambda_0) + \gamma_0 log(t)] - [log(\lambda_1) + \gamma_1 log(t)] \\ &= log(\lambda_0) - log(\lambda_1) + (\gamma_0 - \gamma_1)\, log(t) \\ &= log(hr) + (\gamma_0 - \gamma_1) log(t) \end{aligned}$$

demonstrates that failure of the shape parameters to be equal causes the hazards ratio (hr) to depend on survival time. In other words, the hazard rates are not proportional, and a single hazards ratio is no longer usefully interpreted as a summary measure of survival times.

The question of equality of the shape parameters in a two-sample analysis is directly addressed by comparing three likelihood statistics. A Weibull model applied to data from one of the compared groups produces estimates $\hat{\lambda}_0$ and $\hat{\gamma}_0$. A Weibull model applied to the data from the other group produces estimates $\hat{\lambda}_1$ and $\hat{\gamma}_1$. A third Weibull model applied to the combined data produces estimates $\hat{\lambda}$ and $\hat{\gamma}$. In addition, each analysis produces a log-likelihood statistic, respectively denoted $log(L_0)$, $log(L_1)$ and $log(L)$.

The test statistic $X^2 = log(L) - [log(L_1) + log(L_0)]$ has an approximate chi-square distribution (degrees of freedom = 1) when the shape parameters do not differ ($\gamma_0 = \gamma_1$).

Specifically for the HIV/AIDS data, these log-likelihood values are presented in Table 18.5.

The summary log-likelihood values are $log(L_0) + log(L_1) = -638.599$ ($\gamma_0 \neq \gamma_1$) and $log(L) = -638.605$ ($\gamma_0 = \gamma_1$). Therefore, the chi-square distributed likelihood ratio test statistic (degrees of freedom = 1)

$$X^2 = 2\,[log(L) - (log(L_0) + log(L_1))] = 2\,[-638.599 - (-638.605)] = 0.013$$

measures the likelihood that the scale parameters are equal by chance alone or $\gamma_0 = \gamma_1 = \gamma$. The associated p-value is 0.908.

An alternative approach compares estimates $\hat{\gamma}_0$ and $\hat{\gamma}_1$. The logarithms of these estimates and their standard errors are usually calculated as part of the computer software used to estimate the Weibull parameters. For the smoking/AIDS data, they are $log(\hat{\gamma}_0) = 0.141$ and $log(\hat{\gamma}_1) = 0.155$ with respective estimated standard errors of $\hat{s}e_0 = 0.091$ and $\hat{s}e_1 = 0.086$. The variance of the difference between the two independent estimated log-scale parameters

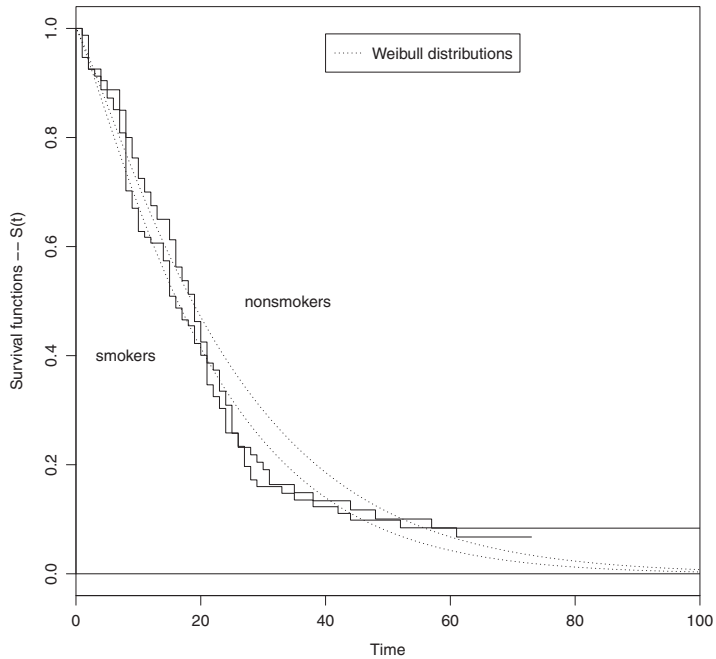


Figure 18.7 Direct Comparison of Parametric Estimated Weibull (Dotted Lines) and Nonparametric Product-Limit (Solid Lines) Survival Probability Functions for Smokers and Nonsmokers

is *variance* $(log[\hat{\gamma}_1] - log[\hat{\gamma}_0]) = 0.016$. A typical approximate two-sample chi-square distributed test statistic (degrees of freedom = 1)

$$X^2 = \left[\frac{log(\hat{\gamma}_1) - log(\hat{\gamma}_0) - 0}{\sqrt{variance(log[\hat{\gamma}_1] - log[\hat{\gamma}_0])}}\right]^2 = \left[\frac{0.155 - 0.141}{\sqrt{0.016}}\right]^2 = 0.013$$

yields a p-value of 0.908.

Goodness-of-fit, as always, is a necessary component of model-based analysis. For the Weibull model, assessing the equality of the shape parameters is a start. Some assurance that the hazard function is likely an increasing function adds to the confidence of the model results. A visual comparison of log-cumulative hazard functions is also a valuable addition to the process of evaluating model accuracy.

To start, product-limit (Kaplan-Meier) nonparametric survival probabilities are estimated for each group (denoted $\hat{P}_i$) based entirely on the data for smokers and for nonsmokers (Figure 18.7, solid lines) (Chapter 17). Also displayed (Figure 18.7, dotted lines) are the corresponding parametric estimated Weibull survival functions. Both estimates account for noninformative censored survival times. A direct graphic comparison of model-free and Weibull model–estimated survival probability functions visually indicates model accuracy.

The transformation $log(-log[\hat{P}_i])$ applied to each model-free estimated survival distribution produces separate data-generated log-cumulative hazard functions for smokers and nonsmokers (Figure 18.8). As noted, the parametric log-cumulative hazard function from

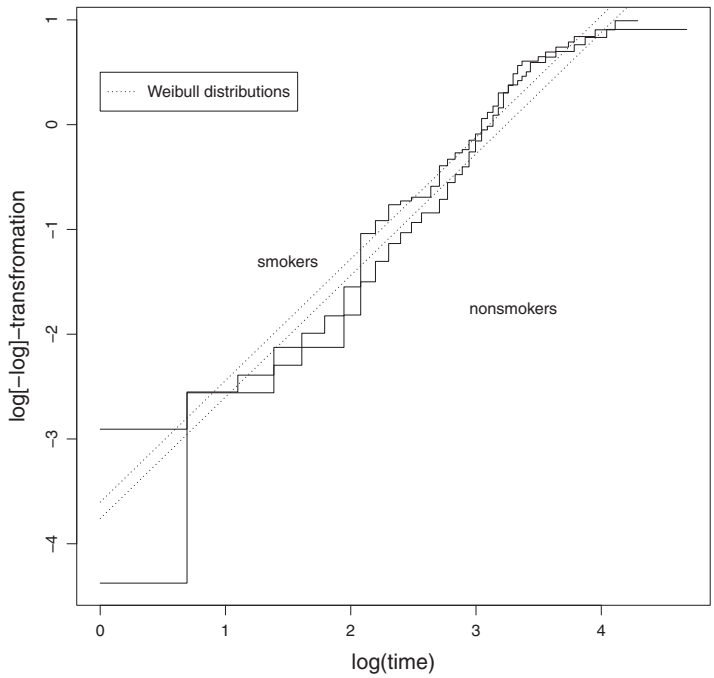


Figure 18.8 A Direct Comparison of Parametric Estimated Weibull (Dotted Lines) and Nonparametric Product-Limit (Solid Lines) Log-Cumulative Distribution Functions for Smokers and Nonsmokers

the Weibull probability distribution is a straight line or $y = log[H(t)] = log(-log[S(t)]) = log(\lambda) + \gamma log(t)$ with intercept $a = log\ (\lambda)$ and slope $b = \lambda$. Specifically, estimated from the AIDS/smoking data, for nonsmokers:

$$log[\hat{H}(t)] = log(-log[S_0(t)]) = log(0.0273) + 1.151log(t) = -3.601 + 1.151log(t)$$

and for smokers:

$$log[\hat{H}(t)] = log(-log[S_1(t)]) = log(0.0233) + 1.168log(t) = -3.240 + 1.168log(t).$$

The comparison to the log-cumulative hazard functions is usually a clearer indication of the correspondence between data and model generated values (Figure 18.8, solid versus dotted lines). As always, a visual comparison of straight lines is ideal (Chapter 12).

Multivariable Weibull Survival Time Model – A Case Study

A multivariable Weibull model is an extension of the previous two-sample model and differs little in principle. A Weibull multivariable regression analysis follows the pattern of most regression models (Chapters 7 and 9). An expression for a k-variable additive Weibull model for survival times t is

$$h(t|x_1,\ x_2, \ldots, x_k) = \lambda_i \gamma t^{\gamma - 1}.$$

Table 18.6 *Description: HIV Data* ($n = 72$) *Time to Diagnosis of AIDS (Weeks)*

	Minimums	1st quartiles	Medians	Means	3rd quartiles	Maximums
*cd*4	48.0	477.5	641.5	657.8	810.0	1468.0
β	0.8	2.0	2.5	2.6	3.0	5.1
Age	16.0	26.0	31.0	30.5	34.0	47.0

The scale parameter is created as $\lambda_i = e^{[b_0+b_1x_{i1}+b_2x_{i2}+\cdots+b_kx_{ik}]}\gamma$ where $i = 1, 2, \ldots, n$ observations and $j = 1, 2, \ldots, k$ variables. Typical of a multivariable analysis, specific risk variables are denoted $x_j$and t represents survival time for a multivariate observation.

The San Francisco Men's Health Study cohort of HIV-positive men consisting of 43 complete (AIDS cases) and 29 censored observations from a 42-month study again illustrates ($n = 72$) (Chapter 17). Three factors potentially related to the prognosis of AIDS are the *cd* 4-lymphocyte count (*cd*4), the serum β-microglobulin level(β) and the study subject's age (*age*). As before, "survival time" is time from diagnosis of HIV to diagnosis of AIDS or the end of the study period (Chapter 17). Table 18.6 briefly summarizes the HIV/AIDS data.

The principle focus of the analysis of the HIV/AIDS data is a clear picture of the role of the *cd*4 count. The Weibull hazard model describing only the influence of *cd*4 counts

$$h(t|cd4_i) = \lambda_i\gamma\, t^{\gamma-1} = e^{[b_0+b_1cd4_i]\gamma} \times \gamma\, t^{\gamma-1}$$

provides a good place to begin. Using AIDS/HIV data, the Weibull model is a straightforward extension of the two-sample model. The only important difference is that the binary variable (F) is replaced by observed *cd*4 counts, as reported (denoted $cd4_i$). The estimated model parameter $\hat{b}_1 = -0.002$ measures the expected increasing risk of AIDS with decreasing *cd*4 counts (Table 18.7).

A few descriptive calculations from the estimated Weibull model are the following:

Shape parameter: $\hat{\gamma} = e^{log(\hat{\gamma})} = e^{0292} = 1.339$,
Hazards ratio per *cd*4-count: $\hat{hr} = e^{b_1\hat{\gamma}} = e^{-0.002(1.339)} = e^{-0.0026} = 0.997$,
Hazards ratio for *cd*4 counts that differ by a count x: $\hat{hr}_x = [\hat{hr}]^x$.

Then for $x = 100$, $\hat{hr}_{100} = 0.997^{100} = 0.772$ or $\hat{hr}_{100} = 1/0.772 = 1.295$ and median survival time for a *cd*4 count of 657.8 (mean value $= \overline{cd4}$):

$$\hat{m} = \left[log(2)e^{-[\hat{b}_0+\hat{b}_1\overline{cd4}]\hat{\gamma}}\right]^{1/\gamma} \text{ and } \left[log(2)e^{-[-3.333-0.002(657.8)]1.339}\right]^{1/1.339} = 75.9 \text{ weeks.}$$

Table 18.7 *Analysis: Weibull Model Applied to the* cd*4 Counts for the HIV/AIDS Data*

Variables	Estimates	*s. e.*	*p*-values
b_0	−3.333	–	–
b_1	−0.002	0.0006	<0.001
$log(\gamma)$	0.292	0.1214	
log-likelihood = −282.085			

Table 18.8 *Analysis: Interaction Weibull Model Estimated Parameters from n = 72 HIV/AIDS Subjects to Describe Time to Diagnosis of AIDS*

Variables	Estimates	*s. e.*	*p*-values
Intercept	1.9225	–	–
*cd*4	−0.0080	0.0042	–
β	−0.1031	0.8290	–
Age	−0.2150	0.1191	–
$cd4 \times \beta$	−0.0004	0.0006	0.525
$cd4 \times age$	0.0002	0.0001	0.039
$\beta \times age$	0.0197	0.0246	0.423
$log(\gamma)$	0.3818	0.1200	0.001
log-likelihood (L) = −275.705			

Additivity is an extremely descriptive statistical property. When a multivariate observation is accurately represented by an additive relationship, it becomes possible to separate a multivariate influence into components to evaluate the extent of influence of each of a series of individual contributions. Not all multivariate observations, as expected, are accurately summarized by an additive relationship. Therefore, it is important to evaluate the adequacy of an additive model to represent the relationships within the observed data.

For the Weibull survival model, the effectiveness of an additive model, like most regression models, can be assessed by a comparison to the corresponding nonadditive model containing all pairwise interaction terms. Applied to the HIV/AIDS data, the two Weibull models are the following:

Nonadditive Hazard Function Model:

$$h(t|cd4_i, \beta_i, age_i) = \left[e^{[b_0+b_1 cd4_i+b_2\beta_i+b_3 age_i+b_4(cd4_i\times\beta_i)+b_5(cd4_i\times age_i)+b_6(\beta_i\times age_i)]\gamma}\right] \times \gamma t^{\gamma-1}$$

and Additive Hazard Function Model:

$$h(t|cd4_i, \beta_i, age_i) = \left[e^{[b_0+b_1 cd4_i+b_2\beta_i+b_3 age_i]\gamma}\right] \times \gamma\, t^{\gamma-1}.$$

The square brackets indicate the respective scale parameters. Interaction and additive models each yield log-likelihood statistics (Tables 18.8 and 18.9).

Table 18.9 *Analysis: Additive Weibull Model Estimated Parameters from n = 72 HIV/AIDS Subjects to Describe Time to a Diagnosis of AIDS*

Variables	Estimates	*s. e.*	*p*-values
intercept	−3.327	–	–
*cd*4	−0.002	0.0006	0.001
β	0.263	0.1130	0.020
age	−0.023	0.0181	0.197
$log(\gamma)$	0.327	0.1186	0.006
log-likelihood (L) = −278.382			

Table 18.10 *Comparison: Regression Coefficients from Additive Weibull and Cox Models from HIV/AIDS Data*

	Coefficients	
	Weibull	Cox
*cd*4	−0.002	−0.002
β	0.263	0.372
age	−0.023	−0.016

The chi-square comparison (degrees of freedom = 3) of log-likelihood values

$$X^2 = 2[-275.705 - (-278.382)] = 5.353$$

produces a *p*-value of 0.148, providing marginal evidence that the additive model is a useful summary of the HIV/AIDS data.

Another indication of model suitability is a direct comparison of the estimated regression coefficients from the additive parametric Weibull model to the same coefficients estimated from the nonparametric (Cox) additive proportional hazards model applied to the identical data (Chapter 17).

From this prospective, the parametric additive Weibull model differs little from the Cox multivariable approach (Chapter 17), providing a further indication that the additive Weibull model likely provides a useful parametric description of the HIV/AIDS data (Table 18.10).

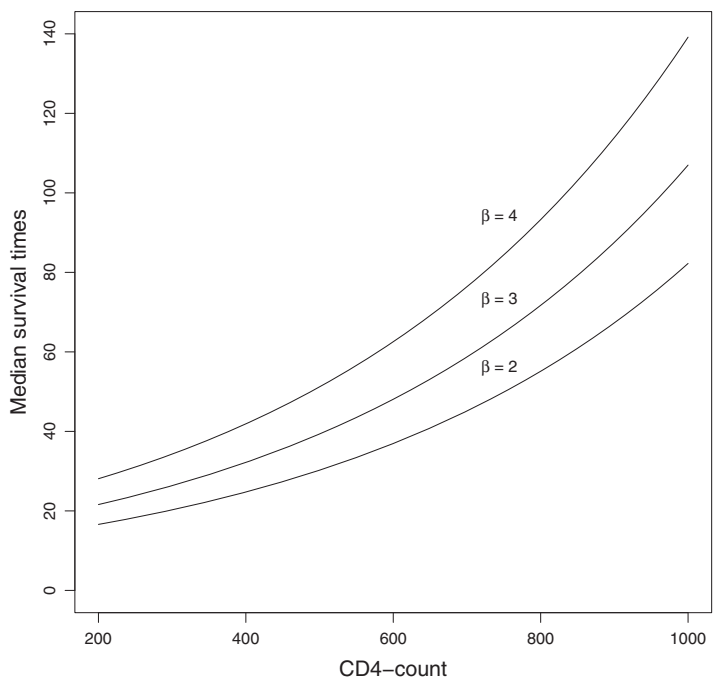


Figure 18.9 Median Survival Time by *cd4*-Counts for Three Levels of β-microglobulin (*age* = 30)

The Weibull model survival analysis, with some evidence of additivity, indicates that the *cd*4 count is a major influence on time from a diagnosis of HIV to a diagnosis of AIDS (Table 18.9). The β-microglobulin level plays a lesser role but remains an important influence. The age of the subject appears to have little or no influence on survival time. The Cox regression analysis of the same data produces essentially the same inferences.

Primary features of a parametric Weibull model analysis are visual and summary descriptions of the survival data using the estimated survival and hazard functions. Among a variety of ways an estimated parametric model can effectively describe relationships within collected data, estimated median values are an intuitive and useful summary statistic. The median values for the HIV/AIDS data based on the estimated Weibull model parameters are given by the expression

$$\text{estimate median value} = \hat{m}_i = \left[\frac{log(2)}{\hat{\lambda}_i}\right]^{1/\hat{\gamma}}$$

where the estimated scale parameter is $\hat{\lambda}_i = e^{[-3.327-0.002cd4_i+0.263\beta_i-0.023age_i]1.387}$ (Table 18.9). A plot displays the estimated relationship of *cd*4 counts to median survival times for three levels of β-microglobulin (Figure 18.9) for subjects age 30. If other values of the variable *age* are used, the result would be to shift the exact same pattern of influence of the three β-measures to slightly different median survival times, a property of additive models. An additive model dictates that the age variable does not affect the model estimated $cd4/\beta$ relationship.

As always, the benefits of a parametric model are a detailed and unambiguous description of the relationships within the data using estimated parameters. The cost is an artificial structure (model) that is the basis of the description.

Epidemiology

19

Prediction, a Natural Measure of Performance

A fundamental and natural evaluation of performance of one analytic model relative to another is a comparison of differences in model prediction. A number of parallel methods exist to measure and explore improvement in model prediction, but the principle is the same: more accurate models produce better predictions. Improvement occurs in two ways: more accurate prediction of an event among individuals observed with the event and more accurate prediction of the absence of an event among individuals observed without the event. The statistical analysis becomes an evaluation of the total extent of improvement.

For a case study, the outcome of interest is the occurrence of a coronary heart disease event (*chd*), and several different statistical approaches are applied to assess the improvement in prediction by comparing models including and excluding a measure of each study subject's reported cholesterol level. The model used is a linear additive logistic regression model. The data come from the Atherosclerosis Risk in Communities study (ARIC) and consist of a cohort of individuals ages 45 to 64 from a defined population. Data collection began in 1987 and includes four extensive examinations of close to 3900 study participants.

The case study illustrates the principles and issues surrounding comparison of two statistical models applied to seven risk variables (sex, systolic blood pressure, diastolic blood pressure, cholesterol level, smoking exposure [three levels], body mass index, and age). The risk variable of interest is chosen to be a subject's cholesterol level. The influence of cholesterol on the likelihood of a *chd* event is extensively documented. For the following case study, this well-studied risk factor provides a focus on the statistical assessment as well as on the question of how the analytic process works.

The analytic results from two additive logistic models applied to these ARIC coronary heart disease data are presented in Table 19.1. The first model contains six risk variables (denoted in "short-hand" as model X), and the second model is identical but additionally includes the subject's cholesterol level (denoted as model $X + C$). Parallel to a frequently used convention, the two models also are referred to as "old" and "new" as if cholesterol level is a "new" risk variable with unknown influence.

An almost always used analytic technique to assess the influence of a variable added to a statistical model is a comparison of log-likelihood statistics with a *likelihood ratio test* (Chapter 10). The log-likelihood values from the two example logistic models are (Table 19.1) the log-likelihood value for the seven variable logistic model $(X + C) = -836.46$ that includes the cholesterol influence and the log-likelihood value for the six variable logistic model $(X) = -845.00$ when the cholesterol variable is not included in the model. The likelihood ratio test statistic measuring the influence on *chd* risk from the cholesterol variable is

$$X^2 = 2[-836.46 - (-845.00)] = 17.03.$$

Table 19.1 *Coronary Heart Disease: Six Variable Additive Logistic Model* (X) *Applied to ARIC Data* ($n = 3854$)

Variables	Estimates	*s. e.*	z-values
intercept	−4.398	–	–
sex	−1.445	0.166	−8.69
sysbp	0.006	0.005	1.21
diabp	−0.032	0.009	−3.56
smk. 1	−0.009	0.165	0.05
smk. 2	−0.685	0.191	−3.58
bmi	0.043	0.014	3.01
age	0.076	0.014	5.62
log-likelihood = −845.00			

Table 19.1 *(Continued) Coronary Heart Disease: Additive Logistic Model* ($X + C$) *Applied to ARIC Data* ($n = 3854$) *with Cholesterol Levels Added as a Seventh Variable*

Variables	Estimates	*s. e.*	z-values
intercept	−5.421	–	–
sex	−1.554	0.169	−9.87
sysbp	0.006	0.005	1.19
diabp	−0.033	0.009	−3.73
smk.1	0.014	0.166	0.08
smk.2	−0.665	0.192	−3.35
bmi	0.043	0.014	3.03
age	0.071	0.014	5.23
cholesterol	0.007	0.002	4.20
log-likelihood = −836.46			

When two log-likelihood values are calculated from identical data and the variable added (cholesterol in this case) causes the two models to differ only because of the natural influence of sampling variation, the test statistic represented by X^2 has an approximate chi-square distribution with degrees of freedom equal to the number of added variables (one in this case). In other words, an added variable such as cholesterol capitalizes only on the always present random variation to create apparent improvement. For the example data, the small significance probability or p-value $= P(X^2 \geq 17.03 \mid \textit{no influence from cholesterol}) < 0.001$ indicates the likely presence of a systematic influence from including a measure of cholesterol level in the model.

This frequently used likelihood ratio test is rigorous, sophisticated, and at the heart of many analyses. However, it is not an intuitive process, provides little insight into how the added variable influences risk, and requires considerable statistical theory to fully understand the application on anything but a mechanical level. Analytic methods, called *reclassification techniques*, are alternatives to likelihood comparisons of "nested" models that are intuitive and statistically simple and produce easily interpreted results (Chapter 10).

Table 19.2 *Example: 20 Probabilities Generated from "New" Model* $(X + C)$ *and "Old" Model* (X) *from ARIC Data*

"New" model $(X + C)$			"Old" model (X)		
	chd	*no-chd*		*chd*	*no-chd*
$\bar{C}^a$	$\hat{p}_{11}$	$\hat{p}_{01}$	$\bar{C}^a$	$\hat{p}_{10}$	$\hat{p}_{00}$
45.3	0.211	0.162	−44.7	0.012	0.016
0.3	0.019	0.009	−7.7	0.004	0.015
8.3	0.211	0.201	46.3	0.057	0.055
3.3	0.011	0.011	−35.7	0.013	0.018
51.3	0.251	0.088	5.7	0.015	0.011
6.3	0.219	0.213	3.7	0.024	0.027
−24.7	0.029	0.040	−10.7	0.132	0.138
−31.7	0.017	0.053	25.3	0.044	0.037
76.3	0.216	0.039	13.7	0.062	0.058
−5.3	0.058	0.062	−94.7	0.007	0.075

[a] $\bar{C} = chol - mean(chol)$ = an individual's deviation from the mean cholesterol level.

When a variable added to a model contributes to predicting risk, two changes measure improvement. In terms of the example analysis, the predicted probabilities associated with a *chd* event go up among subjects who have had a *chd* event and go down among subjects without a *chd* event, parallel to the expression, "The rich get richer and the poor get poorer." The statistical analysis then becomes an assessment of the changes in prediction accuracy sometimes referred to as sensitivity and specificity (Chapters 6 and 22).

A small example of 20 model-estimated probabilities from the ARIC analysis is displayed in Table 19.2. The symbol $\hat{p}_{ij}$ represents a model-estimated probability of a *chd* event. The first subscript (i) denotes the presence or absence of a coronary heart disease event ($chd = 1$ and $no\text{-}chd = 0$). The second subscript (j) denotes the model used to calculate the probability of a *chd* event ("new" model $(X + C) = 1$ and "old" model $[X) = 0$). For example, the model-estimated value $\hat{p}_{11} = 0.211$ is the probability of a *chd* event calculated from individuals known to have had a *chd* event using the model including cholesterol levels (Table 19.2). Note that improvement in model prediction is indicated by $\hat{p}_{11} > \hat{p}_{10}$ and $\hat{p}_{00} > \hat{p}_{01}$.

The two logistic models ($X + C$ and X) each yield $n = 3854$ probabilities from the ARIC data, denoted $\hat{p}$ from the "new" model $(X + C)$ and $\hat{p}_0$ from the "old" model (X).

These 3854 pairs of model-estimated probabilities graphically display the difference between these two models in terms of predicted values (Figure 19.1). For the probabilities above the diagonal line from (0, 0) to (1, 1), estimates $\hat{p}$ are greater than $\hat{p}_0$. For values below the diagonal line, estimates $\hat{p}$ are less than $\hat{p}_0$. Thus, the spread of these points reflects the extent of changes in prediction from the added cholesterol variable. The diagonal line $\hat{p} = \hat{p}_0$ represents exactly no change.

Net Reclassification Index (Binary Case)

To further explore differences between model-estimated probabilities, the estimates $\hat{p}$ and $\hat{p}_0$ are classified into 2×2 tables – specifically, a table containing counts of changes

Table 19.3 *Model Results: Binary Classification of Model-Predicted Probabilities Classified into Four Categories*[a]

	Up	*Down*	Total
chd	133	126	259
no-chd	1431	2164	3595
Total	1564	2290	3854

[a] $up = \hat{p} > \hat{p}_0$ and $down = \hat{p} < \hat{p}_0$

in probabilities $\hat{p}$ and $\hat{p}_0$ among subjects with *chd* events and counts of the changes in probabilities $\hat{p}$ and $\hat{p}_0$ among subjects without *chd* events. The increase in correct predictions measures improvement from the influence of adding a "new" variable to the regression model. The ARIC data create Table 19.3.

Among the 259 *chd* event subjects:

$$up = 133 \text{ } chd \text{ subjects, then } P(up|chd) = \frac{133}{259} = 0.514$$

and

$$down = 126 \text{ } chd \text{ subjects, then } P(down|chd) = 1 - P(up|chd) = \frac{126}{259} = 0.486.$$

Among the 3595 *non-chd* subjects:

$$up = 1431 \text{ } no\text{-}chd \text{ subjects, then } P(up|no\text{-}chd) = \frac{1431}{3595} = 0.398$$

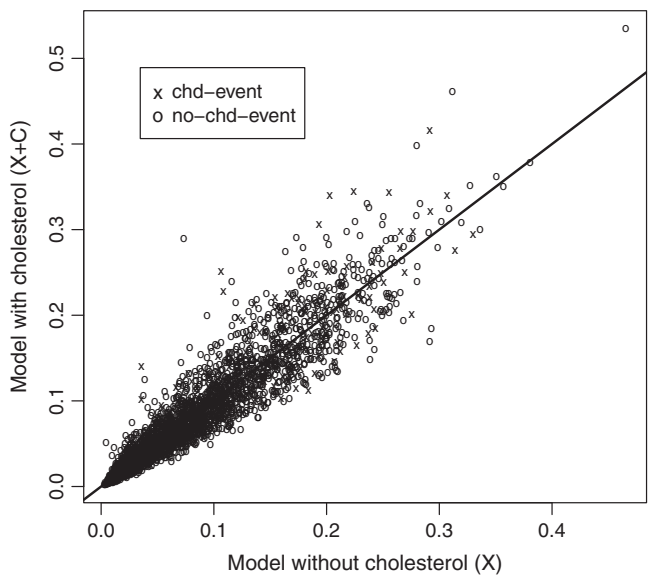


Figure 19.1 Probabilities of a *chd* Event $\hat{p}$ and $\hat{p}_0$ Estimated from Two Additive Logistic Regression Models, One Containing and the Other Not Containing Subject's Cholesterol Level ($n = 3854$)

and

$$down = 2164 \text{ } no\text{-}chd \text{ subjects, then } P(down|no\text{-}chd) = 1 - P(up|no\text{-}chd) = \frac{2164}{3595} = 0.602.$$

The net improvement associated with adding the cholesterol variable to the model is summarized by combining these four probabilities into a single summary value called the *net reclassification index* (*nri*). Two expressions produce the same value:

$$\begin{aligned} net\ reclassification\ index(nri) = \text{improvement} &= [P(up|chd) - P(down|chd)] \\ &+ [P(down|no\text{-}chd) - P(up|no\text{-}chd)] \end{aligned}$$

or

$$\text{improvement} = [P(up|chd) - P(up|no\text{-}chd)] + [P(down|no\text{-}chd) - P(down|chd)].$$

For the ARIC data, the value of this statistical measure of improvement is

$$\hat{nri} = (0.514 - 0.486) + (0.602 - 0.398) = 0.027 + 0.204 = 0.231.$$

An expression for the estimated variance of the distribution of an estimated *nri*-summary statistic is

$$variance(\hat{nri}) = \frac{1}{n_1} + \frac{1}{n_0},$$

where n_1 represents the number of *chd* subjects (259) and n_0 represents the number of *no-chd* subjects (3595). Using the *nri* estimate, a normal distribution–based test statistic and confidence interval are routinely calculated (Chapter 2).

To evaluate the conjecture that the observed prediction improvement from adding cholesterol to the model is entirely due to capitalizing on random variation, using the variance estimate $variance(\hat{nri}) = 1/259 + 1/3595 = 0.0041$, an approximate chi-square test statistic (degrees of freedom = 1) is

$$X^2 = \left[\frac{0.231 - 0}{0.0643}\right]^2 = 12.883.$$

The observed improvement in prediction accuracy is unlikely to have occurred by chance alone (p-value < 0.001). A normal distribution approximate 95% confidence interval based on the estimate $\hat{nri} = 0.231$ is

$$\hat{nri} \pm 1.960\sqrt{variance(\hat{nri})} = 0.231 \pm 1.960(0.0643) \rightarrow (0.105, 0.357).$$

Net Reclassification Index (Extended Case)

The *nri* approach potentially becomes a more sensitive statistical tool by increasing the number of classification categories. To illustrate, changes in the model-estimated probabilities $\hat{p}$ and $\hat{p}_0$ are classified into a two-way table with more than four cells. Instead of 2×2 summary tables, interval boundaries create extended 4×4 tables, one containing subjects with *chd* events and the other containing subjects without *chd* events.

Table 19.4 *Model Outcomes: Classification of Model-Estimated Probabilities* ($\hat{p}$ *and* $\hat{p}_0$) *for* $n_0 = 259$ *Subjects with chd Event*

		$\hat{p} = P(chd\text{-}event\|X + C)$				
		0–0.05	0.05–0.10	0.10–0.20	0.20–1.0	Total
$\hat{p}_0 = P(chd\text{-}event\|X)$	0–0.05	43	7	2	0	52
	0.05–0.10	0	43	13	0	56
	0.10–0.20	0	10	82	20	112
	0.20–1.0	0	0	6	33	39
	Total	43	60	103	53	259

Using the four categories (0–0.05, 0.05–0.10, 0.10–0.20, and 0.20–1.0) and the $n_1 = 259$ ARIC subjects with a *chd* event, the first step is to construct a 4×4 table from the 259 model-generated probabilities $\hat{p}$ and $\hat{p}_0$ (Table 19.4).

Figure 19.2 displays the classification of these model probabilities (*chd-present* table) in terms of logarithms that do not affect the classification frequencies but improves graphical visibility. From the reclassification *chd* table, then

$$\hat{p}_u = P(up|X + C) = \frac{7 + 2 + 13 + 20}{259} = 0.162$$

and

$$\hat{p}_d = P(down|X) = \frac{10 + 6}{259} = 0.062.$$

The observed net improvement from adding the cholesterol variable to the model in terms of classification accuracy of *chd* events among the subjects with a *chd* event is then

$$n\hat{r}i[chd\ event] = \hat{p}_u - \hat{p}_d = 0.162 - 0.062 = 0.100.$$

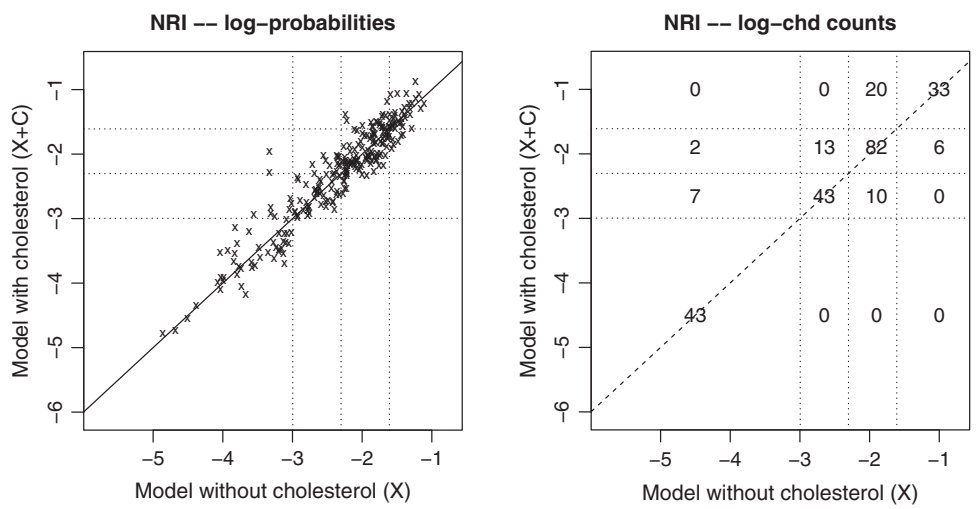


Figure 19.2 Schematic View of Reclassification of Model-Predicted Log-Probabilities among 259 Subjects with *chd* Event from Comparison of Two Logistic Regression Models (Table 19.1)

Table 19.5 *Model Outcomes: reclassification of Model-Estimated Probabilities* ($\hat{p}$ *and* $\hat{p}_0$) *for* $n_0 = 3595$ *subjects Without a* chd *Event*

		$\hat{p} = P(chd\text{-}event\|X + C)$				
		0–0.05	0.05–0.10	0.10–0.20	0.20–1.0	Total
$\hat{p}_0 = P(chd\ event\|X)$	0–0.05	1933	94	2	0	2029
	0.05–0.10	125	595	85	1	806
	0.10–0.20	0	114	442	64	620
	0.20–1.0	0	0	37	103	140
	Total	2058	803	566	168	3595

An important feature of a reclassification table is the possibility to better understand the role of the variable or variables used to create the "new" model. For example, a pattern of improvement reflected by the reclassification *chd*-present table is the following:

Intervals	Improvement
0–0.05	$100 \times (9/52) = 17.3\%$
0.05–0.10	$100 \times (13/56) = 20.0\%$
0.10–0.20	$100 \times (20/112) = 17.8\%$
Total	$100 \times (42/220) = 19.1\%$

The inclusion of cholesterol in the model appears to equally improve prediction for all three levels of cholesterol. Certainly other patterns are possible and likely become apparent from a similar investigation of marginal frequencies or other properties of reclassification tables.

The reclassification table created for those subjects without a *chd* event follows the same pattern (Table 19.5).

Figure 19.3 displays the model probabilities (*chd*-absent table) again in terms of logarithms of the 3595 model prediction probabilities. Model reclassification probabilities from subjects without a *chd* event are

$$\hat{p}_{0d} = P(down|X + C) = \frac{125 + 114 + 37}{3595} = 0.077$$

and

$$\hat{p}_{0u} = P(up|X) = \frac{94 + 2 + 85 + 1 + 64}{3595} = 0.068.$$

The observed net improvement in classification of subjects without a *chd* events is then

$$\hat{nri}[no\text{-}chd\ event] = \hat{p}_{0d} - \hat{p}_{0u} = 0.077 - 0.068 = 0.009.$$

Therefore, the overall summary of improvement in model predicted probabilities resulting from adding the cholesterol variable to the model is

$$\hat{nri} = \hat{nri}[chd\ event] + \hat{nri}[no\text{-}chd\ event] = 0.100 + 0.009 = 0.109.$$

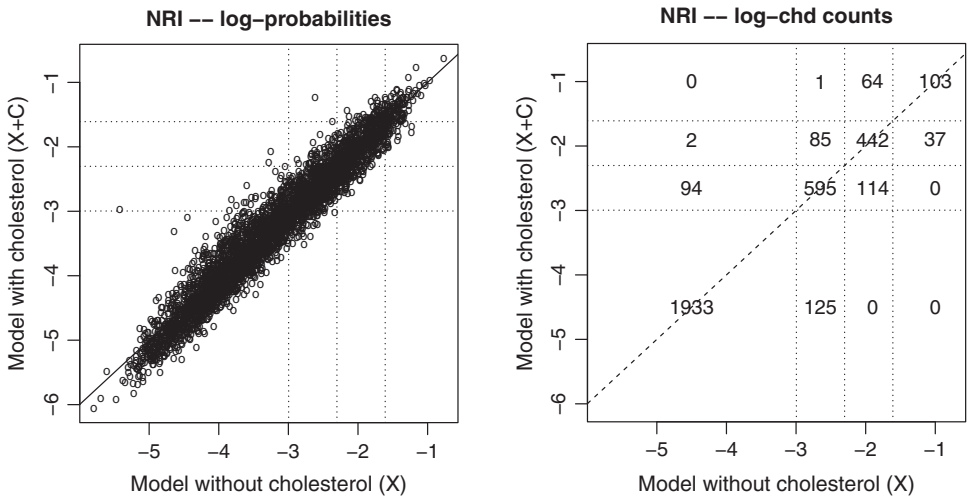


Figure 19.3 Schematic View of Reclassification of Model-Predicted Log-Probabilities among 3595 Subjects without a *chd* Event from Comparison of Two Logistic Regression Models (Table 19.1)

Again, a typical normal distribution–based test statistic and confidence interval follow directly. The estimated variance of the distribution of the estimated $n\hat{r}i$ measure is given by the expression

$$variance(n\hat{r}i) = \frac{\hat{p}_u + \hat{p}_d}{n_1} + \frac{\hat{p}_{0u} + \hat{p}_{0d}}{n_0}.$$

A chi-square test statistic to evaluate the extent of influence of cholesterol on the prediction probabilities is

$$X^2 = \left[\frac{n\hat{r}i - 0}{\sqrt{variance(n\hat{r}i)}}\right]^2 = \left[\frac{0.109 - 0}{0.031}\right]^2 = 13.063$$

and yields a p-value of $P(X^2 \geq 13.063|nri = 0) < 0.001$. A normal distribution–derived approximate 95% confidence interval based on the estimate $n\hat{r}i = 0.109$ is

$$n\hat{r}i \pm 1.960\sqrt{variance(n\hat{r}i)} = 0.109 \pm 1.960(0.031) \rightarrow (0.048, 0.170).$$

Central issues are the following:

1. The pattern of reclassification probabilities can produce natural insights into the model behavior associated with a specific risk factor or factors based on easily understood tables.
2. Reclassification techniques use simple and intuitive concepts that do not require sophisticated statistical theory to understand the analytic process. For example, everyday plots and tables create useful and directly interpreted contrasts between model performances.
3. To accurately estimate improvement between compared models, the models must adequately represent the data. As with all analyses based on a statistical model, the degree of correspondence between model-generated values and data is a critical issue (goodness-of-fit) (Chapter 12).

Table 19.6 *Summary: "New" and "Old" Model-Estimated Mean Predicted Probabilities for Subjects with and without Coronary Heart Disease Event* ($n = 3854$)

	Models	
	$X + C$	X
chd	$\bar{p}_{11} = 0.1281$	$\bar{p}_{10} = 0.1227$
no-chd	$\bar{p}_{01} = 0.0628$	$\bar{p}_{00} = 0.0632$

4. Categories selected to classify predicted probabilities influence the analytic results adding an arbitrary element to inferences and interpretation (Chapter 10). On the other hand, situations arise when the categories created have clinical or biologic significance, enhancing the understanding of the role of added variable or variables.
5. A likelihood approach addresses the question: Did observed differences between compared models arise strictly by chance? A reclassification approach addresses the question: Did the predictions from the "new" model ($X + C$) substantially improve relative to those from the "old" model (X)? Neither approach addresses the question of the accuracy of the model (sometimes called *calibration*) (Chapter 12).

Integrated Discrimination Improvement

A comparison based on probabilities $\hat{p}$ and $\hat{p}_0$ is a more efficient measure of model improvement when the estimated probabilities are directly analyzed without creating sometimes inefficient and arbitrary table categories. Such a summary statistic consists of four mean values calculated from probabilities estimated from the compared models, namely,

"New" model: $(X + C)$: $P(\textit{chd-present}|X + C) = \bar{p}_{11}$ and $P(\textit{chd-absent}|X + C) = \bar{p}_{01}$

and

"Old" model: (X): $P(\textit{chd-present}|X) = \bar{p}_{10}$ and $P(\textit{chd-absent}|X) = \bar{p}_{00}$.

From the ARIC data, these mean probabilities are $\bar{p}_{11} = 0.1281$, $\bar{p}_{01} = 0.0628$, $\bar{p}_{10} = 0.1227$ and $\bar{p}_{00} = 0.0632$ based on the model estimates $\hat{p}$ and $\hat{p}_0$ from all 3854 subjects (*chd* = 259 and *no-chd* = 3595) participating in the study (Table 19.6).

A summary measure of model improvement can be expressed in two ways with identical results:
"new" model (X) versus "old" model $(X + C)$, then

$$\text{rows: } \hat{idi} = (\bar{p}_{11} - \bar{p}_{10}) - (\bar{p}_{01} - \bar{p}_{00})$$

or chd-present versus chd-absent, then

$$\text{columns: } \hat{idi} = (\bar{p}_{11} - \bar{p}_{01}) - (\bar{p}_{01} - \bar{p}_{00})$$

and is called the *integrated discrimination improvement* (*idi*) index. The *idi* index primarily differs from of the previous *nri* statistic in that it treats the model probabilities as continuous

values summarized by mean values as an alternative to counts of subjects classified into tables. Again, the index simply measures net improvement in terms of model prediction.

For the comparison of the two ARIC models (Table 19.1), the estimated net increase associated with the "new" model ($X + C$) is

$$\hat{idi} = (0.1281 - 0.1227) - (0.0628 - 0.0632) = 0.0054 + 0.0004 = 0.0057.$$

or associated with a coronary event (present/absent)

$$\hat{idi} = (0.1281 - 0.0628) - (0.1227 - 0.0632) = 0.0653 - 0.0595 = 0.0057.$$

The *idi* index can be viewed as a measure of the combined gains in sensitivity and specificity of the predicted values (Chapter 6); that is,

$$\text{gain in sensitivity} = P(\textit{chd-present}|X + C) - P(\textit{chd-present}|X) = \bar{p}_{11} - \bar{p}_{10} = 0.0054$$

and

$$\text{gain in specificity} = P(\textit{chd-absent}|X) - P(\textit{chd-absent}|X + C) = \bar{p}_{01} - \bar{p}_{00} = 0.0004.$$

The probabilities $\bar{p}_{11}$ and $\bar{p}_{10}$ are the mean probabilities of a *chd* event estimated from the subjects with an observed *chd* event from both models (model X and model $X + C$), and the difference reflects increased sensitivity. Similarly, the probabilities $\bar{p}_{00}$ and $\bar{p}_{01}$ are mean probabilities among subjects without a *chd* event from both models (model $X + C$ and model X); the difference reflects increased specificity. The net reclassification measure is the sum of these two quantities and, to repeat, is $\hat{idi} = 0.0057$.

An expression to estimate the variance of the *idi*-summary statistic is

$$\textit{variance}(\hat{idi}) = \frac{\textit{variance}(\hat{p}_{11} - \hat{p}_{10})}{n_1} + \frac{\textit{variance}(\hat{p}_{01} - \hat{p}_{00})}{n_0}$$

calculated directly from the estimated model probabilities, where again n_1 represents the number of *chd* subjects (259) and n_0 represents the number of *no-chd* subjects (3595). For the ARIC data, the estimated standard error of the distribution of the estimate $\hat{idi}$ is $\sqrt{\textit{variance}(\hat{idi})} = 0.00213$ and a normal distribution based test statistic yields the chi-square value

$$X^2 = \left[\frac{\hat{idi} - 0}{\sqrt{\textit{variance}(\hat{idi})}}\right]^2 = \left[\frac{0.0057 - 0}{0.00213}\right]^2 = 7.188.$$

The probability the increase in prediction accuracy from adding cholesterol to the model occurred by chance alone is p-value $= 0.007$. A parallel normal distribution based approximate 95% confidence interval is

$$\hat{idi} \pm 1.960\sqrt{\textit{variance}(\hat{idi})} = 0.0057 \pm 1.960(0.00213) \rightarrow (0.0015, 0.0099).$$

Summary Lines as Measures of Improvement in Model Prediction

The comparison of two straight lines allows a direct analysis and, in addition, provides a visual comparison of model improvement. This approach is not fundamentally different

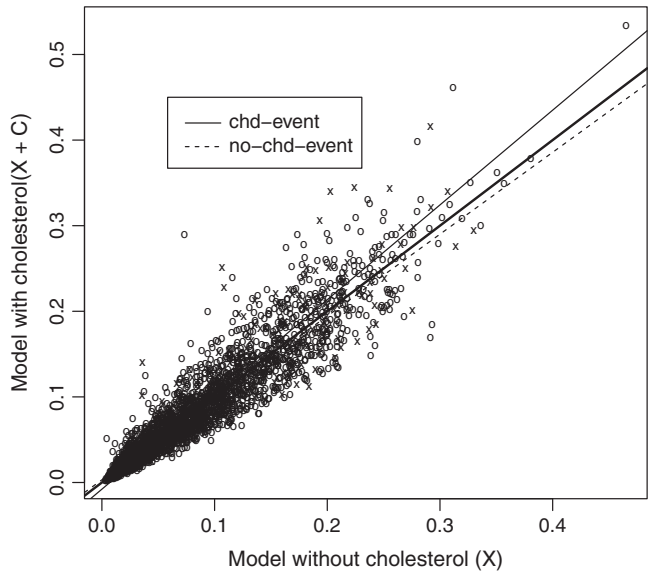


Figure 19.4 Model-Generated Probabilities of a Coronary Heart Disease Event $\hat{p}$ and $\hat{p}_0$ (Figure 19.1) Summarized by Perpendicular Least Squares Estimated Straight Lines with Cholesterol Levels Included and Not Included in Model

from reclassification methods. It too employs summary measures reflecting the increases in correctly predicting cases and noncases from adding to a model an additional variable or variables.

A summary line estimated from the probabilities $\hat{p}$ and $\hat{p}_0$ generated by subjects with the event and a similar summary line estimated from probabilities $\hat{p}$ and $\hat{p}_0$ generated by subjects without the event simply reflect improvement in model prediction. The slopes of these lines reflect the degree of improvement in model prediction. Figure 19.4 again displays the ARIC data and includes these two estimated summary lines.

The slopes of the summary lines are estimated by minimizing the squared perpendicular distances between observed probabilities and the estimated lines. Because the model-estimated probabilities are pairs of sampled values with a joint distribution, ordinary least squares estimation that requires specific independent and dependent variables be specified is not appropriate (details at the end of the chapter). The estimated slope of the line measuring the change in model predicted probabilities ($\hat{p}$ and $\hat{p}_0$) among subjects with *chd events* is $\hat{b}_{chd} = 1.107$. The difference $= 1.107 - 1.0 = 0.107$ summarizes the degree of improvement in the prediction of *chd* events among individuals with *chd* events. In a parallel fashion, the slope of the estimated line summarizing the change in model predicted probabilities ($\hat{p}$ and $\hat{p}_0$) from subjects without *chd* is $\hat{b}_{no\text{-}chd} = 0.957$. The difference $= 1.0 - 0.957 = 0.043$ measures the degree of improvement in the prediction of the *no-chd* events among individuals with *no-chd* events. A natural summary of overall improvement from adding the cholesterol variable to the model is $\hat{B} = \hat{b}_{chd} + \hat{b}_{no\text{-}chd} = 0.107 + 0.043 = 0.150$.

An approximate normal distribution based test statistic and confidence interval follow the usual pattern. A bootstrap estimated standard error of the distribution describing the overall estimated change in slopes is $\sqrt{variance(\hat{B})} = 0.0416$ (2000 replicate samples).

The approximate chi-square test statistic

$$X^2 = \left[\frac{\hat{B} - 0}{\sqrt{variance(\hat{B})}}\right]^2 = \left[\frac{0.150 - 0}{0.0416}\right]^2 = 13.002$$

yields a p-value of $P(X^2 \geq 13.002|B = 0) < 0.001$. An approximate 95% confidence interval is

$$\hat{B} \pm 1.960\sqrt{variance(\hat{B})} = 0.150 \pm 1.960(0.0416) \rightarrow (0.068, 0.232).$$

Covariance and Correlation

Covariances and correlation coefficients relating the influence of a binary observation (*chd* and *no-chd*) on the prediction probabilities can be used to describe improvement. For example, the *covariance* (*chd*, $\hat{p}$) from subjects with *chd events* and the *covariance* (*chd*, $\hat{p}_0$) from subjects without *chd* events are

$$covariance(chd, \hat{p}) = (\bar{p}_{11} - \bar{p}_{01})\hat{P}(1 - \hat{P}) = 0.00409$$

and

$$covariance(chd, \hat{p}_0) = (\bar{p}_{10} - \bar{p}_{00})\hat{P}(1 - \hat{P}) = 0.00373,$$

where $\hat{P} = 259/3854 = 0.067$ is an estimate of the overall probability of coronary heart disease event.

Estimates of the correlation coefficients between *chd* (a binary variable) and $\hat{p}$ or $\hat{p}_0$ (continuous variables), called a *point biserial correlation coefficient* (Chapter 5), are given by the expressions

$$cor(chd, \hat{p}) = (\bar{p}_{11} - \bar{p}_{01})\sqrt{\frac{\hat{P}(1 - \hat{P})}{variance(\hat{p})}} = 0.251$$

and

$$cor(chd, \hat{p}_0) = (\bar{p}_{00} - \bar{p}_{10})\sqrt{\frac{\hat{P}(1 - \hat{P})}{variance(\hat{p}_0)}} = 0.241.$$

Although these correlation coefficients add little to evaluating model differences in classification accuracy, the difference between correlations coefficients identifies interrelationships among model improvement measures. Specifically, the difference in estimated correlation

$$cor(chd, \hat{p}) - cor(chd, \hat{p}_0) = \left[\frac{\bar{p}_{11} - \bar{p}_{01}}{\sqrt{variance(\hat{p})}} - \frac{\bar{p}_{10} - \bar{p}_{00}}{\sqrt{variance(\hat{p}_0)}}\right]\sqrt{\hat{P}(1 - \hat{P})} = 0.010.$$

is a distance between two standardized differences between model-generated mean probabilities. These two correlation coefficients also can be calculated directly from the corresponding presence or absence of a *chd* event and model-estimated probabilities ($\hat{p}$ and $\hat{p}_0$). Of note: When exactly no improvement occurs or $\hat{p} = \hat{p}_0$ for all estimates, then, as expected, $\bar{p}_{11} - \bar{p}_{01} = 0$, $\bar{p}_{10} - \bar{p}_{00} = 0$, $idi = 0$, $variance(\hat{p}) = variance(\hat{p}_0)$, and $cor(chd, \hat{p}) - cor(chd, \hat{p}_0) = 0$.

Table 19.7 *Classification: Estimates of* nri *Based on Seven Selected Categories of Model-Estimated Probabilities from the Cholesterol Analysis (Percentiles:* 50%, 33%, 25%, 20%, 16%, *and* 14%*)*

Intervals/categories	$n\hat{r}i$	Chi-square	p-values
Percentiles (50%): 0.043	0.004	0.067	0.796
Percentiles (33%): 0.025 and 0.074	0.046	4.234	0.040
Percentiles (25%): 0.019, 0.043, and 0.095	0.052	4.092	0.043
Percentiles (20%): 0.016, 0.031, 060, and 0.111	0.073	5.778	0.016
Percentiles (16%): 0.014, 0.025, 0.043, 0.074, and 0.125	0.074	0.125	0.556
Percentiles (14%): 0.013, 0.022, 0.034, 0.055, 0.085, and 0.136	0.051	2.274	0.132

Epilogue

Two important choices should be examined when reclassification techniques are applied to identify differences between models. First, calculation of a net reclassification measure requires a selection of categories to classify the model-generated prediction probabilities. Second, a selection has to be made among a variety of possible measures of improvement.

Situations arise where the categories selected to create the *nri* statistic are natural choices or a consensus exists on the number and range of categories based on clinical or biological or traditional considerations. Thus, the analysis of *nri* summary values potentially describes important reasons for differences observed between compared models. When no persuasive reason exists to choose classification boundaries, results based on arbitrary choices are less rigorous and, therefore, are more difficult to accurately interpret (Chapter 10). For the ARIC case study of the cholesterol variable, different choices for the number of categories used to classify the predicted probabilities illustrate different results (Table 19.7).

The selection of a measure of improvement also brings an unwanted arbitrariness to the analysis. Moderate variation appears to exist among different measures of influence of each risk variable from the ARIC cholesterol data (Table 19.8). The corresponding chi-square statistics ([*estimate/standard error*]^2) provide commensurate comparisons. The summary measures used for the case study of *chd* events show a rough consistency, but, for analyses based on 3854 subjects, substantial differences remain.

Least Squares Estimation – Details

Data consisting of pairs of observations are frequently summarized by a straight line. Two basic estimation methods exist.

Ordinary least squares estimates of the intercept a and slope b of a line $y = a + bx$ are the values that produce the smallest possible value of L where

$$L = \sum (y_i - [a + bx_i])^2 \quad i = 1, 2, \ldots, n = \text{number of pairs.}$$

The distances $| y_i - (a + bx_i)|$ are measures parallel to the vertical axes (Figure 19.5, top). The outcome variable y is a value sampled from a distribution generated by a value of variable x. Formally, the x-values are viewed as fixed and the y-values as sampled from a distributions with different mean values and the same variance. Estimated mean values of these x-generated distributions are summarized by a line with slope and intercept estimated

Table 19.8 *Summary Measures: Improvement Analyses for Seven Variables (One-at-a-Time) Using a Logistic Model Analysis Applied to ARIC Data (Model X versus Model X + C)*

	NRI	X_1^2	*nri*	X_2^2	*idi*	X_3^2	$\hat{B}$	X_4^2	$\hat{b}$	X_5^2	X_6^2
sex	0.697	117.5	0.382	57.1	0.031	83.9	1.14	69.4	−1.55	85.1	100.4
sbp	0.026	0.2	0.015	0.9	0.001	1.4	0.02	1.9	0.01	1.7	1.4
dbp	0.103	2.5	0.135	18.1	0.006	10.8	0.22	20.5	−0.04	16.8	13.6
smk	0.452	49.4	0.135	16.2	0.006	15.4	0.12	12.5	–[a]	–[a]	18.1
bmi	0.238	13.6	0.074	9.2	0.003	6.6	0.05	4.7	0.04	9.3	8.8
age	0.323	28.5	0.184	25.9	0.009	13.3	0.13	4.5	0.07	24.5	28.5
chol	0.231	12.9	0.109	13.1	0.006	7.2	0.15	13.0	0.01	17.6	17.0

[a]A three-level categorical variable.
Key:
1. *NRI* = nri measure (binary)
2. *nri* = nri measure
3. *idi* = idi index
4. $\hat{B}$ = least squares perpendicular slope
5. $\hat{b}$ = logistic model-estimated coefficient
6. X_6^2 = chi-square likelihood ratio test.

(Chapter 3) by

$$\hat{b} = \frac{S_{XY}}{S_X^2} \quad \text{and} \quad \hat{a} = \bar{y} - \hat{b}\bar{x}.$$

Using 50 pairs of blood pressure observations consisting of diastolic (*dbp*) and systolic (*sbp*) values, an ordinary least estimated line summarizes the relationship between pairs of values (Figure 19.6, dashed line). However, these two blood pressure measures are sampled from a joint distribution of values and no natural choice for x (fixed) and y (variable) exists.

From another point of view, these is no natural choice of the dependent (y) and independent (x) variables. Different choices produce different analytic results.

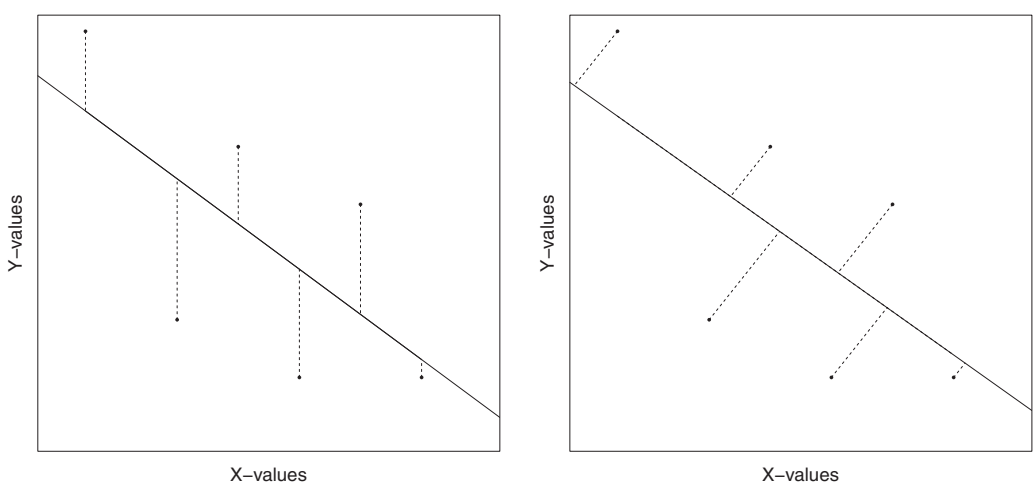


Figure 19.5 Two Views of Least Square Estimation of Summary Straight Line for Same Six x/y-Values

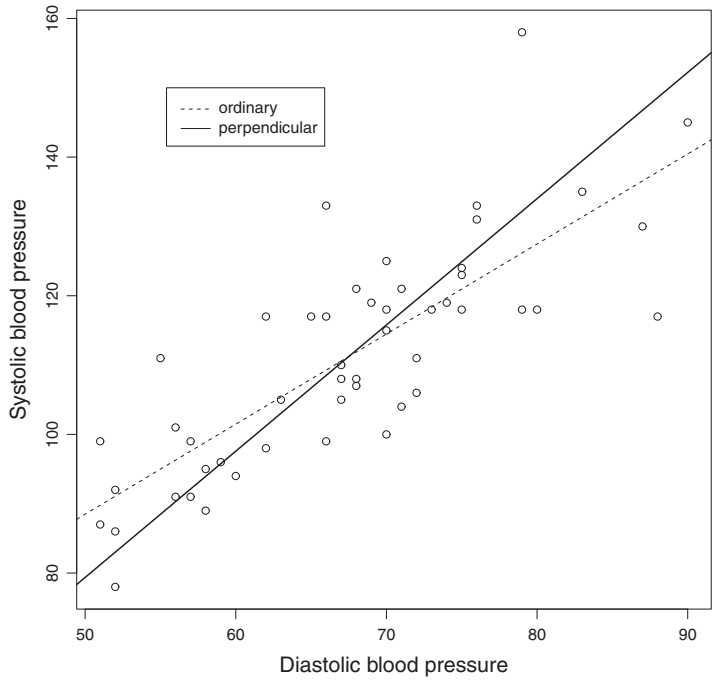


Figure 19.6 Example Joint Distribution of Diastolic and Systolic Blood Pressure Data Summarized by Straight Lines

Both blood pressure variables have distributions with different mean values and variances and both variables play symmetric roles in the description of their relationship. Because neither systolic or diastolic blood pressure measures are naturally associated with the vertical or horizontal axes, the estimates of slope B and intercept A are values that produce the smallest possible value of L' where

$$L' = \frac{\sum(y_i - [A + Bx_i])^2}{\sqrt{1 + B^2}} \quad i = 1, 2, \ldots, n = \text{number of pairs.}$$

The distances $|(y_i - [A + Bx_i])/\sqrt{1 + B^2}|$ are perpendicular measures of the distance from the value y to the estimated line (Figure 19.5, bottom). The expression to estimate slope $\hat{B}$ is

$$\hat{B} = -D_{X,Y} + \sqrt{1 + D_{X,Y}^2},$$

where

$$D_{X,Y} = \frac{1}{2}\left[\frac{S_X^2 - S_Y^2}{S_{XY}}\right]$$

and the estimated intercept is, as usual (Chapters 3 and 15),

$$\hat{A} = \bar{y} - \hat{B}\bar{x}.$$

The expression for the estimated slope, particularly the form of the value D, generates symmetric roles for both x and y variables.

Table 19.9 *Data: Pairs of Systolic (*sbp*) and Diastolic (*dbp*) Blood Pressures* ($n = 50$)

	Blood pressure data										
sbp	108	117	108	96	91	131	117	130	145	98	118
dbp	67	88	68	59	56	76	65	87	90	62	80
sbp	111	86	94	125	91	133	105	133	110	99	118
dbp	72	52	60	70	57	66	63	76	67	51	79
sbp	99	87	115	95	124	118	121	123	121	78	107
dbp	66	51	70	58	75	75	71	75	68	52	68
sbp	119	89	100	158	117	106	101	105	118	118	104
dbp	74	58	70	79	66	72	56	67	73	70	71
sbp	119	135	92	111	99	117					
dbp	69	83	52	55	57	62					

Values to estimate ordinary and perpendicular least squares straight lines from the blood pressure data are (Table 19.9) the following:

Mean Values		Variances		Covariance	
$\overline{dbp}$	$\overline{sbp}$	S^2_{dbp}	S^2_{sbp}	$S^2_{dbp,sbp}$	$D^2_{sbp,dbp}$
67.480	111.200	97.071	257.510	126.106	−0.636

Ordinary least squares estimates are

$$\text{slope: } \hat{b} = \frac{S_{dbp,sbp}}{S^2_{dbp}} = \frac{126.106}{97.071} = 1.299$$

and

$$\text{intercept: } \hat{a} = \overline{sbp} - \hat{b}(\overline{dbp}) = 111.200 - 1.299(67.480) = 23.536.$$

The ordinary least squares estimated line is $s\hat{b}p_i = 23.536 + 1.299(dbp_{\text{i}})$ (Figure 19.6, dashed line).

Perpendicular least squares estimates are

$$\text{slope: } \hat{B} = -D_{sbp,dbp} + \sqrt{1 + D^2_{sbp,dbp}} = 0.636 + \sqrt{1 + 0.636^2} = 1.821$$

and

$$\text{intercept: } \hat{A} = \overline{sbp} - \hat{B}(\overline{dbp}) = 111.20 - 1.821(67.480) = -11.702.$$

The perpendicular least squares estimated line is $s\hat{b}p_i = -11.702 + 1.821(dbp_i)$ (Figure 19.6, solid line). Note that because

$$\hat{B} = -D + \sqrt{1 + D^2} \quad \text{and} \quad \hat{B}' = -D - \sqrt{1 + D^2}, \quad \text{then} \quad \hat{B} = -\frac{1}{\hat{B}'}.$$

Thus, the estimated slope $\hat{B}$ is perpendicular to the slope $\hat{B}'$ (x and y values reversed). Consequently, an analysis based on a perpendicular least squares estimated line does not depend on a choice of specific independent and dependent variables, unlike ordinary least squares estimation. Therefore, statistical assessment of $\hat{B}$ for pairs (x, y) and $-\frac{1}{\hat{B}}$ for pairs (y, x) gives the same analytic result.

20

The Attributable Risk Summary

Many names exist (16 claimed by one author) for the epidemiologic measure most often called *attributable risk*. To list a few:

Population attributable risk
Etiologic fraction
Attributable risk percentage
Fraction of etiology and
Excess incidence.

Similarly a variety of definitions of attributable risk summary exist. A few examples are the following:

1. The definition from the original paper that introduced the attributable risk (Levin, 1953).
   The index [attributable risk] of the maximum proportion of lung cancer cases attributable to smoking. This index is based on the assumption that smokers, if they had not become smokers, would have had the same incidence of lung cancer found in nonsmokers.
2. The proportion of the total load of disease that is attributable to a given factor is the faction of the disease that would not have occurred had the factor been absent from the population.
3. The term population attributable risk has been described as the reduction in incidence that would be observed if the population were entirely unexposed, relative to its current exposure pattern.
4. Attributable risk is that fraction of the exposed cases that would not have occurred if the exposure had not occurred.
5. Attributable risk is that fraction of all cases d in the population that can be attributed to exposure E.
6. A definition of attributable risk (denoted ar), in symbols, is

$$\text{attributable risk } = ar = \frac{P(D) - P(D|\bar{E})}{PD} \qquad (0 \leq ar \leq 1)\,,$$

where D represents a binary measure of disease (present or absent) and E represents a binary measure of exposure (present or absent). Despite the many names and definitions, the statistical structure is unequivocal. Issues, however, arise with interpretation.

Attributable risk applies to binary disease and exposure variables. Thus, an attributable risk summary is conveniently estimated from counts contained in a 2×2 table (Table 20.1).

Table 20.1 *Notation: Counts Contained in a* 2×2 *Table Used to Define an Attributable Risk Summary and Its Properties (D = Disease and E = Exposure) (Chapter 6)*

	D	$\bar{D}$	Total
E	a	b	$a+b$
$\bar{E}$	c	d	$c+b$
Total	$a+c$	$b+d$	n

The most common expression to estimate an attributable risk measure from cohort or cross-sectional data is

$$\hat{ar} = \frac{\hat{P} - \hat{p}}{\hat{P}},$$

where $\hat{P} = (a+c)/n$ estimates the probability $P(D)$, $\hat{p} = c/(c+d)$ estimates the probability $P(D|\bar{E})$, and $n = a + b + c + d$ represents the total number of observations. Consider data describing coronary heart disease event (D) and smoking exposure (E) (Table 20.2). A routine chi-square analysis leaves little doubt that smoking and coronary heart disease are associated ($X^2 = 22.099$ with p-value < 0.001) (Chapter 6).

From the *chd* data, the estimated attributable risk associated with smoking exposure is

$$\hat{P} = \frac{a+c}{n} = \frac{257}{3153} = 0.082 \quad \text{and} \quad \hat{p} = \frac{c}{c+d} = \frac{98}{1651} = 0.059,$$

then

$$\text{estimated attributable risk } = \hat{ar} = \frac{\hat{P} - \hat{p}}{\hat{P}} = \frac{0.082 - 0.059}{0.082} = 0.272$$

or, directly from the 2×2 table,

$$\text{estimated attributable risk } = \hat{ar} = \frac{ad - bc}{(a+c)(c+d)} = \frac{(159)(1553) - (1343)(98)}{(257)(1651)} = 0.272.$$

Several other equivalent expressions to estimate attributable risk are useful. For example,

$$\hat{ar} = \frac{\hat{P}(E)(\hat{rr} - 1)}{\hat{P}(E)(\hat{rr} - 1) + 1}$$

Table 20.2 *Data* 2×2 *Table: Coronary Heart Disease Event* $(D/\bar{D})$ *and Smoking Exposure Status* $(E/\bar{E})$ *from* $n = 3153$ *High-Risk Men*

	chd (D)	*no-chd* ($\bar{D}$)	Total
Smoker (E)	159	1343	1502
Nonsmoker ($\bar{E}$)	98	1553	1651
Total	257	2896	3153

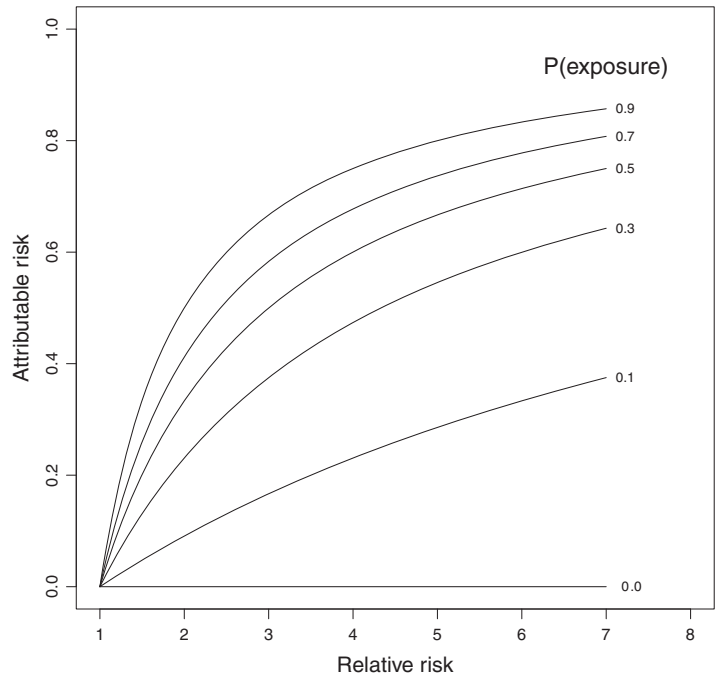


Figure 20.1 Relationship between Relative Risk and Attributable Risk for Six Probabilities of Exposure

when $\hat{rr}$ represents the relative risk of disease (Chapter 6). Relative risk and the probability of exposure estimated from the *chd* data are

$$\text{relative risk} = \hat{rr} = \frac{\hat{P}(D|E)}{\hat{P}(D|\bar{E})} = \frac{159/1502}{98/1651} = 1.783 \text{ and } \hat{P}(E) = \frac{a+b}{n} = \frac{1502}{3153} = 0.476.$$

The estimated attributable risk becomes

$$\hat{ar} = \frac{0.476(1.783-1)}{0.476(1.783-1)+1} = 0.272.$$

An estimate of the attributable risk summary depends directly on the probability of exposure. Thus, when exposure is rare, the attributable risk is small regardless of the magnitude of the risk. For example, when the probability of exposure is $P(E) = 0.02$ and the associated relative risk is $\hat{rr} = 5$, the attributable risk is $ar = 0.07$. More fundamentally, an unbiased interpretation of the attributable risk summary requires that the exposure status under study be accurately determined, sometimes a difficult task. A detailed picture of the exposure-risk relationship in terms of relative risk is displayed in Figure 20.1.

A fourth version of an estimate of attributable risk results from a weighted average of estimated relative risk ratios (rr_i) using the probabilities of exposure (e_i) as weights. For the *chd* data, the estimate from the 2×2 table is

$$\hat{ar} = 1 - \frac{1}{\sum e_i \hat{rr}_i} = 1 - \frac{1}{0.476(1.783) + 0.524(1.0)} = 0.272,$$

where $e_1 = (a + b)/n = 0.476$ $[P(E)]$, $e_2 = (c + d)/n = 0.524$ $[P(\bar{E})]$, $rr_1 = \hat{rr} = 1.783$, and $rr_2 = 1.0$. Like the other estimated values, this weighted average produces exactly the same estimate (algebraically identical). Unlike the other expressions, the weighted average can be extended to estimate attributable risk from more than two levels of exposure.

Yet one more expression to estimate an attributable risk is

$$\hat{ar} = \frac{a}{a+c}\left[\frac{\hat{rr}-1}{\hat{rr}}\right] = 0.619\left[\frac{0.783}{1.783}\right] = 0.272,$$

based directly on the probability of exposure among diseased individuals $P(E|D)$, estimated by $a/(a + c)$. Again $\hat{rr}$ represents the estimated relative risk of disease.

For case/control data, probabilities $P(D)$ and $P(D|E)$ cannot be accurately estimated. Nevertheless, when the "rare-disease assumption" applies, an estimate of an approximate attributable risk is possible (Chapter 6). In other words, the odds ratio reflects risk; that is, when the estimated odds ratio approximately equals the relative risk ($\hat{or} \approx \hat{rr}$) or, in more practical terms, when $a + b \approx b$ and $c + d \approx d$, the estimated odds ratio can be treated in the same way as an estimated relative risk (Chapter 6). This approximate measure of risk allows calculation of an approximate attributable risk based on an odds ratio in this special case.

As is often the case, the logarithm of the estimated attributable risk ratio produces a more accurate evaluation when applying a normal distribution approximation. An approximate 95% confidence interval can be constructed based on the normal distribution, the transformed estimate $log(1 - \hat{ar})$ and its variance. The estimated variance of the distribution of the estimated value $log(1 - \hat{ar})$ is

$$variance\,(log\,[1 - \hat{ar}]) = \hat{V} = \frac{b + \hat{ar}(a+d)}{nc}.$$

From the *chd*-smoking data,

$$\text{lower bound} = \hat{A} = -0.317 - 1.960\,(0.076) = -0.467$$

and

$$\text{upper bound} = \hat{B} = -0.317 + 1.960\,(0.076) = -0.167,$$

where the estimate $log(1 - \hat{ar}) = log(1 - 0.272) = -0.317$ and the estimated $\hat{V} = variance\,(1 - \hat{ar}) = 0.076$. Therefore, the approximate 95% confidence interval for the population attributable risk becomes

$$\text{lower bound} = 1 - e^{\hat{B}} = 1 - e^{-0.167} = 0.154$$

and

$$\text{upper bound} = 1 - e^{\hat{A}} = 1 - e^{-0.467} = 0.373,$$

and, as required, the estimated attributable risk is

$$1 - e^{log(1-\hat{ar})} = 1 - e^{-0.317} = 0.272 = \hat{ar}.$$

An approach based on a Poisson loglinear regression model produces once again the identical estimated attributable risk (Chapter 9). The regression model is

$$log\,[P(D)] = a + bF,$$

Table 20.3 *Estimates: Poisson Regression Analysis of* chd-*Smoking Data*

	Estimates	*s. e.*	*z*-value	*p*-value
Intercept ($\hat{a}$)	−2.824	–	–	–
Coefficient ($\hat{b}$)	0.579	0.128	4.505	<0.001

where $F = 0$ indicates unexposed and $F = 1$ indicates exposed. The model yields an estimate of $P(D|\bar{E})$, denoted again $\hat{p}$. For the example smoking exposure data, the estimated regression model coefficients are $\hat{a} = -2.824$ and $\hat{b} = 0.579$ (Table 20.3).

Therefore, estimated from the model

$$\hat{p} = \hat{P}(D|\bar{E}) = \hat{P}(D|F = 0) = e^{\hat{a}+\hat{b}F} = e^{-2.824+0.579(0)} = 0.059.$$

As before, the estimate $\hat{P} = 257/3153 = 0.082$ and, identically, the estimated attributable risk is

$$\hat{ar} = \frac{\hat{P} - \hat{p}}{\hat{P}} = \frac{0.082 - 0.059}{0.881} = 0.272.$$

The estimate $\hat{p}$ from the regression model is the same as the direct data estimate because a two parameter model applied to a 2×2 table exactly replicates any calculation made from the original data. A Poisson model to estimate the probability $P(D|\bar{E})$ becomes useful as part of a general regression model strategy (to be discussed).

Incidence Rates

The pattern described to estimate an attributable risk from a 2×2 table also applies to incidence rates. Consider two incident rates:

r_0 represents a rate from an unexposed population size P_0

and

r_1 represents a rate from an exposed population size P_1

for a specific time period. An estimated attributable risk is then

$$\hat{ar} = \frac{r - r_0}{r},$$

where $r = e\ r_1 + (1 - e)\ r_0$ estimates the overall population rate when represents the proportion of exposed individuals. Specifically, the probability of exposure is $e = P_1/(P_1 + P_0)$.

For example, using data from a recent study, the incidence rate of lung cancer is $r_0 =$ 41.1 per 100,000 persons at risk among nonsmokers estimated from $P_0 = 202{,}776$ individuals. The incidence rate is $r_1 = 91.3$ per 100,000 persons at risk among smokers estimated from $P_1 = 262{,}663$ individuals. The estimated proportion of exposed individuals is

Table 20.4 *Artificial Data: Calculation of Summary Attributable Risk Estimate for Four Age Strata*

	Unexposed			Exposed			Combined		
Age	d_{i0}[a]	P_{i0}	$\hat{r}_{i0}$	d_{i1}[a]	P_{i1}	$\hat{r}_{i1}$	D_i	$r\hat{r}_i$	$\hat{ar}_i$
30–40	5	100	0.050	10	100	0.10	0.07	2.0	0.33
40–50	15	200	0.075	36	200	0.18	0.26	2.4	0.41
50–60	21	350	0.060	54	300	0.18	0.39	3.0	0.49
60–70	15	600	0.025	40	500	0.08	0.29	3.2	0.52

[a]The symbols d_{i0} and d_{i1} represent the observed number of deaths in unexposed and exposed populations for $i = 1, 2, 3$, and $4 =$ age strata.

$e_1 = P_1/(P_0 + P_1) = 262{,}663/465{,}439 = 0.564$ and unexposed $e_0 = 1 - 0.564 = 0.436$. Thus, the estimated population lung cancer rate ignoring smoking status is

$$r = 0.564\,(91.3) + 0.436\,(41.1) = 69.430 \text{ per } 100{,}000 \text{ persons-at-risk.}$$

The estimated attributable risk from smoking is then

$$\hat{ar} = \frac{r - r_0}{r} = \frac{69.430 - 41.1}{69.430} = 0.408$$

or, alternatively,

$$\hat{ar} = \frac{e(\hat{R} - 1)}{e(\hat{R} - 1) + 1} = \frac{0.564(2.221 - 1)}{0.564(2.221 - 1) + 1} = 0.408$$

based on the estimated rate ratio $\hat{R} = r_1/r_0 = 91.3/41.1 = 2.221$.

The fact that calculation of an attributable risk estimated from counts in a 2×2 table and one estimated from two incidence rates do not differ in principal can be mathematically demonstrated. More simply, in human populations the probability of disease or death frequently is approximately equal to a rate multiplied by the length of the time interval considered (denoted δ). In symbols, the probability $P(\textit{disease}) = P(D) \approx \textit{rate} \times \textit{interval length} = r\delta$ (Chapter 16). From this intuitive point of view, an attributable risk estimate then becomes

$$ar = \frac{P(D) - P(D|E)}{P(D)} \approx \frac{r\delta - r_0\delta}{r\delta} = \frac{r - r_0}{r}.$$

Situations arise where an estimate of a summary attributable risk is desired from stratified incidence rates.

The following artificial data provide an illustrations of one approach (Table 20.4).

A number of summary attributable risk estimates exist that are simply extensions of an expression already described. For example, one of several versions is

$$\overline{ar} = \sum D_i \hat{ar}_i,$$

where $D_i = d_{i1}/\Sigma d_{i1}$ represents the probability of death in the ith strata of exposed individuals. The summary attributable risk from the example data becomes $\overline{ar} = 0.466$ (Table 20.4).

Table 20.5 *Coronary Heart Disease Data: Four Combinations of Two Binary Exposure Variables*[a] Yielding Estimates of Exposure ($\hat{e}$) *and Relative Risk* ($\hat{rr}$)

	smk	*bp*	*chd*	*no-chd*	$\hat{e}_i$	$\hat{rr}_i$
1	1	1	59	271	0.105	3.778
2	0	1	37	325	0.115	2.160
3	1	0	100	1072	0.372	2.803
4	0	0	61	1228	0.409	1.000

[a]Risk factor present = 1 and risk factor absent = 0.

Two Risk Factors

Three binary attributable risk estimates generated from the same *chd* data displayed in Table 20.2:

smoking (exposed = smoker): $\hat{ar}_{smk} = 0.272$
blood pressure (exposed = systolic blood pressure $\geq$ 140): $\hat{ar}_{bp} = 0.470$ and
body weight (exposed = weight $\geq$ median): $\hat{ar}_{wt} = 0.212$.

One might expect the summary estimated attributable risk to be $\hat{ar}_{combined} = 0.272 + 0.470 + 0.212 = 0.935$. Clearly, eliminating three sources of exposure does not almost entirely eliminate coronary heart disease risk.

It is instructive to estimate attributable risk for smokers including the influence of elevated blood pressure (Table 20.5).

In this case, the estimate of attributable risk is

$$\hat{ar} = \frac{\hat{P}(D) - \hat{P}(D|\bar{E}_1, \bar{E}_2)}{\hat{P}(D)}.$$

For $\hat{P} = \hat{P}(D) = 257/3153 = 0.082$ and $\hat{p} = \hat{P}(D/\bar{E}_1, \bar{E}_2) = 61/1289 = 0.047$, the estimated attributable risk from both risk factors is $\hat{ar} = (\hat{P} - \hat{p})/\hat{P} = (0.082 - 0.047)/0.082 = 0.419$ (Table 20.5).

In addition, the exposure weighed sum of the four relative risk estimates $\Sigma \hat{e}_i \hat{rr}_i = \hat{P}/\hat{p} = 1.722$ directly produces the same estimate

$$\hat{ar} = 1 - \frac{1}{\sum e_i rr_i} = 1 - \frac{1}{P/p} = 1 - \frac{1}{1.722} = 1 - 0.581 = 0.419.$$

Thus, attributable risk is again a function of relative risk measures.

When several risk factors are considered, the probability of disease among unexposed individuals is compared to the probability of disease ignoring exposure status. In symbols, the attributable risk from the k exposure groups is

$$ar = \frac{P(D) - P(D|\bar{E}_1, \bar{E}_2, \bar{E}_3, \ldots, \bar{E}_k)}{P(D)}.$$

Of particular importance is the property that the attributable risk estimate depends entirely on complete absence of all relevant exposures. Thus, the pattern of risk among exposure categories does not influence the estimated value. For the example, the number of individuals

Table 20.6 *Coronary Heart Disease Data: Attributable Risk Calculation for Two Binary Exposed/Unexposed Variables (Smoking and Blood Pressure)*

	chd (D)	*no-chd* ($\bar{D}$)	Total
E – one or more exposures	196	1668	1864
$\bar{E}$ – no exposures	61	1228	1289
Total	257	2896	3153

at risk among those who do not smoke but have high blood pressure only contribute to the total of at-risk individuals. The 2×2 table formed by combining all exposure categories, creating a single exposure category, is all that is required to estimate the attributable risk (Table 20.6). For the example among the *chd* subjects, exposed $= 59 + 37 + 100 = 196$ and unexposed $= 61$ (Tables 20.5 and 20.6).

and

$$\hat{ar} = \frac{\hat{P} - \hat{p}}{\hat{P}} = \frac{257/3153 - 61/1289}{257/3153} = 0.419.$$

Thus, the estimated attributable risk is identical to the previous weighted average estimate $\hat{ar} = 0.419$ (Table 20.5). For an attributable risk calculated from a number of exposure categories, the reduction in risk applies to only those individuals who have no exposure relative to the probability of disease ignoring exposure.

Adjustment by Stratification

To illustrate an estimate of attributable risk adjusted for the possible confounding influence of cholesterol on the smoking/*chd* relationship, consider the data in Table 20.7. One possibility

Table 20.7 *Attributable Risk: Estimates from Smoking and Coronary Heart Disease Data Stratified into Four Categories of Cholesterol Levels* ($n = 3153$ – *High-Risk Study Participants*)

	chd	*no-chd*	D_i[a]	$\hat{p}_i$	$\hat{P}_i$	$\hat{ar}_i$	$\hat{V}_i$[b]
Cholesterol < 200							
Smoker	14	333					
Nonsmoker	17	467	0.121	0.035	0.037	0.058	0.023
Cholesterol 200–220							
Smoker	17	248					
Nonsmoker	12	343	0.113	0.034	0.047	0.277	0.024
Cholesterol 220–250							
Smoker	47	381					
Nonsmoker	27	374	0.289	0.067	0.089	0.246	0.012
Cholesterol > 250							
Smoker	81	381					
Nonsmoker	41	369	0.477	0.100	0.140	0.285	0.007

[a] D_i represents the proportion of strata-specific coronary events and variances
[b] $\hat{V}_i = variance(\hat{ar}_i)$(Chapter 6).

Table 20.8 *Poisson Loglinear Regression Model: Results from Analysis of Coronary Heart Disease and Smoking (Exposure) Adjusted for Four Levels of Cholesterol*

	Estimates	*s. e.*	*z*-values	*p*-values
Intercept ($\hat{a}$)	−3.521	0.193	–	–
Exposure ($\hat{b}$)	0.486	0.129	3.772	<0.001
Coefficient ($\hat{c}_1$)	0.0	–	–	–
Coefficient ($\hat{c}_2$)	0.254	0.256	0.993	0.321
Coefficient ($\hat{c}_3$)	0.825	0.214	3.848	<0.001
Coefficient ($\hat{c}_4$)	0.202	6.287	6.287	<0.001

for a summary attributable risk adjusted for influences from cholesterol is the weighted average

$$\overline{ar} = \frac{\sum w_i \hat{a}r_i}{\sum w_i} = 0.241 \quad i = 1, 2, 3, \text{ and } 4,$$

where the weights w_i are the reciprocal of the estimated variances of each attributable risk estimate V_i (Table 20.7). In symbols, the weights are $w_i = 1/\hat{V}_i$. An estimate of the variance of this summary attributable risk estimate is then *variance* $(\overline{ar}) = 1/\Sigma\, w_i = 1/303.68 = 0.0033$ (Chapter 3). Therefore, an approximate 95% confidence interval for the value estimated by $ar = 0.241$ becomes

$$\overline{ar} \pm \sqrt{variance(\overline{ar})} = 0.241 \pm 1.960\,(0.057) \rightarrow (0.128, 0.353).$$

Choosing the weights as the proportion of diseased individuals in each strata (D_i) produces a summary estimate called the *case-load estimate.* Adjusted for the influence of cholesterol, this summary attributable risk value is the weighted average $\overline{ar'} = \Sigma\, D_i \hat{a}r_i = 0.245$ (weights = $w_i = D_i$ and $\Sigma D_i = 1.0$ – Table 20.7).

Adjustment by Model

To adjust for the influence of differing levels of cholesterol, an additive Poisson model

$$\text{log-probability of disease} = log\,[P(D)] = a + bE + c_1C_1 + c_2C_2 + c_3C_3 + c_4C_4$$

can be applied where the C_i-variables represent binary components of a design variable identifying four cholesterol strata and E represents a binary smoking/nonsmoking exposure variable (Table 20.7). The regression model estimated coefficients from the example coronary heart disease and smoking data produce an adjusted estimated relative risk (Table 20.8) (Chapter 9).

Adjusted for the influence of cholesterol, model estimated relative risk associated with smoking exposure is

$$\hat{rr} = e^{\hat{b}} = e^{0.486} = 1.626,$$

making the adjusted estimate of attributable risk

$$\hat{ar} = \frac{\hat{P}(E)(\hat{rr} - 1)}{\hat{P}(E)(\hat{rr} - 1) + 1} = \frac{0.476(1.626 - 1)}{0.476(1.626 - 1) + 1} = 0.230,$$

where again the estimated probability of exposure is $\hat{P}(E) = 1502/3153 = 0.476$ (Table 20.2).

A model approach is naturally extended to include other possibly risk variables. For example, to include a binary measure of systolic blood pressure risk (blood pressure < 140, coded $F = 0$, and blood pressure ≥ 140, coded $F = 1$), an extended Poisson model becomes

$$\text{log-probability of disease} = log\,[P(D)] = a + bE + cF + c_1C_1 + c_2C_2 + c_3C_3 + c_4C_4.$$

The Poisson model estimated relative risk increases to $\hat{rr} = \hat{b} = e^{0.630} = 1.878$, and the adjusted estimate of a summary attributable risk becomes $\hat{ar} = 0.295$ based once again on the model adjusted estimated relative risk.

Issues of Interpretation

An estimated attributable risk $\hat{ar}$ can be occasionally negative. Sampling variation can cause the estimate of the probability of disease $\hat{P}$ to be less than the estimate $\hat{p}$. Or, exposure could reduce risk, again causing $P(D)$ to be less than $P(D/E)$. In both cases an estimate less than zero ($\hat{ar} < 0$) has no useful interpretation. It is worth noting that to estimate attributable risk the measure of exposure is required to be binary or made to be binary. Furthermore, many exposure variables are not discrete, and even truly discrete exposure variables rarely measure dose accurately, such as a smoking/nonsmoking exposure variable based on reported cigarette consumption.

The probability of disease among exposed individuals is not a simple quantity; that is, the probability $P(D|E)$ is potentially a combination of several different exposure influences. In symbols, a simple example is

$$P(D \mid E \text{ and } E \text{ is the cause of the disease})$$

or

$$P(D \mid E \text{ and } E \textit{ is the not the cause of the disease}).$$

For example, individuals with coronary heart disease can be individuals who acquired the disease from smoking exposure or can be smokers who acquired the disease from another causes. Both kinds of smokers would be components of the estimated probability of disease among exposed individuals $P(D|E)$ and cannot be separated without additional and usually unavailable information.

A simple model illustrates a specific example of bias when unexposed individuals with the disease are misclassified as exposed. A 2×2 table displays the model (Table 20.9).

Table 20.9 *Illustration: Simple Model to Illustrate Potential Bias in Interpretation of an Attributable Risk Summary*

	Disease (D)	No disease ($\bar{D}$)
Exposed (E)	$P(DE) + P(D\bar{E})(1-\alpha)$	$P(\bar{D}E)$
Unexposed ($\bar{E}$)	$P(D\bar{E})\alpha$	$P(\bar{D}\bar{E})$

The symbol α represents the probability of a correct classification of exposure among the individuals with the disease. Then, the probability $P(D|\bar{E})$ is

$$P(D \mid \bar{E}) = \frac{P(D\bar{E})\alpha}{P(D\bar{E})\alpha + P(\bar{D}\bar{E})}$$

and attributable risk (denoted ar_α) is

$$ar_\alpha = \frac{P(D) - P(D\bar{E})\alpha \big/ \left[P(D\bar{E})\alpha + P(\bar{D}\bar{E})\right]}{P(D)}.$$

For example, for $\alpha = 1.0$, then $\hat{a}r = 0.272$ from the *chd*/smoking data (Table 20.2). When $\alpha \neq 0$, then $\hat{a}r_{0.95} = 0.306$, $\hat{a}r_{0.90} = 0.341$ and $\hat{a}r_{0.80} = 0.410$. Any misclassification ($\alpha \neq 1$) influences the observed level of attributable risk. The factor α is rarely considered and almost always assumed to be equal to one.

Theory and unbiased estimation of attributable risk require the assumption or knowledge that eliminating exposure does not affect other relationships. For example, when the contribution to coronary heart disease risk from smoking is eliminated, it is assumed that all other risk factors continue to have exactly the same influences. When the elimination of smoking causes increase food consumption resulting in weight gain (a risk factor for coronary heart disease), the estimate of the attributable risk becomes biased, likely too small. On the other hand, when elimination of smoking is accompanied by adopting a healthier lifestyle, the estimate is again biased, likely too large. Thus, individuals who are exposed must not differ from unexposed individuals in any way related to the exposure or disease under study to produce an accurately interpreted attributable risk estimate.

A Few Less Specific Issues of Interpretation

A statistical summary is most easily interpreted when applied to describe the properties of sampled data. Such issues as choice of scale, existence of interactions, and decisions about confounding variables are frequently manageable. When statistical measures derived from observational data are used to describe population relationships, their accuracy is often suspect because useful interpretation is considerably more difficult. The correspondence between a statistical summary based on data and actual population relationships is a complex and often ignored issue. Statistical analysis is most effective for understanding the properties of the data sampled but is only one element in describing corresponding properties in a population. For an estimate, such as an attributable risk of $\hat{a}r = 0.272$, dealing with population accuracy or validity issues is not simple. If data subject to statistical analysis were highly accurate, horse racing and sports gambling would disappear.

A useful interpretation of the consequences of elimination of a risk factor, measured by an estimated attributable risk, almost always depends on important unmeasured variables and unverified assumptions about the population sampled. In general, approximate or even precise statistical measures based on observational data are typically influenced by a variety of unmeasured and uncontrolled factors. Extrapolating statistical inferences beyond sample data is always at best difficult and certainly is vulnerable to misleading interpretation.

21

Time-Space Analysis

A property of infectious disease is a time-space association. Illnesses arise at similar times among individuals in proximity. This expected relationship is due to the fact that underlying causes are transmitted by contact, for example, bacteria or viruses. For noninfectious diseases, underlying mechanisms of transmission are certainly more complex, but interest remains in evaluating the extent to which observed patterns are similar to those of infectious diseases. For example, childhood leukemia and Hodgkin lymphoma have a long and well-documented history of apparent clustering in time and location. It is not known, however, whether these clusters of cases are rare occurrences of striking but random variation or reflect common underlying causes producing the kinds of spatial patterns observed in infectious disease data.

The risk of disease varies by location. Also, risk of disease varies by time of occurrence. Spatial statistical methods to evaluate evidence of time-space association are particularly efficient because they do not require the data to be a sample from an identified at-risk population; that is, most time-space methods require only that time of onset and location of observed cases be recorded, sometimes referred to as "case only" data.

A statistical analysis typically begins with creating a few summary measures to reflect the relationships within collected data and exploring the properties of these values. Thus, a large number of frequently unmanageable observations are reduced to a few values to address the issues at hand. Time-space analysis begins in the opposite direction. The first step is to increase the number of observations and complexity of the data to be analyzed. Specifically, collected times of occurrence and locations of the cases are transformed into all possible $N = n(n-1)/2$ time-distance pairs where n represents the number of cases observed. For a specific disease, case i located at location (x_i, y_i) occurring at time t_i and case j located at location (x_j, y_j) occurring at time t_j, create the time-distance pair

$$\text{time difference between cases} = t_{ij} = |t_i - t_j| \text{ and}$$
$$\text{distance between cases} = d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}.$$

These "new" data become N slightly correlated time-distance pairs, denoted (t_{ij}, d_{ij}), for $i > j = 1, 2, \ldots, N = n(n-1)/2$ pairs and are the basis of case only time-space analysis.

Knox's Time-Space Method

A statistical approach, frequently referred to as the *Knox method,* for identifying time-space clustering is simple and often effective and utilizes standard statistical methods. The summary statistic used to identify possible nonrandom clustering of disease cases is a 2×2 table. To

Table 21.1 *Notation: Description of the Knox Analytic Time-Distance 2 × 2 Table*

	Time		
Distance	$t \le t_0$	$t > t_0$	Total
$d \le d_0$	$n_{11} = X + m$	$n_{12} = a$	$X + m + a$
$d > d_0$	$n_{21} = a$	$n_{22} = N - (X + m + 2a)$	$N - (X + m + a)$
Total	$X + m + a$	$N - (X + m + a)$	$N = n(n-1)/2$

start, a distance is also chosen (denoted d_0) that divides the distribution of N distances (D) into "close" ($D \le d_0$) and "not close" ($D > d_0$). A time is chosen (denoted t_0) that divides the distribution of N time differences (T) into "close" ($T \le t_0$) and "not close" ($T > t_0$). For the following, values t_0 and d_0 are selected to be the same quantile/percentile value of their respective distributions of time differences and distances. Based on t_0 and d_0, two binary variables "close and not close" in time and distance are created and classified into a 2 × 2 Knox analysis table (Table 21.1).

The symbol n again represents the number of original observed cases of disease, N represents the number of time-distance pairs created, and n_{ij} represents counts of the four kinds of time-distance pairs. The variable X represents the count of "close" but unrelated pairs, and m represents the number of nonrandom "close" pairs in the sample of time-space data. The central element of the Knox analysis is the observed number of "close" pairs (denoted $\hat{n}_{11}$). Thus, a 2 × 2 table describes four kinds of time-distance pairs and is characterized by a single summary value. More specifically, the statistical focus becomes a comparison of the theoretical count $X + m$ to the observed count $\hat{n}_{11}$ or, in symbols

	"Close"	"Not Close"	Total
Theoretical counts	$X + m$	$N - (X + m)$	N
Observed counts	$\hat{n}_{11}$	$\hat{n}_{12} + \hat{n}_{21} + \hat{n}_{22}$	N

The Statistical Question Becomes: Is $m = 0$?

Consider artificial time-space data ($n = 126$) made up of $X = 120$ randomly distributed observations and a cluster of six nonrandom observations ($m = 6$) creating $N = 7875$ time-distance pairs (Figure 21.1). The resulting 2 × 2 Knox table, where $t_0 = 3.072$ and $d_0 = 1.303$ are the fifth quantiles of their respective time and distance distributions, is a tabulation of these $N = 7875$ time-distance pairs created from $n = 126$ observations (Table 21.2). The Knox summary statistic is the observed number of "close" pairs, namely, $\hat{n}_{11} = 23$.

Or, more succinctly,

"close"	"not close"	Total
23	7852	7875

Table 21.2 *Data: Knox Analytic* 2×2 *Table Summarizing Spatial Time* (t) *and Distance* (d) *Data (Figure 21.1)*

	$t \le 3.072$	$t > 3.072$	Total
$d \le 1.303$	23	371	394
$d > 1.303$	371	7110	7481
Total	394	7481	$126(125)/2 = 7875$

The expected number of "close" pairs, when no spatial pattern exists ($m = 0$), is estimated from the Knox table marginal frequencies. An estimate of this value is

$$\hat{X} = \text{estimated number of "close" pairs} = \frac{(\hat{n}_{11} + \hat{n}_{12})^2}{N} = \frac{(\hat{n}_{11} + \hat{n}_{21})^2}{N}$$

and, for the example data,

$$\hat{X} = 7875\left[\frac{394}{7875}\right]^2 = 19.713$$

when no spatial pattern exists (Chapters 6 and 17). Thus, the estimated count $\hat{X}$ is 19.713 "close" pairs calculated as if no spatial pattern exists. The number of observed corresponding "close" pairs is 23.

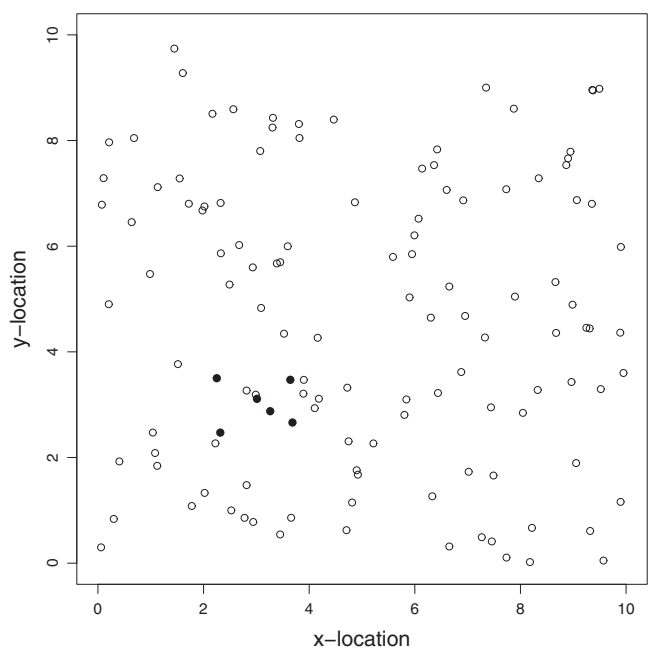


Figure 21.1 Artificial Random Data (Circles) with a Cluster of Six Nonrandom Values (Dots) to Illustrate the Analysis of Time-Space Data ($n = 126$)

The observed number of "close" pairs $\hat{X}$ frequently has an approximate Poisson probability distribution when no spatial pattern exists (Chapter 4). This assumption is realistic for time-space data that describe rare occurrences of a spatial outcomes among a large number of independent at-risk individuals. In this situation, a p-value calculated from a Poisson distribution (parameter $= \lambda$) is a direct evaluation of the probability that the number of "close" pairs arose by chance alone. In symbols, the Poisson probability is $P(\text{close} \geq \hat{n}_{11} \mid \text{no spatial pattern}) = p\text{-value}$. For the example spatial data ($n = 126$ and $\hat{n}_{11} = 23$), using the expected number of "close" pairs $\hat{\lambda} = \hat{X} = 19.713$, the directly calculated Poisson probability is $p\text{-value} = P(\text{close} \geq 23 \mid \text{no spatial pattern}) = 0.194$. The Poisson assumption is slightly flawed. The application of a Poisson distribution requires the observations to be independent observations, and time-space pairs are weakly associated.

Bootstrap sampling of the data (Table 21.2 $-$ 2000 replicate samples of each $N = 7875$ pairs) produces an empirical p-value of $P(\text{close} \geq 23) = 0.179$; that is, the bootstrap-estimated distribution of "close" pairs produces 358 replicate 2×2 tables with number of "close" pairs that equal or exceed the observed number of "close" pairs of $\hat{n}_{11} = 23$ yielding an assumption-free estimated p-value of $358/2000 = 0.179$. Thus, the example data (Figure 21.1) indicate, without the assumption that the observed time-distance pairs have a Poisson distribution and are uncorrelated, the observed number of "close" pairs is again marginally unlikely a result of only random variation.

The distribution of the square root of the number of "close" pairs $\hat{n}_{11}$ has an approximate normal distribution with estimated mean value $= \sqrt{\hat{n}_{11}}$ and variance $= 0.25$ when $\hat{n}_{11}$ has a Poisson probability distribution (Chapters 12 and 27). Thus, the test statistic

$$z = \frac{\sqrt{\hat{n}_{11}} - \sqrt{\hat{X}}}{\sqrt{0.25}}$$

has an approximate standard normal distribution when no time-space association exists ($m = 0$). This test statistic provides a significance probability of $p\text{-value} = 0.238$ from the example data based on the approximate normally distributed value $z = 0.712$. In addition, this approximation is effectively used to explore theoretical statistical issues associated with the Knox approach to the analysis of time-space data based on a Poisson distribution (Chapters 12 and 27).

The Knox method for identifying nonrandom clusters of disease is simple, has an approximate normal distribution, and, of particular importance, provides protection from extremely large time and distance pairs likely to disproportionately influence analytic results. Large noninformative values are present in time-space data because creating all possible N time-distance pairs necessarily produces the largest possible time and distance differences as part of the pairs data, which potentially obscure the identification of nonrandom "close" pairs.

The success of the Knox analysis depends on the choice of table bounds t_0 and d_0. When the chosen bounds are too large, "close" pairs are likely diluted by random pairs of observations. When the bounds are too small, nonrandom pairs are likely omitted from the count of "close" pairs. In the absence of a rigorous criteria for selection of t_0 and d_0, some investigators try a number of bounds and select an "optimum" result. This strategy destroys the interpretation of the estimated p-value as a measure of randomness, which is the purpose of the analysis.

Table 21.3 *Notation: Time Strata Created for a Test Equality of Variance Applied to Time-Space Data*

	$<t_1$	t_1–t_2	t_2–t_3	...	$>t_k$	Total
Observations	n_1	n_2	n_3	...	n_k	N
Variances	S_1^2	S_2^2	S_3^2	...	S_k^2	S_{pooled}^2

Test of Variance Applied to Spatial Data

A statistical tool useful for a variety of tasks applies to the assessment of time-space data. Extending the Knox time-distance table creates an opportunity to apply a test of the equality of variances called *Bartlett's test of variance* (Chapter 13). To start, the observed distances are stratified into k categories based on the observed time differences. Example strata and notation are displayed in Table 21.3.

The interval bounds (denoted t_i) used to create this table can be chosen, for example, as quantiles each containing equal or close to equal numbers of distance measures. In symbols, each stratum then contains $n_j \approx N/k$ distance observations. When no spatial pattern exists among the time-space data, the k within strata-estimated variances calculated from the n_j distances (denoted S_j^2) differ only because of random variation. A test statistic constructed to evaluate formally the observed differences in variability among these strata-specific estimates can effectively identify nonrandom variation caused by spatial clustering.

From the example spatial data (Figure 21.1), five time strata ($k = 5$) yield a table that displays the elements of a test of homogeneity of variance (Table 21.4 – again, $n = 126$ and $N = 126(125)/2 = 7875$ distances).

Bartlett's test of variance is then applied to the five estimated variances (S_j^2) created from the distance measures classified into $k = 5$ time percentiles (Chapter 13). Bartlett's test statistic is

$$X^2 = \frac{(N-k)\,log\left(S_{pooled}^2\right) - \sum (n_j - 1)\,log\left(S_j^2\right)}{1 + \frac{1}{3(k-1)}\left[\sum \frac{1}{n_j - 1} - \frac{1}{n-k}\right]} \qquad j = 1, 2, \ldots, k$$

where $S_{pooled}^2 = \Sigma(n_j - 1)S_j^2/(N - k)$ and $N = \Sigma n_j$ (Chapter 13). Each of the k within-time-strata–estimated variances is $variance(d_{ij}) = S_j^2 = \Sigma(d_{ij} - \bar{d}_j)^2/(n_j - 1)$ for $i = 1, 2, \ldots, n_j$ (Table 21.4). In addition, Bartlett's test requires the observed distances to have at least an approximate normal distributions. The test statistic X^2 then has an approximate chi-square distribution with $k - 1$ degrees of freedom when the time-space pattern of disease is

Table 21.4 *Summary: Estimated Within-Strata Variances of spatial Distances Classified into Five Time Intervals ($k = 5$ and $N = 7875$ Distances)*

	Quintiles of time (t_{ij})					
	<13.0	13.0–28.0	28.0–45.9	45.9–68.0	>68.0	Total
Observations (n_j)	1575	1575	1575	1575	1575	7875
Variances (S_j^2)	6.190	5.767	5.907	6.032	5.399	5.878

random, causing the estimated variances within each time interval to differ by chance alone. Bartlett's test is a comparison of the variance calculated as if no clustering exists (S^2_{pooled}) to a summary measure of variability that combines the k observed variances (S^2_j) that possibly reflects clustering within the strata. For the example observations (Figure 21.1), the test statistic $X^2 = 8.457$ yields an approximate p-value of $P(X^2 \geq 8.457 \mid \text{equal variances}) = 0.076$ (degrees of freedom $= k - 1 = 4$).

Bartlett's chi-square test statistic critically depends on independent and normally distributed distances (Table 21.4). A nonparametric randomization approach yields a parallel assumption-free analysis potentially producing a more accurate p-value (Chapter 11). For the spatial observations classified into five strata, a randomization analysis yields the p-value of $P(X^2 \geq 8.457 \mid \text{equal variances}) = 246/2000 = 0.123$ based on 2000 replicate tables of randomized distances (no spatial pattern).

Similar to the Knox approach to time-space observations, the test of variance requires a table created from arbitrary choices of bounds to form the time strata. Again, too many strata potentially distribute nonrandom distances among several strata, reducing the effectiveness of the analysis, or too few strata potentially dilute the nonrandom observations, also reducing the likelihood of identifying spatial patterns. In addition, the choice of time categories can substantially influence the statistical assessment.

Mantel's Time-Space Regression Method

Methods that require time-distance pairs to be classified into categories typically depend on the selection of more or less arbitrary interval boundaries. Also of importance, the N actual values of the time-distance pairs are not directly analyzed. Mantel suggests a more efficient approach based on "generalized regression" techniques that do not require classifying time-distance pairs into categories but apply directly to the observed values of the N time-distance pairs.

The essence of Mantel's method is that the association within time and distance pairs is summarized by a test statistic with the general form

$$\text{test statistic} = \Sigma\Sigma X_{ij} Y_{ij} \qquad i > j = 1, 2, \ldots, N = \text{number of time-distance pairs}$$

where X_{ij} and Y_{ij} are functions of time and distance measures calculated from the N time-distance pairs created from the n observed values. Thus, the analysis of these N directly measured values avoids the need to specify time or distance categories.

One version of Mantel's test statistic is created when time-distance measures X_{ij} and Y_{ij} are set to the value 1 when both $t_{ij} \leq t_0$ and $d_{ij} \leq d_0$ and set to 0 otherwise. This special case then becomes the Knox count of "close" pairs.

Another possibility is to define $X_{ij} = t_{ij} - \bar{t}$ and $Y_{ij} = d_{ij} - \bar{d}$. Mantel's test statistic then becomes an estimate of the association between time and distance. Technically the test statistic is a measure of covariance between time and distance (Chapter 27). A nonrandom association within even a moderately large number of time-distance pairs is unlikely to be identified by calculating a time-distance covariance. The large number of random time-distance pairs likely overwhelms the influence created by a small number of nonrandomly distributed cases (in symbols, $m \ll n \ll N$). Specifically, the example of $n = 126$ individuals

creates $N = 7875$ time-distance pairs and the influence of $m = 6$ nonrandom pairs is unlikely to be identified by a time-distance covariance. Also of importance, large noninformative time-distance observations likely influence the test statistic.

An effective version of the Mantel statistic, based on a change of scale, directly measures "closeness." Mantel's scale is created by setting X_{ij} to $1/t_{ij}$ and Y_{ij} to $1/d_{ij}$. The summary time-distance test statistic

$$\text{Mantel's test statistic} = \Sigma\Sigma X_{ij}Y_{ij} = \Sigma\frac{1}{t_{ij}}\frac{1}{d_{ij}} \qquad i > j = 1, 2, \ldots, N$$

becomes a more sensitive measure of time-distance proximity. This measure increases as the distance and time differences within pairs decrease. In addition, the unwanted influences of noninformative large distances are minimal; that is, the values $1/d_{ij}$ emphasize close distances and become small and inconsequential for noninformative large distances. The reciprocal of the time differences has the same property. The product of these two measures even more strongly emphasizes "closeness" of pairs, when both time and distance pairs are simultaneously small. The statistical distribution for the Mantel summary is not simple but randomization methods directly apply (Chapter 11).

A few individuals may have the same date of disease onset or diagnosis. The problem of dividing by zero is eliminated by selecting a small lower bound of time-distance differences for all observations or removing these usually few pairs from the analysis.

For the example data (Figure 21.1), Mantel's test statistic is

$$\text{test statistic} = \sum\sum\frac{1}{t_{ij}}\frac{1}{d_{ij}} = 2096.8 \qquad i > j = 1, 2, \ldots, N = 7875.$$

The corresponding mean test statistic when the cluster of six observation is replaced with six random values is 2032.7 based on 5000 randomized values (estimated standard error = 131.56). The normal distributed test statistic

$$z = \frac{2096.8 - 2032.7}{131.56} = 0.487$$

yields a more extreme result by chance alone with an approximate probability of 0.313 or a p-value = P(test statistic $\geq 0.487 \mid$ no spatial pattern) = 0.313. The nonparametric assessment of the same difference yields 1587 values less than 2032.7 or an empirical p-value of $1587/5000 = 0.317$.

Performance and Time-Space Methods

The biostatistical-epidemiologic literature contains a number of theoretical descriptions of the properties of time-space methods. These discussions address important issues to be considered in a rigorous time-space analysis but provide little guidance for a specific application. To assess the likely performance of a time-space analysis, typically a number of assumptions have to be made about the properties of the data collected and the kinds of nonrandom spatial patterns that may be detected. These assumptions are usually untestable speculations. Such things as the number of clusters, the shape of clusters, the number and influence of extreme observations, and the occurrence of confounding factors typically play

critical roles in the effectiveness of a specific analytic technique. If reliable descriptions of the influence of these properties were available, a statistical analysis would likely be unnecessary.

Of critical importance, identification of a suspected time-space cluster frequently motivates collection of spatial data. A statistical analysis of these data is not useful. The question as to whether a nonrandom cluster likely exists has been answered. Nevertheless, frequently post hoc statistical analyses are conducted as if this crippling bias did not exist.

22

ROC Curve and Analysis

An ROC Curve

A technique developed during World War II, called a *receiver operating characteristic (ROC) analysis*, is today a popular analytic approach in a variety of fields. The central purpose of this statistical tool is a detailed description of the accuracy of a "new" method relative to an established method used to classify a binary outcome. To simplify terminology, the binary variable is referred to as the presence or absence of a disease determined by a "new" relative to a "standard" method of detection. An ROC analysis is, from another point of view, a study of the *sensitivity* and *specificity* of a "new" approach to classifying subjects as disease present or absent (Chapter 6).

Sensitivity, as usual, is measured by the probability of identifying individuals with disease among those with the disease (Chapter 6). Another term for sensitivity is true-positive fraction (*tpf*). Similarly, specificity is measured by the probability of identifying individuals without the disease among those without the disease (Chapter 6). Another term for specificity is true-negative fraction (*tnf*). An ROC analysis primarily describes the unavoidable trade-off between true-positive fractions (benefits) and false-positive fractions (costs) of a method for declaring a disease as present or absent. The accompanying ROC curve is a visual display of this cost/benefit trade-off. For example, a ROC curve constructed from two samples of arterial tissue data displays the relationship between true-positive and false-positive fractions associated with the presence or absence of an arterial pathologic condition using a computer diagnosis compared to a color deconvolution method for all possible classifications strategies.

For the following, a positive test is denoted T^+ and a negative test as T^-. Similarly, the presence of a disease is denoted D^+ and absence as D^-. These two binary variables create four outcomes. Two outcomes are correct and two are incorrect. A display of the four outcomes is called a *performance table* (Table 22.1).

Classifying a disease present or absent based on a positive or negative test creates four conditional probabilities:

$$\text{true-positive fraction (sensitivity)} = P(T^+|D^+) = tpf = \frac{tp}{tp+fn},$$

$$\text{false-negative fraction} = P(T^-|D^+) = fnf = \frac{fn}{tp+fn},$$

$$\text{false-positive fraction} = P(T^+|D^-) = fpf = \frac{fp}{fp+tn},$$

Table 22.1 *Notation: Performance Table for Four Outcomes of a Binary Test Used to Declare a Disease as Present or Absent*

	Disease present D^+	Disease absent D^-
Declared positive T^+	True and positive $= tp$	False and positive $= fp$
Declared negative T^-	False and negative $= fn$	True and negative $= tn$

$$\text{true-negative fraction (specificity)} = P(T^-|D^-) = tnf = \frac{tn}{fp + tn}, \text{ and}$$

$$\text{to summarize, the accuracy of classification} = \frac{tp + tn}{tp + fp + fn + tn}.$$

From another perspective, these performance measures are four probabilities of a specific test result conditional on the presence or absence of a disease, called *fractions* in the context of a ROC analysis.

A small numeric example clarifies the definitions of these probabilities and their notation (Table 22.2).

$$\text{true-positive fraction (sensitivity)} = tpf = \frac{tp}{tp + fn} = \frac{80}{100} = 0.80,$$

$$\text{false-negative fraction} = fnf = \frac{fn}{tp + fn} = \frac{20}{100} = 0.20,$$

$$\text{false-positive fraction} = fpf = \frac{fp}{fp + tn} = \frac{25}{100} = 0.25,$$

$$\text{true-negative fraction (specificity)} = tnf = \frac{tn}{fp + tn} = \frac{75}{100} = 0.75 \text{ and}$$

$$\text{to summarize, the accuracy of classification} = \frac{tp + tn}{tp + fp + fn + tn} = \frac{155}{200} = 0.775.$$

A useful statistical framework to describe the properties of ROC classification is the two-sample normal distribution model (Figure 22.1); that is, the data are analyzed as if they consist of two independent samples from normal distributions with possibly different mean values, represented by $\mu_t (t =$ "new" test method) and by $\mu_s (s =$ standard method). Furthermore, both distributions are required to have the same variance (denoted σ^2). Such a comparison

Table 22.2 *Example: Artificial Performance Table Describing Accuracy of a Binary Disease Classification (Present/Absent)*

	Disease	
Test	D^+	D^-
T^+	$tp = 80$	$fp = 25$
T^-	$fn = 20$	$tn = 75$

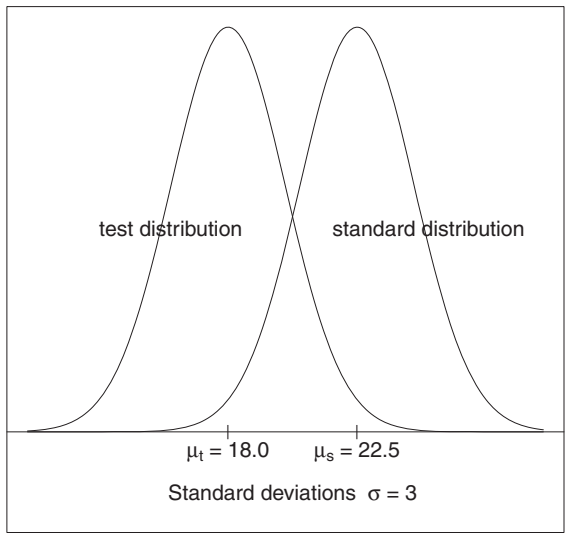


Figure 22.1 Normal Distributions Displaying Two-Sample "Shift Model" with Different Mean Values and Same Variance

is frequently called a *shift model* because the compared normal distributions only differ in location determined by their mean values. This model is the theoretical basis of Student's two-sample *t*-test.

A ROC analysis explores the consequences of declaring a measurement from a specific subject as positive when it is greater than a predetermined value denoted c and negative when it is less than the value c (c = cut-point). Thus, every value c produces a different 2×2 performance table. For the two-sample shift model, the results for a cut-point $c = 20$ are illustrated with two normal distributions: mean value $\mu_t = 18.0$ for a "new" classification method and mean value $\mu_s = 22.5$ for a standard classification method. The variance of both distributions is $\sigma^2 = 9$. Figures 22.2 and 22.3 display two versions of the same ROC comparison for the specific choice of a cut-point $c = 20$. For the example, the value $c = 20$ produces false-positive fraction $fpf = P(T \geq 20 \mid \mu_t = 18) = P(Z \geq (20 - 18)/3) = P(Z \geq 0.667) = 0.25$ and a true-positive fraction $tpf = P(T \geq 20 \mid \mu_s = 22.5) = P(Z \geq (20 - 22.5)/3) = P(Z \geq -0.833) = 0.80$ (Chapter 1).

A plot of the pairs of true-positive fractions (*tpf*) and the false-positive fractions (*fpf*) for all possible choices of the value c produces a ROC curve. This curve is a display of the classification accuracy of the "new" method compared to the standard method for detecting a disease.

Figure 22.4 illustrates three ROC curves. When the test is perfectly accurate, then $fpf = 0$ and $tpf = 1$ and the ROC "curve" becomes the single point (0, 1). From a slightly different point of view, perfect classification occurs, when $tpf = 1$ (sensitivity) and $tnf = 1$ (specificity). On the other hand, when for all comparisons the true-positive and the false-positive probabilities are equal ($fpf = tpf$), then the ROC "curve" is a straight line from the point (0, 0) to (1, 1); that is, these two performance fractions *fpf* and *tpf* are equal for all values of c because no difference exists between the "new" and standard classification methods ($\mu_s = \mu_t$). Between these two extremes, an ROC curve displays the entire range of values of two

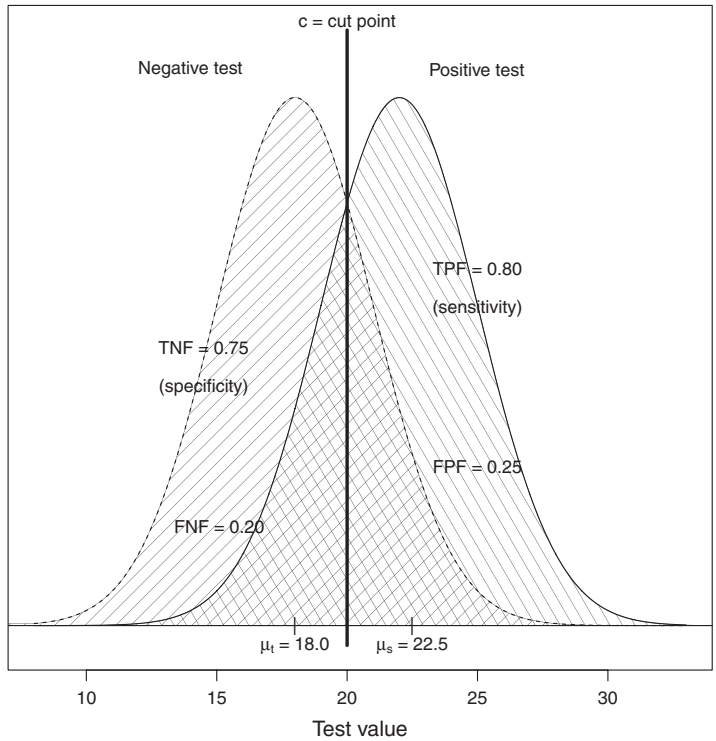


Figure 22.2 Normal Distribution–Based ROC Performance Fractions

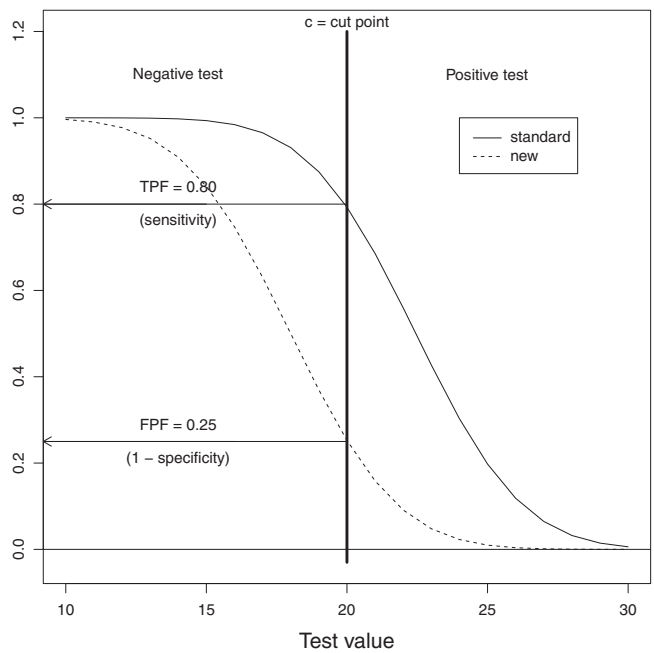


Figure 22.3 Normal Cumulative Distribution–Based ROC Performance Fractions

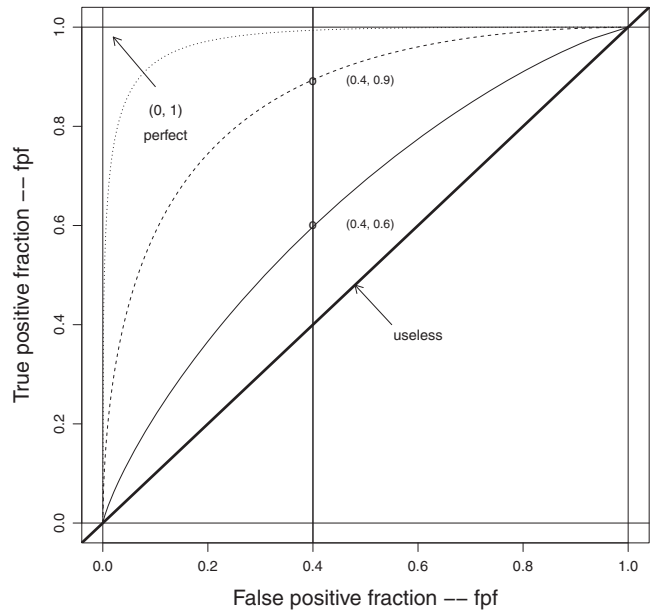


Figure 22.4 Three Example ROC Curves

performance fractions, namely, *tpf* (vertical axis) and *fpf* (horizontal axis), forming a complete picture (Figure 22.4) of the trade-off in accuracy between a "new" method compared to a standard method as a curve from point (0, 0) to (1, 1). As the accuracy of a test increases, the ROC curve moves away from the "useless" diagonal line toward "perfect" discrimination, the upper left-hand corner (Figure 22.4). Thus, the area under the ROC curve reflects the performance of a "new" method relative to a standard method as a single value (details to be discussed). For two ROC curves, the upper curve indicates better performance (right-tail probabilities). For the example, at the value of *fpf* equal to 0.4, the *tpf*-value is 0.6 on the lower curve and increases to 0.9 on the higher curve (the vertical line – Figure 22.4). Similarly, the higher ROC curve has the lower *fpf*-value at the same *tpf*-value (not illustrated).

The ROC curve applied to the two-sample shift model consists of the two cumulative probability functions from normal distributions given by the expressions

$$fpf = 1 - \Phi\left[\frac{c - \mu_t}{\sigma}\right] \quad \text{and} \quad tpf = 1 - \Phi\left[\frac{c - \mu_s}{\sigma}\right].$$

Again μ_s represents the mean value of the normal distribution of the standard measures and μ_t represents the mean value of the normal distribution of the "new" measures ($\mu_s \geq \mu_t$ – Figures 22.2 and 22.3). Both distributions are required to have the same variance (denoted σ^2). The symbol $\Phi(z) = P(Z \leq z)$ represents the value of the cumulative probability from a standard normal distribution at the value z (percentile/quantile) (Chapter 1). Mechanically, a ROC curve is no more than a typical x/y-line plot of $x = fpf$ against $y = tpf$ based on the right-tail cumulative probabilities calculated from two normal distributions with the same variance for all values of c. The ROC curve is a picture of the properties of the top row of the 2×2 performance table for all values of c (Table 22.1).

Table 22.3 *Example: Relationship for 10 Selected Cut-points c in Terms of Sensitivity (tpf) and Specificity (tnf) for an ROC Curve ($\mu_t = 18.0$, $\mu_s = 22.5$ and $\sigma^2 = 9$ –Figure 22.5)*

Cut-point	*tpf*	*tnf*	*fpf*
16	0.98	0.25	0.75
17	0.97	0.37	0.63
18	0.93	0.50	0.50
19	0.88	0.63	0.37
20[a]	**0.80**	**0.75**	**0.25**
21	0.69	0.84	0.16
22	0.57	0.91	0.09
23	0.43	0.95	0.05
24	0.31	0.98	0.02

[a]Example calculation shown in Figures 22.2 and 22.3.

Figure 22.3 displays two performance fractions generated from a single cut-point (again $c = 20$). In general, a ROC curve displays two performance fractions ($x = fpf$ and $y = tpf$) for all possible values of c (Figure 22.5). To further illustrate, two performance fractions (fpf, tpf) for 10 selected cut-points c are specifically displayed (Figure 22.5 and Table 22.3).

For the normal distribution model, the point on a ROC curve closest to perfect discrimination ($x = fpf = 0$ and $y = tpf = 1$) occurs at the cut-point $c = (\mu_t + \mu_s)/2$.

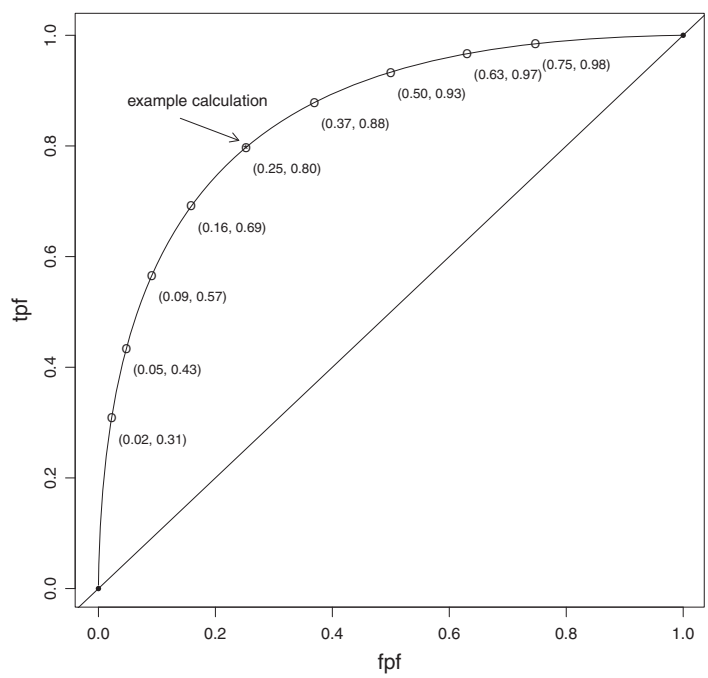


Figure 22.5 ROC for All Possible Cut-Points c for Two Normal Distributions with Mean Values $\mu_s = 22.5$ and $\mu_t = 18$ ($\sigma^2 = 9$)

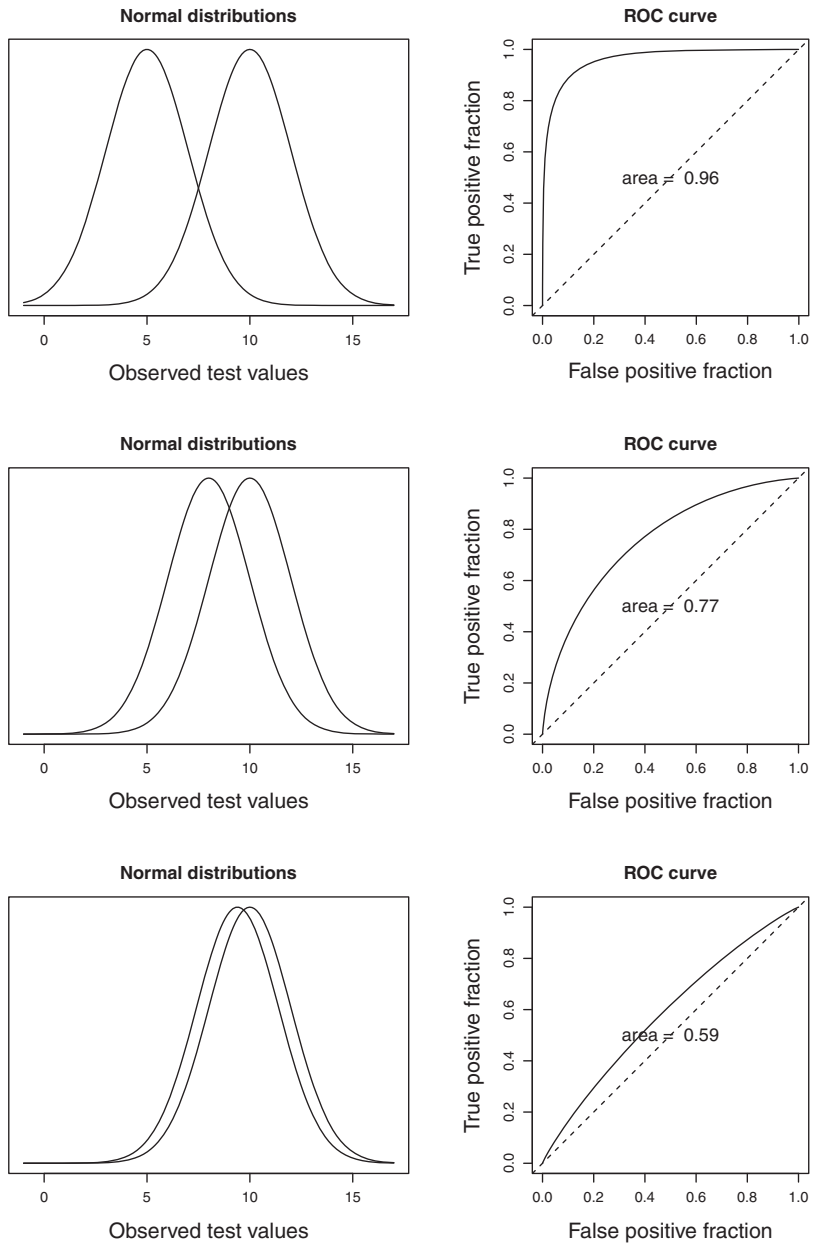


Figure 22.6 Three Graphic Descriptions of Normal Distributed Data Compared with Shift Model and the Same Comparison with a ROC Curve

A key relationship between a ROC curve and the difference between mean values generated by the comparison of "new" and standard distributions is seen in Figure 22.6. The left-hand column displays two normal distributions with same variance as the difference between mean values decreases toward equality ($\mu_t - \mu_s \rightarrow 0$). The parallel ROC curves in the right-hand column display the same relationships as the ROC curve deceases toward the

Table 22.4 *Data: Standard* (x_s) *and Test* (x_t) *Measurements from Two Methods for Detection of Carotid Artery Disease* ($n_s = n_t = 30$)

Carotid Artery Disease Data												
Test (x_t)	20.0	30.0	40.0	45.0	50.0	50.0	55.0	55.0	57.5	66.7	67.5	67.5
Standard (x_s)	39.5	55.9	59.5	60.0	63.0	64.9	66.8	67.2	68.0	70.6	72.4	73.6
Test (x_t)	73.6	75.0	75.0	75.0	77.5	77.5	80.0	80.0	80.0	80.0	82.5	82.5
Standard (x_s)	75.3	76.3	77.4	77.8	78.2	78.4	78.5	78.6	80.8	85.6	86.3	87.4
Test (x_t)	85.0	87.5	87.5	92.5	85.0	85.0						
Standard (x_s)	88.5	88.6	88.7	91.5	87.4	88.3						

straight line (*tpf* = *fpf*). The area enclosed by the ROC curve ($0.5 \leq$ area ≤ 1.0) provides a single quantitative measure of the classification performance reflected by the two sampled distributions. An explicit estimate is given by the area under the ROC curve (denoted *auc*) where

$$\text{area} = auc = 1 - \Phi\left[\frac{\mu_t - \mu_s}{\sigma\sqrt{2}}\right] \text{ (Figure 22.6).}$$

An ROC Analysis Example

Two methods for diagnosing artery disease status provide an illustration of a ROC curve as a technique to statistically describe classification accuracy. Using measured intensity of histochemical staining to characterize disease status of carotid artery tissue produces two sets of data from $n = 60$ sampled individuals. A method called *color picking* is a computer-based scanning process that reads color intensities pixel by pixel from slide images made from tissue microarrays and is selected as the "standard." This computer process is compared to a visual method, called *color deconvolution*. Example data for the standard and "new" method ($n = n_s + n_t = 30 + 30 = 60$ observations) are a small subset of a large study (Table 22.4).

Smoothed estimates based on these measurements of artery health are displayed as two continuous distributions (Figure 22.7) (Chapter 3). A sample of 30 individuals from each of these two distributions that appear close to symmetric makes it likely that the properties of the respective estimated mean values are accurately approximated by normal distributions with the same variance (Chapter 1).

The fundamental summary of a ROC analysis is the area under the ROC curve. As noted, this summary statistic directly reflects the difference in accuracy between two compared classification methods. This area is a measure of the average sensitivity (*tpf*) and, in addition, provides a summary value with a probabilistic interpretation. An *auc*-value between 0.5 and 1.0 is an estimate of the probability that a random observation from the "new" method is less than a random observation from the standard. In symbols, the probability is represented as

$$\text{area} = auc = P(\text{"new"} \leq \text{standard} = P([\text{"new"} - \text{standard}] \leq 0).$$

Therefore, for the normal distribution two-sample comparison, the area under the ROC curve is estimated by

$$\hat{auc} = 1 - \Phi(\hat{R}) \quad \text{where} \quad \hat{R} = \frac{\bar{x}_t - \bar{x}_s}{S\sqrt{2}}.$$

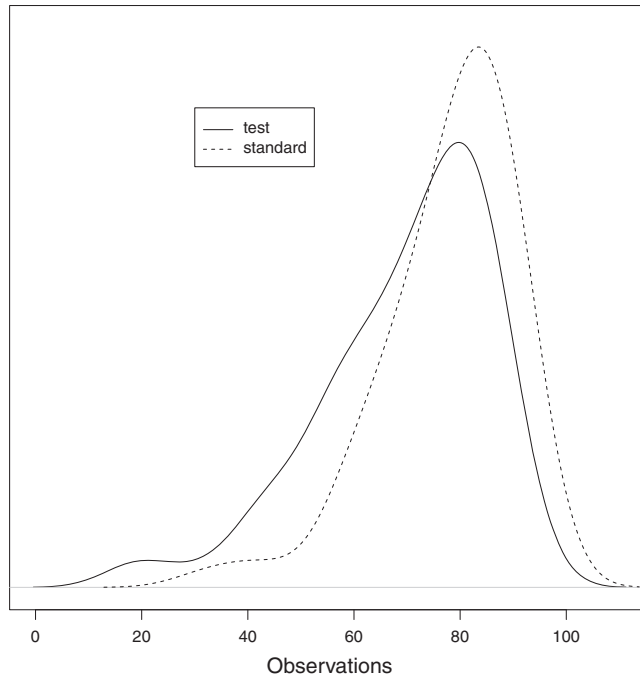


Figure 22.7 Test Data: Smoothed Distributions of Histochemical Measurements Characterizing Carotid Artery Disease, Computer versus Visual Assessment ($n_t = n_s = 30$)

The symbol S^2 represents the usual sample size weighted average estimate of a common variance created by pooling the estimated variances from each sample or

$$S^2 = \frac{(n_t - 1)variance(x_t) + (n_s - 1)variance(x_s)}{n_t + n_s - 2}$$

and again n_t and n_s represent the number of "new" and standard sampled observations. Thus, the $a\hat{u}c$ value is an estimate of the probability that a random test value is less than a random standard value and is geometrically the area under the estimated ROC-curve (Figure 22.8). Note that for the artery disease data, the point $(fpf_0, tpf_0) = (0.419, 0.581)$ is closest to the point (0, 1) and is indicated on the plot.

The estimated value of R from the example data is

$$\hat{R} = \frac{\bar{x}_t - \bar{x}_s}{S\sqrt{2}} = \frac{68.843 - 75.167}{15.486\sqrt{2}} = -0.289$$

using the estimated pooled variance of $S^2 = 239.819$ (Table 22.4). The estimated area under the ROC curve becomes $a\hat{u}c = 1 - \Phi(R) = 1 - \Phi(-0.289) = P(\text{deconvolution} \leq \text{color picking}) = P(x_t \leq x_s) = 0.614$. Thus, the parametric estimated probability that a random observation using the "new" method is less than a random value using the standard method is $\hat{P}(\text{"new"} \leq \text{standard}) = 0.614$.

Two approaches describing the influence of sampling variation on the estimated $a\hat{u}c$-value are, as usual, a statistical test and a confidence interval. Using an evaluation based on the estimate $\hat{R}$ is identical to the two-sample comparison of mean values $\bar{x}_t$ and $\bar{x}_s$. The conjecture

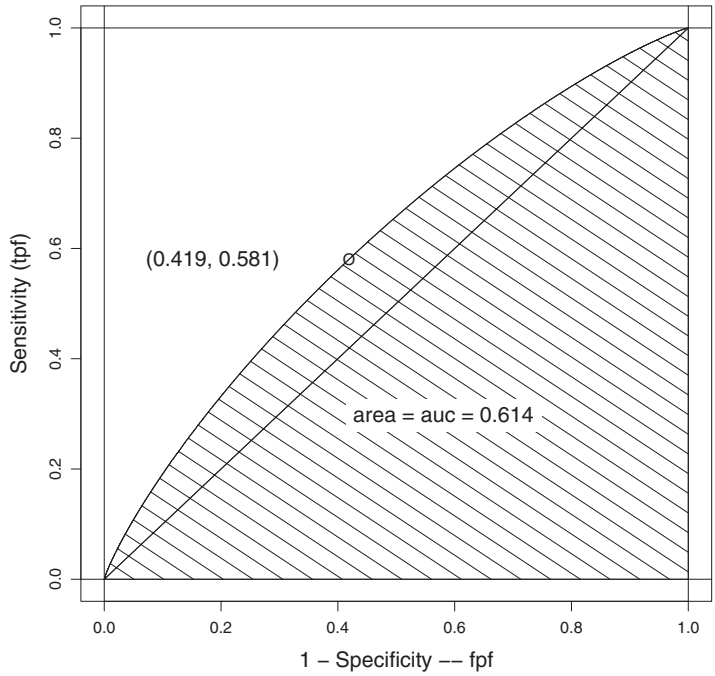


Figure 22.8 ROC Comparison of Color Picking versus Visual Determination for Identifying Carotid Artery Disease – Normal Distribution Two-Sample Model

that $R = 0$ is evaluated in light of the estimated value $\hat{R}$. An estimate of the variance of $\hat{R}$ is given by the expression

$$\hat{V} = variance(\hat{R}) = \frac{1}{2}\left[\frac{1}{n_s} + \frac{1}{n_t}\right]$$

The associated z-statistic for the two sets of $n = 30$ example observations (Table 22.4) is $z = \hat{R}/\sqrt{\hat{V}} = -0.289/\sqrt{0.033} = -1.581$ using the estimated variance $\hat{V} = 0.033$. For the two-sample comparison, the test statistic is

$$z = \frac{\bar{x}_t - \bar{x}_s}{S_{\bar{x}_t - \bar{x}_s}} = \frac{68.843 - 75.167}{4.0} = -1.581$$

with a p-value $= P(Z \leq -1.581 \mid \text{no difference}) = P(Z \leq -1.581 \mid R = 0) = P(Z \leq -1.581 \mid auc = 0.5) = 0.057$.

Confidence intervals constructed from either the estimate $\hat{R}$ or $a\hat{u}c$ usefully reflect the influence of sampling variation. For the estimate $\hat{R} = -0.289$ and its associated variance of 0.033, an approximate 95% confidence interval is $-0.289 \pm 1.960(0.181) \rightarrow (-0.647, 0.069)$. The confidence interval for R can be directly transformed into a confidence interval for the underlying area under the ROC-curve. The corresponding confidence interval has lower bound $= \hat{A} = 1 - \Phi(\hat{R}_{upper})$ and upper interval bound $= \hat{B} = 1 - \Phi(\hat{R}_{lower})$ where the bounds $\hat{R}_{lower}$ and $\hat{R}_{upper}$ are constructed from the estimate $\hat{R}$.

Specifically, for the carotid artery data (Table 22.4), the lower and upper bounds

$$\hat{A} = 1 - \Phi(0.093) = 0.472 \quad \text{and} \quad \hat{B} = 1 - \Phi(-0.671) = 0.741$$

form the 95% confidence interval based on the estimate $a\hat{u}c = 1 - \Phi(-0.289) = 0.614$.

One last comment: The two-sample parametric comparison of mean values has a parallel nonparametric and an often effective approach based on the *Wilcoxon (Mann-Whitney)* two-sample rank test (Chapter 8). It is not a surprise, therefore, that a nonparametric ROC analysis exists with many properties in common with a nonparametric two-sample analysis.

Nonparametric Estimation of an ROC Curve

A nonparametric estimated *auc*-value results from recognizing that the nonparametric estimated ROC analysis is an alternate version of the Mann-Whitney test statistic applied to describe the difference between two independent samples of observations. More specifically, the nonparametric estimated $a\hat{u}c$ statistic is identical to the Mann-Whitney estimate of the probability that a randomly selected value from one sample is smaller than a randomly selected value from another sample (Chapter 8). In symbols, the Mann-Whitney probability ($\hat{P}$) in the context of a ROC analysis is

$$\hat{P} = P(\text{test} \leq \text{standard}) = a\hat{u}c$$

where again *test* represents a random test measurement and *standard* represents a random standard measurement. Therefore, the area under the nonparametric ROC curve is estimated by

$$\hat{P} = a\hat{u}c = \frac{\sum u_i}{n_s n_t} = \frac{U}{n_s n_t} \quad i = 1, 2, \ldots, n_t$$

where u_i represents the Mann-Whitney count of the number of values in the test sample that are less than the ith value in the standard sample (Chapter 8).

Geometrically, the counts u_i are proportional to the heights of the nonparametric ROC curve at each *fpf*-measure (Figures 22.9 to 22.15 below). The total area calculated from these heights is the area enclosed by the ROC plot ($a\hat{u}c$) and, to repeat, is exactly the value of the Mann-Whitney two-sample test statistic ($\hat{P}$).

The variance of the nonparametrically estimated area enclosed by the ROC curve, therefore, is the estimated variance of the Mann-Whitney test statistic given by the lengthy expression

$$variance(a\hat{u}c) = \frac{1}{n_s n_t}[a\hat{u}c\,(1 - a\hat{u}c) + (n_s - 1)(p_1 - a\hat{u}c^2) + (n_t - 1)(p_2 - a\hat{u}c^2)],$$

where

$$P_1 = \frac{a\hat{u}c}{2 - a\hat{u}c} \quad \text{and} \quad P_2 = \frac{2a\hat{u}c^2}{1 + a\hat{u}c}.$$

For the value $auc = 0.5$ making $p_1 = p_2 = 1/3$, a considerably simpler estimate of the variance of the distribution of the estimate $a\hat{u}c$ becomes

$$variance(a\hat{u}c) = \frac{n_s + n_t + 1}{12 n_s n_t} \approx \frac{1}{12}\left[\frac{1}{n_s} + \frac{1}{n_t}\right].$$

Table 22.5 *Example: Small Artificial Sample of Test* ($n_s = 12$) *and Standard* ($n_t = 8$) *Observations to Illustrate Nonparametric and Parametric Estimates of the Area under the ROC Curve*[a]

	1	2	3	4	5	6	7	8	9	10
x	11.49	10.61	11.51	10.60	12.64	11.62	11.20	12.11	10.79	10.62
y	0	0	0	0	0	0	0	0	0	0
Ranks	15	6	16	5	20	17	14	18	10	7
	11	12	13	14	15	16	17	18	19	20
X	10.70	11.05	8.12	9.97	12.31	10.97	10.96	9.46	10.67	10.50
Y	0	0	1	1	1	1	1	1	1	1
Ranks	9	13	1	3	19	12	11	2	8	4

[a]Standard observations are coded = 0 and test observations are coded = 1.

This special case ($auc = 0.5$) of the variance is used to assess the likelihood that the estimated value $\hat{auc}$ is a random deviation for 0.5 (exactly no systematic difference between test and standard methods).

The Mann-Whitney test statistic (denoted U) directly relates to the Wilcoxon test statistic (denoted W) (Chapter 8). The Wilcoxon test statistic is

$$W = U + \frac{n_s(n_s + 1)}{2},$$

where W represents the sum of the n_s ranks of the standard observations (Chapter 8). This relationship provides a computational form for the nonparametric estimation of the area under the ROC,

$$\hat{auc} = \hat{P} = \frac{\bar{W} - \frac{1}{2}(n_s + 1)}{n_t},$$

where $\bar{W}$ is the mean value of the ranks of the standard observations.

An Example – Nonparametric ROC Analysis

For an artificial sample of $n_s = 12$ standard and $n_t = 8$ test observations, the $n = n_s + n_t = 20$ values illustrate calculation of the area under the ROC curve (Table 22.5).

The Mann-Whitney U-statistic is $U = 72$, the estimated auc value is then

$$\hat{P} = \hat{auc} = \frac{U}{n_s n_t} = \frac{72}{12(8)} = 0.75,$$

or, equivalently,

$$\hat{auc} = \frac{\bar{W} - \frac{1}{2}(n_s + 1)}{n_t} = \frac{150/12 - 13/2}{8} = \frac{12.5 - 6.5}{8} = 0.75,$$

where the sum of the ranks of the standard observations is the Wilcoxon test statistic $W = 150$ or $W = U + n_s(n_s + 1)/2 = 72 + 78 = 150$ making $\bar{W} = 150/12 = 12.5$ (Table 22.5 – last row) (Chapter 8).

As before, the value $auc = 0.5$ indicates that exactly no difference exists between the standard and test sampled distributions. A statistical comparison of the estimated value $a\hat{u}c = 0.75$ to the postulated value $auc = 0.5$ follows the usual pattern; that is, to evaluate the influence of sampling variation, the approximately normal distribution test statistic

$$z = \frac{a\hat{u}c - 0.5}{\sqrt{variance}(a\hat{u}c)} = \frac{0.75 - 0.5}{\sqrt{\frac{1}{12}\left[\frac{1}{12} + \frac{1}{8}\right]}} = 1.897$$

produces the associated p-value of $P(|Z| \geq 1.897 \mid auc = 0.5) = 0.029$. To repeat, the key summary is the estimated Mann-Whitney statistic $= \hat{P} = a\hat{u}c =$ area under the ROC curve.

How to Draw a Nonparametric ROC Curve

Learning to draw a ROC-curve is hardly necessary with the variety of computer systems available to create an estimate with little effort. However, it is empowering to know the mechanics of the process and, of more importance, detailed knowledge of the construction leads to a deeper understanding of the properties of an ROC-curve analysis.

As noted, the ROC curve is a comparison of two cumulative distribution functions (again denoted *cdf*), one based on data from a "new" method and the other based on the data from a standard method of classification. The effectiveness of the "new" relative to a standard method to correctly classifying a binary variable, such as disease absent or disease present, begins with two cumulative distribution functions (Chapter 12). The two nonparametric estimated *cdf*-functions are

$$\text{"new" method: } cdf_t = F(t) = P(T \leq t) = \frac{\text{number of ordered values} \leq t}{n_t},$$

where n_t represents the number of "new" test observations, and

$$\text{standard method: } cdf_s = F(s) = P(S \leq s) = \frac{\text{number of orderd values} \leq s}{n_s},$$

where n_s represents the number of standard observations (Chapter 12).

Construction of the ROC Curve

The small set of artificial data ($n = 16$) serves to illustrate the construction of nonparametric and parametric ROC curves from two estimated cumulative distribution functions (Table 22.6).

Figures 22.9 display two estimated cumulative distribution functions $F(t)$ and $F(s)$ that ultimately produce the ROC-curve, one from "new" test observations (n_t) and the other from standard method observations (n_s). These representations of the cumulative probabilities, for obvious reasons, are sometimes called *step-functions*.

Figures 22.10 display $1 - F(s)$ and $1 - F(t)$ that directly relate to classification probabilities; that is, the two cumulative distributions functions $F(s)$ and $F(t)$ are transform to reflect classification performance. The x/y-coordinate values for the transformed *cdf*-functions are located on the plots.

Table 22.6 *Example: Artificial Data and Their Estimated Cumulative Distribution Functions to Illustrate Construction of Nonparametric and Parametric ROC curves* ($n_t = 6$ *and* $n_s = 10$)

"New" Method	1	2	3	4	5	6				
Observations	2.800	5.200	5.600	6.500	8.800	9.900				
$F(t)$	0.167	0.333	0.500	0.667	0.833	1.000				
$1 - F(t)$	0.833	0.677	0.500	0.333	0.167	0.000				
Standard Method	1	2	3	4	5	6	7	8	9	10
Observations	3.4	5.0	6.0	6.1	7.0	7.3	7.7	8.4	9.0	9.2
$F(s)$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$1 - F(s)$	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0

Figures 22.11 display the relationship between the cumulative distribution functions and the classification probabilities $1 - F(t) = fpf$ and $1 - F(s) = tpf$ for the two values $c = 5.6$ and $c = 7.0$. The symbol c again represents a cut-point where the binary test outcome is declared negative (less than c) or positive (greater than c).

Figures 22.12 display again the two sets of values $1 - F(t) = fpf\,(X)$ and $1 - F(s) = tpf\,(Y)$ plotted on the same set of axes (Table 22.7 – last two columns). The total number of observations $n = n_s + n_t = 10 + 6 = 16$ produces 16 comparisons between these two cumulative functions (16 lines connecting two step functions – lower plot).

Figures 22.13 are again plots of pairs of values $(X, Y) = (fpf, tpf)$ consisting of 16 points that constitute the estimated ROC-curve. From a more mechanical point of view, these plots are the 16 pairs of values given in columns labeled X and Y (Table 22.7).

Figures 22.14 display the estimated ROC-curve plotted on a grid of 60 rectangles. All possible nonparametric ROC-curves are contained in a rectangle that is $n_t = 6$ equal intervals wide and $n_s = 10$ equal intervals high. The nonparametric ROC-curve divides these 60 rectangles into those above and below the estimated step-function. The sum $2 + 6 + 8 + 8 + 10 = 34$ is the number of rectangles below the example ROC-curve (area $= \hat{auc}$) and summarizes the classification accuracy of the "new" method relative to the standard method.

Figures 22.15 are the completed ROC-plot with the area enclosed in terms of the proportion of the total area or $\hat{auc} = 34/60 = 0.567$, reflecting the extent of difference in accuracy between the "new" to standard classification methods.

Figure 22.16 contrasts the parametric estimated ROC-curve (solid line) to the nonparametric ROC-curve (dashed line) based on the same data (Table 22.6). The parametric estimated X and Y values are

$$X = fpf = 1 - \Phi(z_t) \quad \text{and} \quad Y = tpf = 1 - \Phi(z_s)$$

where the value $z_x = (c - mean_x)/S$. The value c represents a cut-point, and S^2 represents the pooled estimated variances of the sampled normal distributions. The symbol $1 - \Phi(z) = P(Z \geq z)$ is again notation for the right-tail cumulative probability that a random value from a standard normal distribution is greater than or equal to the value z. From the example data, the sample mean values are $mean_s = \bar{x}_s = 6.91$ and $mean_t = \bar{x}_t = 6.47$. The estimated pooled standard error of the sampled distributions is $S = 2.123$. A plot over the range of the values of the cut-point c produces the parametric ROC-curve comparison of "new" and standard methods of classification (Table 22.7).

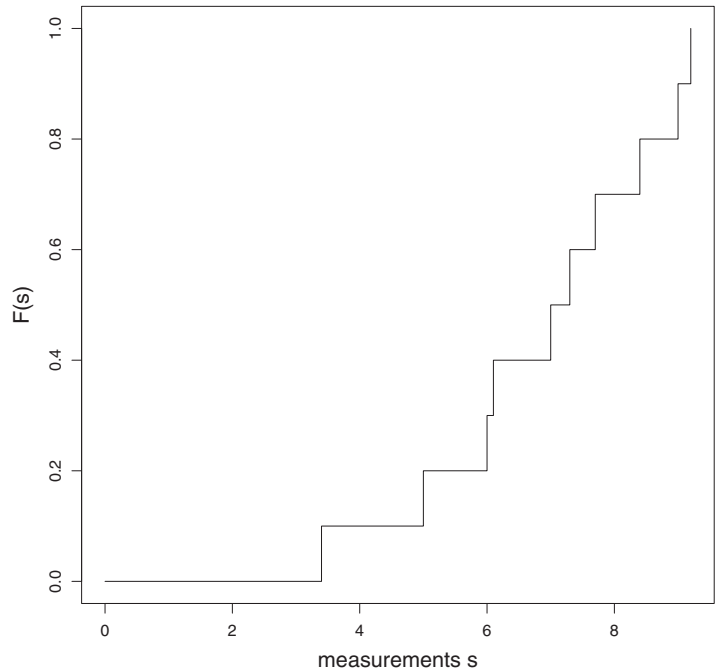


Figure 22.9(a) CDF – standard.

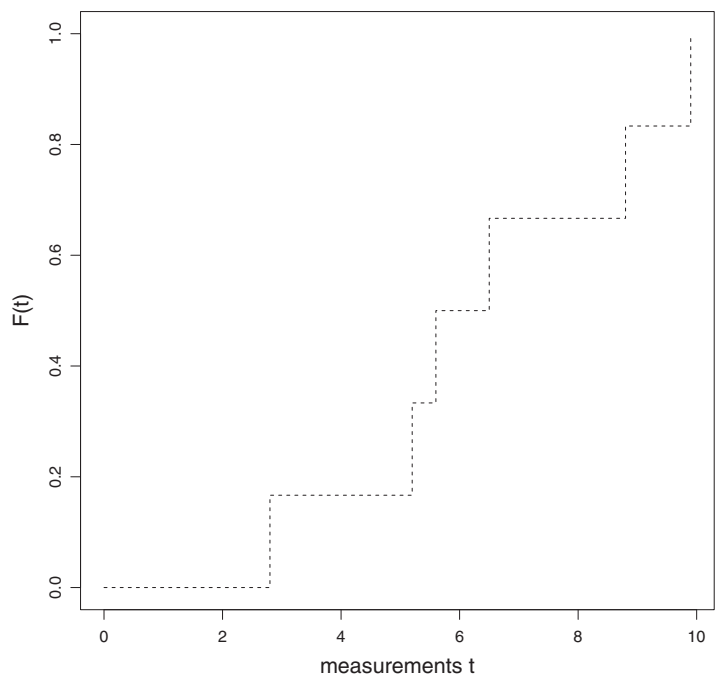


Figure 22.9(b) CDF – new.

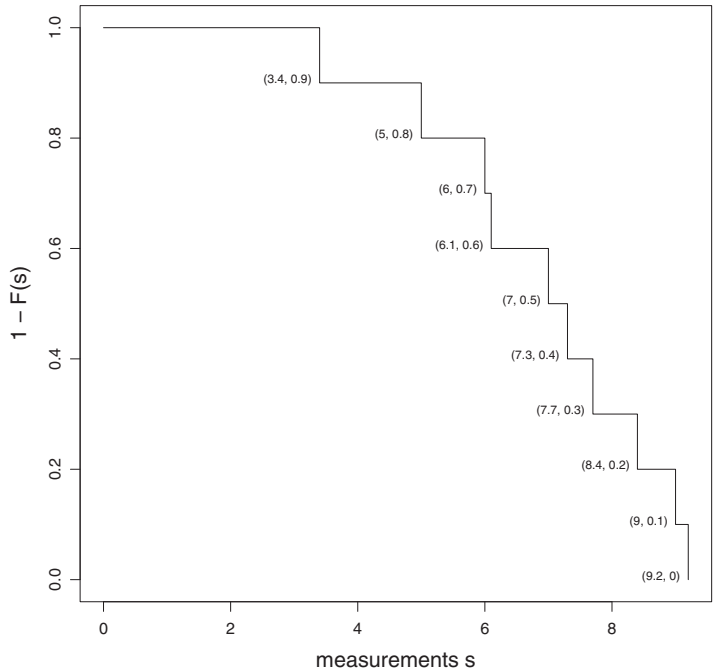


Figure 22.10(a) Complementary $= 1 -$ CDF – standard.

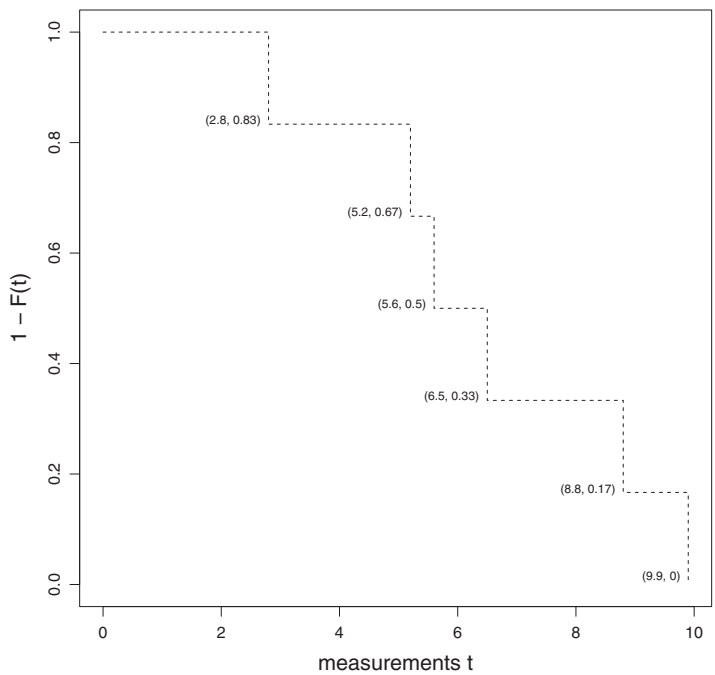


Figure 22.10(b) Complementary $= 1 -$ CDF – new.

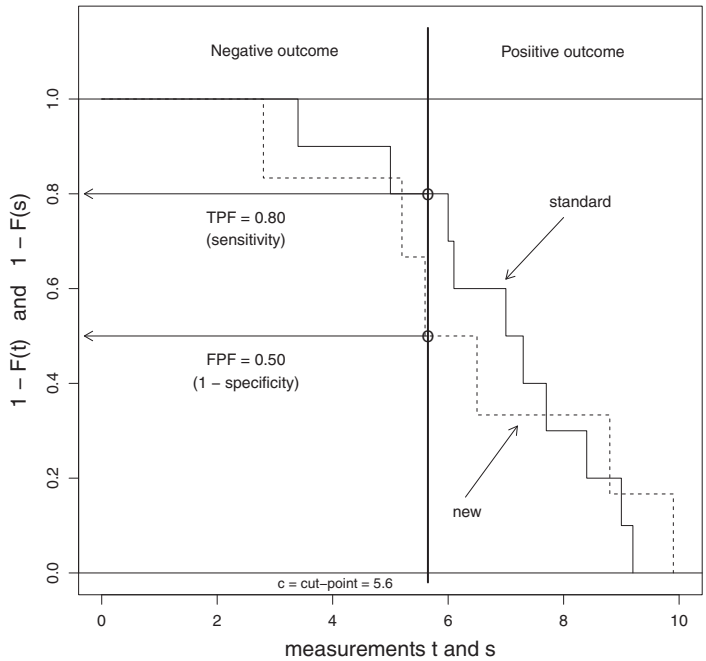


Figure 22.11(a) Cumulative distributions – new/standard.

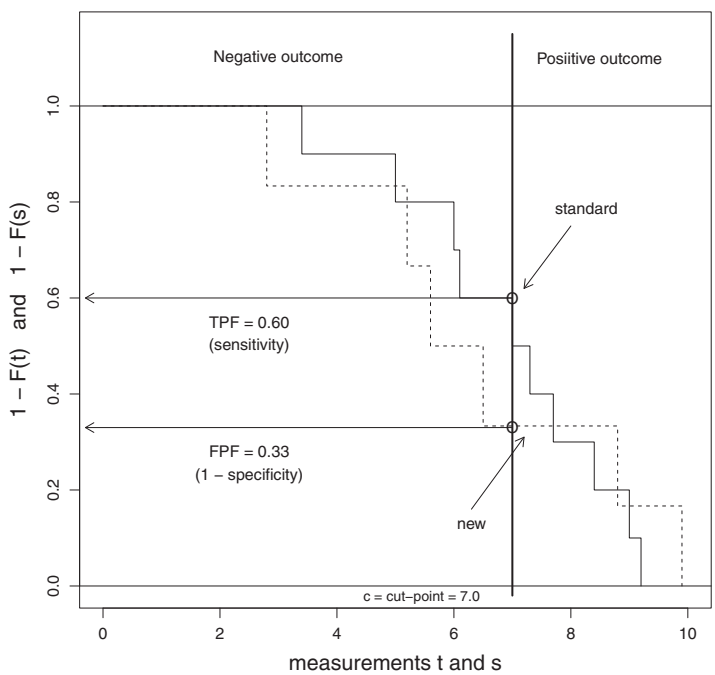


Figure 22.11(b) Cumulative distributions – new/standard.

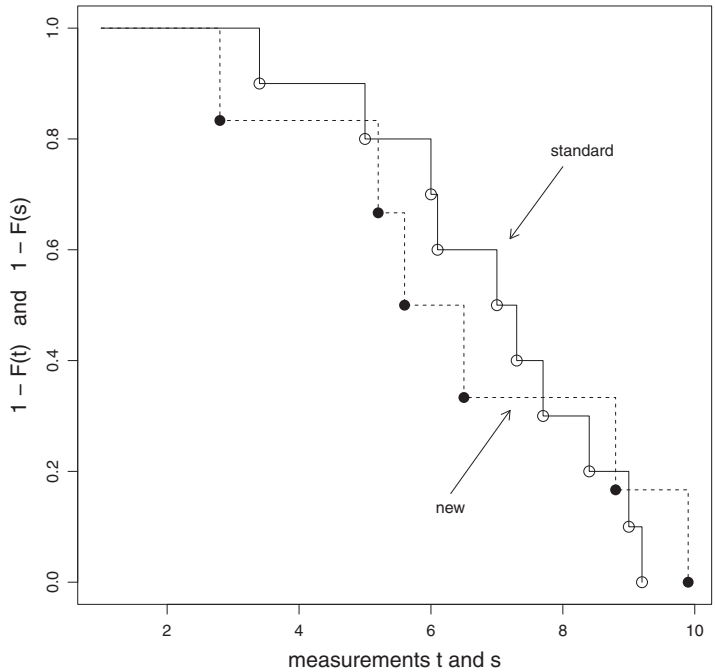


Figure 22.12(a) Complementary 1 − CDF − new/standard.

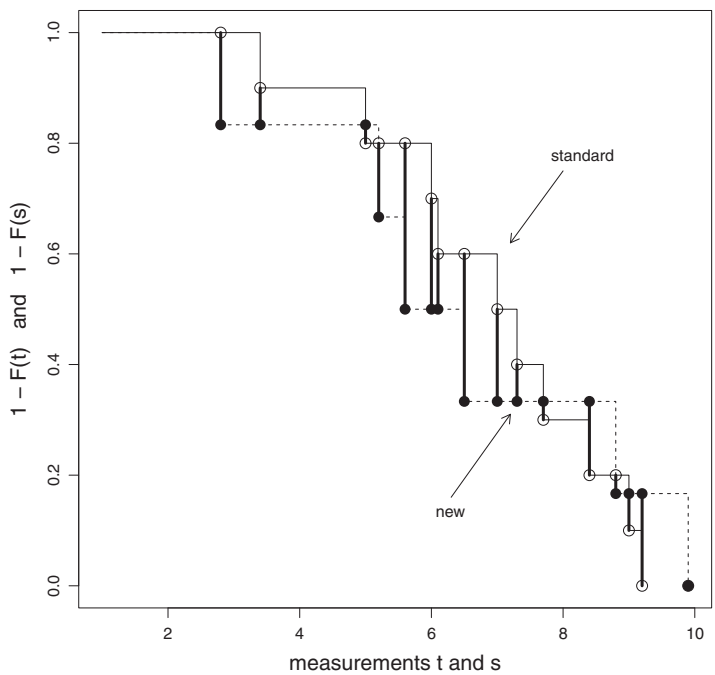


Figure 22.12(b) Complementary 1 − CDF − new/standard.

Table 22.7 *Data: Summary Table for Construction of an ROC Curve – Two Cumulative Probability Functions and Their Comparison, Labeled X and Y*

	New t	*std* s	*cdf* (*new*) $F(t)$	*cdf* (*std*) $F(s)$	1-*cdf* (*new*) $1-F(t)$	1-*cdf* (*std*) $1-F(s)$	id^a	*fpf* X	*tpf* Y
1	2.8	–	0.17	–	0.83	–	1	0.83	1.0
2	–	3.4	–	0.1	–	0.9	0	0.83	0.9
3	–	5.0	–	0.2	–	0.8	0	0.83	0.8
4	5.2	–	0.33	–	0.67	–	1	0.67	0.8
5	5.6	–	0.50	–	0.50	–	1	0.50	0.8
6	–	6.0	–	0.3	–	0.7	0	0.50	0.7
7	–	6.1	–	0.4	–	0.6	0	0.50	0.6
8	6.5	–	0.67	–	0.33	–	1	0.33	0.6
9	–	7.0	–	0.5	–	0.5	0	0.33	0.5
10	–	7.3	–	0.6	–	0.4	0	0.33	0.4
11	–	7.7	–	0.7	–	0.3	0	0.33	0.3
12	–	8.4	–	0.8	–	0.2	0	0.33	0.2
13	8.8	–	0.83	–	0.17	–	1	0.17	0.2
14	–	9.0	–	0.9	–	0.1	0	0.17	0.1
15	–	9.2	–	1.0	–	0.0	0	0.17	0.0
16	9.9	–	1.00	–	0.0	–	1	0.00	0.0

[a] $new = id = 1$ and $std = id = 0$.

The parametric estimate of the area under the ROC curve

$$\text{area under ROC-curve} = a\hat{u}c = 1 - \Phi\left[\frac{\bar{x}_t - \bar{x}_s}{S\sqrt{2}}\right]$$
$$= 1 - \Phi\left[\frac{6.47 - 6.91}{2.123\sqrt{2}}\right] = 1 - \Phi(-0.148) = 0.559$$

is similar to the nonparametric estimate $a\hat{u}c = 0.567$ (Figure 22.17).

Implementation

The essence of the construction of a nonparametric ROC-curve in terms of a computer-like description is

start
$x = y = 0$
for $i = 1$ to $n_t + n_s$
if $id = 1$, then $x = x + 1/n_t$
set $X_i = 1 - x$
if $id = 0$, then $y = y + 1/n_s$
set $Y_i = 1 - y$
return

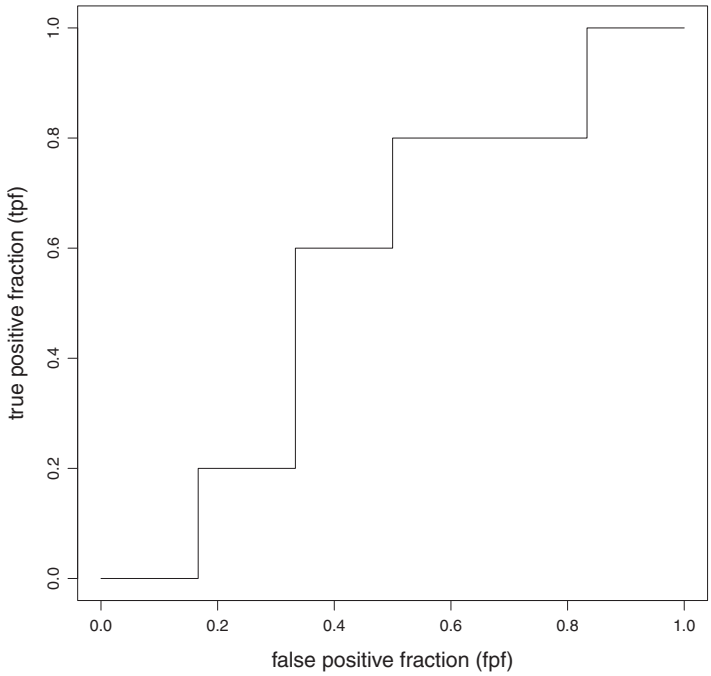


Figure 22.13(a) ROC−curve.

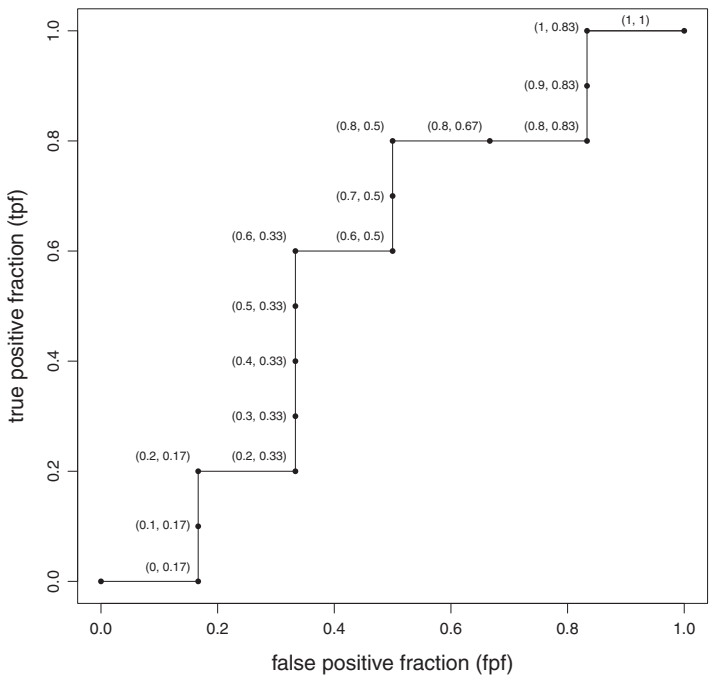


Figure 22.13(b) ROC−curve.

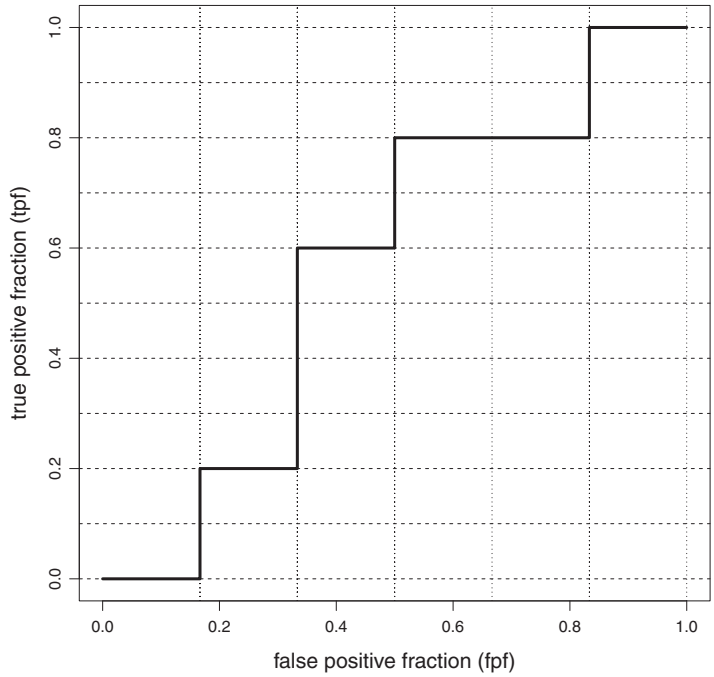


Figure 22.14(a) ROC−curve.

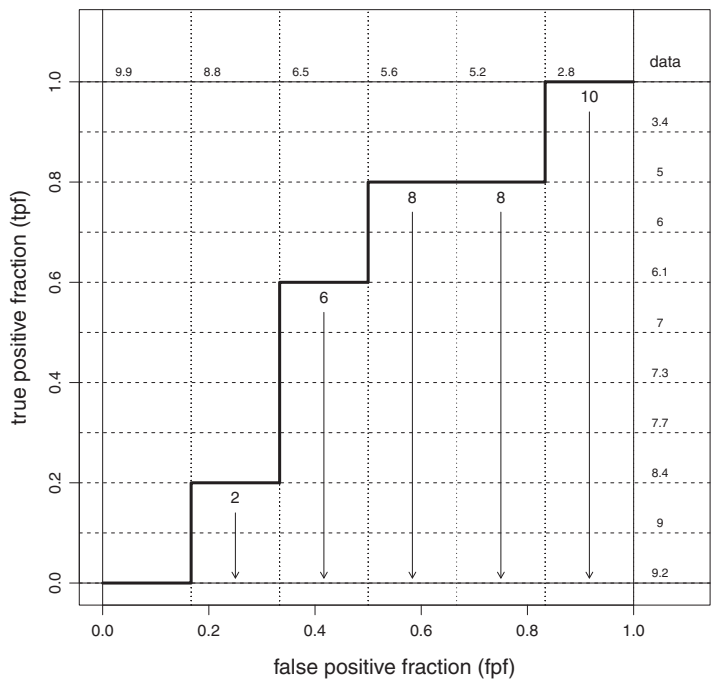


Figure 22.14(b) ROC−curve – count.

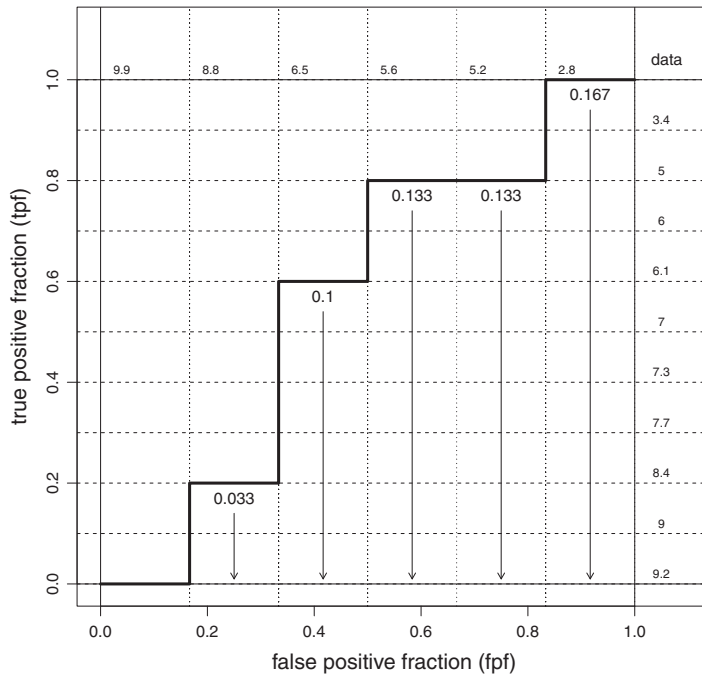


Figure 22.15 ROC−curve – area.

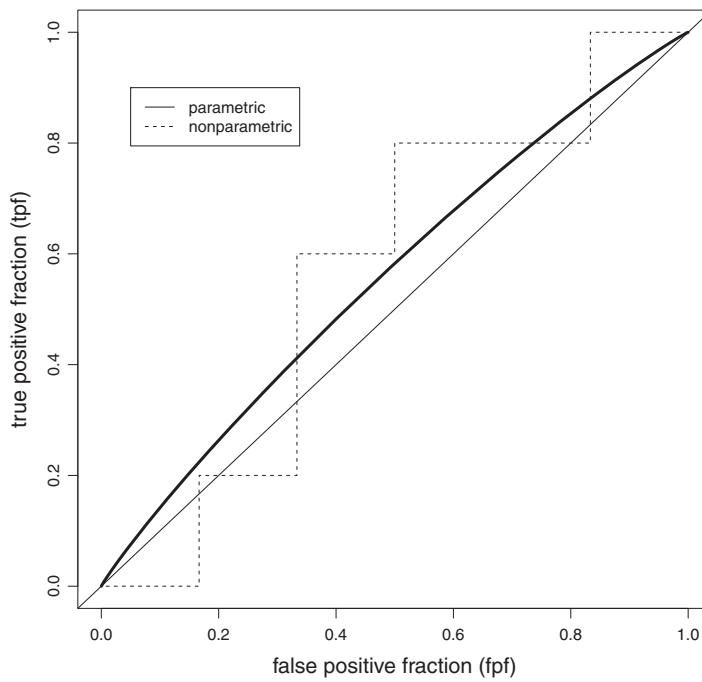


Figure 22.16 Parametric ROC−curve.

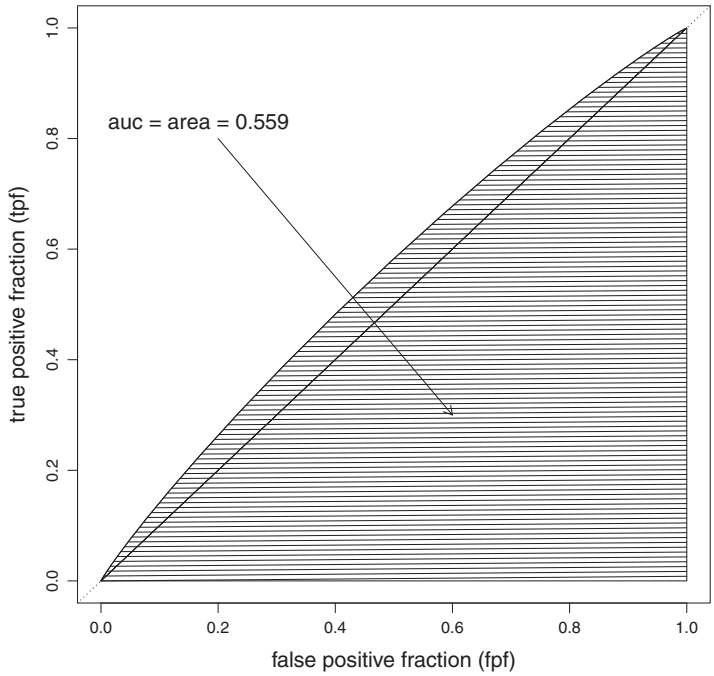


Figure 22.17 ROC−curve − normal parametric model.

The symbol *id* represents a binary variable where $id = 1$ indicates a "new" method and $id = 0$ indicates a standard method among the $n_t + n_s = n$ ordered observations (Table 22.7). Plotting the pairs of values (X_i, Y_i) creates the nonparametric-estimated ROC-curve (for example, Figures 22.13).

Genetics

23

Selection: A Statistical Description

A statistical description of the dynamics of genetic inheritance starts with the distribution of allele frequencies. Therefore, it is necessary to define the notation and properties and understand the estimation of an allele frequency from genotype data. The most elementary and certainly the most important case is simple Mendelian inheritance of two alleles, denoted A and a, producing three genotypes AA, Aa, and aa. An example of specific genotype counts and frequencies among 80 individuals is the following:

	Genotype Data			
	AA	Aa	aa	Total
n_i	40	20	20	80
$\hat{P}_i$	0.50	0.25	0.25	1.0

The symbol n_i represents the number of each of the observed genotypes ($n_1 = AA$-types, $n_2 = Aa$-types, and $n_3 = aa$-types), $\hat{P}_i$ represents the corresponding proportions of observed genotypes ($\hat{P}_i = n_i/n$) and $n_1 + n_2 + n_3 = n =$ the total number of genotypes observed. An estimate of the frequency of the A-allele (denoted $\hat{p}$) is the direct count of the A-alleles among $2n$ total alleles or

$$\hat{p} = \frac{\text{number of alleles } A}{\text{total number of alleles}}$$

and, from the example data, then

$$\hat{p} = \frac{2(40) + 20}{2(40) + 2(20) + 2(20)},$$

$$\hat{p} = \frac{2(0.50) + 0.25}{2(0.50) + 2(0.25) + 2(0.25)}, \text{ or}$$

$$\hat{p} = \frac{0.50 + \frac{1}{2}(0.25)}{0.50 + 0.25 + 0.25} = 0.50 + 0.125 = 0.625,$$

making $\hat{p} = 0.625$ the estimated frequency of the A-allele and $\hat{q} = 1 - \hat{p} = 0.375$ the frequency of the a-allele. In general, the expression for the estimated frequency of the A-allele is

$$\hat{p} = \frac{n_1 + \frac{1}{2}n_2}{n_1 + n_2 + n_3} = \frac{\hat{P}_1 + \frac{1}{2}\hat{P}_2}{\hat{P}_1 + \hat{P}_2 + \hat{P}_3} = \hat{P}_1 + \frac{1}{2}\hat{P}_2.$$

Table 23.1 *Family Laws: Probability Distributions of Offspring Genotypes from Six Parental Matings*

	Offspring Probabilities		
Parents	$P(AA)$	$P(Aa)$	$P(aa)$
$AA \times AA$	1	0	0
$AA \times Aa$	1/2	1/2	0
$AA \times aa$	0	1	0
$Aa \times Aa$	1/4	1/2	1/4
$Aa \times aa$	0	1/2	1/2
$aa \times aa$	0	0	1

Incidentally, the estimate $\hat{p}$ is the maximum likelihood estimate and, therefore, has an approximate normal distribution with minimum variance for large sample sizes. Frequently a simple and natural estimate such as $\hat{p} = 100/160 = 0.625$ is also the maximum likelihood estimate (Chapter 27).

Stable Equilibrium

The next step in describing the selection-fitness relationship is a demonstration that random mating produces constant allele frequencies, and, as a result, no change occurs in genotype frequencies over subsequent generations. This property of genetic inheritance is referred to as the *Hardy-Weinberg equilibrium*, named after an English mathematician and a German physician of the early twentieth century who first described the phenomenon.

Probabilities of three possible different offspring from six kinds of parental matings are referred to as the *family laws* (Table 23.1). For example, for $Aa \times Aa$ parents, the probability of an Aa-offspring is $P(Aa\ offspring|Aa \times Aa) = 0.5$. Applying the family laws to a random mating population produces the Hardy-Weinberg equilibrium genotype frequencies (Table 23.2). The basic requirement underlying this fundamental property of the dynamics of allele transmission from one generation to the next is that mating is strictly random; that is, nonrandom genetic forces such as mutation, selection, random drift, and associative mating do not influence the random inheritance patterns of the alleles under consideration (Chapter 26).

Consider the specific case where the A-allele has frequency $p = 0.7$ making the a-allele frequency $q = 1 - p = 0.3$, and all mating is random with respect to these alleles (generation 1 – Table 23.2). Then, based on the family laws, the existing genotype frequencies produce the distribution of genotypes in the next generation (generation = 2 – Table 23.2). The result, basic to many genetic analyses, is the allele frequency p in the next generation is identical to the value in the previous generation ($p = 0.7$).

Of central importance, the allele frequency p remains unchanged in the following generations of random mating. Therefore, the genotype frequencies remain unchanged. For example, for the AA-genotype (bottom – Table 23.2), the frequency of AA genotypes in the next generation is

$$P(AA) = p^4 + 2p^3q + p^2q^2 = p^2(p^2 + 2pq + q^2) = p^2,$$

which is exactly the same value as the previous generation and remains p^2 in all future generations. Specifically, for the example, the equilibrium frequency of the A-allele is $p^2 + pq = 0.49 + 0.21 = 0.7$, and the A-allele frequency p remains 0.7 as long as mating remains random. Hardy-Weinberg equilibrium (constant allele/genotype frequencies) is achieved in a single generation of random mating.

The Hardy-Weinberg equilibrium is an elegant and tractable description of the dynamics of a gene pool. The equilibrium results from four conditions:

1. A large population
2. All mating is random
3. The population is closed – for example, no migration in or out, and
4. Mutation or selection or other genetic forces do not change the allele frequencies.

A Description of Recombination

Forty six chromosomes contain all human genes (≈27,000). It might be expected that alleles on the same chromosome would be inherited together. In fact, this is not the case. The distribution of alleles in a random mating population are themselves randomly distributed. This remarkable property requires a unique biological mechanism called *crossover* or *recombination*. For example, when one chromosome of a pair contains two alleles A and B and the corresponding alleles a and b reside on the homologous chromosome, if crossovers did not exist, the A/B-allele pair and the a/b-allele pair would always be inherited together. However, homologous chromosomes "break" between the locations of A and B and a and b alleles and reconnect forming a "new" pairing on the homologous chromosomes where now A and b are on one chromosome and a and B are on the other (Figure 23.1). The allele pairs Ab and aB created in the next generation are called recombinants.

In general, A-allele frequency in the next generation is again $p' = p^2 + pq = p$.

Table 23.2 *Equilibrium: Demonstration That Random Mating Produces a Constant A-Allele Frequency of* $p = 0.7$ *Generation 1:* $p = 0.7$

Matings	Frequencies	*AA*	*Aa*	*aa*
$AA \times AA$	0.2401	0.2401	0	0
$AA \times Aa$	0.2058	0.1029	0.1029	0
$AA \times aa$	0.0441	0	0.0441	0
$Aa \times AA$	0.2058	0.1029	0.1029	0
$Aa \times Aa$	0.1764	0.0441	0.0882	0.0441
$Aa \times aa$	0.0378	0	0.0189	0.0189
$aa \times AA$	0.0441	0	0.0441	0
$aa \times Aa$	0.0378	0	0.0189	0.0189
$aa \times aa$	0.0081	0	0	0.0081
Generation 2	1.0	0.49	0.42	0.09

Note: The A-allele frequency $p = 0.49 + \frac{1}{2}(0.42) = 0.7$ and remains 0.7 for all future generations.

Generation 1: Gene Frequency = p				
Matings	Frequencies	AA	Aa	aa
$AA \times AA$	p^4	p^4	0	0
$AA \times Aa$	$2p^3q$	p^3q	p^3q	0
$AA \times aa$	p^2q^2	0	p^2q^2	0
$Aa \times AA$	$2p^3q$	p^3q	p^3q	0
$Aa \times Aa$	$4p^2q^2$	p^2q^2	$2p^2q^2$	p^2q^2
$Aa \times aa$	$2pq^3$	0	pq^3	pq^3
$aa \times AA$	p^2q^2	0	p^2q^2	0
$aa \times Aa$	$2pq^3$	0	pq^3	pq^3
$aa \times aa$	q^4	0	0	q^4
Next generation	1.0	p^2	$2pq$	q^2

Consider the four pairs of alleles ab (frequency $= w$), AB (frequency $= x$), Ab (frequency $= y$), and aB (frequency $= z$). A rate of recombination (denoted r) produces over a number of generations frequencies of the alleles A, a, B, and b that are randomly distributed in the gene pool, regardless of the original pair frequencies of w, x, y, and z. A bit of algebra demonstrates the dynamics of this fundamental recombination process (Table 23.3).

Table 23.4 is a specific example of the convergence to equilibrium of two simple polymorphic genetic systems, again represented as A/a and B/b, where the initial allele frequencies are $p_A = 0.3$, $p_a = 0.7$, $p_B = 0.7$, and $p_b = 0.3$. In generation 0, the allele pair frequencies are $x_0 = 0.10$, $y_0 = 0.20$, $z_0 = 0.60$, and $w_0 = 0.10$ making a measure of nonrandomness $\delta_0 = 0.011$ (to be described). In addition, the recombination rate is $r = 0.10$. After 45 generations, these allele frequencies converge to an equilibrium of $p_{AB} = x = 0.21$, $p_{Ab} = y = 0.09$, $p_{aB} = z = 0.49$, and $p_{ab} = w = 0.21$, making $\delta = (0.09)(0.49) - (0.21)(0.21) = 0$ (Table 23.4).

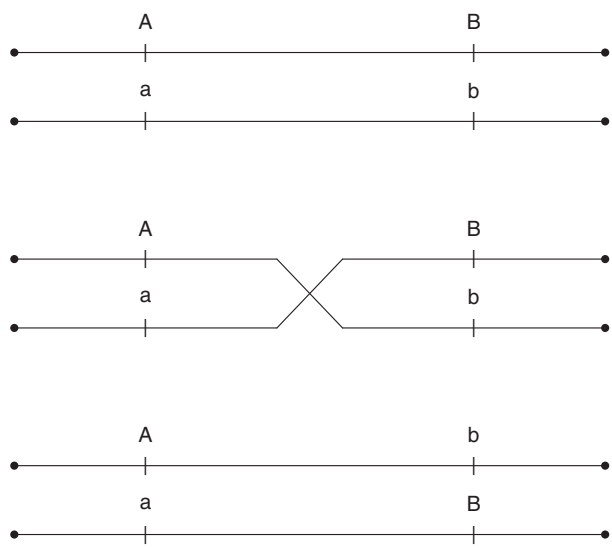


Figure 23.1 Schematic Representation of Crossover between Two Alleles on Homologous Chromosomes

Table 23.3 *Allele Frequencies: Demonstration of Convergence to Independence of Allele Frequencies within a Random Mating Population (r = Rate of Recombination)*

		Generation 0:			
	Frequency	*AB*	*Ab*	*aB*	*ab*
AB/*AB*	x_0^2	x_0^2	0	0	0
AB/*Ab*	$2x_0y_0$	$\frac{1}{2}\ (2x_0y_0)$	$\frac{1}{2}\ (2x_0y_0)$	0	0
AB/*aB*	$2x_0z_0$	$\frac{1}{2}\ (2x_0z_0)$	0	$\frac{1}{2}\ (2x_0z_0)$	0
AB/*ab*	$2x_0w_0$	$\frac{1}{2}(2x_0w_0)(1-r)$	$\frac{1}{2}(2x_0w_0)r$	$\frac{1}{2}(2x_0w_0)r$	$\frac{1}{2}(2x_0w_0)(1-r)$
Ab/*Ab*	y_0^2	0	$y^2{}_0$	0	0
Ab/*aB*	$2y_0z_0$	$\frac{1}{2}\ (2y_0z_0)r$	$\frac{1}{2}\ (2y_0z_0)(1-r)$	$\frac{1}{2}\ (2y_0z_0)(1-r)$	$\frac{1}{2}\ (2y_0z_0)$r
Ab/*ab*	$2y_0w_0$	0	$\frac{1}{2}(2y_0w_0)$	0	$\frac{1}{2}(2y_0w_0)$
aB/*aB*	z_0^2	0	0	z_0^2	0
aB/*ab*	$2z_0w_0$	0	0	$\frac{1}{2}(2z_0w_0)$	$\frac{1}{2}(2z_0w_0)$
ab/*ab*	w_0^2	0	0	0	w_0^2
	1.0	x_1	y_1	z_1	w_1

Then, for generation 1:

$$AB : x_1 = x_0^2 + x_0y_0 + x_0z_0 + x_0w_0(1-r) + y_0z_0r = x_0 + r(y_0z_0 - x_0w_0),$$
$$Ab : y_1 = x_0y_0 + x_0w_0r + y_0^2 + y_0z_0(1-r) + y_0w_0 = y_0 - r(y_0z_0 - x_0w_0),$$
$$aB : z_1 = x_0z_0 + x_0w_0r + y_0z_0(1-r) + z_0^2 + z_0w_0 = z_0 - r(y_0z_0 - x_0w_0),$$
$$ab : w_1 = x_0w_0(1-r) + y_0z_0r + y_0w_0 + z_0w_0 + w_0^2 = w_0 + r(y_0z_0 - x_0w_0),$$

then, for $\delta_0 = (y_0z_0 - x_0w_0)$,

$$AB : x_1 = x_0 + r\delta_0\ Ab :\ y_1 = y_0 - r\delta_0\ aB :\ z_1 = z_0 - r\delta_0\ ab :\ w_1 = w_0 + r\delta_0,$$
$$\delta_1 = (y_1z_1 - x_1w_1) = (y_0 - r\delta_0)(z_0 - r\delta_0) - (x_0 + r\delta_0)(w_0 + r\delta_0) = \delta_0(1-r);$$

Table 23.4 *Illustration: Specific Example of Random Mating and Recombination Causing Alleles to Become Distributed at Random within a Gene Pool (r = 0.10)*

Generation	*x*	*y*	*z*	*w*	*δ*
	AB	*Ab*	*aB*	*ab*	
0	0.100	0.200	0.600	0.100	0.011
5	0.145	0.155	0.555	0.145	0.065
10	0.172	0.128	0.528	0.172	0.038
15	0.187	0.113	0.513	0.187	0.023
20	0.197	0.103	0.503	0.197	0.013
25	0.202	0.098	0.498	0.202	0.008
30	0.205	0.095	0.495	0.205	0.005
35	0.207	0.093	0.493	0.207	0.003
40	0.208	0.092	0.492	0.208	0.002
45	0.210	0.090	0.490	0.210	0.000

Table 23.5 *Summary: Convergence from Arbitrary Distribution of Four Allele Pairs to Equilibrium State of Random Distribution of Alleles Determined Only by Allele Frequencies in the Population ($p_A = 0.3$ and $p_B = 0.7$)*

	Allele Pairs				
	AB	*Ab*	*aB*	*ab*	δ
Generation 0	0.10	0.20	0.60	0.10	0.11
Equilibrium	0.21	0.09	0.49	0.21	0.00
Independence	0.3(0.7)	0.3(0.3)	0.7(0.7)	0.7(0.3)	0.00
In symbols	$p_A(p_B)$	$p_A(p_b)$	$p_a(p_B)$	$p_a(p_b)$	–

therefore,

$$\begin{aligned}
&\text{generation } 1: \ \delta_1 = \delta_0(1-r), \text{ then}\\
&\text{generation } 2: \ \delta_2 = \delta_1(1-r) = \delta_0(1-r)^2, \text{ then}\\
&\text{generation } 3: \ \delta_3 = \delta_2(1-r) = \delta_1(1-r)^2 = \delta_0(1-r)^3, \text{ then}\\
&---\\
&\text{generation } k: \ \delta_k = \delta_{k-1}(1-r) = \delta_{k-2}(1-r)^2 = \cdots = \delta_0(1-r)^k.
\end{aligned}$$

When the number of generations k is large, δ_k becomes essentially zero ($\delta \to 0$).

The symbol δ represents a measure of the extent the allele frequencies are not randomly distributed. Specifically, the expression for δ is $(yz - xw)$. Thus, the value $\delta = 0$ occurs when alleles are distributed independently in the gene pool. In symbols, for the example $\delta_0 = 0.20(0.60) - (0.10)(0.10) = 0.11$ becomes

$$\delta = yz - xw = p_{Ab}p_{aB} - p_{AB}p_{ab} = (p_A q_b)(p_a q_B) - (p_A q_B)(p_a q_b) = 0,$$

where again the allele frequencies are p_A (the frequency of the A-allele) and p_B (the frequency of the B-allele).

Three issues of note:

First, the allele frequencies are the same for every generation ($p_A = x + y = 0.3$ and $p_B = z + w = 0.7$). Hardy-Weinberg equilibrium guarantees this property as long as the mating is random.

Second, as illustrated (Table 23.5), the allele frequencies become distributed so that the alleles A, a, B, and b are statistically independent ($\delta = 0$). For the example, the probability of the occurrence of an AB allele pair after 45 generations is $p_{AB} = p_A(p_B) = 0.3(0.7) = 0.21$ (Tables 23.4 and 23.5).

Third, the rate of recombination (r) does not influence the ultimate distribution of allele/genotype frequencies. It determines the rate random mating, and recombination produces equilibrium random allele frequencies.

To summarize, regardless of the original genotype frequencies, crossover and random mating eventually produce a random assortment of alleles in a gene pool.

The fact that a gene pool resulting from random mating individuals can be viewed as a random assortment of alleles allows simple determinations of many general properties of a

random mating population based on the single allele frequency p. Seven examples describe a few properties generated from randomly assorting alleles in a gene pool with A-allele frequency p and a-allele frequency $q = 1 - p$:

probability of an $aa \times aa$ mating $= q^4$,
probability of an $Aa \times Aa$ mating $= 4p^2q^2$,
probability of an AA or $Aa = A^+$ individual $= p^2 + 2pq = 1 - q^2$,
probability of an Aa individuals among individuals with A-allele $= P(Aa \mid A) = q$,
probability of a mother daughter pair $AA/Aa = p^2q$,
probability of an AA offspring among $A^+ \times A^+$ mating $= (1 - q^2)^2$ and
probability of a sib/sib pair $Aa/Aa = pq(1 + pq)$.

Random distribution of alleles in the gene pool is evolutionary protection against catastrophic outside pressures for change. Casino gambling provides an excellent analogy. All gambling games in a casino produce random results; that is, no patterns exist. Therefore, players cannot discover a pattern and take advantage to win large sums of money. Randomness protects a casino from outside destructive influences much as the random distribution of alleles protects an established gene pool.

Two Examples of Statistical Analyses

The following two examples demonstrate typical analytic patterns that capitalize on the properties of a Hardy-Weinberg equilibrium to explore specific genetic issues with statistical tools.

Mother-Child Disequilibrium

An important genetic technique is the search for deviations from Hardy-Weinberg expected frequencies. The identification of disequilibrium associated with a specific allele, for example, presents evidence of a nonrandom genetic influence potentially leading to a better understanding of a disease. Data consisting of $n = 120$ mother-child pairs where the child has been diagnosed with acute lymphatic leukemia present the opportunity to search their genetic make-up for nonrandom allele mother-child assortment that potentially indicates a genetic role in the risk of the disease. A specific genotype (allele labels A/a) illustrates one of the huge number of possible analyses involved in a search for substantial disequilibrium, one for each observed allele pair.

The seven possible different mother-child pairs are displayed in Table 23.6 (row 1). The frequencies of these pairs (row 2) are derived under the premise that mating is random, and, therefore, the genotype frequencies conform to those expected from a Hardy-Weinberg equilibrium (Table 23.2).

Like the previous case of a single allele, estimation of the frequency of a specific allele from mother-child data is the count of independent occurrences of that allele divided by the total number of alleles. Mother-child pairs always have a common allele, so to correctly count the independent alleles it is necessary not to double count the duplicated allele. For

Table 23.6 *Seven Possible Mother-Child Pairs, Their Observed (o_i) and Hardy-Weinberg–Generated Frequencies (e_i) (Rows 3 and 5)*

	Mother-Child Genotype Pairs							
Genotypes Frequencies	AA/AA p^3	AA/Aa p^2q	AA/Aa p^2q	Aa/Aa pq	Aa/aa pq^2	aa/Aa pq^2	aa/aa q^3	Total 1.0
Data/notation (o_i)	$a = 9$	$b = 14$	$c = 16$	$d = 25$	$e = 18$	$f = 15$	$g = 23$	120
Frequencies	0.081	0.106	0.106	0.245	0.139	0.139	0.182	1.0
Expected (e_i)	9.73	12.75	12.75	29.46	16.71	16.71	21.89	120

example, the mother-child pair AA/Aa contributes two A-alleles to the count, not three. Following this pattern the count of independent A-alleles is

$$A\text{-allele count} = 3a + 2b + 2c + d + e + f = 145.$$

Similarly, the a-allele count is

$$a\text{-allele count} = b + c + d + 2e + 2f + 3g = 190.$$

The total number of alleles becomes

$$N = 3n - d = 335 \quad \text{where} \quad n = a + b + c + d + e + f + g = 120 \text{ pairs (Table 23.6).}$$

The estimated A-allele frequency is then

$$\hat{p} = \frac{A\text{-allele count}}{N} = \frac{145}{335} = 0.433 \quad \text{and} \quad 1 - \hat{p} = \hat{q} = \frac{a\text{-allele count}}{N} = \frac{190}{335} = 0.567$$

and are, in addition, maximum likelihood estimates (Chapter 27).

Using the two estimates $\hat{p}$ and $1 - \hat{p} = \hat{q}$, the expected count of the mother-child pairs is directly calculated as if the alleles assort randomly (Hardy-Weinberg equilibrium – Table 23.6). The comparison of the observed counts to theoretical counts likely identifies the existence a genetic disequilibrium. Mechanically a chi-square statistic $X^2 = \Sigma(o_i - e_i)^2/e_i$ is a comparison of the seven observed counts (row 3 – o_i) to the Hardy-Weinberg estimated counts (row 5 – e_i). The resulting measure of disequilibrium in the example data measured by a chi-square statistic $X^2 = 2.011$ (degrees of freedom = 6) yields no evidence of a nonrandom allele distribution within the 120 mother-child pairs (p-value = 0.918).

Like all estimates, the estimate $\hat{p} = 0.433$ is subject to random variation. For this mother-child data, the estimated value $\hat{p}$ has a binomial distribution estimated variance of

$$variance(\hat{p}) = \frac{\hat{p}(1 - \hat{p})}{N} = 0.00073.$$

The corresponding approximate 95% confidence interval is $0.433 \pm 1.960(0.0271) \rightarrow (0.380, 0.486)$.

Mendelian Inheritance – Synder's Ratio

Hardy-Weinberg genotype frequencies again make it possible to explore genetic issues based on the properties of random mating within a gene pool with statistical tools. An important

Table 23.7 *Genetic Model: Phenotype Frequencies That Result from Random Matings between Dominant (Denoted* A^+*) and Recessive (Denoted* aa*) Parents*

Matings	Frequency	A^+	aa	Total
$A^+ \times A^+$	$(1-q^2)^2$	$p^2(1+2q)$	p^2q^2	$p^2(1+q)^2$
$A^+ \times aa$	$2q^2(1-q^2)$	$2pq^2$	$2pq^3$	$2pq^2(1+q)$
$aa \times aa$	q^4	0	q^4	q^4
Total	1.0	$1-q^2$	q^2	1.0

case occurs when only two kinds of offspring are observed; one is a combinations of dominant AA or Aa genotypes, represented by phenotype $A^+ = AA$ or Aa, and the other is the recessive genotype aa. In a random mating population the phenotype frequencies are then combinations of genotype frequencies. For example, the frequency of recessive aa-offspring from parents that are A^+ and aa is $pq^3 + pq^3 = 2pq^3$ (Table 23.2).

Combinations of the phenotype frequencies, such as Table 23.7, are created under the frequently realistic assumption of random mating that allows genetic data to be statistically analyzed from a variety of perspectives.

The probability of recessive aa-genotypes from an $A^+ \times A^+$ mating is

$$P(aa|A^+ \times A^+) = S_2 = \frac{p^2q^2}{p^2(1+2q)+p^2q^2} = \left[\frac{q}{1+q}\right]^2,$$

and similarly the probability of aa-genotypes from an $A^+ \times aa$ mating is

$$P(aa|A^+ \times aa) = S_1 = \frac{2pq^3}{2pq^2+2pq^3} = \frac{q}{1+q}.$$

Thus, when mating is random and data reflect a simple Mendelian inheritance pattern, the ratio S_2/S_1^2 is likely to be in the neighborhood of 1.0, called *Synder's ratio*. Thus, an observed ratio S_2/S_1^2 close to 1.0 serves as an indication of Mendelian mode of inheritance of a two-allele genetic trait.

Consider the data describing the inheritance of the ability to taste a specific substance called phenylthiocarbamide (PTC). The data in Table 23.8 were collected to assess the conjecture that PTC individuals who can taste this substance have a Mendelian dominant phenotype (A^+) and those who cannot have a recessive genotype (aa).

The PTC genotype data (Table 23.8) yield the values $S_2 = 19/213 = 0.089$ and $S_1 = 28/145 = 0.193$ making Synder's ratio $S_2/S_1^2 = 0.089/(0.193)^2 = 2.392$. Such a large ratio likely indicates that inheritance is more complicated than the basic random two-allele mating Mendelian pattern (Table 23.2). The estimate of the a-allele frequency q comes from the proportion of nontasters (aa) where $\hat{q}^2 = 73/384 = 0.190$ yields estimate $\hat{q} = \sqrt{0.190} = 0.436$ and $1 - \hat{q} = \hat{p} = 0.564$.

A straightforward statistical comparison of the observed PTC counts (o_i) to the expected counts (e_i) that conform exactly to a pattern of simple Mendelian inheritance (counts in parentheses – Table 23.8) is achieved with Pearson's chi-square goodness-of-fit test statistic (Chapter 1). The chi-square test statistic $X^2 = 8.400$ (degrees of freedom = 2) produces a p-value of 0.015 providing substantial evidence that the observed phenotype frequencies are

Table 23.8 *Example Data: Observed Counts of the Present* (A^+) *and Absent (aa) of Phenylthiocarbamide (PTC) Taste Deficiency and Expected Mendelian Frequencies (in Parentheses*[a]*)*

Type	Frequency	A^+	aa	Total
$A^+ \times A^+$	$(1-q^2)^2$	194 (193.4)	19 (19.6)	213
$A^+ \times aa$	$2q^2\,(1-q^2)$	117 (101.0)	28 (44.0)	145
$aa \times aa$	q^4	(18)[b] (0)	26 (26)	26

[a]Expected values when genotypes result from the Hardy-Weinberg equilibrium (Tables 23.2 and 23.3).
[b]Treated as random misclassification and excluded from the chi-square analysis.

not likely random deviations from a Mendelian inheritance pattern. In addition, the fact that 18 A^+-individuals are observed from the mating pairs $aa \times aa$ and Mendelian inheritance dictates none is either substantial misclassification or additional evidence of failure of the expected Mendelian pattern to accurately describe the observed relationships. These recent data using more advanced measures of PTC tasting do not support the previous belief that PTC tasting is governed by a two-allele simple Mendelian inheritance. Note that key to both statistical analyses is the random association of alleles created by a Hardy-Weinberg equilibrium.

Selection – Recessive Lethal

The following genetic model illustrates extreme genetic selection in a random mating population. In every generation, the *aa*-genotype does not survive. Statistically, the probability that an *aa*-type individual survives is zero. Therefore, *a*-alleles are lost from the gene pool each generation and proportion of *A*-alleles correspondingly increases. Consider an example of a recessive lethal genotype in generation 0 with *A*-allele frequency 0.3 ($p = 0.3$) that produces genotype frequencies $P(AA) = 0.09$, $P(Aa) = 0.42$, and $P(aa) = 0.49$. When mating between individuals is random and *aa*-genotypes do not survive, then the allele frequency p increases in the next generation and

Generation 1	*AA*	*Aa*	*aa*
Genotypes	$(0.3)^2$	2(0.3)(0.7)	$(0.7)^2$
Frequency	0.09	0.42	0

$$p = \frac{0.09 + \frac{1}{2}(0.42)}{0.09 + 0.42} = 0.588$$

Generation 2	*AA*	*Aa*	*aa*
Genotypes	$(0.588)^2$	2(0.588)(0.412)	$(0.412)^2$
Frequency	0.346	0.485	0

$$p = \frac{0.346 + \frac{1}{2}(0.485)}{0.346 + 0.485} = 0.708$$

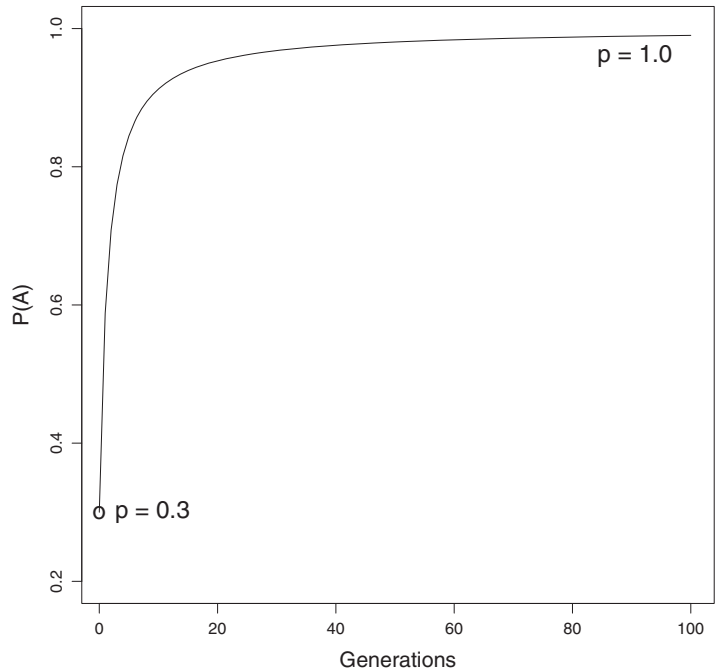


Figure 23.2 Increase in Frequency of A-Alleles When All aa-Genotypes Do Not Survive

Generation 3	AA	Aa	aa
Genotypes	$(0.708)^2$	$2(0.708)(0.292)$	$(0.292)^2$
Frequency	0.502	0.413	0

$$p = \frac{0.502 + \frac{1}{2}(0.413)}{0.502 + 0.413} = 0.774$$

$$---$$

$$\text{generation} = 10$$

$$p = \frac{833 + \frac{1}{2}(0.160)}{0.833 + 0.160} = 0.913.$$

For the A-allele with frequency p, the allele frequency in the next generation p' is

$$p' = \frac{p^2 + pq}{p^2 + 2pq} = \frac{p}{1 - q^2} > p.$$

Because p' is greater than p for every subsequent generation, the frequency of the A-allele increases ultimately to $p = 1.0$ (Figure 23.2).

In the first half of the twentieth century it was thought that the quality of a gene pool could be improved by sterilization or execution of individuals who possessed "defective genes." A simple statistical illustration shows the folly of this belief. Notice that even under

the strongest possible selection against the recessive genotype, the rate of elimination of *aa*-genotypes dramatically slows after a few generations (Figure 23.2).

Specifically, after 50 generations or so the allele frequency of the deleterious *a*-allele $q = 0.7$ ($p = 0.3$) is reduced to $q = 0.01$ (Figure 23.2). The genotype frequencies are then

$$\begin{aligned} aa: \ q^2 &= 0.01^2 = 0.0001, \\ Aa: \ 2pq &= 2(0.99)(0.01) = 0.0198, \text{ and} \\ AA: \ p^2 &= (0.99)^2 = 0.9801. \end{aligned}$$

The *Aa*-genotypes contain the vast majority of the *a*-alleles (99.5%). Therefore, the *aa*-genotypes become an extremely rare source of the *a*-allele. The proportion of *aa* genotypes eliminated is approximately $q/2$ in each generation; for the example, this proportion is $0.01/2 = 0.005$. Therefore, the vast majority of *a*-alleles are not expressed (visible) and, therefore, not subject to artificial selection. The example illustrates that after a few generations the elimination of "defective" individuals has at best a tiny affect on the frequency of the *a*-allele. Furthermore, many harmful alleles have frequencies considerably less than 0.01. In addition, a variety of other forces such as mutation, migration, or possibly selection advantages of the *Aa*-genotype are among a large number of possible sources that produce replacement *a*-alleles with frequency of considerably more than $q/2$ (0.5% in the example) each generation. Thus, in practical terms, elimination *aa*-genotypes has no genetic effect.

A More General Selection Pattern

A more general pattern of selection allows the probability of elimination to vary from $s = 0$ (no selection) to lethal $s = 1$ (no surviving *aa*-types – previous case). In symbols, the resulting random mating genotypes for any generation produces frequencies where *s* denotes

Generation *k*	*AA*	*Aa*	*aa*	Total
Genotypes	p^2	$2pq$	$q^2(1-s)$	$1-q^2s$

the probability of failure to survival and $(1 - s)$ denotes the probability of survival. Again, allele frequency *p* increases each generation ($s \neq 0$). Specifically, the increase is

$$p' = \frac{p^2 + pq}{p^2 + 2pq + q^2(1-s)} = \frac{p}{1-q^2} \geq p,$$

where p' is the frequency of the *A*-allele in the next generation. As before, allele frequently *p* increases (*q* decreases) every generation, and the rate of increase is governed by the probability of *aa*-genotype survival, $1 - s$ (Figure 23.3). For a selection probability other than $s = 0$, the *a*-allele ultimately disappears from the gene pool when no source of replacement exists. Note that when $s = 0$, then $p = p'$ each generation.

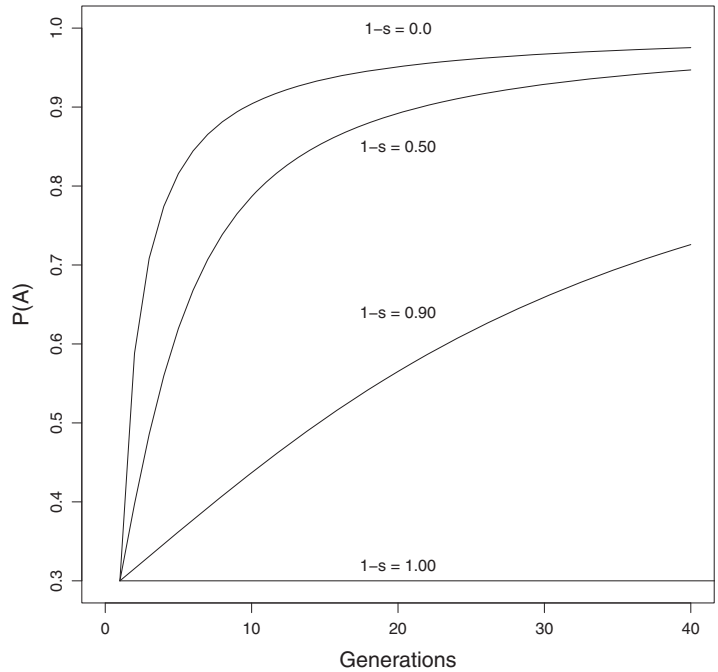


Figure 23.3 Increase in Frequency of the *A*-Allele Due to Selection against the *aa*-Genotype for Values of $s = \{1.0, 0.10\ 0.5$ and $0.0\}$

Selection – A Balanced Polymorphism

Under specific conditions, selection produces a stable allele frequency after a number of generations of random mating. In other words, the allele frequency p converges to a specific value and remains at that value. When an allele frequency is not zero or one, selection is then said to have created a "balance" of the three genotype frequencies forming a *balanced polymorphism*. For example, when selection begins (generation 0) with an *A*-allele frequency at $p_0 = 0.2$, with *AA*-genotype survival probability of $1 - S = 0.8$ ($S = 0.2$) and an *aa*-genotype survival probability of $1 - s = 0.4$ ($s = 0.6$), the genotype frequencies are the following:

Generation 0	*AA*	*Aa*	*aa*
Genotypes	$(0.2)^2$	2(0.2)(0.8)	$(0.8)^2$
Selection	0.8	1.00	0.4
Frequency	0.032	0.320	0.256

$$p = \frac{0.032 + \frac{1}{2}(0.320)}{0.032 + 0.32 + 0.256} = 0.316.$$

For the next generations:

Generation 1	*AA*	*Aa*	*aa*
Genotypes	$(0.316)^2$	$2(0.316)(0.684)$	$(0.684)^2$
Selection	0.8	1.00	0.4
Frequency	0.080	0.432	0.187

$$p = \frac{0.080 + \frac{1}{2}(0.432)}{0.080 + 0.432 + 0.187} = 0.423,$$

Generation 2	*AA*	*Aa*	*aa*
Genotypes	$(0.423)^2$	$2(0.423)(0.577)$	$(0.577)^2$
Selection	0.8	1.00	0.4
Frequency	0.143	0.488	0.133

$$p = \frac{0.143 + \frac{1}{2}(0.488)}{0.143 + 0.488 + 0.133} = 0.507$$

$$---$$

Generation 10	*AA*	*Aa*	*aa*
Genotypes	0.389	0.432	0.037

$$p = 0.708$$

$$---$$

Generation 20	*AA*	*Aa*	*aa*
Genotypes	0.442	0.380	0.026

$$p = 0.747$$

$$---$$

Generation 40	*AA*	*Aa*	*aa*
Genotypes	0.450	0.375	0.025

$$p = 0.75.$$

After a number of generations of random mating (40 in the example), the A-allele frequency converges to $p = s/(S + s) = 0.6/(0.2 + 0.6) = 0.75$ (Figure 23.4) and remains $p = 0.75$ in the absence of changes in the selection probabilities S or s.

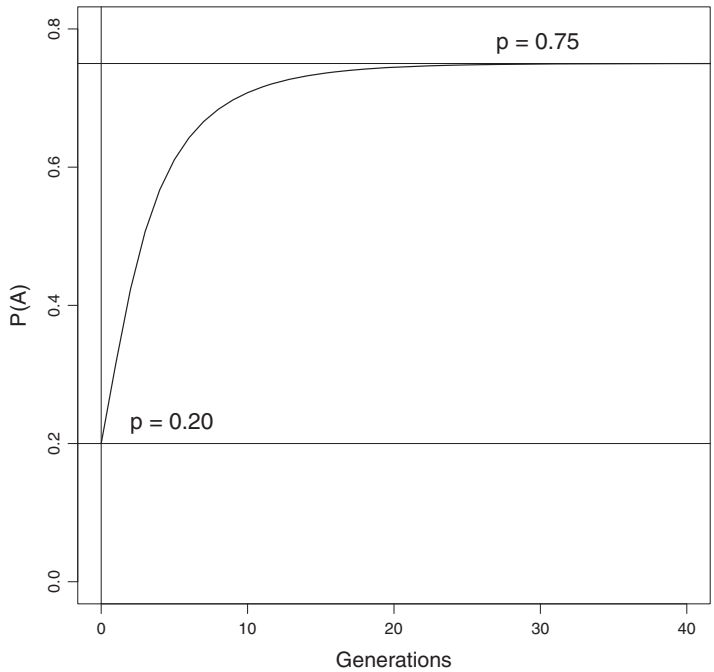


Figure 23.4 Display of Convergence of the A-Allele Frequency of $p = 0.2$ (Generation 0) to Constant Value of 0.75 in 40 Generations Creating a Balanced Polymorphism ($p_0 = 0.2$, $1 - S = 0.8$ and $1 - s = 0.4$)

A demonstration of the convergence to a specific allele frequency (0.75 in the example) requires a little algebra. The A-allele frequency p in the next generation, denoted again p', is

$$p' = \frac{p^2(1-S) + pq}{p^2(1-S) + 2pq + q^2(1-s)} = \frac{p - p^2S}{1 - p^2S - q^2s}.$$

When consecutive generations produce a constant allele frequency, then $p' = p$ or $p' - p = 0$; that is, the genotypic frequencies have reached a stable equilibrium yielding a constant value of p. Specifically, in symbols,

$$p' - p = \frac{p - p^2S}{1 - p^2S - q^2s} - p = 0 \quad \text{or} \quad p' - p = p - p^2S - p(1 - p^2S - q^2s) = 0,$$

and this expression simplifies to

$$-pS + p^2S - (1-p)^2s = -p\,(1-p)\,S + (1-p)^2s = -pS + s - ps = 0,$$

making

$$p\,(S+s) = s.$$

Thus, allele frequencies that result from selection become constant and are

$$p = \frac{s}{S+s} \quad \text{and} \quad 1 - p = q = \frac{S}{S+s};$$

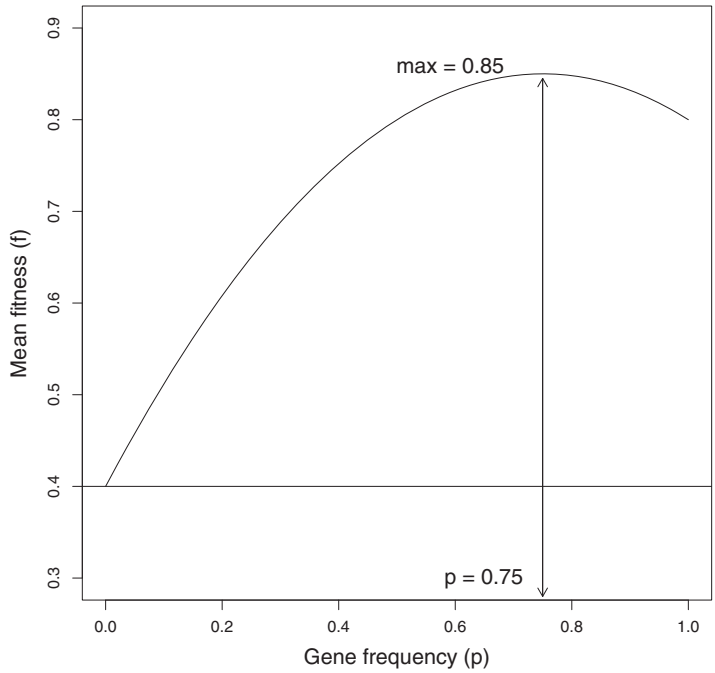


Figure 23.5 Frequency of A-Allele as Function of Mean Fitness, Again $1 - s = 0.4$ and $1 - S = 0.8$

that is, allele frequencies are ultimately determined only by the values of the selection probabilities s and S regardless of their original allele frequencies. The genotype frequencies have formed a balanced polymorphism.

Fitness

Fitness in a genetic context is measured only by the degree of success in passing specific alleles from the present generation to the next. For two alleles A and a producing genotypes AA, Aa, and aa, a genetic definition of mean fitness is the weighted average of selection frequencies $(1 - S, 1, 1 - s)$ with genotypic frequencies as weights $\{p^2, 2pq, q^2\}$ or

$$\begin{aligned}\text{mean fitness} &= \text{average survival} = \bar{f} = (1 - S)p^2 + (1)2pq + (1 - s)q^2 \\ &= (1 - S)p^2 + 2p(1 - p) + (1 - s)(1 - p)^2 \\ &= -(S + s)p^2 + 2sp + (1 - s),\end{aligned}$$

where, to repeat, $1 - S$ and $1 - s$ are probabilities of survival to the next generation. The expression for mean fitness $\bar{f}$ is a quadratic function of the allele frequency p with a maximum value between $1 - s$ and $1 - S$.

Values of mean fitness ($\bar{f}$) ranging from $1 - s$ to $1 - S$, calculated for selected frequencies of the A-allele for survival probabilities again $1 - s = 0.4$ and $1 - S = 0.8$ illustrate (Table 23.9 and Figure 23.5). If $s = S = 0$, then the mean fitness is 1.0 for all values of p.

Table 23.9 *Example: Values of Mean Fitness for* $1 - s = 0.4$ *and* $1 - S = 0.8$ *for Selected Allele Frequencies p*

p	0.0	0.2	0.4	0.6	0.75[a]	0.8	1.0
$\bar{f}$	0.400	0.608	0.752	0.832	0.85	0.848	0.800

[a] p at the stable equilibrium.

Consider again the previous balanced selection described by the following:

	AA	*Aa*	*aa*
Symbols	$1 - S$	1	$1 - s$
Values	0.8	1	0.4

For the example, the maximum mean fitness occurs at $\bar{f} = 0.85$ when $p = 0.75$. Therefore, the value of mean fitness is maximized ($\bar{f} = 0.85$) and occurs at the gene frequency that results from a polymorphism created by balanced selection ($p = 0.75$).

In general, the quadratic equation describing mean fitness has a maximum value that occurs at $p = s/(S + s)$, which is the allele frequency at the equilibrium created by balanced selection. Algebraically, the most extreme value of a quadratic function $ax^2 + bx + c$ occurs at $x = -b/2a$ (Chapter 27). From the quadratic expression describing mean fitness $\bar{f}$, where $a = -(s + S)$ and $b = 2s$, the maximum value then occurs at $p = s/(S + s)$.

Therefore, selection either essentially removes an allele from the population or creates a frequency that yields a stable maximum fitness. Of course, in reality the situation is far more complex, but this simple description of Mendelian inheritance gives a sense of the dynamics of selection in the determining the composition of a random mating gene pool with maximum fitness.

24

Mendelian Segregation Analysis

The distribution of genotypes among siblings within a family offers an opportunity to investigate the pattern of inheritance of a specific genetic trait. When heterozygotic parents ($Aa \times Aa$) produce a recessive offspring (aa) with probability $p = 0.25$, the trait potentially has a simple Mendelian mode of inheritance (Chapter 23). As they say in mathematics, "It is a necessary but not a sufficient condition." From a random sample of offspring of heterozygotic parents, it is straightforward to calculate the proportion of a recessive genotypes, called the *segregation probability*. Gergor Mendel's (b. 1822) sweet pea experiments in the 19th century established segregation probabilities for a number of genetic traits by calculating the proportion of *aa*-recessive offspring from an extremely large number of plant-breeding $Aa \times Aa$ experiments and comparing the value to 0.25 (binomial distribution; Chapter 4). The same frequency of recessive offspring is also called the *segregation ratio* referring to the fact that the expected ratio of nonrecessive to recessive genotypes within a family is (3/4):(1/4) or 3:1 (Chapter 23). Unlike Mendel's sweet pea experiments, genetic traits of interest in human populations are frequently rare, and a sample of families would have to be large, usually very large, to obtain an adequate number of recessive individuals with heterozygotic parents to produce a stable estimate of the segregation ratio.

When occurrence of human disease is rare, data collection starts with locating individuals with disease and selecting appropriate controls. This case/control strategy allows the study of even extremely rare diseases. The statistical analysis, however, requires an approach that accounts for the way the data were collected.

Study of a segregation ratio of a specific genetic trait using collected sibships has a similar property. Data collection begins with identifying individuals who exhibit the potentially recessive trait (aa) whose parents are both heterozygotic (Aa), called *carriers for the recessive allele*. The estimation of the frequency of recessive individuals is no longer simply an estimate of a binomial probability. Like case/control data, the statistical analysis must account for the way the data are collected. Thus, two data collection issues, *ascertainment* and *truncation*, need to be addressed to produce an unbiased estimate of the segregation probability to be compared to the value of $p = 0.25$, the value that provides "necessary but not sufficient" evidence of a simple Mendelian inheritance pattern.

Ascertainment

To study the distribution of a genetic trait among sibs, subjects with the trait are located from lists, surveys, registries, or any high-frequency source. Such individuals are sometimes called *index cases*. In statistical language, the collection of study subjects is conditional

Table 24.1 *Illustration: Binomial Probabilities for* $n = 700$ *Family Size* $s = 4$ *and* $p = 0.25$ *and Expected Counts* ($n = 700$ *Families with a Total of* $sn = 2800$ *Offspring)*

	Probability/number of recessive *aa* offspring					
R	0	1	2	3	4	Total
p_r	q^4	$4q^3p$	$6q^2p^2$	$4qp^3$	p^4	1.0
p_r ($p = 0.25$)	0.316	0.422	0.211	0.047	0.004	1.0
n_r	221	295	148	33	3	700

Note: This kind of random sample used to estimate the frequency of a rare genetic trait is usually not practical.

on at least one recessive individual in every collected sibship, namely, the index case. The ascertainment is said to be *complete* when not only recessive individuals but also their entire sibship is identified and included in the collected data. Technically, complete ascertainment means that the probability of including all sibs within an index family is 1.0. Ascertainment remains an issue when the probability is less than 1.0. One such case arises when the probability of ascertaining a complete family is proportional to the number of recessive members in the sibship.

Truncation

Collecting subjects from lists or other relevant records causes families that could have produced a recessive offspring but did not to be excluded from the collected data. These "missing" families consist of parents who are carriers of the recessive allele without recessive offspring. The distribution of sampled individuals is then said to be *truncated* (Chapter 4). For example, when the probability of a recessive offspring is 0.25 and inheritance has a simple Mendelian pattern, then a sizeable proportion of families with heterozygotic parents consists entirely of *AA* or *Aa* offspring. These relevant families are not identified when the data collection strategy is based on identified recessive individuals (index case = *aa*). For example, for a sibship size of $s = 4$, about a third of families with heterozygotic parents would not be included (exactly $0.75^4 = 0.316$). Without accounting for truncation, estimates from index case ascertained sibships are biased (too large).

Estimation of a Segregation Probability: Complete Ascertainment

Consider an example consisting of n families of sibship size denoted s. Then Mendelian inheritance produces observations each with a binomial distribution with parameters s and $p = P(aa \mid Aa \times Aa) = 0.25$ (Chapters 4 and 23). These binomial probabilities, calculated from the expression $p_r = \binom{s}{r} p^r q^{s-r}$, illustrate a sibship $s = 4$ for $p = 0.25$ ($q = 1 - p = 0.75$ – Table 24.1); that is, the symbol p_r represents the binomial probability of r recessive individuals in a family of size s. Furthermore, the probability of an *aa*-recessive family member is directly estimated by the observed proportion of recessive offspring or $\hat{p} = R/ns = 700/2800 = 0.25$ where $R = \Sigma r n_r = 700$ is the total number of recessive offspring among the $ns = 700(4) = 2800$ collected individuals (Table 24.1).

Table 24.2 *Example: Theoretical Probabilities for Family Size* $s = 4$ *for* $p = 0.25$ *from a Truncated Binomial Distribution and Expected Counts from* $n = 700$ *Complete Ascertainment Families*

	Probability/number of recessive *aa* offspring				
R	1	2	3	4	Total
p_r	0.617	0.309	0.068	0.006	1.0
n_r	432	216	48	4	700

For data created by locating index cases where all sibships contain at least one recessive individual, the expression for these binomial probabilities becomes $p_r = \binom{s}{r} p^r q^{n-r}/(1 - q^s)$ when ascertainment is complete. Again for a sibship $s = 4$, this probability distribution is described in Table 24.2, called a *truncated binomial probability distribution* (Chapter 4). The term $1 - q^s$ represents the probability of one or more recessive offspring in a family of size s from *Aa*-genotype parents. Specifically, for $q = 0.75$, then for $s = 4$, $1 - q^s = 1 - 0.75^4 = 0.684$. Note that, as required, $\Sigma p_r = 1.0$. Probabilities estimated from a truncated binomial distribution account for the index case data collection strategy and produce unbiased estimates of p.

From the example, for family size $s = 4$, the probability an index case identified sibship contains two recessive individuals is

$$P(r = 2) = p_2 = \frac{6p^2q^2}{1 - q^4} = \frac{6(0.25)^2(0.75)^2}{1 - (0.75)^4} = \frac{0.211}{0.684} = 0.309.$$

For $n = \Sigma n_r = 700$ families, then $n_2 = np_2 = 700(0.309) = 216$ families sampled would expected to contain two recessive offspring. The total number of recessive offspring is $R = \Sigma r n_r = 1024$. Table 24.3 presents the distribution of two example truncated binomial distributions for families of sizes $s = 2$, again $s = 4$ and $p = 0.25$ illustrating exact Mendelian segregation when the data are truncated and ascertainment is complete.

As noted, sibship analysis is the comparison of a data estimated segregation probability (denoted $\hat{p}$) to the theoretical value $p = 0.25$. Therefore, the estimated segregation probability

Table 24.3 *Summary: Two Truncated Binomial Distributions of Families with Recessive Offspring* ($n = 700$ *Sibships*) *for Sibship Sizes* $s = 2$ *and* $s = 4$ *Illustrating Exact Mendelian Segregation* ($p = 0.25$)

	Number of recessive offspring					
	1	2	3	4	R^a	Total
$s = 2$	600	100	–	–	800	700
$s = 4$	432	216	48	4	1024	700
Total	1032	316	48	4	1824	1400

[a] $R = \Sigma r n_r$ = total number of recessive in a family of size s ($r = 1, 2, \ldots, s$).

is the central element of the genetic analysis. Notation to describe its estimation is the following:

R = total number of recessive offspring,
S = number of families with a single recessive offspring,
s = size of the family, and
n = number of families.

Then an estimate of the segregation probability p is

$$\text{estimated segregation probability} = \hat{p} = \frac{R - S}{ns - S}.$$

A justification for this estimate comes from a truncated binomial distribution. Noting that the observed number of recessive offspring R is an estimate of

$$R \to \frac{nsp}{1 - q^s},$$

and the observed number families with a single offspring S is an estimate of

$$S \to \frac{nspq^{s-1}}{1 - q^s}.$$

Then the expressions $R - S$ and $ns - S$ estimate

$$R - S \to \frac{nsp(1 - q^{s-1})}{1 - q^s} \quad \text{and} \quad ns - S \to \frac{ns(1 - q^{s-1})}{1 - q^s}.$$

Therefore, the ratio $(R - S)/(ns - S) = \hat{p}$ is an unbiased estimate of the segregation probability p from truncated binomial data. The estimate is similar to the estimate of the Poisson parameter λ from truncated data (Chapter 4).

From Table 24.3, when $s = 2$, $R = 800$, $S = 600$ from $n = 700$ families, then

$$\hat{p} = \frac{R - S}{ns - S} = \frac{800 - 600}{1400 - 600} = 0.25.$$

For sibship size $s = 4$, similarly $R = 1024$, $S = 432$, and again from $n = 700$ families, then

$$\hat{p} = \frac{1024 - 432}{2800 - 432} = 0.25.$$

Of particular importance, this same expression for the estimate of p can be used to estimate the segregation probability from all sibships combined. Continuing the example data (Table 24.3, last row), combining sibship sizes $s = 2$ and $s = 4$ gives $R = 800 + 1024 = 1824$ and $S = 600 + 432 = 1032$, making the estimated segregation probability

$$\hat{p} = \frac{1824 - 1032}{4200 - 1032} = 0.25,$$

where $1400 + 2800 = 4200$ is total number of collected offspring.

An alternative estimate of the segregation probability is achieved by solving the equation

$$R = \frac{nsp}{1 - q^s}$$

Table 24.4 *Data/Estimates: 14 Sibships (Sizes = s) Containing at Least One Case of Human Albinism to Estimate the Segregation Probability*

	Data										Estimation					
s	1	2	3	4	5	6	7	8	9	10	R	S	n	ns	$\hat{p}$	$\hat{p}'$
2	31	9	–	–	–	–	–	–	–	–	49	31	40	80	0.367	0.367
3	37	15	3	–	–	–	–	–	–	–	76	37	55	165	0.305	0.308
4	22	21	7	0	–	–	–	–	–	–	85	22	50	200	0.354	0.348
5	25	23	10	1	1	–	–	–	–	–	110	25	60	300	0.309	0.309
6	18	13	18	3	0	1	–	–	–	–	116	18	53	318	0.327	0.333
7	16	10	14	5	1	0	0	–	–	–	103	16	46	322	0.284	0.290
8	4	8	7	6	1	0	1	0	–	–	77	4	27	216	0.344	0.344
9	10	4	9	4	1	0	1	0	0	–	73	10	29	261	0.251	0.261
10	6	3	7	2	1	1	0	0	0	0	52	6	20	200	0.237	0.244
11	0	2	4	6	2	0	0	0	0	0	50	0	14	154	0.325	0.320
12	2	0	0	4	2	0	0	0	0	0	28	2	8	96	0.277	0.287
13	0	0	1	1	1	0	1	0	0	0	19	0	4	52	0.366	0.364
14	0	1	1	1	0	0	1	0	0	0	16	0	4	56	0.286	0.283
15	0	0	0	0	0	0	0	0	0	1	10	0	1	15	0.667	0.500

Note: $\hat{p}$ represents the singles removed estimate, and $\hat{p}'$ represents the maximum likelihood estimate of the segregation probability.

for the value $p = 1 - q$. This expression is created by setting the observed number of recessive individuals (R) equal to the theoretical number expected from a truncated binomial distribution. The estimate $\hat{p}$ is the solution to this equation and is the maximum likelihood estimate of the segregation probability p. The solution requires an iterative numerical process. For the example, the solution for the "perfect data," where $R = 800$, $s = 2$, and $n = 700$, is $\hat{p} = 0.25$ (Table 24.3). Incidentally, the singles removed method and the maximum likelihood estimate are identical when family size $s = 2$, otherwise the maximum likelihood estimate is more precise (Chapters 17 and 23).

Classic and extensive data collected by the famous biologist J. B. Haldane (b. 1892) allow the estimation of segregation probabilities associated with the trait of human albinism. Fourteen sibships sizes (s) each yield both kinds of estimates of the segregation probability assuming complete ascertainment (Table 24.4 – last two columns).

The estimate from the entire data set directly follows from the total number of recessive genotypes R_0, the total number families with one recessive offspring S_0, and the total number of individuals in the collected data set N_0. Specifically the summary values $R_0 = 864$, $S_0 = 171$, and $N_0 = 2435$ yield an estimated segregation probability

$$\hat{P} = \frac{R_0 - S_0}{N_0 - S_0} = \frac{864 - 171}{2435 - 171} = 0.306.$$

A weighted average segregation ratio of the maximum likelihood estimates (denoted $\hat{p}'$) produces an overall estimate of $\bar{p} = 0.309$ where the weights are the total number of members in each sibship (ns – Table 24.4).

Table 24.5 *Example: Properties of Single Ascertainment for sibship of Size* $s = 5$ *for a Segregation Probability Exactly* $p = 0.25$ *Based on* $n = 1024$ *Family Members*

	Probability/Number Recessive Family Members						
R	0	1	2	3	4	5	Total
p_r	q^5	$5pq^4$	$10p^2q^3$	$10p^3q^2$	$5p^4q$	p^5	1.0
p_r $(p = 0.25)$	0.237	0.396	0.264	0.088	0.015	0.001	1.0
np_r	243	405	270	90	15	1	1024
$w_r = r/(sp)$	$0/(5p)$	$1/(5p)$	$2/(5p)$	$3/(5p)$	$4/(5p)$	$5/(5p)$	–
$w_r \times p_r$	0	q^4	$4pq^3$	$6p^2q^2$	$4p^3q$	p^4	1.0

Estimation of a Segregation Probability: Single Ascertainment

Unbiased estimates of a segregation probability from incomplete ascertained data require the family size to be reduced by one member resulting in a complete binomial probability distribution that is then used to estimate the probability of recessive offspring. In this case, the term *single ascertainment* frequently is used. Technically, unbiased estimated probabilities from a sibship of size s result from applying weights proportional to the probability of ascertaining a family with r recessive offspring. The weights are $w_r = \frac{r}{sp}$ making the probability of r recessive offspring from a family of size s

$$p_r = \frac{r}{sp}\left[\binom{s}{r} p^r q^{s-r}\right].$$

This weighting results in a binomial probability of r recessive offspring from single ascertained families of

$$p_r = \binom{s-1}{r-1} p^{r-1} q^{(s-1)-(r-1)}.$$

Thus, accounting for the way the data were collected, the analysis becomes an application of a complete binomial distribution. The unbiased estimated probability of a recessive offspring directly follows from the number of recessive individuals divided by the number of individuals collected but based on a family size reduced by one or "sibship size" $= s - 1$ (denoted $\tilde{s}$).

Tables 24.5 and 24.6 present a detailed illustration of single ascertainment for a sibship of size $s = 5$ and $p = 0.25$.

Table 24.6 *Example: Properties of the Binomial Distribution for a Sibship of Size* $s = 5$ *Adjusted for Single Ascertainment Sampling* ($p = 0.25$ *and* $\tilde{s} = 4$)

	Probability/Number Recessive Family Members					
R	0	1	2	3	4	Total
$\tilde{p}_r$	q^4	$4pq^3$	$6p^2q^2$	$4p^3q$	p^4	1.0
$\tilde{p}_r$	0.316	0.422	0.211	0.045	0.004	1.0
$n\tilde{p}_r$	324	432	216	48	4	1024

Note: $\tilde{p}_r = \binom{\tilde{s}}{r} p^r(1-p)^{\tilde{s}-r}$ where $\tilde{s} = s - 1 = 5 - 1 = 4$ and $r = 0, 1, 2, 3,$ and 4.

Table 24.7 *Illustration: "Perfect" Data Illustrating Single Ascertainment of a Recessive Trait* ($p = 0.25$)

		Number of Recessive Family Members						
s	$\tilde{s}$	0	1	2	3	4	5	n
3	2	9	6	1	–	–	–	16
4	3	27	27	9	1	–	–	64
5	4	81	108	54	12	1	–	256
6	5	243	405	270	90	15	1	1024

Table 24.7 *(Continued) Summary: Values for Estimation of the Probability of Recessive Individual from All Five Sibships*

	Number of Recessive Family Members			
$\tilde{s}$	R	n	$n\tilde{s}$	$\hat{p} = R/n\tilde{s}$
2	8	16	32	0.25
3	48	64	192	0.25
4	256	256	1024	0.25
5	1280	1024	5120	0.25
Total	1592	1360	6368	0.25

Table 24.8 *Data: Single Ascertainment Using the Sibships Containing Human Albinism to Estimate the Segregation Ratio from Haldane's Original Data* ($\tilde{s} = s - 1$ – *Table 24.4)*

	Single Ascertainment			
$\tilde{s}$	R	n	$n\tilde{s}$	$\hat{p}$
1	9	40	40	0.225
2	21	55	110	0.191
3	35	50	150	0.233
4	50	60	240	0.208
5	63	53	265	0.238
6	57	46	276	0.207
7	50	27	189	0.265
8	44	29	232	0.190
9	32	20	180	0.178
10	36	14	140	0.257
11	20	8	88	0.227
12	15	4	48	0.312
13	12	4	52	0.231
14	9	1	14	0.643
Total	453	411	2024	0.224

The binomial distribution–based estimate of the segregation probability is routinely the number of recessive sibs (R) divided by the total number of sampled sibs ($n\tilde{s}$). The estimated segregation probability is $\hat{p} = R/n\tilde{s}$. As noted, when the data are generated from single ascertainment, the unbiased estimate of p is the usual binomial estimate ($\hat{p} = R/n\tilde{s}$) but with the family size s reduced by one (Chapter 4). Specifically, the estimate, from a sibship of size $s = 5$, using $\tilde{s} = 4$, the estimated probability of a recessive offspring becomes $\hat{p} = R/(n\tilde{s}) = 1024/(1024 \times 4) = 0.25$, where $R = \Sigma r p_r = 1024$ ($r = 1$, 2, 3, and 4 – Table 24.6).

Parallel to the estimate in the complete ascertainment case, the number of recessive offspring and the number of individuals from each sibship can be combined to estimate the segregation probability from the entire data set. Table 24.7 displays artificial single ascertained data with "perfect" segregation probabilities (0.25) for family sizes $s = 3$ to $s = 6$.

Four segregation probabilities are estimated from a binomial distribution or, specifically, $\hat{p} = R/n\tilde{s}$ because trait ascertainment is proportional to the number of number of recessive individuals r within a family. For example, a sibship $s = 6$ yields $\hat{p} = R/n\tilde{s} = 1280/5120 = 0.25$ from an adjusted binomial distribution based on a "family size" of $\tilde{s} = 5$ (Table 24.7). Overall, the total number of recessive individuals is 1592 from a total of 6368 collected offspring directly producing the estimated segregation probability of $\hat{p} = 1592/6368 = 0.25$ from the entire "data" set (Table 24.7, last row).

Treating the albinism data (Table 24.4) as single ascertained samples produces alternative estimated segregation probabilities ($\hat{p}$ – Table 24.8). The data are the same as in Table 24.4 but are treated as 13 binomial distributions with reduced family sizes $s - 1 = \tilde{s}$. The last line provides an estimated segregation ratio $\hat{p}$ for single ascertainment from the entire data set of $\hat{p} = 453/2024 = 0.224$. For the complete ascertainment case, the corresponding estimated segregation probability is $\hat{P} = 0.309$.

25

Admixed Populations

Confounding, admixture, disequilibrium, and population stratification are different terms applied to describe diversity in genetic data. All four terms refer to the fact that the relationships within a single population are often not the same as those found within separate strata of the same population. In the following discussion, the term *admixture* describes the distribution of alleles that results when members of two or more populations interbreed. For example, the current African American gene pool is a mixture of original African alleles and alleles that originate from white populations. Such populations are said to be *admixed.* Or, sometimes, this lack of genetic homogeneity is called *confounding* or *population stratification* due to ethnicity and is occasionally an issue in studies of the role of genetics in disease and mortality.

Consider an example. In populations labeled *I* and *II*, the frequencies of two alleles, denoted *A* and *B*, in population *I* are

$$\text{population I: } P_A^{(I)} = 0.6 \quad \text{and} \quad P_B^{(I)} = 0.2,$$

and in population *II*, the frequencies of the same two alleles are

$$\text{population II: } P_A^{(II)} = 0.7 \quad \text{and} \quad P_B^{(II)} = 0.9.$$

Under the conditions that such factors as mutations, genetic drift, and selection within random mating populations have inconsequential influences, the frequency of the *AB*-genotype is the product of the *A* and *B* allele frequencies (Chapter 23). The *AB*-frequencies would be $P_{AB}^{(I)} = (0.6)(0.2) = 0.12$ in population *I* and $P_{AB}^{(II)} = (0.7)(0.9) = 0.63$ in population *II.*

A definition of a mixture is necessary to estimate and describe genetic admixture. An admixed proportion (denoted p_a) is a special case of the mixture of two proportions when

$$p_a = Mp_1 + (1 - M)p_0 = p_0 + M(p_1 - p_0),$$

where M represents the rate of mixture of the frequencies p_0 and p_1. In a different form, the expression for the rate of mixture is

$$M = \frac{p_a - p_0}{p_1 - p_0}.$$

Geometrically the rate M is the proportion of the distance from p_0 to p_a relative to the distance from p_0 to p_1 (Figure 25.1). An admixture of populations *I* and *II* yields frequencies of alleles *A* and *B* that is a result of this relationship.

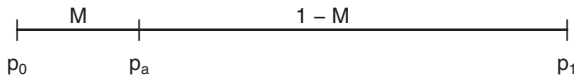


Figure 25.1 Mixture Rate M within Interval p_0 to p_1

For the example, when the populations I and II are admixed at a rate represented by M, the allele frequencies become

$$P_A = Mp_A^{(I)} + (1-M)p_A^{(II)} = 0.2(0.6) + 0.8(0.7) = 0.68 \quad \text{and}$$
$$P_B = Mp_B^{(I)} + (1-M)p_B^{(II)} = 0.2(0.2) + 0.8(0.9) = 0.76, \quad \text{rate} = M = 0.2$$

in the next generation. Therefore, the frequency of the *A/B*-genotype in the admixed population becomes

$$P'_{AB} = P_A P_B = (0.68)(0.76) = 0.517.$$

When no admixture occurs, then the frequency of the *A/B*-genotype in the combined population, directly calculated from the population specific allele frequencies, is

$$\begin{aligned} P_{AB} &= MP_{AB}^{(I)} + (1-M)P_{AB}^{(II)} = Mp_A^{(I)}p_B^{(I)} + (1-M)p_A^{(II)}p_B^{(II)} \\ &= 0.2(0.12) + 0.8(0.63) = 0.528. \end{aligned}$$

That is, the allele frequencies of A and B in each population do not change from generation to generation (Hardy-Weinberg equilibrium), and the frequency within the combined population of the *A/B*-genotype remains $P_{AB} = 0.528$. The difference in *A/B*-genotype frequencies between the admixed and simply mixed populations is $P'_{AB} - P_{AB} = 0.517 - 0.528 = -0.011$ reflecting the influence of one generation of admixture. The magnitude of this difference in *A/B*-allele frequencies is

$$\text{difference} = P'_{AB} - P_{AB} = M(1-M)\left(p_A^{(I)} - p_A^{(II)}\right)\left(p_B^{(I)} - p_B^{(II)}\right),$$

and, from the example, again is

$$\text{difference} = 0.517 - 0.528 = 0.2(1-0.2)(0.6-0.7)(0.2-0.9) = -0.011.$$

Two properties of genetic admixture are clear from this expression, namely,

1. If either $M = 1$ or $M = 0$, no admixture occurs and
2. If either $p_A^{(I)} = p_A^{(II)}$ or $p_B^{(I)} = p_B^{(II)}$, no admixture occurs.

The difference $P'_{AB} - P_{AB}$ measures a specific kind of genetic disequilibrium and, as noted, can be an important consideration in the analysis of genetic data. The influence of admixture is small when either $p_A^{(I)} - p_B^{(II)}$ or $p_B^{(I)} - p_B^{(II)}$ is small. In fact, when both differences are moderately large the product is considerably reduced, producing less admixture influence than might be expected.

A Model Describing Admixture

A model describing the pattern of change in frequency of a single allele from continuous admixture over generations provides a simple example of the dynamics of the admixture

Table 25.1 *Example: Admixture between a Population with Allele Frequency* $P = 0.8$ *and Another Population with Original Allele Frequency* $p_0 = 0.1$ *at Rate* $M = 0.1$, 80 *Generations*

Generation 1	$p_1 = (1 - M)\,p_0 + MP = 0.9(0.100) + 0.1(0.8) = 0.170$
Generation 2	$p_2 = (1 - M)\,p_1 + MP = 0.9(0.170) + 0.1(0.8) = 0.233$
Generation 3	$p_3 = (1 - M)\,p_2 + MP = 0.9(0.233) + 0.1(0.8) = 0.290$
Generation 4	$p_4 = (1 - M)\,p_3 + MP = 0.9(0.290) + 0.1(0.8) = 0.341$
Generation 5	$p_5 = (1 - M)\,p_4 + MP = 0.9(0.341) + 0.1(0.8) = 0.387$
– – –	– – – – – – – – – –
– – –	– – – – – – – – – –
Generation 10	$p_{10} = (1 - M)\,p_9 + MP = 0.9(0.529) + 0.1(0.8) = 0.556$
– – –	– – – – – – – – – –
– – –	– – – – – – – – – –
Generation 80	$p_{80} = P \approx 0.8$

process. The model traces an allele frequency over k nonoverlapping generations of unidirectional admixture of an original allele frequency in one population (denoted p_0) with a second population with a different allele frequency (denoted P). The symbol M represents a unidirectional rate of allele flow from the second population, causing the original population to become admixed. It is assumed that the value M remains constant and the allele frequency P does not change during a sequence of discrete nonoverlapping generations.

To start, when a parent population with an allele frequency $P = 0.8$ is admixed with another population with the original frequency of the same allele of $p_0 = 0.1$, then the allele frequency in the first generation increases from $p_0 = 0.1$ to

$$\text{generation 1: } p_1 = (1 - M)p_0 + MP = 0.9(0.1) + 0.1(0.8) = 0.170,$$

where $M = 0.1$ is the admixture rate. Thus, the new frequency of the allele p in the admixed population increases $100 \times (0.17 - 0.10)/0.17 = 41\%$ after the first generation. The admixed allele frequency similarly increases in the next generation. The frequency of p in subsequent generations is always larger than the value from the previous generation (Table 25.1). This change in the allele frequency is displayed for 35 generations in Figure 25.2.

In general, the allele frequency in the kth generation is

$$\text{generation } k\text{: } p_k = (1 - M)^k p_0 + [1 - (1 - M)^k]P.$$

For the example, the 20th generation ($k = 20$) allele frequency p_{20} in the admixed population is

$$\begin{aligned}\text{generation 20: } p_{20} &= (1 - M)^{20} p_0 + [1 - (1 - M)^{20}]P \\ &= (1 - 0.1)^{20}(0.1) + [1 - (1 - 0.1)^{20}](0.8) = 0.715.\end{aligned}$$

As the number of generations increases, the allele frequency p approaches the allele frequency P. Ultimately the original allele frequency $p_0 = 0.1$ becomes $p = P = 0.8$ (Figure 25.2). Technically, the allele frequency p becomes P because the term $(1 - M)^k$ eventually becomes

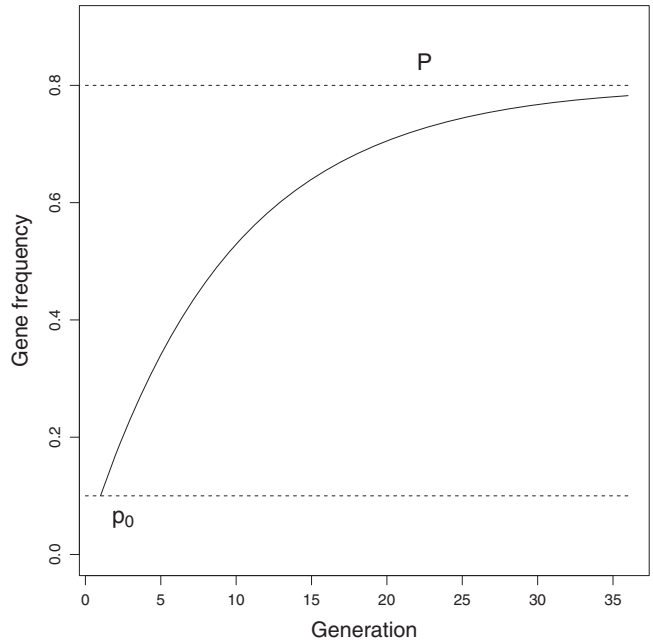


Figure 25.2 Increase in Allele Frequency from Admixture (Generation 0: $p_0 = 0.1$, $P = 0.8$, and $M = 0.1$)

indistinguishable from zero for large values of k (generations). Another version of the general relationship between admixed allele frequencies and passage of time (k generations) is

$$(1 - M)^k = \frac{P - p_k}{P - p_0}.$$

Thus, after $k = 20$ generations, a measure of the level of admixture is

$$\frac{p_k - p_0}{P - p_0} = 1 - (1 - M)^{20} = 1 - (0.9)^{20} = 0.878.$$

The gene pool of the admixed population then consists of 12% of the original alleles and 88% of alleles from the parent population due to continuing admixture. As noted, for large values of k, then $(1 - M)^k \approx 0$, and $p_k = P \approx 0.8$ (Figure 25.2).

Typically, in more realistic situations, the number of generation k is not known. Furthermore, nonoverlapping generations do not occur in human populations unlike some animal and annual plant populations. Therefore, the model produces a rigorous demonstration of admixture that is not a relevant description for human populations.

Estimation of the Admixture Rate

Analogous to the previous admixture model rate M, a definition that usefully applies to human populations is

$$m = \frac{P - p_a}{P - p},$$

where p_a represents allele frequency in the admixed population, P the allele frequency in a parent population, and p the allele frequency in a second relevant population. The admixture rate $(1 - M)^k$ from the unidirectional model is replaced by a general summary rate represent by m. The admixture rate of m can be estimated from allele frequencies P, p, and p_a.

An Example: Estimation of the Extent of Admixture

An estimate of admixture (denoted m) between U.S. white and African American populations requires an estimate of the frequency of a specific allele from a genetic locus such as the *Rh*-system allele *cDe* from a white parent population (denoted p) and an estimate of the frequency of the same allele from an African American population (denoted p_a). Both values estimated from a sample of residents of South Carolina (whites = 2090 and African Americans = 1287) are $\hat{p} = 0.037$ and $\hat{p}_a = 0.535$. The original population allele frequency of the *cDe*-allele estimated from a black West African population is $\hat{P} = 0.605$, based on a sample of 2400 individuals. This specific *Rh*-allele is chosen because of the large difference between the parent West African population and the American white population allele frequencies ($difference = \delta = \hat{P} - \hat{p} = 0.605 - 0.037 = 0.568$). This difference is sometimes called the "δ-distance." Larger δ-distances produce more precise estimates of the admixture rate (m). The estimated black/white *Rh*-allele admixture rate is

$$\hat{m} = \frac{\hat{P} - \hat{p}_a}{\hat{P} - \hat{p}} = \frac{0.605 - 0.535}{0.605 - 0.037} = 0.123;$$

that is, about 12% of the *Rh*-allele frequency in the admixed African American population originates from the white population.

An estimate of the variance of the distribution of this estimated value is derived from binomial probability distributions where the variance of the estimated allele frequency is given by the expression $variance\ (\hat{p}) = p(1 - p)/2n$ when p represents an allele frequency and n the number of sampled individuals (Chapter 4). For the *Rh*-allele data, the binomial distribution estimated standard errors are $se(\hat{P}) = 0.017$ (West African), $se(\hat{p}) = 0.004$ (whites), and $se(\hat{p}_a) = 0.014$ (African Americans). The estimated variance of the distribution of the estimated admixture rate such as $\hat{m} = 0.123$ is approximately

$$variance(\hat{m}) = \frac{variance(\hat{p}_a) + m^2 variance(\hat{p}) + (1 - m)^2 variance(\hat{P})}{(\hat{P} - \hat{p})^2},$$

where each sample size, as usual, plays an important role in the precision associated with the estimate. Because the squared δ-distance $\hat{P} - \hat{p}$ is the denominator of the estimate of the variance of $\hat{m}$, it too substantially influences the precision of the admixture estimate.

For the estimate $\hat{m} = 0.123$, the estimated variance for the example data is $variance(\hat{m}) = 0.00084$ and yields an approximate 95% confidence interval of $0.123 \pm 1.960(0.030) \rightarrow (0.066, 0.179)$.

Data collected to estimate the rate of admixture in African Americans from three populations (South Carolina, Detroit, and Oakland, CA) are displayed in Table 25.2. The three admixed alleles frequencies ($\hat{p}_a$) are from a two-allele red blood cell genetic determined trait called the Duffy system (traditionally denoted Fy^a and Fy^b) (Chapter 27). This simple Mendelian inherited trait is particularly effective for estimating the admixture proportion m

Table 25.2 *Allele Frequencies: Three African American Populations Used to Estimate Level of Admixture Rate* (m) *Based on the Duffy Red Blood Cell System* ($\delta = 0.429$)

Location	$\hat{p}_a$	n	$\hat{m}$	*s. e.*	Weights (w)
South Carolina	0.045	304	0.106	0.020	2567.3
Detroit	0.111	404	0.260	0.026	1479.6
Oakland (CA)	0.094	3146	0.220	0.009	12481.4

because the frequency among black Africans is essentially zero ($\hat{P} = 0$). In addition, the white population in Oakland (CA) serves to estimate the allele frequency of the white gene pool (p). The white Duffy allele frequency estimate is $\hat{p} = 0.429$ from $n = 5064$ white individuals making the δ-distance $= |\hat{P} - p| = 0.429$. The admixture rates $\hat{m}$ estimated from the three populations are presented in Table 25.2.

Assuming estimates from these three locations differ only because of sampling variation, a single summary value based on combining individual estimates is unbiased and certainly more precise. A weighted average, using as weights the reciprocal of the estimated variances, produces a summary estimate (Chapter 3). Specifically, the summary estimated admixture rate is $\bar{m} = \Sigma w_i \hat{m}_i / \Sigma w_i = 0.205$, where $i = 1$, 2, and 3 populations (Table 25.2). The variance of a weighted average, when the weights are reciprocal values of the estimated variances, is the reciprocal of the sum of the weights (Chapter 3). For the Duffy system allele data (Table 25.2), the estimated standard error is

$$s.e. = \sqrt{variance(\bar{m})} = \frac{1}{\sqrt{\Sigma w_i}} = \frac{1}{\sqrt{16,\,528.2}} = 0.0078 \text{ where } w_i = \frac{1}{variance(\hat{m}_i)},$$

yielding an approximate 95% confidence interval of

$$\bar{m} \pm 1.960\sqrt{variance(\bar{m})} = 0.205 \pm 1.960(0.0078) \rightarrow (0.190, 0.221)$$

based on the normal distribution and the estimated summary admixture rate $\bar{m} = 0.205$.

The estimation of an admixture rate associated with a single population can be extended to three or more parent populations. To illustrate, Table 25.3 contains admixed frequencies of the *ABO*-alleles from the ABO-blood group in a northern Brazil population (Nordestinos) and three parent populations (African, Indian, and Portuguese). A simple expression for admixture and ordinary least squares estimation produce estimates of the three admixture

Table 25.3 *Allele Frequencies: Four Ethnic Groups Illustrate Estimation of Three Rates of Admixture Denoted* m_1, m_2, *and* m_3

Allele	Admixed ($\hat{p}_a$)	African ($\hat{p}_1$)	Indian ($\hat{p}_2$)	Portuguese ($\hat{p}_3$)
A_0	0.167	0.097	0.0	0.236
A_1	0.043	0.081	0.0	0.068
B	0.083	0.114	0.0	0.066
O	0.707	0.708	1.0	0.630
n	295	858	1622	3048

rates [denoted: m_1 (African), m_2 (Indian), and m_3 (Portuguese)]. The admixture relationship is

$$p_a = m_1 p_1 + m_2 p_2 + m_3 p_3,$$

where m_1, m_2, and m_3 represent the admixture rates to be estimated; that is, the allele admixed frequency (p_a) is the dependent variable, the three allele frequencies (p_i) are the independent variables, and the coefficients (m_i) are the admixture rates combined to form a linear multivariable regression equation. In addition, the estimated coefficients m_i are required to add to 1.0. The regression equation can be directly modified to yield estimates with this property. The three-variable regression equation

$$p_a = m_1 p_1 + m_2 p_2 + m_3 p_3$$

becomes the two-variable regression equation

$$p_a = m_1 p_1 + m_2 p_2 + (1 - m_1 - m_2) p_3$$

when $1 - m_1 - m_2$ replaces m_3 and then

$$p_a - p_3 = m_1(p_1 - p_3) + m_2(p_2 - p_3) \quad \text{or} \quad P_a = m_1 P_1 + m_2 P_2.$$

The two estimated admixture rates $\hat{m}_1$ and $\hat{m}_2$ follow from applying least squares estimation to the two-variable regression equation. The third admixture rate m_3 is then estimated by $\hat{m}_3 = 1 - \hat{m}_1 - \hat{m}_2$. For the data in Table 25.3, the least squares estimates of the two admixture rates are $\hat{m}_1 = 0.274$ and $\hat{m}_2 = 0.146$ making the third estimated value $\hat{m}_3 = 1 - 0.274 - 0.146 = 0.579$ so that $\hat{m}_1 + \hat{m}_2 + \hat{m}_3 = 1$.

Note that the mechanical process of least squares estimation of the regression coefficients is assumption-free. The more usual application of least squares estimation that produces estimates of the variance and accompanying inferences require a statistical structure (Chapters 3 and 7).

26

Nonrandom Mating

Present-day geneticists use extremely advance molecular methods to explore the properties of thousands of genes, sometimes tens of thousands of genes. The human genome has been completely described at a molecular level as well as the genomes of numerous animals, bacteria, and viruses. Before the availability of molecular techniques, most genetic knowledge was based on experiments and observations. The first notable experiments were Gregor Mendel's (b. 1822) breeding of sweet peas later followed by extensive experimental use of *Drosophila melanogaster* (a fruit fly). In human genetics, observations made on traits such as eye color, albinism, and sickle cell anemia led to important contributions to genetic knowledge. The dynamics of genetic inheritance in human populations, however, cannot be studied with molecular techniques, experiments, or direct observations. A major source of genetic theory at the population level comes from statistical concepts, such as variance, covariance, and correlation, applied to mathematical/genetic models (Chapter 27). A starting point of this approach was the Hardy-Weinberg (1908) description of the consequences of random mating (Chapter 23). Since then the results of random and nonrandom mating have been described and studied with statistical tools from a large variety of perspectives. The following descriptions of statistical thinking applied to a few issues that arise in nonrandom mating populations give a sense of this use of statistical methods. The property that nonrandom mating decreases population diversity illustrates several ways that statistical arguments describe the forces that influence change in the patterns of genetic variation in human populations. Topics described are correlation between alleles, Wahlund's model of subgroup variation, consequences of random genetic drift, balance between selection and mutation, and two basic properties of assortative mating. As they say, "This is only the tip of the iceberg."

Genotype Frequencies – Correlation and Variance

Genotypes AA, Aa, and aa are made up of two alleles. The population genetics of these three genotypes can be viewed statistically as the properties of two binary variables. Specifically the binary outcome of allele A that occurs with frequency $P(X = 1) = p$ and an allele a that occurs with frequency $P(X = 0) = q = 1 - p$. The statistical properties of the distribution of genotypes that result from the joint occurrence of these not necessarily independent binary variables denoted X and X' are described in Tables 26.1 and 26.2.

The results of the joint occurrence of two binary variables X and X' are statistically described in a number of ways (Table 26.2). A value denoted F has several interpretations

Table 26.1 *Notation: Joint Distribution of Alleles A and a That Generate Frequencies of Genotypes AA, Aa, and aa*

		Distribution X	
		$X = 1$	$X = 0$
Distribution X'	$X' = 1$	AA	Aa
	$X' = 0$	Aa	Aa

in this genetic context. Some examples are that the value of F indicates the degree of population admixture, reflects the extent of inbreeding, plays a role in genetic variation, and indicates the frequency of heterozygotic Aa-individuals in the make-up of a population. In pure statistical terms, the symbol F represents the Pearson product-moment correlation coefficient measuring the association between the two binary variables X and X' (Chapter 5). In symbols,

$$0 \leq correlation(X, X') \leq 1 \text{ or } 0 \leq F \leq 1.$$

For example, when alleles A and a are independently distributed in a gene pool, then $F = 0$.

Figure 26.1 displays the influence of selected values of F on the frequency of heterozygotic individuals for the range of A-allele frequencies p. The expressions in Table 26.2 indicate that the highest possible frequency of heterozygotic individuals occurs when F is zero. In general, because $(1 - F) \leq 1$, then $H = 2\,pq(1 - F) \leq 2\,pq$.

Table 26.2 *Joint Distributions: Four Expressions of Distribution of Genotypes AA, Aa, and aa or, in Statistical Terms, Joint Distribution of Alleles A and a*

	1	2	3[a]	4
AA	$p^2 + Fpq$	$pF + p^2(1 - F)$	$p - \frac{1}{2}H$	$p - pq(1 - F)$
Aa	$2pq - 2pqF$	$2pq(1 - F)$	H	$2pq(1 - F)$
aa	$q^2 + Fpq$	$qF + q^2(1 - F)$	$q - \frac{1}{2}H$	$q - pq(1 - F)$
Total	1.0	1.0	1.0	1.0

[a] H represents the frequency of heterozygotic individuals (Aa).

Definitions:

Expression 1.0 displays F as a measure of deviation from a random distribution of allele frequencies.

Expression 2.0 displays F as the frequency of pairs of identical alleles and $1 - F$ as the frequency of random pairs of alleles.

Expression 3.0 displays the role of heterozygotic genotypes (denoted H) in the joint distribution.

Expression 4.0 displays F as a measure of the deviation from fixed allele frequencies p and q.

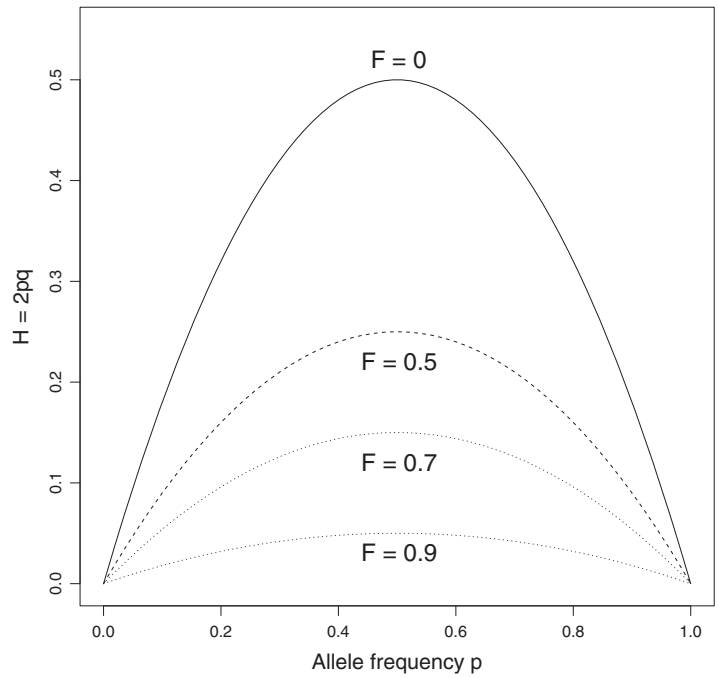


Figure 26.1 Distributions of Heterozygotic Genotypes for Selected Values of F over the Entire Range of A-Allele Frequency p

The notation for data to estimate genetic parameters p and F is

	Genotypes			
	AA	*Aa*	*aa*	Total
Counts	*a*	*b*	*c*	*n*

and yields estimates of the frequency of allele A (p) and correlation (F) (Chapters 4 and 27) where

$$\hat{p} = \frac{2a + b}{2n} \quad \text{and} \quad \hat{F} = 1 - \frac{b}{2n\hat{p}\hat{q}}.$$

The respective estimated variances are

$$variance(\hat{p}) = \frac{\hat{p}\hat{q}(1 + \hat{F})}{2n} \quad \text{and}$$

$$variance(\hat{F}) = (1 - \hat{F})\frac{1 - 2\hat{p}\hat{q}(1 - \hat{F}) - (1 - 4\hat{p}\hat{q})(1 - \hat{F})^2}{2n\hat{p}\hat{q}}.$$

Table 26.3 *Data: Artificial Sample from Genotype Distribution in Terms of Two Binary Variables X and X′ (n = 30 Individuals) Describing Genotypes AA, Aa, and Aa*

	AA	*Aa*	*aa*
X	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 0 0	0 0 0 0 0 0 0 0 0 0
X′	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	0 0 1 1	0 0 0 0 0 0 0 0 0 0

Table 26.3 *(Continued) Data: Genotype Distribution as Counts*

	Genotypes			
	AA	*Aa*	*aa*	Total
Symbols	*a*	*b*	*c*	*n*
Counts	17	4	9	30

A popular alternative version of the estimate of the correlation coefficient F based directly on counts of genotypes (Chapter 6) is

$$\hat{F} = \frac{ac - b^2/4}{(a + b/2)(c + b/2)}.$$

Notice that when $F = 0$, the estimated variance of the distribution of $\hat{F}$ is $variance(\hat{F}) = 1/n$. Thus, a test statistic to assess the statistical hypothesis that $F = 0$ (random mating only) is

$$X^2 = \left[\frac{\hat{F} - 0}{1/\sqrt{n}}\right]^2 = n\hat{F}^2$$

and has an approximate chi-square distribution (degrees of freedom = 1) when $F = 0$ (Chapter 6).

As noted, genotype data are a special case of a joint distribution of two binary distributed random variables coded A-allele = 1 and a-allele = 0. For example, the Aa-genotype would be represented as $X = 1$ and $X' = 0$ or $X = 0$ and $X' = 1$ (Table 26.1). A small artificial set of two such identically distributed binary variables produces an illustration of the Pearson product-moment correlation coefficient (labeled F) applied to measure the association within allele pairs (Table 26.3). Direct estimation of the sample mean value from the $n = 30$ binary variables X and X' yields

$$\text{mean of } x = \bar{x} = \text{mean of } x' = \bar{x}' = \hat{p} = (34 + 4)/60 = 38/60 = 0.633.$$

The sample variance of X and X' is

$$S_X^2 = S_{X'}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 = \hat{p}(1 - \hat{p}) = \hat{p}\hat{q} = 0.633(0.367) = 0.232,$$

and the estimated sample covariance is

$$\text{sample covariance} = S_{XX'} = \frac{1}{n}\Sigma(x_i - \bar{x})(x_i' - \bar{x}') = \hat{p}\hat{q}\hat{F} = 0.166.$$

Then the Pearson product-moment correlation coefficient measuring the association between X and X' is (Chapter 5)

$$\hat{F} = \text{correlation of } x \text{ and } x' = \textit{correlation}(x, x') = \frac{S_{XX'}}{\sqrt{S_X^2 S_{X'}^2}} = \frac{0.166}{0.232} = 0.713.$$

Alternatively, the estimation of the parameter F from a table of genotype counts yields the identical value. Using the more common summary count data (Table 26.3), the identical values are

$$\hat{p} = \frac{2a + b}{2n} = \frac{38}{60} = 0.633$$

and

$$\text{correlation} = \hat{F} = 1 - \frac{b}{2n\hat{p}\hat{q}} = 1 - \frac{4}{2(30)(0.633)(0.367)} = 0.713.$$

Genetic Variance

The allele frequency p is the key component of genetic variability, and the correlation coefficient F, as might be expected, plays an important role. To rigorously calculate and interpret the genetic variance, note that, in general (Chapter 27),

$$\textit{variance}(X + Y) = \textit{variance}(X) + \textit{variance}(Y) + 2\textit{covariance}(XY).$$

Then, specifically the variance of the probability $p_1 + p_2$ is

$$\begin{aligned}\textit{variance}(p_1 + p_2) &= \frac{1}{n}[(p_1 + p_2)(1 - [p_1 + p_2])] \\ &= \frac{1}{n}[p_1(1 - p_1) + p_2(1 - p_2) - 2p_1p_2] \\ &= \textit{variance}(p_1) + \textit{variance}(p_2) - 2\textit{covariance}(p_1, p_2),\end{aligned}$$

making the expression for the $\textit{covariance}(p_1, p_2) = -\frac{p_1 p_2}{n}$.

Genotypic frequencies represented as three proportions (denoted $P_i - i = 1, 2$, and 3) are the following:

	Data			
	AA	*Aa*	*aa*	Total
Probabilities	P_1	P_2	P_3	1.0

The frequency of allele A is $p = P_1 + \frac{1}{2} P_2$. Properties of the genetic variance of the allele frequency A follow from the expression for the $\textit{variance}(p)$ (Chapters 23 and 27).

Specifically, the *variance*(p) is

$$\begin{aligned} variance\left(P_1 + \frac{1}{2}P_2\right) &= variance(P_1) + \frac{1}{4}variance(P_2) + covariance(P_1, P_2) \\ &= \frac{1}{n}\left[P_1(1-P_1) + \frac{1}{4}P_2(1-P_2) - P_1P_2\right] \\ &= \frac{1}{n}\left[P_1 + \frac{1}{4}P_2 - \left(P_1 + \frac{1}{2}P_2\right)^2\right] = \frac{1}{n}\left[p - \frac{1}{4}P_2 - p^2\right] \\ &= \frac{1}{n}\left[pq - \frac{1}{4}P_2\right] = variance(p). \end{aligned}$$

The frequency of heterozygotic individuals (Table 26.2) in a nonrandom mating population ($F \neq 0$) and its variance are

$$P_2 = 2pq(1-F), \text{ then } variance(p) = \frac{1}{n}\left[pq - \frac{1}{2}pq(1-F)\right] = \frac{pq(1+F)}{2n}.$$

In a random mating population ($F = 0$), then

$$P_2 = 2pq, \text{ then } variance(p) = \frac{1}{n}\left(pq - \frac{1}{2}pq\right) = \frac{pq}{2n}.$$

Note that random assortment of alleles A and a ($F = 0$) produces the minimum genetic variance of the distribution of the allele frequency p.

Wahlund's Model

Hardy-Weinberg equilibrium guarantees that allele frequencies in a large random mating population do not change over generations (Chapter 23). When a population is divided into random mating subgroups, allele frequencies in the original undivided population are not the same as the frequencies that result from combining values calculated from within subgroups. The mean frequency of heterozygotic individuals calculated from combining subgroups, for example, will always be less than the frequency from the original undivided population.

The simplest case occurs when a population is divided into two separate random mating groups that have different frequencies of an allele A denoted p_1 and p_2. The mean frequency of heterozygotic individuals calculated from combining the two groups (denoted $\bar{H}$) is less than the frequency the original population frequency (denoted H). In symbols, the mean frequency calculated by combining the observed values from each group is

$$\bar{H} = \frac{1}{2}(2p_1q_1 + 2p_2q_2) = \overline{2pq},$$

and the frequency in the undivided population is

$$H = 2pq = 2\left\{\frac{p_1 + p_2}{2}\left[1 - \frac{p_1 + p_2}{2}\right]\right\} = p_1q_1 + p_2q_2 + \frac{1}{2}(p_1 - p_2)^2;$$

therefore,

$$H = \bar{H} + \frac{1}{2}(p_1 - p_2)^2 \quad \text{or} \quad H \geq \bar{H},$$

Table 26.4 *Notation: Description of k Random Mating Subpopulations with Different A-Allele Frequencies* ($p_i - i = 1, 2, \ldots, k$)

		Genotype frequencies		
Populations	Probabilities	*AA*	*Aa*	*Aa*
1	p_1	p_1^2	$2p_1q_1$	q_1^2
2	p_2	p_2^2	$2p_2q_2$	q_2^2
3	p_3	p_3^2	$2p_3q_3$	q_3^2
–	–	–	–	–
–	–	–	–	–
–	–	–	–	–
k	p_k	p_k^2	$2p_kq_k$	q_k^2
	Mean values	$\frac{1}{k}\sum p_i^2$	$\frac{1}{k}\sum 2p_iq_i$	$\frac{1}{k}\sum q_i^2$
	Notation	$\overline{p^2}$	$\overline{2pq}$	$\overline{q^2}$

where $p = \frac{1}{2}(p_1 + p_2)$ and $q = 1 - p$. The result shows a reduction in the mean level of heterozygotic individuals relative to the original population (in symbols, $H = 2pq \geq \bar{H} = \overline{2pq}$). Furthermore, the reduction is proportional to difference between allele frequencies $|p_1 - p_2|$.

A general demonstration of the decrease in heterozygotic *Aa*-individuals, frequently referred to as Wahlund's model, begins with a population consisting of k groups with different frequencies of the allele labeled *A*. The mating within separated subpopulations is required to be random. Table 26.4 provides the notation and details.

The relationship among the subgroups and the population as a whole is simply described as partitioning the total genetic variance of the *A*-allele frequency ($p = 1 - q$). As before, the total variation divides into within-subgroup variation and between-subgroup variation (Chapters 13 and 15). In general, for k subgroups,

$$variance(total) = variance(within) + variance(between) \text{ or, in symbols, } \sigma_t^2 = \sigma_w^2 + \sigma_b^2.$$

The specific partition of the genetic variation is

$$pq = \frac{1}{k}\Sigma p_i(1 - p_i) + \frac{1}{n}\Sigma(p_i - p)^2 = \overline{pq} + \sigma_b^2 \quad i = 1, 2, \ldots, k,$$

where $\{p_1, p_2, \ldots, p_k\}$ represents the k group-specific *A*-allele frequencies and $p = \frac{1}{k}\Sigma p_i$ is the frequency of the *A*-allele in the original population (Table 26.4) (Chapter 13).

Rearrangement of the within and between partitioned variances yields

$$2pq - 2\sigma_b^2 = \overline{2pq} \quad \text{or} \quad 2pq \geq \overline{2pq} \text{ and again, } H \geq \bar{H}.$$

Therefore, the original population frequency of heterozygotic individuals (H) is greater than the mean frequency in the k subpopulations ($\bar{H}$) (Table 26.4, last line); that is, dividing a population into subgroups decreases the mean frequency of heterozygotic *Aa*-individuals.

Table 26.5 *Illustration: Four Subpopulations with Allele Frequencies* $p_i = \{0.8, 0.6, 0.4,$ *and* $0.2\}$

Subdivided subpopulations				
Genotypes	p_i	AA p_i^2	Aa $2p_i\, q_i$	aa q_i^2
1	0.8	0.64	0.32	0.04
2	0.6	0.36	0.48	0.16
3	0.4	0.16	0.48	0.36
4	0.2	0.04	0.32	0.64
Subgroup means	–	0.30	0.40	0.30
Population means	–	0.25	0.50	0.25
Differences	–	0.05	−0.10	0.05

Note: $\sigma_t^2 = 0.25$, $\sigma_w^2 = 0.20$ and $\sigma_b^2 = 0.05$ and $\sigma_t^2 = \sigma_w^2 + \sigma_b^2$.

Following a similar pattern, the AA-genotype frequency relationship is

$$\overline{p^2} = p^2 + \sigma_b^2 \quad \text{or} \quad \overline{p^2} \geq p^2,$$

and the aa-frequency relationship is

$$\overline{q^2} = q^2 + \sigma_b^2 \quad \text{or} \quad \overline{q^2} \geq q^2.$$

Thus, the mean frequencies of homozygotic individuals increase when the population is subdivided. As expected, for either the original population and mean frequencies from the subgroups,

$$\overline{p^2} + \overline{2pq} + \overline{q^2} = p^2 + 2pq + q^2 = 1.0.$$

In purely statistical terms, the Wahlund model is a special case of partitioning the total variation into within and between components (Chapter 13).

A small worked example of subdividing an original random mating population with an A-allele frequency of 0.5 illustrates Walhund's model applied to four subgroups (Table 26.5 – $p = q = 0.5$).

The population between variance is $\sigma_b^2 = 0.05$. Therefore, the subgroup mean genotypic frequencies are

$$\overline{2pq} = 2pq - 2\sigma_b^2 = 0.50 - 0.10 = 0.40,$$

$$\overline{p^2} = p^2 + \sigma_b^2 = 0.25 + 0.05 = 0.30,$$

and

$$\overline{q^2} = q^2 + \sigma_b^2 = 0.25 + 0.05 = 0.30.$$

Thus, the frequency of Aa-individuals from the original random mating frequency of $H = 2pq = 0.50$ is reduced to $\bar{H} = \overline{2pq} = 0.40$, a difference of $2\sigma_b^2 = 2(0.05) = 0.10$.

These example expressions and results are based on known population allele frequencies for simplicity and clarity. For subpopulations containing different numbers of individuals, the same relationships emerge using estimated gene frequencies weighted by the size of the subgroups. The decrease in genetic diversity demonstrated by Wahlund's model is the

Table 26.6 *Illustration: History of a Series of Random Samples Starting with a Population Consisting of* 10 *A-Alleles and* 10 *a-Alleles*

Sample	*A*-alleles	*a*-alleles	p^a
0	10	10	0.50
1	13	7	0.65
2	11	9	0.55
3	12	8	0.60
4	11	9	0.55
5	15	5	0.75
6	14	6	0.70
7	16	4	0.80
8	19	1	0.95
9	20	0	1.00

[a] p represents the proportion of *A*-alleles after each sampling of 20 alleles with replacement.

beginning of the description of the influence of genetic isolation on evolutionary changes in a population gene pool.

Random Genetic Drift

An important determinant of genetic variability in human populations comes from random variation. When a population is small, random variation has a particularly strong influence on allele frequency over a sequence of generations. Generation to generation, random variation increases or decreases allele frequency by chance alone, potentially producing an allele frequency of zero (completely eliminated) or one (completely identical). The process is given the name *random genetic drift*. If this process is not countered by other genetic forces, heterozygotic individuals will cease to exit. That is, the allele frequencies eventually become *fixed*. Agreement exists that random genetic drift has important evolutionary consequences, but the details are complicated and debated.

A Description: Consecutive Random Sampling of a Genetic Population

Key to describing the influence of random genetic drift on the make-up of a gene pool is a concrete and unequivocal definition of the term *random sampling* in this genetic context. Consider the artificial situation that starts with a population of 20 alleles consisting of 50% *A*-type and 50% *a*-type individuals; that is, 10 alleles are *A* and 10 alleles are *a*. Say that a random sample of 20 alleles from this distribution yields 13 *A*-alleles and 7 *a*-alleles ($p = 0.65$), sampled *with replacement* (Chapter 11). Sampled with replacement means that the composition of the population sampled does not change. More simply, the frequencies of the alleles sampled are not changed by the sampling process. The identical population is sampled 20 times. The sampling process is not different from the previously described bootstrap replicate sampling (Chapter 11). A second sample of 20 is then selected from the new population where the proportion of *A*-alleles (denoted p) has increased to $p = 13/20 = 0.65$. Sampling this second population with replacement yields 11 *A*-alleles and a new value $p = 11/20 = 0.55$. Yet a third sample is selected with replacement from the second population with frequency $p = 0.55$, yields 12 *A*-alleles, and the value p becomes $p = 12/20 = 0.60$. Table 26.6 describes a specific

series of sampling with replacement where the distribution of the A-alleles and, therefore, the proportion p in the next generation are entirely determined by the frequency in the previous generation.

The consecutive random samples of 20 alleles produce a population that consists entirely of A-alleles by chance alone in nine generations. Additional sampling yields the same result, 20 A-alleles, because $p = 1.0$. The A-allele frequency is fixed. Random sampling variation has turned a heterogeneous population ($p = 0.5$) into an entirely homogeneous population ($p = 1.0$). In other words, the a-allele has completely disappeared.

Consecutive sampling with replacement of any 20 alleles will require more or less generations, but, sooner or later, the allele frequency becomes fixed. A process analogous to this artificial example creates the phenomenon called *random genetic drift*.

Random Genetic Drift – A Description

Results of four computer simulations based on different sized populations (denoted N) demonstrate the dynamics of random genetic drift. The plots display changing allele frequencies in 50 populations as lines (Figure 26.2), each illustrating the results of random genetic drift on the A-allele frequency over 100 generations starting at an original frequency of $p_0 = 0.5$. The plot (upper left) displays the changing A-allele frequency p over 100 generations for each of 50 populations made up of $N = 15$ individuals ($2N = 30$ alleles), one line for each history of allele frequency p. The allele frequency in each generation results from random sampling alleles from the previous generation with replacement. The A-allele becomes fixed in 53% of the simulations and the a-allele fixed in 47% of the simulations for populations of size 15 after 100 generations. The parallel results for population sizes $N = \{30, 60, \text{and } 100\}$ are plotted, and the distributions of an allele frequencies after 100 generations are given in Table 26.7. The fundamental property of random drift is clear. As populations increase in size, the mean time required for the allele frequency to become fixed increases. Specifically, after n generations, the expected frequency of heterozygotic individuals is

$$H_n = \left[1 - \frac{1}{2N}\right]^n H_0.$$

The decrease in the original frequency of heterozygotic individuals H_0 to a frequency of H_n over n generations for a population of size $N = 15$ is illustrated in Figure 26.3 (simulated values – dotted line). The theoretical frequency H_n is included (solid line – to be discussed).

The Statistics of Random Genetic Drift

For A-allele frequency in the initial generation 0 denoted p_0 ($q_0 = 1 - p_0$) and the estimate of p_0 denoted $\hat{p}$, the following are true (Chapter 27):

1. $E[\hat{p}] = p_0$,
2. $\hat{p} = \frac{2a+b}{2N}$ and $\hat{q} = 1 - \hat{p}$,
3. $variance(\hat{p}) = \sigma^2_{\hat{p}} = E[\hat{p} - p_0]^2 = E[\hat{p}^2] - p_0^2$, and
4. $variance(\hat{p}) = \sigma^2_{\hat{p}} = \frac{p_0 q_0}{2N}$.

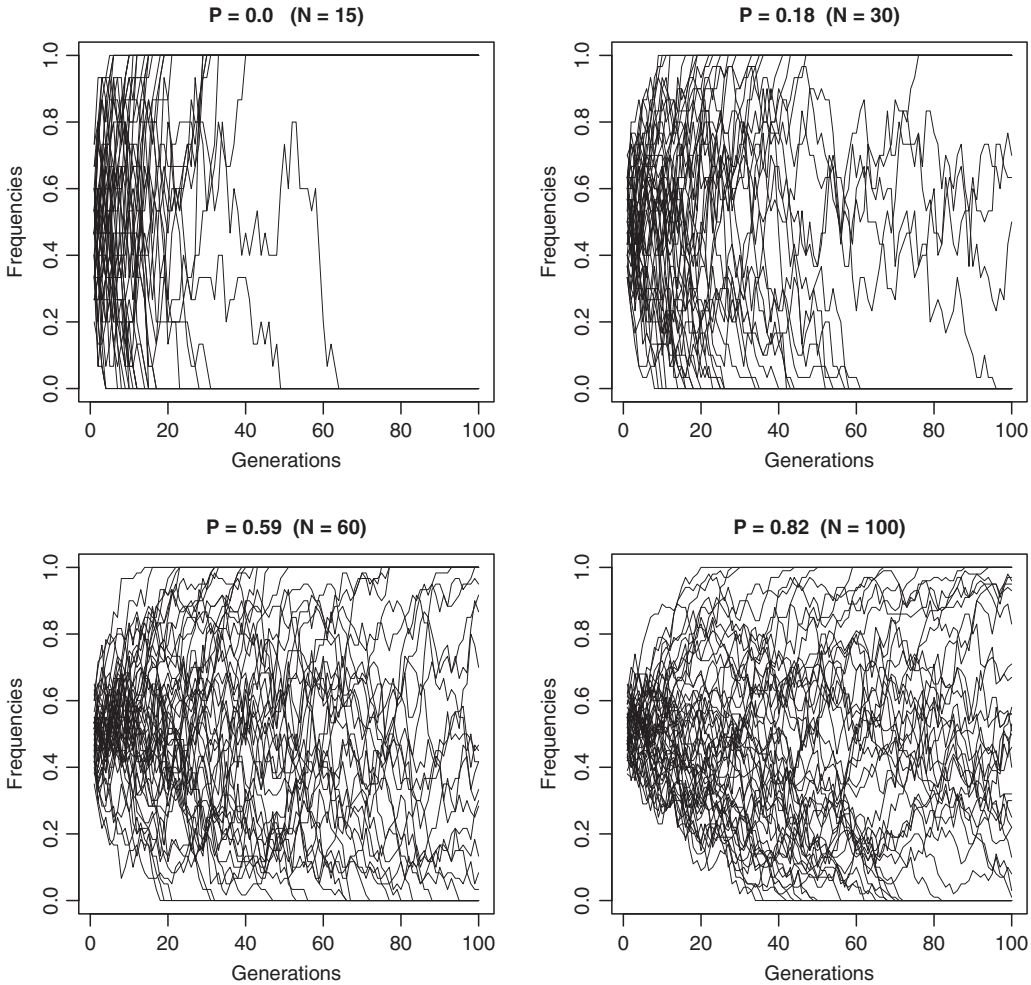


Figure 26.2 Four Examples of 50 Computer Simulations (Lines) Representing the History of Random Genetic Drift for Each Population Size = $N = \{15, 30, 60, \text{and } 100\}$ (P = Proportion of Allele Frequencies That Are Not Fixed after 100 Generations)

Therefore, starting with the frequency of heterozygotic individuals $H_0 = 2p_0q_0$, in the next generation the expected frequency of heterozygotic individuals H_1 is

$$\begin{aligned} H_1 &= E[2\hat{p}\hat{q}] = 2\{E[\hat{p}] - E[\hat{p}^2]\} = 2\left[E[\hat{p}] - E[\hat{p}^2] - p_0^2 + p_0^2\right] \\ &= 2\left[p_0 - p_0^2 - \left\{E[\hat{p}^2] - p_0^2\right\}\right] = 2\left[p_0 - p_0^2 - \sigma_p^2\right] \\ &= 2\left[p_0q_0 - \frac{p_0q_0}{2N}\right] = 2p_0q_0\left[1 - \frac{1}{2N}\right] = H_0\left[1 - \frac{1}{2N}\right]. \end{aligned}$$

The general expression describing influence of random genetic drift is then

$$H_{i+1} = H_i\left[1 - \frac{1}{2N}\right];$$

Table 26.7 *Proportion of Allele Frequencies after 100 Generations of Random Mating for Population Sizes* $N = \{15, 30, 60, \text{ and } 100\}$

	Generations = 100		
N	A-allele	a-allele	Not fixed
15	0.53	0.47	0.00
30	0.41	0.41	0.18
60	0.18	0.23	0.59
100	0.08	0.10	0.82

that is, in each generation the expected decrease in heterozygotic individuals is $[1 - \frac{1}{2N}]$. Therefore, for consecutive generations

$$H_2 = H_1\left[1 - \frac{1}{2N}\right] = H_0\left[1 - \frac{1}{2N}\right]^2 \rightarrow H_3 = H_2\left[1 - \frac{1}{2N}\right]^2 = H_0\left[1 - \frac{1}{2N}\right]^3 \rightarrow \cdots,$$

and, in general, for n generations

$$H_n = H_0\left[1 - \frac{1}{2N}\right]^n.$$

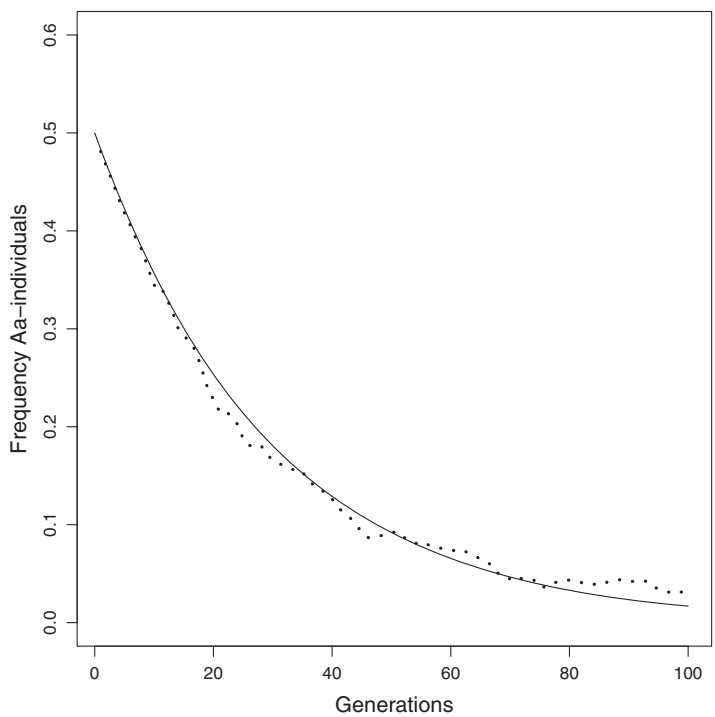


Figure 26.3 Decrease in Heterozygotic Individuals over 100 Generations for $N = 15$ Individuals – Theoretical (Solid Line) and Computer Simulation (Dotted Line)

Thus, the expression for the degree of heterogeneity H_n describes the decrease from random genetic drift in a population of size N (Figure 26.3, solid line) after n generations. Clearly, when the number of generations is large and eliminated heterozygotic individuals are not replaced, the population allele frequency sooner or later becomes fixed.

Mutation/Selection Equilibrium

Genetic mutation is usually a naturally occurring change of one kind of allele to another and is critically important for maintaining diversity in a gene pool. Typically, A-alleles mutate to a-alleles ($A \rightarrow a$) or the reverse ($a \rightarrow A$). In addition, mutation can produce entirely new alleles, although these "new" alleles can be alleles lost in the past and arise anew. Changes in allele frequencies from mutation appear to be without cause in many situations creating important sources of genetic variation as well as replacing alleles eliminated by other forces such as selection or random loss. It is said that mutation produces the raw material for natural selection.

The Story of Mr. A and Mr. B

Mr. A has 100 dollars, and Mr. B has 10 dollars. In moment of generosity, Mr. A decides to give Mr. B 80% of his money ($80) if Mr. B gives him 20% of his money ($2). After the exchange Mr. A has a total of $22 ($100 − $80 + $2 = $22) and Mr. B has a total of $88 ($10 + $80 − $2 = $88). Mr. B then says, "Let's do that again." To their surprise, after the second exchange, they both have the same amounts of money, again $22 and $88. Such exchanges produce stable values after one exchange. In this case, after the first exchange, the amount of money Mr. A gives Mr. B becomes the same amount Mr. B gives Mr. A, namely, $\$22(0.8) = \$88(0.2) = \$17.60$. Their total amounts after the first exchange remain the same for any subsequent exchange. In general, when the exchange rates, such as 0.8 and 0.2, add to 1.0, the amount traded back and forth becomes the same after one exchange, and the total amounts remain unchanged for any additional exchanges.

In fact, for any pair of exchange rates, an equilibrium value is reached after a series of exchanges. Consider another example. Again Mr. A starts with 100 dollars and Mr. B starts with 10 dollars, but the exchange rates are $p_a = 0.50$ ($A \rightarrow B$) and $p_b = 0.10$ ($B \rightarrow A$). The exchanges continue until neither participant's total amount changes. A history of this not very realistic situation is presented in Table 26.8. After 12 exchanges, equilibrium occurs.

The reason the series of exchanges reaches equilibrium can be viewed two ways. First, Mr. A gives less and less money to Mr. B, and Mr. B gives more and more money to Mr. A until they exchange the same amount and at that point the amounts exchanged no longer change. In the example, the amount is $\$18.34(0.50) = \$9.17$ from Mr. A and $\$91.66(0.10) = \$9.17$ from Mr. B, making exchanged amounts equal. Thus, Mr. A has a total amount $18.34 and Mr. B has $91.66 and further exchanges produce exactly the same amounts.

A second more general point of view follows. For any pair of exchange rates, the increase/decrease pattern is described by

$$\text{Mr. A: } A_n = A_{n-1} - A_{n-1}(p_a) + B_{n-1}(p_b)$$

Table 26.8 *Example: Mr. A with $100 and Mr. B with $10 and Exchange Rates of* $p_a = 0.50$ $(A \to B)$ *and* $p_b = 0.10$ $(B \to A)$

	Exchanges				
	A	B	$A \to B$	$B \to$ A	Ratio[a]
1	100.00	10.00	50.00	1.00	10.00
2	51.00	59.00	25.50	5.90	0.86
3	31.40	78.60	15.70	7.86	0.40
4	23.56	86.44	11.78	8.64	0.27
5	20.42	89.58	10.21	8.96	0.23
6	19.17	90.83	9.58	9.08	0.21
7	18.67	91.33	9.33	9.13	0.20
8	18.47	91.53	9.23	9.15	0.20
9	18.39	91.61	9.19	9.16	0.20
10	18.35	91.65	9.18	9.16	0.20
11	18.34	91.66	9.17	9.17	0.20
12	18.34	91.66	9.17	9.17	0.20

[a] A/B-ratio.

and

$$\text{Mr. B: } B_n = B_{n-1} + A_{n-1}(p_a) - B_{n-1}(p_b)$$

for the nth exchange. At equilibrium, for stable amounts $A_n = A_{n-1} = A$ and $B_n = B_{n-1} = B$ (no change), then

$$A = A - A(p_a) + B(p_b) \quad \text{or} \quad A(p_a) - B(p_b) = 0.$$

That is, again the amount money exchanged becomes the same, specifically, $A(p_a) = B(p_b)$.

The basic property of this "give and take" pattern is that the division of money at equilibrium is determined entirely by the exchange rates p_a and p_b. Thus, after a number of exchanges, sooner or later,

$$A(p_a) = B(p_b) \quad \text{or} \quad \frac{A}{B} = \frac{P_b}{p_a}$$

regardless of the distribution of money at the initial exchange. For the example, at equilibrium, the division of the original $\$100/\$10$ becomes $\$18.34/\$91.66 = 0.2$ determined entirely by the ratio of exchange rates $p_b/p_a = 0.1/0.5 = 0.2$.

To summarize, at equilibrium two properties generally apply:

1. The total amounts to be distributed at the first exchange do not affect the final distribution and, therefore,
2. The ratio at equilibrium is entirely determined by the exchange rates.

These two properties in a genetic context, over a sequence of generations, cause increases from mutation of an allele ("Mr. B") to become equal to the same alleles eliminated by selection ("Mr. A") producing a stable allele frequency determined entirely by the mutation and selection rates.

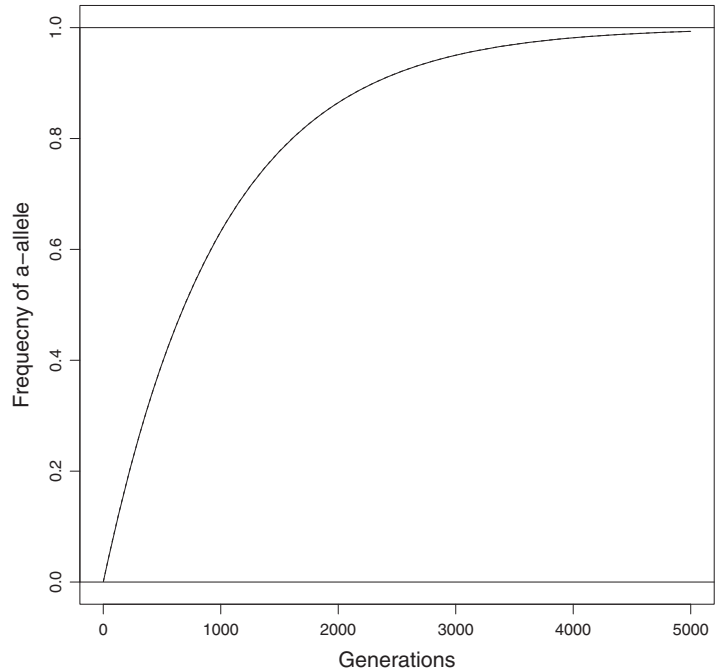


Figure 26.4 Rate of Elimination of the A-allele by Mutation Causing the Corresponding Increase in the a-Allele for a Mutation Rate of $\mu = 0.001$ over 5000 Generations

One-Way Mutation

As an introduction, consider of the dynamics of mutation only (no selection). The mutation rate of an A-allele to a-allele ($A \rightarrow a$) is denoted μ, and the initial frequency of the A-allele is denoted p_0. In the first generation, the A-allele frequency p_0 decreases to $p_1 = (1 - \mu)p_0$. The decrease of A-allele frequency p_0 over a sequence of generations continues:

$$p_1 = (1-\mu)p_0 \rightarrow\ p_2 = (1-\mu)^2 p_0 \rightarrow\ p_3 = (1-\mu)^3 p_0 \rightarrow\ \text{and}\ \rightarrow \cdots.$$

After n generations, the A-allele frequency is reduced to $p_n = (1 - \mu)^n\ p_0$. Regardless of mutation rate μ, it is only a question of time until the allele A is eliminated from the population, and only the a-allele remains ($q \rightarrow 1$). Specifically, the increase in the a-allele is $q_n = 1 - p_n = 1 - (1 - \mu)^n\ p_0$ after n generations (Figure 26.4).

The a-allele frequency after n generations is approximately

$$q_n = 1 - (1 - q_0) \times e^{-n\mu}.$$

Solving this expression for n yields an expression for the approximate number of generations needed to produce an increase of the a-allele frequency from $q_0 = 0$ to an a-allele frequency of q and is

$$n \approx -\frac{1}{\mu}\log(1-q).$$

Table 26.9 *Mutation/Selection: Elimination Rate s of Aa-Genotypes*

	Genotypes frequency			
	AA	*Aa*	*aa*	Total
Frequency	p^2	$2pq$	q^2	1.0
Survival Rate	1.0	$1-s$	1.0	–
Next Generation	p^2	$2pq(1-s)$	q^2	$1-2pqs$

If the mutation rate is $\mu = 0.001$, then after $n = 4605$ generations, the frequency of $q_0 = 0$ becomes $q = 0.99$ in approximately 100,000 years, only "minutes" on an evolutionary time scale.

Mutation/Selection Balance

The dynamics of mutation replacing the same allele eliminated by selection is illustrated by two kinds of selection patterns: selection eliminating *aa*-genotypes and selection eliminating *Aa*-genotypes.

For rate of elimination by selection of *Aa*-genotypes denoted s, the next generation produces a decrease in the frequency of the *a*-allele (Table 26.9).

The decrease in the *a*-allele frequency q to q' $(a \to A)$ is approximately

$$q' - q = \frac{q^2 + pq(1-s)}{1-2pqs} - q = \frac{-pqs(1-2q)}{1-2pqs} \approx -sq \text{ per generation,}$$

where q' represents the *a*-allele frequency in the next generation. Mutation then replaces the lost alleles at a rate of μ $(A \to a)$ establishing a "Mr. A/Mr. B type give and take" exchange and producing an equilibrium that depends only on the two "exchange rates," μ and s. Regardless of the original allele frequency, the *a*-allele ultimate frequency is determined entirely by the ratio of rates μ and s, and eventually a stable frequency becomes $q = \mu/s$.

A similar mutation/selection balance occurs when selection eliminates *aa* genotypes again at a rate of s, and mutation replaces the lost *a*-alleles at again a rate of μ (Table 26.10).

The decrease in the *a*-allele frequency q to q' is approximately

$$q' - q = \frac{q^2(1-s) + pq}{1-q^2s} - q = \frac{-pq^2s}{1-q^2s} \approx -sq^2 \text{ per generation,}$$

Table 26.10 *Mutation/Selection: Elimination Rate s of aa-Genotypes*

	Genotype frequencies			
	AA	*Aa*	*aa*	Total
Frequencies	p^2	$2pq$	q^2	1.0
Survival rate	1.0	1.0	$1-s$	–
Next generation	p^2	$2pq$	$q^2(1-s)$	$1-q^2s$

where again q' represents the a-allele frequency in the next generation. Therefore, a "give and take" exchange equilibrium produces $\mu/s = q^2$. The stable frequency of a-allele is again entirely determined by rates μ and s, eventually yielding stable a-allele frequency of $q = \sqrt{\mu/s}$.

When selection is complete ($s = 1$), the a-allele frequency becomes $q = \mu$ for Aa-selection and $q = \sqrt{\mu}$ for aa-selection. The relationship between mutation rate (μ) and selection rate (s) indicates the reason that fatal genetic diseases are rare ($q \ll small$) but continue to exist at low levels determined primarily by a stable balance between mutation and selection rates μ and s.

Assortative Mating

About a Harmonic Series

The details and properties of a harmonic series not only describe the results of assortative mating, they illustrate the often simple and sometimes magical descriptive powers of mathematics.

First, consider the most fundamental numeric series, one with equally spaced values. For example, the values {0, 2, 4, 6, 8, . . . } is a series with equal differences between any two consecutive values. A general expression such a series is represented by

$$\{a_0, a_0 + d, a_0 + d(2), a_0 + d(3), \ldots, a_0 + d(k), \ldots\}$$

with a constant difference d between any two consecutive values. The expression for the constant difference is, therefore,

$$a_k - a_{k-1} = [a_0 + dk] - [a_0 + d(k-1)] = d.$$

In addition, an expression for the nth term in the sequence is $a_n = a_0 + d(n)$.

A harmonic series is defined as a sequence of values such that the reciprocal values create a sequence of equally spaced values. For example, consider the harmonic series denoted by $\{b_0, b_1, b_2, \ldots, b_k\}$ or, for example, {1/1, 1/4, 1/7, 1/10, . . . }, then the following applies:

Index (i)	0	1	2	3	4	5	6	7	8	· · ·
a_i	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_9	· · ·
Harmonic series (b_i)	1/1	1/4	1/7	1/10	1/13	1/16	1/19	1/22	1/25	· · ·
Reciprocal $\left(a_i = \frac{1}{b_i}\right)$	1	4	7	10	13	16	19	22	25	· · ·

For any two consecutive terms the constant difference is $d = a_k - a_{k-1} = 3$. The property that equally spaced sequences produce a specific expression for the nth term applies to the reciprocal values of a harmonic series. Therefore, from the example, for the common difference of three ($d = 3$), then $a_n = \frac{1}{b_n} = a_0 + (3)n$. The nth term in the harmonic series becomes the reciprocal value or

$$\frac{1}{a_n} = b_n \quad \text{and} \quad b_n = \frac{1}{a_0 + (d)n} = \frac{1}{1 + (3)n}.$$

Some examples are

$$\text{for } n = 3 \text{ then, } b_3 = \frac{1}{1+(3)3} = \frac{1}{10}, \text{ for } n = 7 \text{ then, } b_7 = \frac{1}{1+(3)7} = \frac{1}{22},$$
$$\text{and for } n = 77 \text{ then, } b_{77} = \frac{1}{1+(3)77} = \frac{1}{231}.$$

In general, the nth term of a harmonic series is $b_n = \frac{1}{1+(d)n}$.

Assortative Mating Model

The Hardy-Weinberg equilibrium of allele frequencies that results from random mating is not always a realistic description of genetic inheritance. In human populations, selection of a mate is rarely completely random with respect to genotype or phenotype. Selection patterns are based on many obvious and sometimes not obvious genetic characteristics. For example, similar race, proximity, body size, socioeconomic status, and intelligence are just a few characteristics that produce nonrandom mating, called *assortative mating*.

A simple model identifies the properties of assortative mating in the context of the genetics of genotypes AA, Aa, and aa and is presented as a start to describing the complex influences of assortative mating. Nevertheless, this elementary description of a specific inheritance pattern demonstrates two fundamental properties: namely, assortative mating does not change the allele frequency and decreases the frequency of heterozygotic individuals in the population.

Two kinds of assortative mating are the following:

Negative assortative matings: Mating is between only A^+ phenotypes and aa genotypes ($A^+ \times aa$) where A^+-phenotype represents either AA or Aa genotypes (Chapter 23).

Positive assortative matings: Mating is between only A^+-phenotypes ($A^+ \times A^+$) or between aa-genotypes ($aa \times aa$).

The simplest case is complete negative assortative mating. The first generation consisting of $AA \times aa$ or $Aa \times aa$ matings produces offspring with only genotypes Aa or aa. Thus, the next generation consists entirely of negative assortative $Aa \times aa$ matings that produce only two genotypes with equal frequencies, one-half Aa and one-half aa individuals. Therefore, no change occurs in future generations.

The details and notation for positive assortative mating are described in Table 26.11 for an A-allele with frequency p ($q = 1 - p$).

The first property to note is that the frequency of the A-allele after one generation of positive assortative mating (denoted p') is the same as in the previous generation (denoted p). From Table 26.11 (last row), after one generation of assortative mating, the frequency of the A-allele becomes

$$p' = \frac{p^2 + p^2 q}{1 - q^2} = \frac{p^2(1+q)}{1-q^2} = p,$$

where $1 - q^2 = p(1 + q)$. Thus, the A-allele frequency remains p over a sequence of generations.

Table 26.11 *Assortative Mating: Probabilities Associated with Positive Phenotypic Matings A^+-Phenotype × A^+-Phenotype and aa-Genotype × aa-Genotype*

		Genotype frequencies		
Matings	Frequencies	*AA*	*Aa*	*aa*
$AA \times AA$	$p^4/(1-q^2)$	$p^4/(1-q^2)$	–	–
$AA \times Aa$	$2p^3q/(1-q^2)$	$p^3q/(1-q^2)$	$p^3q/(1-q^2)$	–
$Aa \times AA$	$2p^3q/(1-q^2)$	$p^3q/(1-q^2)$	$p^3q/(1-q^2)$	–
$Aa \times Aa$	$4p^2q^2/(1-q^2)$	$p^2q^2/(1-q^2)$	$2p^2q^2/(1-q^2)$	$p^2q^2/(1-q^2)$
$aa \times aa$	q^2	–	–	q^2
Next generation	1.0	$p^2/(1-q^2)$	$2p^2q/(1-q^2)$	$p^2q^2/(1-q^2)+q^2$

The frequency of heterozygotic individuals in the next generation (denoted H'), from Table 26.11 (last row), becomes

$$H' = \frac{2p^2q}{1-q^2} = \frac{2p^2q}{p^2-2pq} = \frac{Hp}{p^2+2pq} = \frac{Hp}{p^2+pq+pq} = \frac{Hp}{p+\frac{1}{2}H} = \frac{2Hp}{2p+H},$$

where $H = 2pq$ in the previous generation. The frequency of heterozygotic individuals (H_n) based on this expression gives rise to the sequence

$$H_1 = \frac{2pH_0}{2p+H_0} \rightarrow H_2 = \frac{2pH_1}{2p+H_1} \rightarrow H_3 = \frac{2pH_2}{2p+H_2} \rightarrow \cdots \rightarrow H_n = \frac{2pH_{n-1}}{2p+H_{n-1}} \rightarrow \cdots.$$

Note that $H_{n+1} < H_n$, indicating a continuous decease in heterozygotic individuals.

Four steps identify that continuous positive assortative mating causes a decrease in heterozygotic individuals over a sequence of generations described by a harmonic series.

Step One

The pattern created by the reciprocal values of H_n creates an equally spaced sequence of values and is, therefore, a harmonic series. For any two consecutive generations n and $n-1$, the difference in reciprocal values is

$$\frac{1}{H_n} - \frac{1}{H_{n-1}} = \frac{2p+H_{n-1}}{2pH_{n-1}} - \frac{1}{H_{n-1}} = \frac{1}{H_{n-1}} + \frac{1}{2p} - \frac{1}{H_{n-1}} = \frac{1}{2p}.$$

Thus, the sequence of values generated by $1/H_n$ for $n = 1, 2, \ldots$ are equally spaced ($p =$ constant value).

Step Two

Using the previous relationship from an equally spaced series of values that $a_n = a_0 + n(d)$, the relationship between generation 0 and generation n from the equally spaced series generated by assortative mating is

$$\frac{1}{H_n} = \frac{1}{H_0} + \left[\frac{1}{2p}\right] n = \frac{2p+nH_0}{2pH_0}.$$

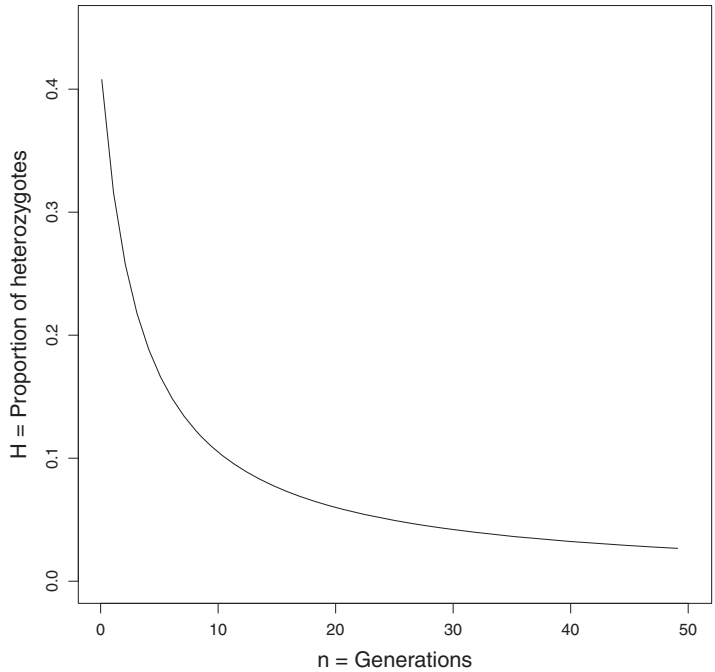


Figure 26.5 Decrease in the Proportion of Heterozygotic Individuals from Positive Assortative Mating (A-Allele Frequency of $p = 0.7$ and Initially $H = 2pq = 2(0.7)(0.3) = 0.42$)

Step Three

The general form for H_n from a harmonic series in terms of H_0 is then the reciprocal value of $1/H_n$ or

$$H_n = \frac{2pH_0}{2p + nH_0}.$$

Step Four

For $H_0 = 2pq$, then

$$H_n = \frac{2p(2pq)}{2p + n(2pq)} = \frac{H_0}{1 + nq}.$$

Figure 26.5 displays the continuous downward frequency of heterozygotic individuals resulting from positive assortative mating described by the harmonic series H_n ($p = 0.7$).

One generation of assortative mating is illustrated by a specific worked example for an A-allele frequency of $p = 0.7$ making $H = 2(0.7)(0.3) = 0.42$. then (Table 26.12, last line)

$$p' = 0.539 + \frac{1}{2}(0.323) = 0.7 = p \text{ and}$$

$$H = 2pq = 2(0.7)(0.3) = 0.42, \text{ then } H' = \frac{H}{1+q} = \frac{0.42}{1 + 0.30} = 0.323.$$

Table 26.12 *Illustration: One Generation of Positive Assortative Mating Where the A-Allele Frequency Is* $p = 0.7$

Mating	Frequency	*AA*	*Aa*	*aa*
$AA \times AA$	0.264	0.264	–	–
$AA \times Aa$	0.226	0.113	0.113	–
$Aa \times AA$	0.226	0.113	0.113	–
$Aa \times Aa$	0.194	0.049	0.097	0.049
$Aa \times aa$	0.090	–	–	0.090
Next generation	1.0	0.539	0.323	0.139

After one generation, the allele frequency $p = 0.7$ has not changed, and the frequency of the *Aa*-genotype decreases from 0.420 to 0.323. This pattern of decreasing heterozygotic individuals, therefore, occurs in each following generation of positive assortative mating while maintaining an *A*-allele frequency $p = 0.7$. The distribution of genotypes resulting from positive assortative mating does not yield a Hardy-Weinberg pattern. A constant allele frequency persists over a sequence of generations, but to repeat, the frequency of heterozygotic individuals continues to decease (harmonic series).

Theory

27

Statistical Estimation

Statistical analysis explores two worlds, properties and application of summary estimates calculated from data and properties and parameters of populations that produce the data. These two worlds are related but it is critical to keep in mind their distinctions to successfully apply statistical methods. The following description of the basics of estimation starts with the data world of the sample. The discussion begins with mean values, variances, and covariances estimated from sampled observations. The discussion then turns to the relationships of these fundamental statistical summaries to their parallel world of population values.

The Sample

The mean value and its variance are routinely estimated by the expressions

$$\textit{sample mean value} = \overline{x} = \frac{1}{n}\sum x_i$$

and

$$\textit{sample variance} = S_X^2 = \frac{1}{n-1}\sum (x_i - \overline{x})^2$$

from a sample of observations $X = \{x_1, x_2, \ldots, x_n\}$ where $i = 1, 2, \ldots, n$ = sample size.

An important property of these two estimates is the change in their values caused by changes in the data. That is, for the "new" variable $ax + b$, the mean value becomes

$$\textit{mean}(aX + b) = \frac{1}{n}\sum (ax_i + b) = a\overline{x} + b,$$

and the variance becomes

$$S_{aX+b}^2 = \frac{1}{n-1}\sum ([ax_i + b] - [a\overline{x} + b])^2 = a^2\frac{1}{n-1}\sum (x_i - \overline{x})^2 = a^2 S_X^2$$

for constant values a and b.

This linear transformation has numerous applications. For example, consider the Wilcoxon (W) and Mann-Whitney ($\hat{P}$) nonparametric test statistics. For sample sizes n_1 and n_2 ($N = n_1 + n_2$), the Mann-Whitney test statistic is

$$U = W + n_1(n_1 + 1)/2$$

and

$$\textit{variance}(W) = \frac{1}{12}n_1 n_2(N + 1).$$

Therefore, the estimated Mann-Whitney probability

$$\hat{P} = \frac{U}{n_1 n_2} = \frac{1}{n_1 n_2}[W + n_1(n_1 + 1)/2]$$

has variance

$$variance(\hat{P}) = \left[\frac{1}{n_1 n_2}\right]^2 variance(W) = \frac{1}{12 n_1 n_2}(N + 1),$$

where the constant values are $a = \frac{1}{n_1 n_2}$ and $b = n_1(n_1 + 1) - 2$.

Covariance

Three artificial examples based on the sum of two variables and its properties introduce the statistical measure called the *sample covariance*. For unrelated variables X and Y (correlation = 0), then

Data	Summaries
$X = \{1, 2, 3, 4, 5\}$	$\overline{x} = \overline{y} = 3.0$
$Y = \{5, 1, 2, 3, 4\}$	$S_X^2 = S_Y^2 = 2.5$
$X + Y = \{6, 3, 5, 7, 9\}$	$\overline{x + y} = \overline{x} + \overline{y} = 6.0$ and $S_{X+Y}^2 = S_X^2 + S_Y^2 = 5.0$.

Of particular note, the mean value of the sum is the sum of the mean values of X and Y and the variance is also the sum of their variances.

When variables X and Y are positively related (correlation > 0), then

Data	Summaries
$X = \{1, 2, 3, 4, 5\}$	$\overline{x} = \overline{y} = 3.0$
$Y = \{1, 2, 3, 4, 5\}$	$S_X^2 = S_Y^2 = 2.5$
$X + Y = \{2, 4, 6, 8, 10\}$	$\overline{x + y} = \overline{x} + \overline{y} = 6.0$ and $S_{X+Y}^2 = 10.0$.

The mean value of the sum remains the sum of the mean values, but the variance of the sum is greater than the sum of the variances.

When X and Y are negatively related (correlation < 0), then

Data	Summaries
$X = \{1, 2, 3, 4, 5\}$	$\overline{x} = \overline{y} = 3.0$
$Y = \{5, 4, 3, 2, 1\}$	$S_X^2 = S_Y^2 = 2.5$
$X + Y = \{6, 6, 6, 6, 6\}$	$\overline{x + y} = \overline{x} + \overline{y} = 6.0$ and $S_{X+Y}^2 = 0.0$.

The mean value of the sum again remains the sum of the mean values but the variance of the sum is less than the sum of the variances.

This simple illustration demonstrates two important properties illustrated by the sum of two variables $X + Y$. The mean value of the sum is the sum of the respective mean values,

but the variance of a sum is influenced by the relationship between the two variables. This influence is measured by the sample covariance, defined as

$$\text{sample covariance} = S_{XY} = \frac{1}{n-1}\sum (x_i - \overline{x})(y_i - \overline{y})$$

from a sample of n pairs of observations (x_i, y_i). For the examples, the three estimated covariances are $S_{XY} = 0$ (unrelated), $S_{XY} = 2.5$ (positive), and $S_{XY} = -2.5$ (negative).

The reason a covariance measures the degree of association between two x/y-valuables follows from four x/y-relationships:

1. $(x_i - \overline{x}) > 0$ and $(y_i - \overline{y}) > 0$ then, $(x_i - \overline{x})(y_i - \overline{y}) > 0$,
2. $(x_i - \overline{x}) < 0$ and $(y_i - \overline{y}) < 0$ then, $(x_i - \overline{x})(y_i - \overline{y}) > 0$,
3. $(x_i - \overline{x}) > 0$ and $(y_i - \overline{y}) < 0$ then, $(x_i - \overline{x})(y_i - \overline{y}) < 0$, and
4. $(x_i - \overline{x}) < 0$ and $(y_i - \overline{y}) > 0$ then, $(x_i - \overline{x})(y_i - \overline{y}) < 0$.

Relationships 1 and 2 occur when values increase together or decrease together. Their product adds positive quantities to the estimated covariance. Relationships 3 and 4 occur when an increase in one variable is associated with a decrease in the other variable. Their product adds negative quantities to the estimated covariance. When the covariance estimate contains approximately equal numbers of positive and negative values, the resulting covariance in the neighborhood of zero indicates the two variables X and Y are close to unrelated. Thus, the sum of cross-product terms measures the strength of association within sampled pairs (x_i, y_i).

The specific relationship of the estimated covariance to the sample variance of the sum $X + Y$ for n pairs of x/y-values is

$$\begin{aligned} S^2_{X+Y} &= \frac{1}{n-1}\sum ([x_i + y_i] - [\overline{x} + \overline{y}])^2 = \frac{1}{n-1}\sum ([x_i - \overline{x}] + [y_i - \overline{y}])^2 \\ &= \frac{1}{n-1}\left[\sum (x_i - \overline{x})^2 + \sum (y_i - \overline{y})^2 + 2\sum (x_i - \overline{x})(y_i - \overline{y})\right] \\ &= S^2_X + S^2_Y + 2S_{XY} \end{aligned}$$

and similarly for the difference $X - Y$, then $S^2_{X-Y} = S^2_X + S^2_Y - 2S_{XY}$.

Parallel to the variance, the covariance is influenced by changes in the variables X and Y; that is, for "new" variables $aX + b$ and $cY + d$, the covariance becomes

$$\begin{aligned} S_{aX+b,cY+d} &= \frac{1}{n-1}\sum ([ax_i + b] - [a\overline{x} + b])([cy_i + d] - [c\overline{y} + d]) \\ &= ac\frac{1}{n-1}\sum (x_i - \overline{x})(y_i - \overline{y}) = acS_{XY}. \end{aligned}$$

Because the value of a sample covariance depends on the units used to measure variables X and Y ($S_{aX+b,cY+d} = acS_{XY}$), a direct interpretation based on the estimated value is rarely useful. Values of the covariance, for example, can be large because of a strong x/y-association or large because the measurement units of either X or Y create a large value. Thus, for a variable measured in feet or centimeters, the covariances measuring the same relationship differ considerably.

A sample covariance takes on concrete meaning in two ways. When the estimated covariance of X and Y is divided by the estimated variance of X, the result is an estimate of the slope of a straight line describing the linear relationship between the x/y-variables. In symbols, the estimate of the slope (denoted $\hat{b}$) is

$$\text{estimated slope} = \hat{b} = \frac{S_{XY}}{S_X^2}.$$

Thus, the transformed sample covariance measures the linear relationship between X and Y as rate of change in Y relative to the rate of change in X (Chapter 3).

When the estimated covariance of X and Y is divided by the estimated standard deviations of X and Y, the result is the sample estimate of the correlation coefficient. In symbols, the estimated correlation coefficient (denoted r_{XY}) is

$$\text{estimated correlation coefficient} = r_{XY} = \frac{S_{XY}}{S_X S_Y}.$$

The transformed covariance measures the extant a straight line summarizes an x/y-relationship as a value between -1 and 1. The value $r_{XY} = 1$ indicates that a straight line with a positive slope perfectly represents the relationship between two variables. Similarly, the value $r_{XY} = -1$ indicates that a straight line with a negative slope perfectly represents the relationship between two variables. As might be expected, a value of r_{XY} close to zero indicates that a straight line has little value as a summary of the x/y-relationship. The variable X is frequently said to be unrelated the variable Y. Technically, "unrelated" in this context means linearly unrelated.

The value of the correlation coefficient estimated by r_{XY} is the same regardless of the measurement units of X and Y. The correlation between $aX + b$ and $cY + d$ remains r_{XY} or, specifically

$$r_{aX+b,cY+d} = \frac{S_{aX+b,cY+d}}{\sqrt{S_{aX+b}^2}\sqrt{S_{cY+d}^2}} = \frac{acS_{XY}}{aS_X cS_Y} = \frac{S_{XY}}{S_X S_Y} = r_{XY}$$

for constant values a, b, c, and d.

The Population

A sample mean value calculated from n observations can be estimated in terms of a weighted sum using the proportions of each of the same observations in the sampled data. The weighted average is

$$\text{sample mean value} = \overline{x} = \sum \left[\frac{n_i}{n}\right] x_i = \sum \hat{p}_i x_i \quad i = 1, 2, \ldots, n$$

when each of n_i observations x_i appears with a proportion of occurrence $\hat{p}_i = n_i/n$.

The mean of a probability distribution, more precisely called the *expected value*, is similarly defined. The expression for an expected value (denoted EX) for a distribution with k different pairs of values p_i and x_i is

$$\text{expected value} = EX = \sum p_i x_i \quad i = 1, 2, \ldots, k = \text{number of values } x_i$$

based on the probability distribution $\{p_1, p_2, p_3, \ldots, p_k\}$. In other words, the expected value is a weighted average of each possible value of x_i where the weighs are the known and fixed probabilities represented by p_i. These probabilities indicate the likelihood of occurrences of each value of x_i ($\sum p_i = 1.0$). They are not generated by sample data but are known and fix values from a theoretical probability distribution. The value represented by EX, as noted, is the mean value of the probability distribution described by the values p_i.

For example, the simplest probability distribution is a binary variable X that takes on the values 0 or 1 with probabilities $P(X = 0) = 1 - p$ or $P(X = 1) = p$. The expected value is

$$EX = (1 - p) \times 0 + p \times 1 = p.$$

A slightly extended example starts with the probability distribution consisting of $k = 5$ different values of x_i each occurring with probabilities p_i or and produces the expected value

X	$x_1 = 0$	$x_2 = 1$	$x_3 = 2$	$x_4 = 3$	$x_5 = 4$	Total
p_i	0.0625	0.25	0.375	0.25	0.0625	1.0

$$EX = \sum p_i x_i = 0.0625(0) + 0.25(1) + 0.375(2) + 0.25(3) + 0.0625(4) = 2.0.$$

The value $EX = 2.0$ is the mean value of the probability distribution of variable X.

The variance of a probability distribution is defined and calculated in much the same way. It too is a weighted average where the weights are again probabilities p_i from a specific probability distribution. The quantities weighted are the squared deviations of each x_i-value from the expected value, a measure of variability. In symbols, the deviations are $(x_i - EX)^2$ and the weights are again the probability of occurrence p_i. For the probability distribution of X, the expression for the variance (denoted σ_X^2) becomes the weighted average

$$\text{expected variance} = E(X - EX)^2 = \sigma_X^2 = \sum p_i (x_i - EX)^2,$$

where again $i = 1, 2, \ldots, k =$ the number of pairs of values of x_i and p_i. The variance is a weighted average of k known and fixed measures of variability.

The variance of the previous binary variable X is then

$$\sigma_X^2 = (1 - p)(0 - p)^2 + p(1 - p)^2 = p(1 - p),$$

where $EX = p$.

Continuing the example for the distribution of $k = 5$ outcomes, the variance is

$$\begin{aligned}\sigma_X^2 &= 0.0625(0 - 2)^2 + 0.25(1 - 2)^2 + 0.375(2 - 2)^2 + 0.25(3 - 2)^2 + 0.0625(4 - 2)^2 \\ &= 1.0,\end{aligned}$$

where $EX = 2$. Important and sophisticated expected values and variances play key roles in the analysis of data. Particularly, important are the expected values and variance that define of the normal, binomial and Poisson probability distributions.

Analogous to the sample mean value and variance, for the expected value and variance of a probability distribution and constant values a and b, then

$$E(aX + b) = \sum p_i (a x_i + b) = a \sum p_i x_i + b \sum p_i = aEX + b \quad \text{where} \quad \sum p_i = 1$$

Table 27.1 *Probabilities: joint Probability Distribution of Two Variables X and Y (X and Y Are Related)*

	$Y=-2$	$Y=0$	$Y=2$	Total
$X=-2$	$p_{11}=0.1$	$p_{12}=0.2$	$p_{13}=0.2$	$p_1=0.5$
$X=2$	$p_{21}=0.2$	$p_{22}=0.2$	$p_{23}=0.1$	$p_2=0.5$
Total	$q_1=0.3$	$q_2=0.4$	$q_3=0.3$	1.0

and

$$\sigma^2_{aX+b} = \sum p_i[(ax_i+b)-E(ax_i+b)]^2 = \sum p_i[ax_i - E(ax_i)]^2$$
$$= a^2 \sum p_i[x_i - E(x_i)]^2 = a^2 E[X-EX]^2 = a^2\sigma_X^2.$$

To summarize, parallel the sample mean and variance, then $E(aX+b) = aEX + b$ and $variance(aX+b) = a^2 variance(x)$.

The sampled population covariance measures the association between variables X and Y and is defined as

$$\text{covariance} = \sigma_{XY} = E([X-EX][Y-EY]) = \sum\sum p_{xy}(x-EX)(y-EY),$$

where p_{xy} represents a joint probability of occurrence of each possible pair of the variables X and Y. Like the expected value and the variance, the covariance is a weighted average of known and fixed values (weights = joint probabilities = p_{xy}). Note that the variance of X (σ_X^2) is a special case of covariance when $X=Y$.

A small artificial example illustrates expectations, variances and covariance calculated from a joint probability distribution. The six probabilities (p_{xy}) of the joint probability distribution of variable $X=\{-2,2\}$ and a variable $Y=\{-2,0,2\}$ are displayed in Table 27.1. For example, the joint probability of $X=2$ and $Y=0$ is $p_{xy}=p_{22}=0.2$.

Based on the joint probability distribution (Table 27.1), then

$$EX = \sum p_i x_i = 0.5(-2)+0.5(2) = 0,$$
$$\sigma_X^2 = E(X-EX)^2 = \sum p_i(x_i-EX)^2 = 0.5(-2)^2+0.5(2)^2 = 4.0 \quad i=1 \text{ and } 2,$$
$$EY = \sum q_j y_j = 0.3(-2)+0.4(0)+0.3(2) = 0,$$
$$\sigma_Y^2 = E(Y-EY)^2 = \sum q_j(y_j-EY)^2$$
$$= 0.3(-2)^2+0.4(0)+0.3(2)^2 = 2.4 \quad j=1,2, \text{ and } 3.$$

Also

$$E(X+Y) = \sum\sum p_{xy}(X+Y)$$
$$= 0.1(-4)+0.2(0)+0.2(-2)+0.2(2)+0.2(0)+0.1(4) = 0.$$

The covariance of the x/y-pairs is

$$\sigma_{XY} = E([x-EX][y-EY]) = \sum\sum p_{xy}([x_i-EX][y_j-EY])$$
$$= 0.1(4)+0.2(-4)+0.2(0)+0.2(0)+0.2(-4)+0.1(4) = -0.8$$

Table 27.2 *Probabilities: joint Probability Distribution of Two Independent Variables X and Y (X and Y Are Unrelated)*

	$Y=-2$	$Y=0$	$Y=2$	Total
$X=-2$	$p_{11}=0.15$	$p_{12}=0.20$	$p_{13}=0.15$	$p_1=0.5$
$X=2$	$p_{21}=0.15$	$p_{22}=0.20$	$p_{23}=0.15$	$p_2=0.5$
Total	$q_1=0.3$	$q_2=0.4$	$q_3=0.3$	1.0

for $i=1, 2$ and $j=1, 2$ and 3. The variance of the sum of values X and Y is

$$\sigma^2_{X+Y}=\sigma^2_X+\sigma^2_Y+2\sigma_{XY}=4.0+2.4+2(-0.8)=4.8.$$

Calculated directly,

$$\begin{aligned}\sigma^2_{X+Y}&=E([x+y]-E[X+Y])^2=\sum\sum p_{xy}([x+y]-E(X+Y])^2\\&=0.1(16)+0.2(0)+0.2(4)+0.2(4)+0.2(0)+0.1(16)=4.8.\end{aligned}$$

For two unrelated variables X and Y ($\sigma_{XY}=0$), an example of a joint probabilities is given in Table 27.2.

Independence means the joint probability of p_{xy} is also p_xp_y (Table 27.2). The expected values and variances of X and Y are not influenced by covariance between X and Y. They remain $EX=0, \sigma^2_X=4.0, EY=0$ and $\sigma^2_Y=2.4$ (Table 27.1). When X and Y are independent, however, the covariance is $\sigma_{XY}=0$. Specifically, based on the example x/y-pairs (Table 27.2)

$$\begin{aligned}\sigma_{XY}&=E([x-EX][y-EY])=\sum\sum p_{xy}([x-EX][y-EY])\\&=0.15(4)+0.15(-4)+0.2(0)+0.2(0)+0.15(-4)+0.15(4)=0.0\end{aligned}$$

and

$$\sigma^2_{X+Y}=\sigma^2_X+\sigma^2_Y=4.0+2.4=6.4.$$

Thus, the variance of the sum $X+Y$ is the sum of the variances when $\sigma_{XY}=0$.

Furthermore, three variables X, Y, and Z follow the same pattern. The variance of the sum $X+Y+Z$ is

$$\sigma^2_{X+Y+Z}=\sigma^2_X+\sigma^2_Y+\sigma^2_Z+2\sigma_{XY}+2\sigma_{XZ}+2\sigma_{YZ}=\sigma^2_X+\sigma^2_Y+\sigma^2_Z$$

when the variables X, Y, and Z are unrelated (*covariances* $=0$).

In general, the expectation and variance of a sum ($\sum x_i$) are

$$E\left(\sum x_i\right)=E(X_1+X_2+\cdots X_k)=EX_1+EX_2+\cdots+EX_k=\sum EX_i$$

and

$$\sigma^2_{X_1+X_2+\cdots+X_k}=\sigma^2_{\sum X_i}=\sum\sigma^2_{X_i}+\sum\sum\sigma_{X_iX_j}\quad i=1,2,\ldots,k \text{ and } j=1,2,\ldots,k,$$

where σ_{X_i,X_j} represents the covariance of variables X_i and X_j ($i\neq j$).

Of particular importance, when the values $\{x_1, x_2, \ldots, x_n\}$ are n independent observations, then the variance of the sum is the sum of the variances. In symbols, for this special and

important case of independence ($covariance(x_i, x_j) = \sigma_{X_i,X_j} = 0$), the variance of the sum ($\sum x_i$),

$$\sigma^2_{X_1+X_2+\cdots+X_k} = \sigma^2_{\sum X_i} = \sum \sigma^2_{X_i},$$

is the sum of the variances because, to repeat, all covariances are zero. The expectation of the sum $E(\sum x_i)$ is always the sum of the expected values $\sum EX_i$.

For example, when the mean value is estimated from n independent observations x_i sampled from a single population with variance represented by σ^2_X, the variance of the sample mean $\overline{x} = \frac{1}{n}\sum x_i$ is

$$\begin{aligned} variance(\overline{x}) = variance\left(\frac{1}{n}\sum x_i\right) &= \frac{1}{n^2} variance\left(\sum x_i\right) \\ &= \frac{1}{n^2}\sum variance(x_i) = \frac{1}{n^2} n\sigma^2_X = \frac{1}{n}\sigma^2_X \end{aligned}$$

$$\text{because } variance\left(\sum x_i\right) = \sum variance(x_i).$$

Two special cases of the mean value and variance useful for both sample and expected values are:

$$\text{if } x_i = a, \text{ then } \overline{x} = \frac{\sum a}{n} = \frac{na}{n} = a \quad \text{and}$$

$$\text{if } x_i = a, \text{ then } EX = \sum p_x a = a\sum p_x = a$$

for sample values represented by $\{x_1, x_2, \ldots, x_n\}$. Also,

$$S^2_X = \frac{1}{n-1}\sum (x_i - \overline{x})^2 = \frac{1}{n-1}\sum (x_i^2 - 2x_i\overline{x} + \overline{x}^2) = \frac{1}{n-1}\left[\sum x_i^2 - n\overline{x}^2\right]$$

and

$$\sigma^2_X = E(X - EX)^2 = E(X^2 - 2XEX + [EX]^2) = EX^2 - [EX]^2.$$

Infinite Series with Statistical Applications

Early Greek mathematicians concluded it was not possible to calculate the sum of an infinite series. After centuries of failed attempts, a solution emerged in 1715 when Brook Taylor and Colin Maclaurin produced methods to calculate the sum of a specific infinite series.

Their solution starts with an infinite power series. An expression for the sum of this series is

$$\begin{aligned} f(x) &= a_0 + a_1(x-c) + a_2(x-c)^2 + a_3(x-c)^3 + \cdots + a_k(x-c)^k + \cdots \\ &= \sum a_k(x-c)^k \quad k = 0, 1, 2, \ldots. \end{aligned}$$

The fact that a function $f(x)$ can be represent a sum of an infinite series of powers for a variable x provides a mathematical relationship that plays an important role in creating several statistical tools. Clearly, a method must be available to calculate the power series coefficients represented by a_k. Such a method produces a Taylor series; that is, the power

Table 27.3 *Details: Coefficients for a Taylor Series Expression for Function* $f(x)$

Derivatives of the power series of $f(x)$	at $x = c$
$f(x) = a_0 + a_1(x-c) + a_2(x-c)^2 + a_3(x-c)^3 + a_4(x-c)^4 + \cdots$	$f(c) = a_0$
$f^{[1]}(x) = (1)a_1 + 2a_2(x-c) + 3a_3(x-c)^2 + 4a_4(x-c)^3 + \cdots$	$f^{[1]}(c) = 1!a_1$
$f^{[2]}(x) = 2(1)a_2 + 3(2)a_3(x-c) + 4(3)a_4(x-c)^2 + \cdots$	$f^{[2]}(c) = 2!a_2$
$f^{[3]}(x) = 3(2)(1)a_3 + 4(3)(2)a_4(x-c) + \cdots$	$f^{[3]}(c) = 3!a_2$
$f^{[4]}(x) = 4(3)(2)(1)a_4 + \cdots$	$f^{[4]}(c) = 4!a_4$
– – –	– – –
– – –	– – –
– – –	– – –

series becomes a Taylor series when the function $f(x)$ is expressed as

$$f(x) = f(c) + \left[\frac{1}{1!}f^{[1]}(c)\right](x-c) + \left[\frac{1}{2!}f^{[2]}(c)\right](x-c)^2 + \left[\frac{1}{3!}f^{[3]}(c)\right](x-c)^3 + \cdots + \left[\frac{1}{k!}f^{[k]}(c)\right](x-c)^k + \cdots,$$

where the coefficients are

$$a_k = \frac{1}{k!}f^{[k]}(c) \quad k = 0, 1, 2, \ldots.$$

The symbol $f^{[k]}(c)$ represents the kth derivative of the function $f(x)$ or $\frac{d^k}{dx^k}f(x)$, evaluated at the value c (Table 27.3 – details). Thus, the kth consecutive derivative of the function $f(x)$, evaluated at the value c, produces the coefficient a_k when divided by $k!$. A Taylor series expression for the function $f(x)$ is then

$$f(x) = \sum a_k(x-c)^k = \sum \left[\frac{1}{k!}f^{[k]}(c)\right](x-c)^k \quad k = 0, 1, 2, \ldots.$$

When c is set to 0, a Taylor series becomes a Maclaurin series and the coefficients are

$$a_k = \frac{1}{k!}f^{[k]}(0).$$

The parallel Maclaurin series expression of the function $f(x)$ is

$$f(x) = f(0) + \left[\frac{1}{1!}f^{[1]}(0)\right]x + \left[\frac{1}{2!}f^{[2]}(0)\right]x^2 + \left[\frac{1}{3!}f^{[3]}(0)\right]x^3 + \left[\frac{1}{4!}f^{[4]}(0)\right]x^4 + \cdots$$
$$= \sum a_k x^k = \sum \left[\frac{1}{k!}f^{[k]}(0)\right]x^k \quad k = 0, 1, 2, \ldots.$$

Examples of Infinite Power Series

For $f(x) = e^x$, then $f^{[k]}(x) = e^x$ for all values of k. Evaluated at $c = 0$, the Maclaurin series derivatives are then $f^{[k]}(0) = e^0 = 1.0$ making the coefficients $a_k = \frac{1}{k!}$. Thus, the Maclaurin infinite series for the function $f(x) = e^x$ becomes

$$f(x) = e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots = \sum \frac{x^k}{k!} \quad k = 0, 1, 2, \ldots.$$

Specifically, for $x = 2$, then, from the Maclaurin series

$$e^2 = 1 + \frac{2}{1!} + \frac{2^2}{2!} + \frac{2^3}{3!} + \frac{2^4}{4!} + \cdots = \sum \frac{2^k}{k!} \quad k = 0, 1, 2, \ldots.$$

For the Poisson probability distribution, for example,

$$EX = \sum x\, p_x = \sum x \left[\frac{\lambda^x e^{-\lambda}}{x!}\right] = \lambda e^{-\lambda} \sum x \left[\frac{\lambda^{x-1}}{x!}\right]$$
$$= \lambda e^{-\lambda} \left[1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \frac{\lambda^4}{4!} + \cdots\right] = \lambda e^{-\lambda}[e^{\lambda}] = \lambda.$$

A second example: for the function $f(x) = \frac{1}{1-x}$, then $f^{[k]}(c) = \frac{k!}{(1-c)^{k+1}}$ and when $c = 0$, the expression $f^{[k]}(0) = k!$, making $a_k = k!$. A Maclaurin series yields the expression

$$f(x) = \frac{1}{1-x} = 1 + x + 1(2)\frac{x^2}{2!} + 1(2)(3)\frac{x^3}{3!} + 1(2)(3)(4)\frac{x^4}{4!} + \cdots + k!\frac{x^k}{k!} + \cdots$$
$$= 1 + x + x^2 + x^3 + x^4 + \cdots + x^k + \cdots = \sum x^k \quad k = 0, 1, 2, \ldots \text{ for } x < 1.$$

To show:

$$\left[1 - \frac{x}{n}\right]^n \to e^{-x}$$

consider $f(x) = log[1 - \frac{x}{n}]$ where the Maclaurin power series coefficients are

$$f(0)\, log(1) = 0,\ f^{[1]}(0) = -\frac{1}{n},\ f^{[2]}(0)\frac{1}{n^2},\ f^{[3]}(0) = -\frac{1(2)}{n^3},$$
$$f^{[4]}(0) = \frac{1(2)(3)}{n^4}, \ldots,\ f^{[k]}(0) = \frac{(-1)^k[1(2)\cdots(k-1)]}{n^k} \cdots,$$

then

$$f(x) = log\left[1 - \frac{x}{n}\right] = -\frac{x}{n} + \frac{1}{2}\left[\frac{x}{n}\right]^2 - \frac{1}{3}\left[\frac{x}{n}\right]^3 + \frac{1}{4}\left[\frac{x}{n}\right]^4 - \frac{1}{5}\left[\frac{x}{n}\right]^5 + \cdots$$

for $n \to \infty$, then $n\, log[1 - \frac{x}{n}] \to -x$ and $e^{n\, log[1-x/n]} = [1 - \frac{x}{n}]^n \to e^{-x}$.

For example, for a probability p ($q = 1 - p$), then $p^n = [1 - \frac{nq}{n}]^n \to e^{-nq}$ for $n \to \infty$.

Note that the symbol $n \to \infty$ reads "when the value of n becomes infinite."

Binomial Theorem

An application of a Taylor/Maclaurin series demonstrates the binomial theorem. For the function $f(x) = (1 + x)^n$, the Taylor series coefficients are

$$f^{[1]}(x) = \frac{n(1+x)^{n-1}}{1!}, \quad f^{[2]}(x) = \frac{n(n-1)(1+x)^{n-2}}{2!},$$
$$f^{[3]}(x) = \frac{n(n-1)(n-2)(1+x)^{n-3}}{3!} \cdots.$$

Then, evaluated at $x = c$

$$(1+x)^n = 1 + \frac{n(1+c)^{n-1}}{1!}(x-c) + \frac{n(n-1)(1+c)^{n-2}}{2!}(x-c)^2 + \frac{n(n-1)(n-3)(1+c)^{n-3}}{3!}(x-c)^3 + \cdots,$$

making the Taylor series $f(x) = (1+x)^n = \sum \binom{n}{x}(1+c)^{n-k}(x-c)^k$.

For the Maclaurin series where $c = 0, f(x) = (1+x)^n$ is

$$(1+x)^n = 1 + \frac{n}{1!}x + \frac{n(n-1)}{2!}x^2 + \frac{n(n-1)(n-2)}{3!}x^3 + \cdots = \sum \binom{n}{k} x^k.$$

Setting $x = p/q$ $(q = 1-p)$, produces

$$(q+p)^n = \left[q\left(1+\frac{p}{q}\right)\right]^n = q^n\left(1+\frac{p}{q}\right)^n = q^n \sum \binom{n}{k}(p/q)^k$$
$$= \sum \binom{n}{k} p^k q^{n-k} \quad k = 0, 1, 2, \ldots, n,$$

which is a statement of the binomial theorem.

Another example of a Taylor/Maclaurin series useful in the description of the expectation of the geometric probability distribution with parameter p $(q = 1-p)$ is

$$\frac{d}{dq}\left[\frac{1}{1-q}\right] = \left[\frac{1}{1-q}\right]^2 = \frac{1}{p^2}$$

and

$$\frac{d}{dq}\left[\frac{1}{1-q}\right] = \frac{d}{dq}[1 + q + q^2 + q^3 \cdots] = \sum \frac{d}{dq} q^k = \sum k q^{k-1};$$

therefore, equating these two expressions yields $\sum k q^{k-1} = \frac{1}{p^2}$ for $k = 0\ 1, 2, 3, \ldots$.

Functions of Estimates

Summary values are at the center of most data analysis strategies. Furthermore, functions of summary values frequently are important for evaluating and describing the properties of sampled data. That is, when x represents an observed value, the focus is often on a function of x, denoted $f(x)$.

Three examples are the following:

1. When x represents a value from an asymmetric distribution, the distribution of the functions $f(x) = log(x)$ or $f(x) = \sqrt{x}$ frequently have a more symmetric distributions.
2. When x represents a measure of a distance x, the function $f(x) = 1/x$ emphasizes small values and minimizes the influence of large values.
3. When x represents represents a probability close to zero, the function $f(x) = log(x/[1-x])$ allows the estimation of a usually accurate confidence interval.

Therefore, for a function $f(x)$, the expectation of the function $E(f[x])$ and variance *variance* $(f[x])$ are key elements of the distribution of $f(x)$.

Functions of estimates from data arise in another context. When an estimated value is a function of an estimated parameter of a known or postulated probability distribution, evaluation frequently requires values for the expectation $E(f[x])$ and *variance*$(f[x])$.

Three examples are the following:

1. When the frequency of a recessive offspring is estimated by $f(\hat{q}) = \hat{q}^2$, then to evaluate the estimate $\hat{q}^2$ requires values for $E(f[\hat{q}])$ and *variance*$(f[\hat{q}])$ and are derived from the properties for the estimate $\hat{q}$.
2. When estimate of a virus rate is $\hat{p} = f(\hat{Q}) = 1 - \hat{Q}^{1/n}$, then to evaluate this estimate, values for $E(f[\hat{Q}])$ and *variance*$(f[\hat{Q}])$ are necessary and are derived from the properties of the estimate $\hat{Q}$.
3. When the estimate of the log-odds is $log[\hat{o}] = f(\hat{p}) = log[\hat{p}/(1-\hat{p})]$, to evaluate this statistic values for $E(f[\hat{p}])$ and *variance*$(f[\hat{p}])$ are necessary and are derived from the properties of the estimate $\hat{p}$.

Approximate Values for the Expectation and Variance of a Function

The Taylor series $f(x) = \sum \frac{1}{k!}[f^{[k]}(c)](x-c)^2$ for $k = 0, 1, 2, \ldots$ provides typically accurate but approximate values for the expectation $E(f[x])$ and *variance*$(f[x])$. The origin of these two approximations is a truncated Taylor series consisting the first two terms ($k = 0$ and 1). Specifically, a Taylors series representation of $f(x)$ is approximately

$$f(x) \approx f(c) + f^{[1]}(c)(x-c),$$

and setting $c = EX$, becomes

$$f(x) \approx f(EX) + f^{[1]}(EX)(x - EX).$$

The symbol $f^{[1]}(EX)$ again represents the first derivative with respect to x of the function $f(x)$ evaluated at the expected value EX, then

result 1: $E(f[x]) \approx f(EX)$ because

$$E[f(x)] \approx E[f(EX) + f^{[1]}(EX)(x - EX)] = f(EX) + f^{[1]}(EX)E(x - EX) = f(EX),$$

where $E[f(EX)] = f(EX)$ and $E(x - EX) = EX - EX = 0$.

result 2: *variance*$(f[x]) \approx [f^{[1]}(EX)]^2 E(x - EX)^2 = [f^{[1]}(EX)]^2 variance(x)$ because

$$\begin{aligned}
&f(x) - f(EX) + f^{[1]}(EX)(X - EX),\\
&[f(x) - f(EX)]^2 \approx [f^{[1]}(EX)(X - EX)]^2,\\
&E[f(x) - f(EX)]^2 \approx E[f^{[1]}(EX)(X - EX)]^2,\\
&E[f(x) - f(EX)]^2 \approx [f^{[1]}(EX)]^2 E(X - EX)^2,
\end{aligned}$$

then

$$variance(f[x]) \approx \left[f^{[1]}(EX)\right]^2 variance(x).$$

Five examples of application of a truncated Taylor series to create approximate values for the expectation and variance of a function $f(x)$ follow:

1. The variance of the logarithm of an estimate denoted $\hat{g}$.

Since the derivative of $log(x)$ with respect of x is

$$\frac{d}{dx}\, log(x) = \frac{1}{x},$$

then, for the estimate $\hat{g}$, the approximate expectation and variance of $f(\hat{g}) = log(\hat{g})$ are

$$E(log[\hat{g}]) \approx log(E[\hat{g}]) \quad \text{and} \quad variance(log[\hat{g}]) \approx \left[\frac{1}{\hat{g}}\right]^2 variance(\hat{g}).$$

Consider a variable denoted x that has a Poisson probability distribution, then the estimate $rate = r = \frac{x}{n}$ has an estimated variance of $variance(r) = \frac{1}{n^2}\, variance(x) = \frac{x}{n^2} = \frac{r}{n}$, where the estimate of the $variance(x) = x$. The truncated Taylor series approximation yields the expression for the approximate expectation and estimated variance of $log(r)$

$$E(log[r]) \approx log(E[r]) \quad \text{and} \quad variance(log[r]) \approx \frac{1}{r^2} variance(r) = \frac{1}{r^2}\left[\frac{r}{n}\right] = \frac{1}{nr} = \frac{1}{x}.$$

Furthermore, for two independent estimated rates $r_1 = x_1/n_1$ and $r_2 = x_2/n_2$, for the logarithm of the rate ratio $\hat{R} = r_1/r_2$, the estimated variance of $log(\hat{R}) = log(r_1/r_2)$ is

$$variance(log[\hat{R}]) \approx variance(log[r_1]) + variance(log[r_2]) \approx \frac{1}{x_1} + \frac{1}{x_2}.$$

2. The variance of the square root of a variable x that has a Poisson probability distribution with parameter λ.

Because the derivative of $f(x) = \sqrt{x}$ with respect to x is

$$\frac{d}{dx}\sqrt{x} = \frac{-1}{2\sqrt{x}},$$

then the estimated variance for $f(x) = \sqrt{x}$ is

$$variance(\sqrt{x}) \approx \left[\frac{-1}{2\sqrt{x}}\right]^2 variance(x) = \frac{1}{4x}x = 0.25.$$

3. Because the derivative with respect to x of $f(x) = x^{1/n}$ is

$$\frac{d}{dx}x^{1/n} = \frac{x^{1/n}}{nx},$$

then, for the estimate $\hat{p} = 1 - \hat{Q}^{1/n}$, the estimated variance is

$$variance(\hat{p}) \approx \left[\frac{\hat{Q}^{1/n}}{n\hat{Q}}\right]^2 variance(\hat{Q}) = \left[\frac{\hat{Q}^{1/n}}{n\hat{Q}}\right]^2 \frac{\hat{Q}(1-\hat{Q})}{n}$$

when the estimate $\hat{Q}$ has a binomial probability distribution with parameters Q and sample size n.

4. Consider an estimate denoted $\hat{p}$ from a binomial distribution with parameters p and n.

Because the derivative of $f(x) = log[x/(1-x)]$ with respect to x is

$$\frac{d}{dx}\left[log\left(\frac{x}{1-x}\right)\right] = \frac{1}{x(1-x)},$$

then, for the estimated odds $\hat{o} = f(\hat{p}) = \hat{p}/(1-\hat{p})$ where $\hat{p} = x/n$, the estimated variance of the logarithm of the estimated odds is

$$\begin{aligned} variance(log[\hat{o}]) &\approx variance\left[log\left(\frac{\hat{p}}{1-\hat{p}}\right)\right] \\ &= \left[\frac{1}{\hat{p}(1-\hat{p})}\right]^2 \frac{\hat{p}(1-\hat{p})}{n} = \frac{1}{n\hat{p}(1-\hat{p})} = \frac{1}{x} + \frac{1}{n-x} \end{aligned}$$

when the estimated variance of $\hat{p}$ is the binomial distribution $variance(\hat{p}) = \hat{p}(1-\hat{p})/n$.

For the odds ratio estimated from independent samples of x_1 and x_2 from n_1 and n_2 observations, then

$$\hat{or} = \frac{\hat{o}_1}{\hat{o}_2} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)}$$

where

$$\hat{p}_1 = x_1/n_1$$

and

$$\hat{p}_2 = x_2/n_2.$$

The estimate of the variance of the logarithm of the odds ratio, denoted $log(\hat{or})$, is then

$$\begin{aligned} variance(log[\hat{or}]) &= variance(log[(\hat{o}_1]) + variance(log[\hat{o}_2]) \\ &\approx \frac{1}{x_1} + \frac{1}{n_1 - x_1} + \frac{1}{x_2} + \frac{1}{n_2 - x_2}. \end{aligned}$$

5. Greenwood's formula to estimate the variance of a survival probability from censored data requires two applications of a truncated Taylor series approximation. The expression for the variance of a product-limit survival probability $\hat{P}_k$ estimated from data possibly containing noninformative censored survival times is:

Application 1

$$variance(log[\hat{P}_k]) \approx \frac{1}{\hat{P}_k^2} variance(\hat{P}_k)$$

because, in general, $\frac{d}{dp} log(p) = \frac{1}{p}$.

Application 2

$$variance(log[\hat{P}_k]) = variance(log[\hat{p}_1 \times \hat{p}_2 \times \hat{p}_3 \times \cdots \times \hat{p}_d]),$$

$$variance\left(\sum log[\hat{p}_i]\right) \approx \sum \frac{1}{\hat{p}_i^2}\left\{\frac{\hat{p}_i(1-\hat{p}_i}{n_i}\right\} = \sum \frac{\hat{q}_i}{n_i \hat{p}_i}$$

when estimated probabilities $\hat{p}_i$ have independent binomial distributions. Then, the expression to estimate Greenwood's variance is

$$variance(\hat{P}_k) \approx \hat{P}_k^2 variance(log[\hat{P}_k]) = \hat{P}_k^2 \sum \frac{\hat{q}_i}{n_i \hat{p}_i} \quad k = 1, 2, \ldots, d,$$

where d represents the number of complete survival times.

Partitioning the Total Sum of Squares

Consider k groups consisting of n_i values of a variable represented by y_{ij}. The partition of the total variability into two parts, the within group and between group variation, is a useful and important statistical tool. That is, the *total variation = within variation + between variation.* In symbols, the partition is

$$\sum\sum (y_{ij} - \overline{y})^2 = \sum\sum (y_{ij} - \overline{y}_j)^2 + \sum n_i (\overline{y}_i - \overline{y})^2,$$

where $i = 1, 2, \ldots, k =$ number of groups and $j = 1, 2, \ldots, n_i =$ number of observations within each group.

Consider, the difference $y_{ij} - \overline{y}$ written as

$$y_{ij} - \overline{y} = y_{ij} - \overline{y}_i + \overline{y}_i - \overline{y},$$

where $\overline{y}_i$ represents the mean from the ith group based on n_i observations, then

$$\begin{aligned}(y_{ij} - \overline{y})^2 &= [(y_{ij} - \overline{y}) + (\overline{y}_i - \overline{y})]^2,\\ (y_{ij} - \overline{y})^2 &= (y_{ij} - \overline{y}_i)^2 + 2(y_{ij} - \overline{y}_i)(\overline{y}_i - \overline{y}) + (\overline{y}_i - \overline{y})^2,\\ \sum\sum (y_{ij} - \overline{y})^2 &= \sum\sum (y_{ij} - \overline{y}_i)^2 + \sum n_i (\overline{y}_i - \overline{y})^2\end{aligned}$$

and

$$\sum\sum (y_{ij} - \overline{y}_i)(\overline{y}_i - \overline{y}) = \sum n_i (\overline{y}_i - \overline{y}) \sum (y_{ij} - \overline{y}_i) = 0, \quad \text{because } \sum n_i (\overline{y}_i - \overline{y}) = 0.$$

A simple example:

For $n = 18$ observations y_{ij} classified into $i = 1, 2$ and $3 = k =$ groups (rows) made up of values classified into six (columns). Then, for these data:

	Columns							
Rows	1	2	3	4	5	6	Sum	Mean
1	1	2	3	4	5	6	21	3.5
2	7	8	9	10	11	12	57	9.5
3	13	14	15	16	17	18	93	15.5
Total	21	24	27	30	33	36	171	9.5
mean ($\overline{y}_j$)	7	8	9	10	11	12		

and

$$\mathrm{T} = \sum\sum (y_{ij} - \overline{y})^2 = 484.5, \quad \mathrm{W} = \sum\sum (y_{ij} - \overline{y}_i)^2 = 52.5 \quad \text{and}$$
$$\mathrm{B} = \sum n_i (\overline{y}_i - \overline{y})^2 = 432.$$

Also note that $\sum n_i(\overline{y}_i - \overline{y}) = 6(3.5 - 9.5) + 6(9.5 - 9.5) + 6(15.5 - 9.5) = 0$ (last row). As required, $T = W + B = 484.5 = 52.5 + 432$. The partition of the total variability (T) into within (W) and between (B) components plays a primary role in describing data and estimates in a variety of contexts.

Expected Value and Variance of the Wilcoxon Signed Rank Test

The following description is extensive but not difficult.

1. The sum of k consecutive integers or $\sum i$ and $i = 1, 2, 3, \ldots, k$.

To start, $(n + 1)^2 = n^2 + 2n + 1$ or $(n + 1)^2 - n^2 = 2n + 1$, then

k	Function
0	$1^2 - 0^2 = 2(0) + 1$
1	$2^2 - 1^2 = 2(1) + 1$
2	$3^2 - 2^2 = 2(2) + 1$
3	$4^2 - 3^2 = 2(3) + 1$
4	$5^2 - 4^2 = 2(4) + 1$
5	$6^2 - 5^2 = 2(5) + 1$
	–
	–
	–

Thus, the sum of k values is

$$(k + 1)^2 = 2\sum i + (k + 1),$$

and solving for $\sum i$, therefore

$$\sum i = \frac{(k + 1)^2 - (k + 1)}{2} = \frac{k(k + 1)}{2} \quad i = 1, 2, 3, \ldots, k.$$

2. The sum of k consecutive squared integers or $\sum i^2 i = 1, 2, 3, \ldots, k$.

To start, $(n + 1)^3 = n^3 + 3n^2 + 3n + 1$ or $(n + 1)^3 - n^3 = 3n^2 + 3n + 1$, then

k	Function
0	$1^3 - 0^3 = 3(0)^2 + 3(0) + 1$
1	$2^3 - 1^3 = 3(1)^2 + 3(1) + 1$
2	$3^3 - 2^3 = 3(2)^2 + 3(2) + 1$
3	$4^3 - 3^3 = 3(3)^2 + 3(3) + 1$
4	$5^3 - 4^3 = 3(4)^2 + 3(4) + 1$
5	$6^3 - 5^3 = 3(5)^2 + 3(5) + 1$
	–
	–
	–

Thus, the sum of k values is

$$(k+1)^3 = 3\sum i^2 + 3\sum i + (k+1)$$

then, solving for $\sum i^2$

$$3\sum i^2 = (k+1)^3 - \frac{3k(k+1)}{2} - (k+1)$$

therefore,

$$3\sum i^2 = \frac{2k^3 + 3k^2 + k}{2} \quad \text{and} \quad \sum i^2 = \frac{k(k+1)(2k+1)}{6} \quad i = 1, 2, 3, \ldots, k.$$

Wilcoxon's signed rank test statistic as a weighted sum is

$$W^+ = (1)I_1 + (2)I_2 + (3)I_3 + \cdots + (n)I_n;$$

that is, the value of W^+ is the sum of ranks 1 to n each multiplied by the binary indicator variable I_i where $p_i = P(I_i = 1)$ is the probability that a signed rank value is positive and included in the sum and $1 - p_i = P(I_i = 0)$ is the probability that a signed rank value is negative and not included in the sum. For the signed rank test $p_i = p_0 = 0.5$ when only random differences exist between compared groups.

The expected value of a sum is the sum of the expected values, therefore,

$$EW = E\left[\sum i\, p_i\right] = \frac{1}{2}\sum i = \frac{1}{2}\left[\frac{n(n+1)}{2}\right] = \left[\frac{n(n+1)}{4}\right].$$

The variance of a sum of n independent values is the sum of the variances of each value, therefore,

$$\begin{aligned} variance(W) &= variance\left(\sum i\, p_i\right) = \sum [i^2 variance(p_i)] \\ &= \sum [i^2 p_i(1-p_i)] = \frac{1}{4}\sum i^2 = \frac{1}{4}\left[\frac{n(n+1)(2n+1)}{6}\right] = \frac{n(n+1)(2n+1)}{24}. \end{aligned}$$

That is, when no systematic differences exist between compared groups, the expected value of p_i is $E(p_i) = p_0 = 1/2$ with variance of $p_i = p_0(1 - p_0) = 1/4$ for each of the k signed rank values.

Maximum Likelihood Estimation

Maximum likelihood estimation is, as might be expected from the name, an estimation technique used to find exactly the estimate that maximizes a likelihood function. To start, consider the estimation of an allele frequency from genotypes made up of alleles A and a and a sample of artificial data consisting of $n = 6$ individuals; that is, the individuals sampled have genotypes: AA, Aa, AA, aa, Aa, and AA.

When the mating is random, producing a Hardy-Weinberg equilibrium, genetic theory requires the genotype probabilities for each observation to be

$$p^2, 2pq, p^2, q^2, 2pq, \text{ and } p^2,$$

where $p = 1 - q$ and p is the proportion of A-alleles in the population sampled. The same probabilities are displayed in Table 27.4.

Table 27.4 *Data: Probabilities from the Six Genotypes to Illustrate Maximum Likelihood Estimation of A-Allele Frequency* ($p = 1 - q$)

	Genotypes			
	AA	*Aa*	*aa*	Total
Probabilities or	$p^2 \times p^2 \times p^2$	$2pq \times 2pq$	q^2	–
Probabilities	p^6	$4p^2q^2$	q^2	–
Count	3	2	1	6

The sample of six observations, like all samples, has a specific probability of occurring. For the genetic data, the probability of occurrence depends on a single parameter value, namely p. The probability of any specific sample occurring is called the *likelihood value*, and the expression of likelihood values for all possible parameter values is called the *likelihood function.* The likelihood function for the genetic data is

$$\text{likelihood} = L = K \times p^8 \times (1-p)^4 = K \times p^8 \times q^4 \quad 0 \le p \le 1.0.$$

The value K represents the number of arrangements that produce the same likelihood value associated with a specific sample and is not relevant of the estimation of p because it does not involve the parameter value p. For statistical reasons, the logarithm of the likelihood function (a sum) becomes the focus of the estimation process. Sums are more intuitive and more mathematically tractable than products. Also important, sums of values have useful statistical properties absent in products. For example, sums frequently have approximate normal distributions and products do not. In addition, the value that maximizes the logarithm of the likelihood function is the same value that maximizes the likelihood function itself. Thus, the logarithm of the likelihood function for the genetic data becomes the basis for estimation of the allele frequency p and is

$$\text{log-likelihood} = log(L) = log(K) + 8\,log(p) + 4\,log(q).$$

Figure 27.1 displays the likelihood function calculated from the genotype data for a range of possible parameter A-allele frequencies p ($0.1 < p < 0.9$). Values from the log-likelihood expression calculated for 10 selected values of parameter p are contained in Table 27.5.

Maximum likelihood estimation reverses the usual statistical logic where the parameters are fixed and the data are influenced by sources of variation. To achieve a maximum likelihood estimate, the sample data are considered fixed and all possible parameters are searched to find the value that maximizes log-likelihood function. For the genetic data, a small value $p = 0.1$ produces a small associated log-likelihood value (Table 27.5 and Figure 27.1). That

Table 27.5 *Illustration: Log-Likelihood Values for Selected Frequencies of the A-Allele(p)*

P	0.10	0.20	0.30	0.40	0.50	0.60	0.667	0.70	0.80	0.90
log(*L*)	−12.64	−7.56	−4.85	−3.17	−2.11	−1.55	−1.434[a]	−1.46	−2.02	−3.85

[a]Maximum value.

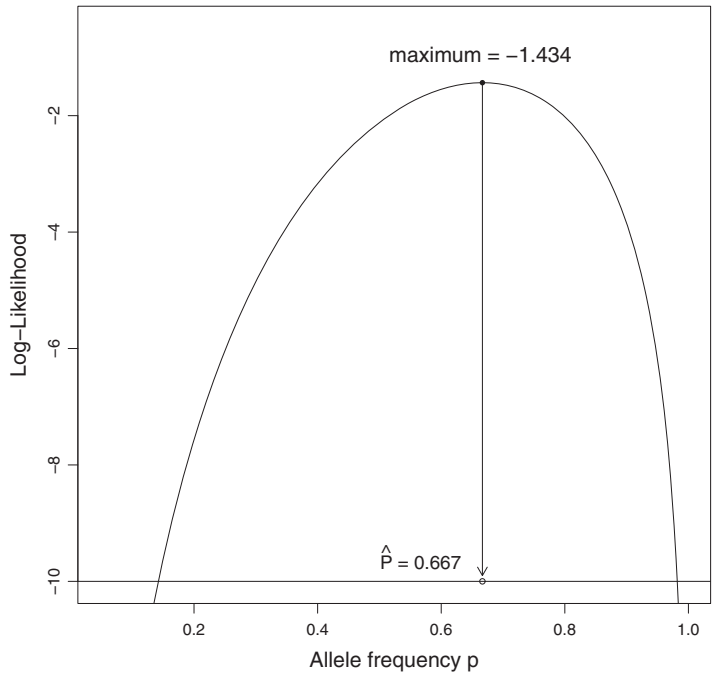


Figure 27.1 Log-Likelihood Function from Sample of Genetic Data Used to Estimate Frequency of Allele A ($0.1 \leq p \leq 0.9$ – Table 27.4)

is, the value for the parameter $p = 0.1$ produces an unlikely log-likelihood value. Thus, the parameter $p = 0.1$ is not a plausible value of the parameter p in light of the observed data. Similarly, a large value of the parameter p, say 0.9, is also an implausible value to have produced the specific sample of data. The two log-likelihood values (-12.64 and -3.85) indicate that the probability of occurrence of the observed sample is extremely small for these two specific parameter values. As small values of p increase and large values of p decrease, the plausibility associated with both parameter values increases (Table 27.5 and Figure 27.1). At the maximum value of the log-likelihood function, these two parameters become the single most plausible value to have produced the observed data, producing the maximum likelihood estimate $\hat{p} = 0.667$ at the maximum value of the log-likelihood function of $log(L) = -1.434$ (Figure 27.1). From another point of view, the choice of the estimate $\hat{p} = 0.667$ is the value of the parameter p most consistent with the observed data among all possible choices, that is, the most plausible value of p.

An elementary result from calculus is at the heart of calculating a maximum likelihood estimate. The maximum or minimum of a single value function is the point x that is the solution to the equation

$$\frac{d}{dx} f(x) = 0.$$

The formal theorem is the following:

> If a function $f(x)$ has a relative extreme value at $x = x_0$, then the first derivative of $f(x)$ is either zero or undefined at the value of x_0.

From a more intuitive point of view, the derivative of a function at any value x is the "instantaneous rate of change" of the function at that point ("a slope"). When the derivative is zero, the rate of change is zero which occurs only at the maximum or minimum value of the function. The "slope" is zero. For example, the derivative of the quadratic polynomial $f(x) = ax^2 + bx + c$ is

$$\frac{d}{dx} f(x) = 2ax + b,$$

and the solution to the equation $2ax + b = 0$ is $x_0 = -b/2a$ and, therefore, is the value of x that produces the maximum/minimum value.

In general, maximum likelihood estimation starts with creating a theoretical likelihood function based on knowledge or assumptions about the properties of the sampled population. In symbols, the likelihood function is expressed as

$$\text{likelihood value} = L(g|x_1, x_2, \ldots, x_n) = \prod f(x_i|g), \quad i = 1, 2, \ldots, n$$

where the parameter to be estimated is represent by g and the values x_i represent n independent observations sampled from the same probability distribution represented by the symbol f. For example, it might be known or assumed that the distribution f is a Poisson probability distribution with the parameter $g = \lambda$ that produced independent observations x_i, then $f(x_i|\lambda) = e^{-\lambda}\lambda^{x_i}/x_i!$. The log-likelihood function follows as

$$\text{log-likelihood function} = log[L(g|x_1, x_2, \ldots, x_n) = \sum log[f(x_i|g)]$$

for the n data independent values $\{x_1, x_2, \ldots, x_n\}$. Therefore, the maximum value of a log-likelihood function necessarily occurs at the value that is the solution to the equation

$$\text{derivative of log-likelihood value} = \frac{d}{dg} \sum log[f(x_i|g)] = 0.$$

The solution produces the maximum and not the minimum value because both extremely small and extremely large parameter values g cause the likelihood function to be small, guaranteeing a maximum value occurs between these two values.

A result from mathematical statistics provides an estimate of the variance of the maximum likelihood estimate (justified in more theoretical texts). The second derivative of the log-likelihood function is called *Fisher's information function*. In symbols, the second derivative of the log-likelihood function is

$$\text{Fisher's information junction} = I(g) = \frac{d^2}{dg^2} \left[\sum log(f(x_i|g) \right].$$

The estimated variance of the distribution of a maximum likelihood estimate $\hat{g}$ becomes

$$variance(\hat{g}) = -\frac{1}{I(\hat{g})},$$

where the parameter value used to evaluate the expression of $I(g)$ is the maximum likelihood estimate $\hat{g}$.

Table 27.6 *Notation: Counts of Three Genotypes (AA, Aa, and aa) to Estimate Frequency of Allele A (Denoted $p = 1 - q$) from n Genotypes*

	Genotype counts			
	AA	*Aa*	*aa*	Total
Count	*a*	*B*	*c*	*N*

For the example genetic data, the maximum likelihood estimate is that value where the likelihood function is maximum, which is the value of the parameter p ($q = 1 - p$) such that

$$\frac{d}{dp} log(L) = \frac{8}{p} - \frac{4}{q} = 0.$$

The maximum value occurs when

$$\frac{8}{\hat{p}} - \frac{4}{\hat{q}} = 0$$

or

$$8\hat{q} - 4\hat{p} = 0$$

and the estimates becomes

$$\hat{p} = \frac{8}{12} = 0.667.$$

The second derivative of the likelihood function is the derivative of the first derivative. For the genetic data, Fisher's information function yields

$$I(\hat{p}) = \frac{d}{dp}\left[\frac{8}{\hat{p}} - \frac{4}{\hat{q}}\right] = -\frac{8}{\hat{p}^2} - \frac{4}{\hat{q}^2} = -\frac{8}{(0.667)^2} - \frac{4}{(0.333)^2} = -54.$$

The estimated variance is then

$$variance(\hat{p}) = -\frac{1}{I(\hat{p})} = -\frac{1}{I(0.667)} = \frac{1}{54} = 0.0185 \quad \text{for } \hat{p} = 0.667.$$

A natural estimate of the A-allele frequency is the count of the number of alleles A (8) divided by the total number of alleles sampled (12), producing the estimate $\hat{p} = 8/12 = 0.667$. Frequently a maximum likelihood estimate is also the intuitive "everyday" estimate. In addition, the estimated variance of $\hat{p}$ is the binomial distribution $variance(\hat{p}) = \hat{p}(1 - \hat{p})/(2n) = 0.667(0.333)/12 = 0.0185$.

An estimated allele frequency in general is similarly calculated. The notation for the sampled data is described in Table 27.6.

The likelihood function is the product

$$L = K \times (p^2)^a \times (2pq)^b \times (q^2)^c = K \times p^{2a+b} \times q^{b+2c},$$

and the log-likelihood function becomes the sum

$$log(L) = log(K) + (2a + b)\, log(p) + (b + 2c)\, log(q).$$

Table 27.7 *Data: Duffy Genotype Frequencies from Sample of* $n = 489$ *African Americans*

	Genotypes – duffy red blood cell types			
	$Fy^a\ Fy^a$	$Fy^a\ Fy^b$	$Fy^b\ Fy^b$	Total
Observed	$a = 8$	$b = 72$	$c = 409$	$n = 489$

Then the maximum value of the likelihood function occurs when

$$\frac{d}{dp}\,log(L) = \frac{2a+b}{\hat{p}} - \frac{b+2c}{\hat{q}} = 0 \quad \text{or} \quad (2a+b)\hat{q} - (b+2c)\hat{p} = 0.$$

The maximum likelihood estimated A-allele frequency is the solution $\hat{p} = (2a + b)/2n$. Again, this estimate is simply the count of the A-alleles divided by the total number of alleles sampled. The symbol K again represents a sometimes complicated value but does not involve the parameter p and, therefore, is not relevant to the estimation. In symbols, the value $\frac{d}{dp}\,log(K) = 0$. Fisher's information function is

$$I(p) = \frac{d}{dp}\left[\frac{2a+b}{p} - \frac{b+2c}{q}\right] = \left[-\frac{2a+b}{p^2} - \frac{b+2c}{q^2}\right] = -\frac{2n}{pq},$$

making the expression for the estimated variance of the distribution of $\hat{p}$

$$variance(\hat{p}) = -\frac{1}{I(\hat{p})} = \frac{\hat{p}\hat{q}}{2n}$$

where

$$\hat{p} = \frac{2a+b}{2n}$$

and

$$\hat{q} = 1 - \hat{p}.$$

Estimation of the frequency of the Duffy allele Fy^a (again denoted p) illustrates (Table 27.7).

The log-likelihood function based the parameter p and the sample of African-American genotype frequencies is displayed in Figure 27.2. The maximum likelihood estimate of the Fy^a-allele frequency is $\hat{p} = (2a + b)/2n = (16 + 72)/[2(489)] = 0.090$ and the estimated variance is $variance(\hat{p}) = \hat{p}\hat{q}/[2n] = 0.090(0.910)/[2(489)] = 0.0000837$.

Properties of a Maximum Likelihood Estimate

Properties of maximum likelihood estimates are many and sometimes complicated, a few basic features indicate the utility of the approach:

1. As the sample size increases, the maximum likelihood estimate converges to the value to be estimated. This theoretical property is called *consistency.* For example, when $\hat{g}$ represents a maximum likelihood estimate of the parameter g, then for large sample sizes, the maximum likelihood estimate $\hat{g}$ becomes indistinguishable from parameter estimated g.

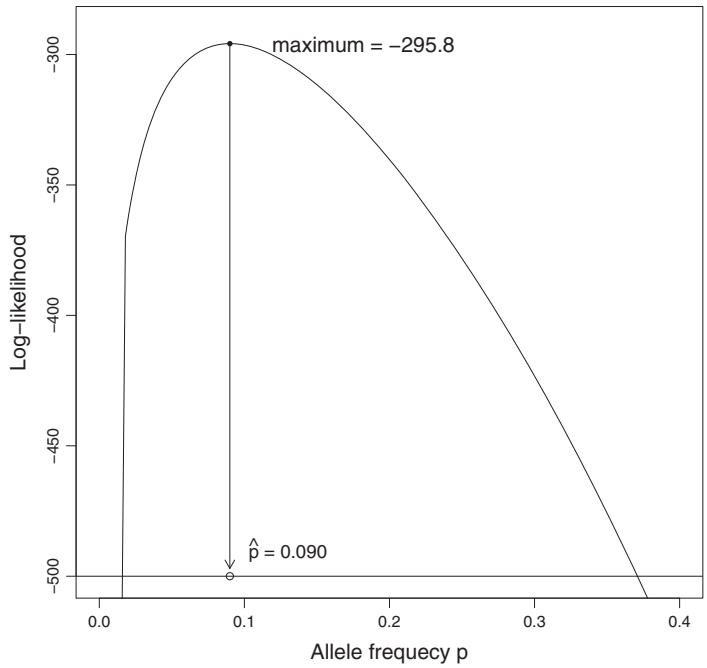


Figure 27.2 Log-Likelihood Function to Estimate the Frequency of the Duffy Allele Fy^a from Genotype Frequencies (Table 27.7)

2. Also for large sample sizes, the maximum likelihood estimate has an approximate normal distribution. In special cases as few as 10 and frequently 30 observations produce accurate but approximate test statistics and confidence intervals based on a normal distribution and maximum likelihood estimates. Thus, a test statistic for the maximum likelihood estimate $\hat{g}$

$$z = \frac{\hat{g} - g_0}{\sqrt{variance(\hat{g})}},$$

sometimes called *Wald's test*, has an approximate standard normal distribution when the value estimated by $\hat{g}$ is g_0. An approximate 95% confidence interval, based on the maximum likelihood estimate $\hat{g}$ and the normal distribution is

$$\hat{g} \pm 1.960\sqrt{variance(\hat{g})}.$$

In both cases, the variance is estimated by Fisher's information function $variance(\hat{g}) = -1/I(\hat{g})$.

3. Maximum likelihood estimates have the smallest variance among consistent estimates, again when the sample size is large. In less technical terms, no consistent estimate is more precise (less variability) for large samples of observations.
4. Functions of maximum likelihood estimates remain maximum likelihood estimates and have properties 1, 2, and 3. For example, when $\hat{g}$ represents a maximum likelihood estimate, then $\hat{g}^2$ or $e^{\hat{g}}$ or $log(\hat{g})$ are also maximum likelihood estimates. For the Duffy

genetic data (Table 27.7), maximum likelihood estimate of the frequency of allele Fy^a is $\hat{p} = 0.090$. Therefore, the estimate of the frequency of the allele Fy^b where $\hat{q} = 1 - \hat{p} = 0.910$ is also a maximum likelihood estimate. The estimated frequency of genotype Fy^bFy^b is $\hat{q}^2 = (0.910)^2 = 0.828$, and it too is a maximum likelihood estimate.

5. Fisher's information function is necessarily calculated as part of the mathematical/computer process that produces maximum likelihood estimates and, therefore, always provides an estimated variance of the estimated value. Thus, maximum likelihood estimation not only produces optimum estimates of parameter values, but also always produces estimates of the variance of the distribution of these estimates.

When more than one parameter is estimated from a likelihood function, the notation and computation become more elaborate but maximum likelihood logic remains the same. Regardless of complexity, maximum likelihood estimate is the unique set of parameter values most likely to have produced the observed data. When k parameters are estimated, the maximum likelihood estimates are the k parameter values that make the likelihood function as large as possible. In symbols, parameters denoted $\{g_1, g_2, g_3, \ldots, g_k\}$ have maximum likelihood estimates denoted $\{\hat{g}_1, \hat{g}_2, \hat{g}_3, \ldots, \hat{g}_k\}$, when the likelihood value (denoted L) evaluated at these k values is larger than the likelihood values calculated from all other possible parameter values $\{g_1, g_2, g_3, \ldots, g_k\}$ or, in symbols,

$$L(\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_k | x_1, x_2, \ldots, x_n) > L(g_1, g_2, \ldots, g_k | x_1, x_2, \ldots, x_n),$$

where again $\{x_1, x_2, \ldots, x_n\}$ represent n independent observations sampled from the same distribution treated for the purposes of estimation as fixed. The computer techniques applied to find maximum likelihood estimates are sophisticated and complex but interpretation of the k estimated parameters remains simple. They are the unique set of k estimates that are most consistent with the observed data. In other words, among all possible sets of parameter values, the maximum likelihood estimates make the probability of occurrence of the observed data most likely. In the same manner of the estimation of a single parameter, the maximum likelihood process produces estimates of the variances and covariances of each of the k estimated parameter values, namely, $variance(\hat{g}_i)$ and $covariance(\hat{g}_i, \hat{g}_j)$, from Fisher's information function.

This discussion lacks a detailed description of the construction of a likelihood function. The process to generate a likelihood function is rather mechanical and can be technically difficult. Maximum likelihood computer software, not only produces estimated values and their variances, it necessarily creates the likelihood function as part of the estimation process.

Example Maximum Likelihood Estimates

The Binomial Distribution

The estimation of the parameter p ($q = 1 - p$) from n independent zero/one binary observations x_i ($\sum x_i = x$) starts with

$$likelihood\ function = L = \prod f(x_i|p) = \binom{n}{x} p^x q^{n-x} \quad i = 1, 2, \ldots, n$$

and

$$\text{log-likelihood function} = log(L) = \left[\binom{n}{x}\right] + x\, log(p) + (n-x)\, log(q),$$

then

$$\frac{d}{dp}\, log(L) = \frac{x}{p} - \frac{n-x}{q}.$$

Therefore, the maximum likelihood estimate of the parameter p is the solution to the equation

$$\hat{q}x - \hat{p}(n-x) = 0$$

or

$$x - n\hat{p} = 0$$

and is

$$\hat{p} = \frac{x}{n}.$$

Estimation of the variance: Fisher's information function (second derivative of the likelihood function) is

$$I(p) = \frac{d^2}{dp^2}\, log(L) = -\frac{x}{p^2} - \frac{n-x}{q^2}.$$

Then, the estimated variance based on the maximum likelihood estimate $\hat{p}$ becomes

$$variance(\hat{p}) = -\frac{1}{I(\hat{p})} = \frac{\hat{p}\hat{q}}{n} \quad \text{where} \quad x = n\hat{p}.$$

The Poisson Distribution

The estimation of the parameter λ from n independent observations x_i sampled from a Poisson distribution starts with

$$\text{likelihood function} = L = \prod f(x_i|\lambda) = \prod \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \quad i = 1, 2, \ldots, n$$

and

$$\text{log-likelihood function} = log(L) = -log(x_i!) - n\lambda + log(\lambda) \sum x_i,$$

then

$$\frac{d}{d\lambda}\, log(L) = -n + \frac{1}{\lambda} \sum x_i.$$

Therefore, the maximum likelihood estimate of the parameter λ is the solution to the equation

$$-n + \frac{1}{\hat{\lambda}} \sum x_i = 0$$

and is

$$\hat{\lambda} = \frac{1}{n} \sum x_i = \overline{x}.$$

Estimation of the variance: the Fisher's information function (second derivative of the likelihood function) is

$$I(\lambda) = \frac{d^2}{d\lambda^2} log(L) = -\frac{1}{\lambda^2} \sum x_i.$$

Then the estimated variance based on the maximum likelihood estimate $\hat{\lambda}$ is

$$variance(\hat{\lambda}) = -\frac{1}{I(\hat{\lambda})} = \frac{\hat{\lambda}}{n}$$

where

$$\sum x_i = n\hat{\lambda}.$$

The Geometric Distribution

The estimation of the parameter p ($q = 1 - p$) from n independent observations x_i sampled from a geometric distribution starts with

$$\text{likelihood function} = L = \prod f(x_i|p) = \prod pq^{x_i} \quad i = 1, 2, \ldots, n$$

and

$$\text{log-likelihood function} = log(L) = n\, log(p) + log(q) \sum x_i,$$

then

$$\frac{d}{dq} log(L) = \frac{n}{p} - \frac{\sum x_i}{q}.$$

Therefore, the maximum likelihood estimate of the parameter q is the solution to the equation

$$n\hat{q} - \hat{p} \sum x_i = 0$$

or

$$\overline{x} = \frac{\hat{q}}{\hat{p}}$$

and is

$$\hat{q} = \frac{\overline{x}}{\overline{x} + 1}.$$

Estimation of the variance: the Fisher's information function (second derivative of the likelihood function) is

$$I(q) = \frac{d}{dq^2} log(L) = -\frac{n}{p^2} - \frac{\sum x_i}{q^2}.$$

Then, the estimated variance based on the maximum likelihood estimate $\hat{q}$ is

$$variance(\hat{q}) = -\frac{1}{I(\hat{q})} = \frac{\hat{p}^2\hat{q}}{n}$$

where

$$\sum x_i = \frac{n\hat{q}}{\hat{p}}.$$

Method of Moments Estimation

When the number of parameter values to be estimated equals the number of independent relationships between data and the parameter values, the maximum likelihood estimate is the solution to the equations formed by equating the expressions describing the model parameters to the corresponding observed data. This approach to estimation is called the *method of moments*. Moment estimates frequently provide easy to derive and intuitive estimates. In special cases these estimates provide maximum likelihood estimates without creating and evaluating a likelihood function, sometimes called *Bailey's method.* Three examples of method of moment estimates that are also maximum likelihood estimates are the following:

1. The estimation of a-allele frequency q from a binary classification of dominant (AA and $Aa = A^+$) and recessive (aa) genotypes illustrates a method of moments estimate. The observed frequencies and theoretical probabilities are:

	Genotypes		
Frequency	A^+	aa	Total
Observed proportions	a/n	b/n	1.0
Theoretical probabilities	$1-q^2$	q^2	1.0

The probability distribution that produces the genotypes frequencies has one independent parameter (q), so equating q^2 to the observed frequency (b/n) produces a maximum likelihood estimate using the method of moments estimate of q and is the solution to the equation

$$q^2 = \frac{b}{n} \text{ and the estimate becomes } \hat{q} = \sqrt{\frac{b}{n}}$$

To contrast the method of moments estimation to the maximum likelihood estimate of q, the likelihood function is

$$\text{likelihood function} = L = (1-q^2)^a (q^2)^b$$

and

$$\text{log-likelihood function} = log(L) = a\, log(1-q^2) + 2b\, log(q),$$

then

$$\frac{d}{dq} log(L) = \frac{d}{dq}[a\, log(1-q^2) + 2b\, log(q)] = -2a\left[\frac{q}{1-q^2}\right] + \frac{2b}{q}$$

therefore, when

$$a\hat{q}^2 - b(1-\hat{q}^2) = 0,$$

Table 27.8 *Data and Model: Values Generated by Random Response Survey Techniques*

	Interview response		
Device	Yes	No	Total
"Yes"	$a/n\ [P\pi]$	$b/n\ [(1-P)\pi]$	π
"No"	$c/n\ [P(1-\pi)]$	$d/n\ [(1-P)(1-\pi)]$	$1-\pi$
	P	$1-P$	1.0

then

$$\hat{q}^2 = \frac{b}{a+b} = \frac{b}{n}$$

making

$$\hat{q} = \sqrt{\frac{b}{n}}.$$

The estimate of the variance of the distribution of $\hat{q}$ is

$$variance(\hat{q}) = \frac{1-\hat{q}^2}{4n}.$$

The expression for the estimated variance equally results from using the truncated Taylor series estimated variance or computing the second derivative of the log-likelihood function to produce the estimated variance from Fisher's information function.

2. The probability of a correct response from the frequencies of the four outcomes from a randomized response interview is a bit tedious to estimate with a maximum likelihood approach. The method of moments estimate, however, is a straight-forward solution to a simple equation.

As usual, a 2×2 table of observed and theoretical values contains only one independent data value, therefore, a single equation produces the method of moments estimate of the probability of a correct response (denoted P) that is also the maximum likelihood estimate (Table 27.8). The data values and the theoretical probabilities [in brackets] are displayed in Table 27.8.

Specifically, the observed frequency of agreement equated to the theoretical frequency is

$$\hat{p} = P\pi + (1-P)(1-\pi),$$

where $\hat{p} = \frac{a+d}{n}$ estimates the frequency of agreement from the data. Solving the expression for the value P (true response probability), the maximum likelihood estimate $\hat{P}$ using method of moments estimation is

$$\hat{P} = \frac{\hat{p} - (1-\pi)}{2\pi - 1}.$$

3. A model describing the joint probability distribution of two binary variables (labeled X and X') creates a probability distribution with three outcomes with correlation between X

and X' represented by r. Specifically, these data values and model parameters (in brackets) are:

$X = X' = 1$	$X = 0, X' = 1$ or $X = 1, X' = 0$	$X = X' = 0$	Total
$a/n\,[p^2 + pqr]$	$b/n\,[2pq(1 - r)]$	$c/n\,[q^2 + pqr]$	1.0

where $a + b + c = n =$ the number of pairs (X, X'). Therefore, the solution of two independent equations

$$P_1 = \frac{a}{n} = p^2 + pqr$$

and

$$P_2 = \frac{b}{n} = 2pq(1 - r)$$

produce the method of moments estimates of p and r. Specifically, the solutions to the equations are

$$p = 2P_1 + P_2$$

making

$$\hat{p} = \frac{2a + b}{2n},$$

where as usual $p = 1 - q$ and

$$P_2 = \frac{b}{n} = 2pq(1 - r)$$

making

$$\hat{r} = 1 - \frac{b}{2n\hat{p}\hat{q}}.$$

These method moments estimates of parameters p and r are also the maximum likelihood estimates.

Appendix: R Code

```
#CHAPTER 1 - Distributions

#probability distributions
#normal distribution
z <- seq(-2,2,.5)
round(rbind(z,pnorm(z),1-pnorm(z)),3)

#t-distribution
p <- c(0.990,0.975,0.950,0.900,0.800)
df <- c(2,10,20,40,60)
df #degrees of freedom
qvalue <- function(p,df) {qt(p,df)}
round(cbind(p,outer(p,df,qvalue)),3)

#chi-square distribution
df <- c(1,2,10,30,50,100)
df #degrees of freedom
qvalue <- function(p,df) {qchisq(p,df)}
round(cbind(p,outer(p,df,qvalue)),3)

#f-distribution
P <- matrix(0,5,16)
df1 <- c(rep(c(1,2,10,30),c(4,4,4,4)))
df2 <- c(rep(c(10,20,30,60),4))
qvalue <- function(p,df1,df2) {qf(p,df1,df2)}
for(i in 1:5) {
P[i,] <- qvalue(p[i],df1,df2)
}
round(rbind(c(0,df1),c(0,df2),cbind(p,P)),3)

#relationship 1
a <- c(.1,.05,.025,.01)
z2 <- qnorm(1-a/2)^2
Z2 <- qchisq(1-a,1)
cbind(a,z2,Z2)

#relationship 2
df <- c(10,20,50,100)
F <- qf(1-a,1,df)
T  <- qt(1-a/2,df)
cbind(a,df,F,T^2)

#relationship 3
df0 <- c(5,10,60,100)
```

```
F <- qf(a,df0,df)
F0 <- qf(1-a,df,df0)
cbind(a,df0,df,F,1/F0)

#relationship 4 - approximate
df <- c(10,60,150,200)
x2 <- df+qnorm(1-a)*sqrt(2*df)
X2 <- qchisq(1-a,df)
cbind(a,df,x2,X2)

#relationship 5 - approximate
F <- qf(1-a,df0,df)*df0
X2 <- qchisq(1-a,df0)
cbind(a,df0,df,F,X2)
```

#CHAPTER 2 - Confidence intervals

```
#r-function for confidence interval
ci <- function(m,v) {
A <- m-1.960*sqrt(v)
B <- m+1.960*sqrt(v)
round(cbind(A,B),5)
}

#ci.r
x <- 16
n <- 100
x <- 60
p <- x/n
v <- p*(1-p)/n
ci(p,v)

#function  - bounds (lower, upper)
A <- ci(p,v)[1]
B <- ci(p,v)[2]
rbind(round(c(100*p,100*c(A,B)),0),
round(c(1-p,1-B,1-A),2),
round(c(1/p,1/B,1/A),2),
round(c(log(p),log(A),log(B)),2),
round(c(log(p/(1-p)),log((A/(1-A))),log((B/(1-B)))),2),
round(c(log(1-p),log(1-B),log(1-A)),2))

#examples of confidence intervals with differing accuracy

p <- c(.1,.2,.5)
n <- 30
x <- n*p

#not adjusted
cbind(p-1.960*sqrt(p*(1-p)/n),p+1.960*sqrt(p*(1-p)/n))
#adjusted
cbind(p-1.960*sqrt(p*(1-p)/n)+1/(2*n),p+1.960*sqrt(p*(1-p)/n)-1/(2*n))
#logistic
v <- ((n+1)*(n+2))/(n*(x+1)*(n-x+1))
l <- log(p/(1-p))
a <- l+1.960*sqrt(v)
```

```
b <- l-1.960*sqrt(v)
cbind(1/(1+exp(-b)),1/(1+exp(-a)))
#exact
binom.test(3,n)$conf.int
binom.test(6,n)$conf.int
binom.test(15,n)$conf.int

#median - confidence interval (n = 15)
#data (ordered)
data  <- c(3.04,3.70,3.80,3.81,3.84,3.89,3.98,4.03,
4.10,4.25,4.26,4.32,4.57,4.84,5.18)
#data <- sort(rnorm(3000,10,5)) #test case
median(data)
n <- length(data)
P <- pbinom(0:n,n,0.5)[-n]
rank.data <- rank(data)
round(rbind(P,data,rank.data),3)

#bounds and confidence intervals
L <- max(which(pbinom(0:n,n,0.5)<=0.025))
U <- min(which(pbinom(0:n,n,0.5)>=0.975))
cbind(L,U,data[L],data[U])

#95% confidence interval based on the estimated mean
ci(mean(data),var(data)/n)

#large sample (ordered) - confidence interval (n = 30)
#data (ordered):
x <- c(1.88,2.48,3.02,4.05,6.87,7.82,7.83,8.40,
8.41,8.48,8.78,9.18,9.86,9.90,10.22,10.81,
10.91,11.71,11.96,12.00,12.04,12.49,12.68,13.86,
13.89,14.44,15.09,15.75,16.55,19.25)
x <- sort(rnorm(2000,10,2)) #test case
median(x)
n <- length(x)
L <- round((n+1)/2-1.960*sqrt(n)/2)
U <- round((n+1)/2+1.960*sqrt(n)/2)
cbind(L,U,x[L],x[U])
ci(mean(x),var(x)/n)
```

#CHAPTER 3 - weighted averages

```
ci <- function(m,v) {
A <- m-1.960*sqrt(v)
B <- m+1.960*sqrt(v)
round(cbind(A,B),5)
}

#least squares versus weighted average
x <- c(1,9,8,6,3,1,3,9)
y <- c(3,8,9,7,2,1,6,4)

#least squares estimation
f <- glm(y~x)
b.lsq <-f$coefficient[2]
#weighted average
```

```
b.wt <- sum((x-mean(x))*(y-mean(x)))/sum((x-mean(x))^2)
cbind(b.lsq,b.wt)

#Childhood leukemia  - confidence interval
data<-c(33,4731047,181,18368181,231,23326046,260,23882120,245,23156490)
deaths <- data[seq(1,length(data),2)]
pop <- data[seq(2,length(data),2)]
rate <- deaths/pop
round(cbind(pop,deaths,rate*10^6),2)
lrate <- log(rate)
lbar <- sum(deaths*lrate)/sum(deaths)
rate <- exp(lbar)*10^6
v <- 1/sum(deaths)
cbind(lbar,ci(lbar,v))
cbind(rate,exp(ci(lbar,v))*10^6)

#weighted.r - rate ratio
data <- c(33,4731047,238,36243470,181,18368181,1118,142332392,231,
23326046,1253,179720009,260,23882120,1505,186897294,245,23156490,1950,
185997993)

#Leukemia data California and US
d <- data[seq(1,length(data),4)]
p <- data[seq(2,length(data),4)]
r <- 10^6*d/p
D <- data[seq(3,length(data),4)]
P <- data[seq(4,length(data),4)]
R <- 10^6*D/P
round(cbind(d,p,r,D,P,R),2)

#confidence interval leukemia California and US
rr <- (d/p)/(D/P)
v <- 1/d+1/D
w <- 1/v
lrr <- sum(w*log(rr))/sum(w)
vrr <- 1/sum(w)
rate <- exp(lrr)
cbind(lrr,vrr,ci(lrr,vrr))
cbind(rate,exp(ci(lrr,vrr)))

#summary odds ratio
a <- c(98,54,11,7)
b <- c(832,227,85,102)
c <- c(169,55,61,90)
d <- c(3520,686,926,1936)
n <- a+b+c+d

#Mantel/Haenszel summary odds ratio
or.mh <- sum(a*d/n)/sum(b*c/n)
or.mh

#weighted average summary odds ratio
or <- a*d/(b*c)
lor <- log(or)
v <- 1/a+1/b+1/c+1/d
w <- 1/v
```

```
round(cbind(or,lor,v,w),3)
lorbar <- sum(lor*w)/sum(w)
vlorbar <- 1/sum(w)

#summary odds ratio and 95% confidence interval
cbind(lorbar,vlorbar,ci(lorbar,vlorbar))
exp(c(lorbar,ci(lorbar,vlorbar)))

#SMR - standard mortality ratio
E <- sum(p*R)/10^6
d <- sum(p*r)/10^6
SMR <- d/E
smr <- sum(p*r)/sum(p*R)
cbind(d,E,SMR,smr)
ci(log(smr),1/sum(d)+1/sum(D))
cbind(exp(ci(log(smr),1/sum(d)+1/sum(D))))

#smooth weighted average
x <- c(0.31,1.27,2.23,3.21,4.17,5.13,6.11,7.07,8.03,9.01)
y <- c(174.07,145.70,160.61,82.36,50.93,35.94,44.26,-
21.78,86.27,242.46)

for(i in 1:10) {
wt <- dnorm(x,x[i],1.0)
Y <- sum(wt*y)/sum(wt)
print(Y)
}
```

#CHAPTER 4 - Binomial/Poisson distributions

```
ci <- function(m,v) {
A <- m-1.960*sqrt(v)
B <- m+1.960*sqrt(v)
round(cbind(A,B),5)
}

#binomial
n <- 10
p0 <- 0.5
x <- 6
rbind(0:10,round(dbinom(0:n,n,p0),3))
rbind(0:10,round(1-pbinom(0:n,n,p0),3))

#three estimates of the same probability - P(x >5)
1-pbinom(x,n,p0)
1-pnorm((x-n*p0+.5)/sqrt(n*p0*(1-p0)))  #approximate
1-binom.test(x,n,p=p0,alternative = "less")$p.value  #exact

#note - binomial coefficients
n <- 9
choose(n,0:n)
cbind(sum(choose(n,0:n)),2^n)

#Viral prevalence
k <- 60
x <- 18
n <- 20
```

```
Q <- x/k
p <- 1-Q^(1/n)
V <- Q*(1-Q)/k
v <- (Q^(1/n)/(n*Q))^2*V
cbind(Q,p,v,V)
ci(Q,V)
rev((1-ci(Q,V)^(1/n)))
ci(p,v)

#randomized response
x <- 127
n <- 200
pi <- .3
p <- x/n
P <- (p-(1-pi))/(2*pi-1)
v <- (p*(1-p)/n)/(2*pi-1)^2
cbind(p,P,v)
ci(P,v)

#geometric
k <- 0:8
p <- 0.4
q <- 1-p
round(p*q^k,3)
#expected value (q/p)
k <- 1:80
cbind(sum(k*p*q^k),q/p)

#Poisson - spatial distribution
p <- c(dpois(0:5,2),1-ppois(5,2))
e <- 100*p
o <- c(16,27,23,17,12,2,3)
round(rbind(p,e,o),3)
X2 <- sum((o-e)^2/e)
pvalue <- 1-pchisq(X2,6)
cbind(X2,pvalue)

#goodness-of-fit
k <- 0:6
x <- log(e)+log(factorial(k))
y <- -2+log(2)*k
plot(x,y,type="b",xlab="Log(expected value)",ylab="Log(observed)")
title("observed/theoretical - Poisson distributions ?")
x <- log(o)+log(factorial(k))
lines(x,y,type="b",pch=16,lty=2)

#truncated Poisson distribution
#data
d <- c(2,2,2,2,1,1,2,2,2,3,2,4,1,2,3,2,5,3,1,1,2,2,1,2,2,2,1,2,2,4,3,2,
2,6,1,5,2,4,1,4,2,6,2,2,2,1,5,4,3,4,1,4,3,3,2,3,1,3,1,2,3,4,3,2,3,3,1,
2,5,1,3,1,2,3,3,3,3,4,3,5,1,3,3,3,3,1,4,1,6,1,2,1,3,2,3,2,4,2,4,2,2,2,
1,3,3,3,6,4,3,6,4,2,3,3,3,4)

tab <- table(d)
tab
mean(d)
```

```
N <- sum(d)
S <- tab[1]
n <- length(d)
cbind(N,S,n)
lambda <- (N-S)/n
cbind(N,S,n,lambda)

#maximum likelihood estimation
f <- function(x) {mean(d)-x/(1-exp(-x))}
uniroot(f,c(1,4))$root

#alternative
x <- seq(1,4,0.0001)
F <- abs(f(x))
k <- which.min(F)
x[k]
```

#CHAPTER 5 - Correlation

```
#data
x <- c(144,154,142,152,250,204,164,259,190,175,167,186,153,166,180,251,
168,225,196,160)
y <- c(77,79,78,74,86,89,84,87,76,90,91,92,73,85,80,88,83,81,71,82)
cbind(mean(x),mean(y),sd(x),sd(y),cov(x,y))

cor.test(x,y)
cor.test(x,y)$estimate
cor.test(x,y,method="spearman")$estimate
cor.test(x,y,method="kendall")$estimate

#estimated x/y-slope
f <- lm(y~x)
summary(f)
summary(f)$coefficient[2,]
cor(x,y)*sd(y)/sd(x) #slope

#ranked data
X <- rank(x)
Y <- rank(y)
cbind(mean(X),mean(Y),sd(X),sd(Y),cov(X,Y))
cov(X,Y)/(sd(X)*sd(Y))
cor.test(x,y,method="spearman")

#point serial correlation coefficient
x <- c(225,177,181,132,255,182,155,140,149,325,223,271,238,189,140,247,
220,176,185,202,227,192,179,237,177,239,210,220,274,225)
y <- c(1,1,0,0,0,0,0,1,0,1,1,1,1,0,0,0,0,1,0,1,1,1,0,1,1,1,1,0,1,0)
cbind(mean(x),mean(y),sd(x),sd(y),cov(x,y))
x0 <- x[y==0]
x1 <- x[y==1]
cbind(mean(x0),mean(x1),
sum((x-mean(x))^2),
sum((y-mean(y))^2),
sum((x-mean(x))*(y-mean(y))))
cor(x,y)
```

```
#short-cut expressions for ssy and sxy
n0 <- sum(y)
n1 <-length(y)-n0
n <- n0+n1
cbind(n0*n1/n,sum((y-mean(y))^2))
cbind(n1*n0*(mean(x[y==1])-mean(x[y==0]))/n,
sum((x-mean(x))*(y-mean(y))))

#equivalent tests
cor.test(x,y,continuity=TRUE)$statistic
t.test(x1,x0,var.equal=TRUE)$statistic

#point biserial correlation
r.pb <- sqrt(n0*n1/n)*(mean(x[y==1])-mean(x[y==0]))/sqrt(sum((x-
mean(x))^2))
cbind(r.pb,cor.test(x,y)$estimate)

#nonparametric correlation - gamma-coefficient

#data 1
#artificial example data (n = 4)
x <- c(1,6,5,3)
y <- c(2,5,3,6)

#data 2
#blood pressure data (n = 20)
x <- c(144,154,142,152,250,204,164,259,190,175,167,186,153,166,180,
251,168,225,196,160)
y <- c(77,79,78,74,86,89,84,87,76,90,91,92,73,85,80,88,83,81,71,82)

S <- sign(outer(x,x,"-")*outer(y,y,"-"))
C <- length(S[S>0])
D <- length(S[S<0])
gamma <- (C-D)/(C+D)
cbind(C,D,gamma)
cor.test(x,y,method="kendall")

#proportional reduction in error criterion (3 by 5 table)
#lambda correlation coefficient

x <- c(78,102,103,124,153,81,149,220,32,221,352,48,343,481,35)
m <- matrix(x,3,5)
m
P0 <- 1-max(apply(m,2,sum))/sum(m)
p.row <- 1-apply(m,1,max)/apply(m,1,sum)
P <- sum(apply(m,1,sum)*p.row)/sum(m)
round(p.row,3)
lambda <- (P0-P)/P0
round(cbind(P0,P,lambda),3)
```

#CHAPTER 6 - two by two tables

```
ci <- function(m,v) {
A <- m-1.960*sqrt(v)
B <- m+1.960*sqrt(v)
round(cbind(A,B),5)
}
```

```
#typical summary statistics: odds ratio, relative risk and
attributable risk
a <- 61
b <- 102
c <- 52
d <- 196
matrix(c(a,c,b,d),2,2)
n <- a+b+c+d
n1 <- a+b
n2 <- c+d
m1 <- a+c
m2 <- b+d

or <- (a*d)/(b*c)
rr <- (a/(a+b))/(c/(c+d))
ar <- ((a+c)/n-(c/(c+d)))/((a+c)/n)
cbind(or,rr,ar)

#association measured by probabilities
#rows
prop.test(c(61,52),c(163,248),correct=FALSE)
#columns
prop.test(c(61,102),c(113,298),correct=FALSE)
chisq.test(matrix(c(61,52,102,196),2,2),correct=FALSE)

#Table 7.0 - text
#standard errors
p1 <- a/(a+b)
p2 <- c/(c+d)
P1 <- a/(a+c)
P2 <- b/(b+d)
vp <- p1*(1-p1)/n1+p2*(1-p2)/n2
vP <- P1*(1-P1)/m1+P2*(1-P2)/m2
vrr <- rr^2*(1/a-1/n1+1/c-1/n2)
lvrr <- 1/a-1/n1+1/c-1/n2
vor <- or^2*(1/a+1/b+1/c+1/d)
lvor <- 1/a+1/b+1/c+1/d
var0 <- (1-ar)^2*(b+ar*(a+d))/(n*c)
lvar0 <- (b+ar*(a+d))/(n*c)
round(cbind(sqrt(vp),sqrt(vP),sqrt(vrr),sqrt(lvrr),
sqrt(vor),sqrt(lvor),sqrt(var0),sqrt(lvar0)),3)

#Table 8.0 - text
#estimates
p1 <- a/(a+b)
p2 <- c/(c+d)
p1.p2 <- p1-p2
P1 <- a/(a+c)
P2 <- b/(b+d)
P1.P2 <- P1-P2
rr <- (a/(a+b))/(c/(c+d))
lrr <- log(rr)
or <- (a/b)/(c/d)
lor <- log(or)
ar <- (a*d-b*c)/((a+c)*(c+d))
```

```
lar <- log(1-ar)
round(cbind(p1.p2,P1.P2,rr,lrr,or,lor,ar,lar),3)

#confidence intervals
rbind(round(c(p1-p2,ci(p1-p2,vp)),3),
round(c(P1-P2,ci(P1-P2,vP)),3),
round(c(exp(lrr),exp(ci(lrr,lvrr))),3),
round(c(lrr,ci(lrr,lvrr)),3),
round(c(exp(lor),exp(ci(lor,lvor))),3),
round(c(lor,ci(lor,lvor)),3),
round(c(1-exp(lar),1-exp(rev(ci(lar,lvar0)))),3),
round(c(lar,ci(lar,lvar0)),3))

#corrected chi-square statistic
p <- (a+c)/(a+b+c+d)
v <- p*(1-p)*(1/n1+1/n2)
z <- (abs(p2-p1)-.5*(1/n1+1/n2))/sqrt(v)
cbind(p,v,z^2)
prop.test(c(a,c),c(n1,n2),correct=TRUE)$statistic
n*(abs(a*d-b*c)-n/2)^2/((a+b)*(c+d)*(a+c)*(b+d))

#corrected estimate
p <- c(1,4,6,4,1)/16
ex <- sum((0:4)*p)
c0 <- 1:5
p0 <- cumsum(p)  #exact
p1 <- pnorm(c0-0.5-ex)  #approximate
round(cbind(c0-1,p0,p1),4)

#Vietnam
a <- 170
b <- 3222
c <- 126
d <- 2912
n <- a+b+c+d
matrix(c(a,c,b,d),2,2)
#relative risk
(a/(a+b))/(c/(c+d))
#odds ratio
(a/b)/(c/d)
#variance of log-odds-ratio
cbind(1/a+1/b+1/c+1/d,1/a+1/c)

#Adjustment  #small sample size
a <- 2
b <- 23
c <- 6
d <- 22
n <- a+b+c+d
m <- matrix(c(a,c,b,d),2,2)
chisq.test(m,correct=FALSE)
chisq.test(m)

#Hypergeometric probability distributions

#tea tasting experiment
round(dhyper(0:4,4,4,4),3)
```

```
#keno
round(dhyper(0:8,20,60,8),6)
#Fisher's exact test
round(dhyper(0:8,8,20,8),4)
round(1-phyper(0:8,8,20,8),4)

#Adjusted/unadjusted - Fisher's exact test
a <- 5
b <- 3
c <- 3
d <- 17
n <- a+b+c+d
m <- matrix(c(a,c,b,d),2,2)
chisq.test(m,correct=FALSE)
chisq.test(m)
fisher.test(m)

#twins
#no birth defects (data set 1)
a <- 18687
b <- 16093
c <- 19188
#birth defects  (data set 2)
a <- 168
b <- 53
c <- 110
n <- a+b+c
p <- (2*a+b)/(2*n)
q <- 1-p
r <- 1-b/(2*p*q*n)
vr <- ((1-r)*(1-2*p*q*(1-r)-(1-4*p*q)*(1-r)^2))/(2*p*q*n)
vp <- (2*p*q*(1+r))/(2*n)
cbind(n,p,r,vr,vp)
ci(r,vr)

#duffy
a <- 8
b <- 72
c <- 409
n <- a+b+c
p <- (2*a+b)/(2*n)
q <- 1-p
chisq.test(c(a,b,c),p=c(p^2,2*p*q,q^2))
r <- 1-b/(2*p*q*n)
X2 <- n*r^2
pvalue <- 1-pchisq(X2,1)
cbind(p,r,X2,pvalue)  #degrees of freedom = 1

#incomplete data - zero in a 2 by 2 table
a <- 33
b <- 44
c <- 14
matrix(c(a,c,b,"d=?"),2,2)
n <- a+b+c
d <- b*c/a
```

```
#estimates - two versions
round(cbind(n+d,(a+b)*(a+c)/a),3)

#false positive - denoted p
d <- c(0.5,0.1,0.05,0.005,0.0005)
t <- (0.95*d+0.1*(1-d))
p <- 0.1*(1-d)/t
round(cbind(d,t,p),3)
```

#CHAPTER 8 - Contingency tables

```
#data
wilcox.test(c(4,35,21,28,66),c(10,42,71,77,90))
U <- wilcox.test(c(4,35,21,28,66),c(10,42,71,77,90))$statistic
n1 <- 5
W <- U+n1*(n1+1)/2
P <- U/(n1*n1)
cbind(n1,U,W,P)

#A/B behavior and cholesterol
x <- c(344,0,246,0,224,1,242,0,252,1,233,1,224,0,
239,1,252,0,202,1,291,1,212,0,239,1,153,0,218,1,
312,1,188,0,254,1,183,0,202,0,185,0,250,0,169,0,
234,1,212,1,250,1,197,1,226,0,137,0,325,1,263,0,
148,0,175,0,181,1,194,0,246,1,268,1,276,1,248,1,213,0)

chol <-x[seq(1,length(x),2)]
type <-x[seq(2,length(x),2)]
t.test(chol[type==0],chol[type==1])
t.test(chol[type==0],chol[type==1],var.equal=TRUE)
wilcox.test(chol[type==1],chol[type==0])
wilcox.test(chol[type==1],chol[type==0],correct=FALSE)
U <- wilcox.test(chol[type==1],chol[type==0])$statistic
n <- length(type)
n1 <- sum(type)
n2 <- n-n1
W <- U+n1*(n1+1)/2
P <- U/(n1*n2)
v <- (n+1)/(12*n1*n2)
z <- (P-0.5)/sqrt(v)
pvalue <- 2*(1-pnorm(z))
cbind(W,U,P,v,z,pvalue)
#check
sum(rank(chol)[type==0])
sum(rank(chol)[type==1])
U+n1*(n1+1)/2
sum(rank(chol)[type==0])+sum(rank(chol)[type==1])
n*(n+1)/2

#analysis 2 by k table - Mann/Whitney
x <- c(7,25,7,34,8,32,20,32)
m <- matrix(x,2,4)
m
y <- rep(1:4,m[1,])
n <- rep(1:4,m[2,])
```

```
M <- outer(y,n,"-")
U <- sum(table(M[M>0]))+length(M[M==0])/2
N <- length(M)
P <- U/N
cbind(U,N,P)

#example data
n1 <- c(18,16,24,24,52)
n2 <- c(20,24,60,35,50)
x <- 0:4

#body weight data (data set 1)
#n1 <- c(25,24,20,46,45,50,47)
#n2 <- c(417,426,327,493,350,398,485)
#x <- 0:6

#x-ray data (data set 2)
n1 <- c(7332,287,199,96,59,65)
n2 <- c(7673,239,154,65,28,29)
x <- 0:5

#analysis
n <- n1+n2
xbar1 <- sum(x*n1)/sum(n1)
xbar2 <- sum(x*n2)/sum(n2)
xbar <- (sum(n1)*xbar1+sum(n2)*xbar2)/sum(n)
cbind(xbar1,xbar2,xbar)
sxx <- sum(n*(x-xbar)^2)
N1 <- sum(n1)
N2 <- sum(n2)
v <- (sxx/(N1+N2))*(1/N1+1/N2)
z <- (xbar1-xbar2)/sqrt(v)
cbind(z^2,1-pchisq(z^2,1))
syy <- sum(n1)*sum(n2)/sum(n)
sxy <- (xbar1-xbar2)*syy
cbind(syy,sxx,sxy)
b <- sxy/sxx
P <- sum(n1)/sum(n)
a <- P-b*xbar
cbind(a,b,P)
p <- n1/n
pp <- a+b*x
round(rbind(p,pp),3)
X2 <- sum(n*(p-P)^2)/(P*(1-P))
XL <- sum(n*(pp-P)^2)/(P*(1-P))
XNL <- sum(n*(p-pp)^2)/(P*(1-P))
cbind(X2,XL,XNL)

#Comparison of mean values
#example data
d1 <- c(18,16,24,24,52)
d2 <- c(20,24,60,35,50)
x <- 0:4

#body weight data (data set 1)
d1 <- c(25,24,20,46,45,50,47)
```

```
d2 <- c(417,426,327,493,350,398,485)
x <- 0:6

#maternal x-ray data (data set 2)
d1 <- c(7332,287,199,96,59,65)
d2 <- c(7673,239,154,65,28,29)
x <- 0:5

#analysis of mean values
n <- d1+d2
xbar1 <- sum(x*d1)/sum(d1)
xbar2 <- sum(x*d2)/sum(d2)
xbar <- sum(x*(d1+d2))/sum(n)
cbind(xbar1,xbar2,xbar)
sxx <- sum(n*(x-xbar)^2)
v <- (sxx/sum(n))*(1/sum(d1)+1/sum(d2))
z <- (xbar1-xbar2)/sqrt(v)
cbind(z,z^2)
```

#CHAPTER 9 – Poisson distribution

```
ci <- function(m,v) {
A <- m-1.960*sqrt(v)
B <- m+1.960*sqrt(v)
round(cbind(A,B),5)
}

#autism – data
cases <- c(202,255,333,442,571,716,803,996,1323)
births <- c(504853,534174,570976,613336,610701,602269,585761,568263,
577113)

year <- 1987:1995
f <- glm(cases~year+offset(log(births)),family=poisson)
summary(f)
rates <- cases*10000/births
RATES <- fitted(f)*10000/births
CASES <- fitted(f)
round(cbind(year,rates,RATES,cases,CASES),2)
X2 <- sum((cases-CASES)^2/CASES)
pvalue <- 1-pchisq(X2,7)
round(cbind(X2,pvalue),3)
round(cbind(RATES[1],RATES[9],RATES[9]/RATES[1]),3)
round(cbind(rates[1],rates[9],rates[9]/rates[1]),3)

#analysis – stomach cancer 45 to 85+
#data
deaths.m <- c(84,136,131,154,76)
pop.m <- c(2014272,1417097,775192,466018,161990)
base <-100000
rate.m <- base*deaths.m/pop.m
deaths.f <- c(52,69,77,121,92)
pop.f <- c(1967034,1474064,891864,642748,325940)
rate.f <- base*deaths.f/pop.f
round(cbind(deaths.m,pop.m,rate.m,deaths.f,pop.f,rate.f,rate.f/rate.m),
3)
```

```
#adjusted comparison
rate.ff <- sum(pop.f*rate.f)/sum(pop.f)
rate.mm <- sum(pop.f*rate.m)/sum(pop.f)
round(cbind(rate.ff,rate.mm,rate.ff/rate.mm),3)

#poisson model
sex <- sort(rep(0:1,5))
age <- c(45,55,65,75,85,45,55,65,75,85)
deaths <- c(deaths.m,deaths.f)
pop <- c(pop.m,pop.f)
cbind(deaths,pop,sex,age)
f <- glm(deaths~age+sex+offset(log(pop)),family=poisson)
summary(f)
round(rbind(deaths,fitted(f)),1)

#ratios
#stomach cancer two young age groups
deaths <- c(44,162,44,135,22,44,12,58)
pop <- c(16569463,17170525,16652229,18283791,2829377,2644142,2981382,
2975071)
base <- 1000000
rate <- base*deaths/pop
age <- c(0,1,0,1,0,1,0,1)
sex <- c(0,0,1,1,0,0,1,1)
race <- c(0,0,0,0,1,1,1,1)
round(cbind(deaths,pop,rate,age,sex,race),2)

#stomach cancer –age,sex and race
f <- glm(deaths~age+sex+race+offset(log(pop)),family=poisson)
summary(f)
DEATHS <- fitted(f)
RATES <- base*DEATHS/pop
#estimates
round(rbind(DEATHS,RATES),1)
b <- summary(f)$coefficients[,1]
v <- (summary(f)$coefficients[,2])^2
round(c(1/122+1/399,1/272+1/249,1/136+1/385),7)
round(v[-1],7)

#log-variances, ratios and confidence intervals
cbind(round(v[-1],3),round(exp(b[-1]),3),round(exp(ci(b[-1],v[-1])),3))

#smoking pattern
#data
m <- matrix(c(98,1554,70,735,50,355,39,253),2,4)
m
smk <- c(0,10,20,30)
n <- m[1,]+m[2,]
p <- m[1,]/n
round(cbind(smk,p),3)
f <- glm(p~smk,weights=n,family=poisson)
summary(f)
round(rbind(log(fitted(f)),fitted(f),p),3)
#risk = 10 cigarettes/day
rr <- exp(10*coef(f)[2])
rr
```

```
#birth defects - vitamin/ethnicity
x <- c(50,84,48,14,18,16,22,30,17,39,32,41)
d <- matrix(x,3,4)
d
chisq.test(d)
e <- outer(apply(d,1,sum),apply(d,2,sum))/sum(d)
round(e,1)
round(log(d),2)
round(log(e),2)

#detailed plot
matplot(1:3,log(d),type="l",xaxt="n",xlab="Categories (vitamin use)",
ylab="Log-frequencies",cex.lab=1.2)
matlines(1:3,log(e),lty=1)
title("Ethnicity by vitamin use")
text(1:3,c(2.6,2.6),"|")
text(c(2.5,2.5,2.5),c(4.4,3.8,3.3,3.0),c("white",
"Asian","Hispanic","African-American"))
text(c(1.15,2.15,2.85),c(2.6,2.6,2.6),c("always","during","never"))
legend(1,4.4,c("model","data"),lty=c(1,3),cex=0.75)

#analysis of incomplete table - mother/daughter education
m <- matrix(c(84,142,43,100,106,48,67,77,92),3,3)
m
chisq.test(m)
rr <- factor(rep(1:3,3))
cc <- factor(sort(rr))
cbind(as.vector(m),rr,cc)
f <- glm(as.vector(m)~rr+cc,family=poisson)
summary(f)
#model estimated counts
round(matrix(fitted(f),3,3),3)
#data estimated counts
m0 <- outer(apply(m,1,sum),apply(m,2,sum))/sum(m)
round(m0,3)

#quasi independence
M <- c(142,43,100,48,67,77)
rr <- factor(c(2,3,1,3,1,2))
cc <- factor(c(1,1,2,2,3,3))
cbind(M,rr,cc)
f <- glm(M~rr+cc,family=poisson)
summary(f)
round(rbind(M,fitted(f)),3)
X2 <- sum((M-fitted(f))^2/fitted(f))
pvalue <- 1-pchisq(X2,1)
round(cbind(X2,pvalue),3)
```

#CHAPTER 10 - Analysis of tables

```
#birthday data
x <- 148
n <- 348
((x/n-0.5)/sqrt(.25/n))^2
```

```
prop.test(x,n,correct=FALSE)
chisq.test(c(x,n-x))$statistic

#birthday - 12 months
x <- c(25,23,25,25,27,23,42,40,32,35,30,21)
p <- x/n
P <- rep(1/12,12)
round(rbind(x,p,P),3)
G2 <- 2*sum(x*log(p/P))
cbind(G2,1-pchisq(G2,11))
chisq.test(x,p=rep(1/12,12))

#loglikelihood
x <- c(16,12,20,28,24)
p0 <- x/sum(x)
L0 <- sum(x*log(p0))
p1 <- c(0.2,0.2,0.2,0.2,0.2)
L1 <- sum(x*log(p1))
p2 <- c(0.14,0.17,0.20,0.23,0.26)
L2 <- sum(x*log(p2))
p3 <- c(0.14,0.14,0.20,0.26,0.26)
L3 <- sum(x*log(p3))
cbind(L0,L1,L2,L3)

#loglikelihood test statistics
cbind(2*(L0-L1),2*(L0-L2),2*(L0-L3))

#vitamin C
x <- c(42,27,87,48,29,4)
n <- sum(x)
m <- matrix(x,2,3)
m
e <- outer(apply(m,1,sum),apply(m,2,sum),"*")/n
e
p <- as.vector(m/n)
P <- as.vector(e/n)
round(rbind(p,P),3)
G2 <-2*sum(x*log(p/P))
cbind(G2,2*(1-pchisq(G2,2)))
chisq.test(m)

#mean polish - parity by education
x <- c(26.35,25.36,22.72,25.57,27.49,26.43,26.92,
28.10,23.58,28.17,26.84,26.10)
m <- t(matrix(x,3,4))
m
#model
rr <- factor(rep(1:4,3))
cc <- factor(sort(rep(1:3,4)))
M0 <- as.vector(m)
cbind(M0,rr,cc)
f <- glm(M0~rr+cc)
summary(f)
M <- matrix(fitted(f),4,3)
M
```

```
#nonparametric – mean polish
mtemp <- sweep(m,1,apply(m,1,mean),"-")
round(mtemp,2)
mm <- sweep(mtemp,2,apply(mtemp,2,mean),"-")
round(mm,2)
m.add <- m-mm
m.add

#parity by maternal age
x <- c(3.281,3.305,3.291,3.258,3.225,3.202,3.278,
3.354,3.371,3.356,3.322,3.303,3.280,3.360,
3.397,3.390,3.374,3.335,3.220,3.345,3.405,
3.422,3.412,3.392,3.202,3.332,3.399,3.434,
3.435,3.417,3.333,3.315,3.398,3.443,3.467,3.482)
m <- t(matrix(x,6,6))
m

#nonparametric – mean polish
mtemp <- sweep(m,1,apply(m,1,mean),"-")
mm <- sweep(mtemp,2,apply(mtemp,2,mean),"-")
#standardized
v <- sum(mm^2)/25
cbind(v,sqrt(v))
round(mm/sqrt(v),3)
ifelse(mm<0,"-","+")

#matched pairs
b <- 225
c <- 132
X2 <- (b-c)^2/(b+c)
X2
prop.test(b,n,correct=FALSE)

#matched – 5 by 5 table
x <- c(11,15,8,8,6,20,20,12,9,16,12,11,12,8,12,
11,14,13,10,38,28,37,33,46,92)
d <- matrix(x,5,5)
d
#expected
e <- (d+t(d))/2
e
#analysis
X2 <- sum((d-e)^2/e)
X2
pvalue <- 1-pchisq(X2,10)
cbind(x2,pvalue)
#alternative version
x2 <- sum((d-t(d))^2/(d+t(d)))/2
x2
pvalue <- 1-pchisq(X2,10)
cbind(x2,pvalue)

#smoking and chd
#data
x <- c(29,155,21,76,7,45,12,43)
```

```
m <- matrix(x,2,4)
m
#expected values - independence
e <- outer(apply(m,1,sum),apply(m,2,sum),"*")/sum(m)
e
#model
rr <- factor(c(1,2,1,2,1,2,1,2))
cc <- factor(sort(c(1,2,3,4,1,2,3,4)))
cbind(x,rr,cc)
f <- glm(x~rr+cc,family=poisson)
summary(f)
#expected values - model and chi-square
matrix(fitted(f),2,4)
p <- as.vector(m/sum(m))
P <- as.vector(fitted(f)/sum(m))
n <- as.vector(m)
round(rbind(n,p,P),3)
G2 <- 2*sum(n*log(p/P))
cbind(G2,1-pchisq(G2,3))

#nonparametric - mean polish
m <- log(m)
mtemp <- sweep(m,1,apply(m,1,median),"-")
mm <- sweep(mtemp,2,apply(mtemp,2,median),"-")
m-mm
exp(m-mm)
#note
chisq.test(exp(m-mm))$statistic  #independent?
```

#CHAPTER 11 - Bootstrap estimation

```
#smoking intervention trial
#data
set.seed(777)
before <- c(131,154,45,89,38,104,127,129,148,181,122,
120,66,122,115,113,163,67,156,191)
after <- c(111,13,16,35,23,115,87,120,122,228,118,121,
20,153,139,103,124,37,163,246)
P <- 100*(before-after)/before
summary(P)
round(rbind(before,after,P),1)

t.test(before,after,paired=TRUE,var.equal=TRUE)

#bootstrap estimation - intervention trial
iter <- 2000
PBAR <- NULL
P <- 100*(before-after)/before
for(i in 1:iter) {
ptemp <- sample(P,length(P),replace=TRUE)
PBAR[i] <- mean(ptemp)
}
pbar <- mean(PBAR)
vbar <- var(PBAR)
```

```
sd <- sqrt(vbar)
cbind(pbar,vbar,sd)
hist(PBAR,50)

#bootstrap test statistics
pvalue.1 <- pnorm((0-pbar)/sd)  #parametric
pvalue.0 <- sum(ifelse(PBAR<0,1,0))/iter  #nonparametric
cbind(pvalue.0,pvalue.1)

#bootstrap confidence intervals
cbind(pbar-1.96*sd,pbar+1.96*sd)  #parametric
cbind(sort(PBAR)[0.025*iter],sort(PBAR)[0.975*iter])  #nonparametric

#example bootstrap for tabled data
iter <- 2000
p <- NULL
d <- rep(1:3,c(10,5,5))
table(d)
n <- length(d)
for(i in 1:iter) {
D <- sample(d,n,replace=TRUE)
P <- table(factor(D,level=1:3))
p[i] <- (P[1]-2*P[2]+P[3])/n
}
cbind(mean(p),var(p))

#Kappa statistic – a bootstrap estimate
a <- 105
b <- 24
c <- 18
d <- 165
n <- a+b+c+d
data <- rep(1:4,c(a,b,c,d))
table(data)
iter <- 2000
K <- NULL
for (i in 1:iter ) {
temp <- sample(data,n,replace=TRUE)
tab <- table(temp)
P <- (tab[1]+tab[4])/n
q1 <- (tab[1]+tab[2])/n
q2 <- 1-q1
p1 <- (tab[1]+tab[3])/n
p2 <- 1-p1
p <- p1*q1+p2*q2
K[i] <- (P-p)/(1-p)
}
m <- mean(K)
v <- var(K)
cbind(m,v)
#parametric - confidence interval
lower <- m-1.96*sqrt(v)
upper <- m+1.96*sqrt(v)
cbind(lower,m,upper)
#nonparametric - confidence interval
```

```
LOWER <- sort(K)[iter*0.025]
UPPER <- sort(K)[iter*0.975]
cbind(LOWER,m,UPPER)

hist(K,50)
```

#CHAPTER 13 – Variance

```
#confidence interval – estimated variance
#data
x <- c(10.92,9.20,11.02,9.20,13.48,11.24,10.41,
12.22,9.59,9.24)
mean(x)
var(x)
df <- length(x)-1
lower <- df*var(x)/qchisq(0.975,df)
upper <- df*var(x)/qchisq(0.025,df)
cbind(lower,mean(x),upper)

#test of variance – Poisson?
#data
x <- c(1,3,1,0,1,0,0,1,0,2,2,0,0,0,1,1,0,1,2,2,0,
1,2,1,1,0,2,2,1,0,2,1,1,1,1,0,0,1,1,3,2,
0,0,4,3,1,0,1,2,2,1,2,1,1,0,0,1,1,2,2)
tab <- table(x)
tab
X2 <- sum((x-mean(x))^2/mean(x))
pvalue <- 1-pchisq(X2,59)
cbind(mean(x),var(x),X2,pvalue)

#birth weight and birth order – analysis of variance
wt <- c(2.48,2.92,3.73,4.16,3.80,3.42,3.62,3.82,3.92,3.34,3.26,
3.87,2.92,3.20,4.10,4.06,3.35,3.40,4.48,4.00,3.42,2.76,
2.98,4.58,4.23,4.26,3.66,4.13,2.80,3.62,3.40,3.70,3.10,
3.56,4.08,2.38,3.74,3.56,3.62,3.52)
order <- c(0,1,3,1,1,2,0,2,2,1,0,1,1,0,1,0,0,1,0,1,1,0,0,2,3,0,0,3,0,1,
1,0,0,0,0,1,0,0,0,1)
#summary statistics
tapply(wt,order,length)
tapply(wt,order,mean)
tapply(wt,order,var)
cbind(length(wt),mean(wt),var(wt))
#analysis of variance
summary(aov(wt~factor(order)))

#five groups...homogeneity/heterogeneity
x <- c(1,5,14,24,72)
n <- c(10,25,35,40,90)
p <- x/n
v <- n*p*(1-p)
N <- sum(n)
P <- sum(x)/N
round(cbind(x,n,p,v),2)
cbind(sum(x),sum(n),sum(n*p)/N,N*P*(1-P))
t <- N*P*(1-P)
```

```
w <- sum(n*p*(1-p))
b <- sum(n*(p-P)^2)
cbind(t,w,b,w+b)

#tests of variances
#f-test
chol <- c(137,148,153,169,175,181,183,185,188,194,197,
202,202,212,212,213,218,224,224,226,233,234,
239,239,242,246,246,248,250,250,252,252,254,
263,268,276,291,312,325,344)
ab <- c(0,0,0,0,0,1,0,0,0,0,1,1,0,0,1,0,1,1,0,0,1,
1,1,1,0,0,1,1,0,1,1,0,1,0,1,1,1,1,1,0)
ranks <- c(1,4,5,8,9,12,13,16,17,20,21,24,25,28,
29,32,33,36,37,40,39,38,35,34,31,30,27,26,23,22,19,18,
15,14,11,10,7,6,3,2)

tapply(chol,ab,length)
tapply(chol,ab,mean)
tapply(chol,ab,var)
var.test(chol[ab==0],chol[ab==1])

#Bartlett test
bartlett.test(chol,ab)
n <- 40
n0 <- 20
n1 <- 20
v.0 <- var(chol[ab==0])
v.1 <- var(chol[ab==1])
v.w <- ((n0-1)*v.0+(n1-1)*v.1)/(n-2)
X2 <- (n-2)*log(v.w)-((n0-1)*log(v.0)+(n1-1)*log(v.1))
#c = correction factor
c <- 1+(1/3)*(2/19-1/38)
X2.c <- X2/c
pvalue <- 1-pchisq(X2.c,1)
#summary
round(cbind(v.w,v.0,v.1,X2,X2.c,pvalue),3)

#Levene test
y.a <- abs(chol[ab==1]-mean(chol[ab==1]))
y.b <- abs(chol[ab==0]-mean(chol[ab==0]))
cbind(mean(y.a),mean(y.b))
t.test(y.b,y.a)

#Siegel/Tukey
W <- sum(ranks[ab==1])
n1 <- length(ranks[ab==1])
W0 <- n1*(2*n1+1)/2
v0 <- n1*n1*(2*n1+1)/12
X2 <- ((W-W0)/sqrt(v0))^2
pvalue <- 1-pchisq(X2,1)
round(cbind(W,W0,v0,X2,pvalue),3)
```

#CHAPTER 14 - Log-normal distribution

```
ci <- function(m,v) {
A <- m-1.960*sqrt(v)
B <- m+1.960*sqrt(v)
```

```
round(cbind(A,B),5)
}

#example (Figure 1.0) -  P(Y>4.48) = P(X>1.5)
1-plnorm(4.48169,1,sqrt(0.7))
1-plnorm(exp(1.5),1,sqrt(0.7))
1-pnorm(1.5,1,sqrt(0.7))

y <- rlnorm(1000,0.5,1.0)
hist(y,50,freq=FALSE)
lines(density(y))
hist(log(y),50,freq=FALSE)
lines(density(log(y)))

#simulated lognorm data - n=30 mean=0.5 and variance=1.0
y <- c(2.69,1.11,2.75,1.11,8.49,3.07,2.02,5.00,1.34,
1.13,1.22,1.74,0.25,1.59,16.63,4.36,4.33,0.96,3.23,
2.72,0.22,2.07,0.75,5.89,6.94,2.53,0.29,1.61,0.37,0.96)

#more simulated lognormal data - mean = 0.5 and variance = 1
n <- 100000
y <- rlnorm(n,0.5,1)

x <- log(y)
n <- length(x)
m <- mean(x)
v <- var(x)
cbind(n,m,v)
ci(m,v/n)

#percentiles
c <- seq(0.2,0.9,0.1)
round(cbind(c,exp(m+qnorm(c)*sqrt(v))),3)

#mean variance | mean, variance, median and mode
round(cbind(m,v,exp(m+0.5*v),
exp(2*m+v)*(exp(v)-1),
exp(m),
exp(m-sqrt(v))),3)
```

#CHAPTER 15 - Nonparametric

```
#smoking intervention trial data
before <- c(131,154,45,89,38,104,127,129,148,181,122,
120,66,122,115,113,163,67,156,191)
after  <- c(111,13,16,35,23,115,87,120,122,228,118,
121,20,153,139,103,124,37,163,246)

xbar0 <- mean(before)
xbar1 <- mean(after)
n <- length(before)
r <- cor(before,after)
round(cbind(n,xbar0,xbar1,xbar0-xbar1,r),3)
t.test(before-after,correct=F)
mean(before-after)/sqrt(var(before-after)/n)

S <- sum(sign(before-after)==1)
p <- S/n
cbind(S,S/n)
```

```
#Sign rank test
#exact
round(1-pbinom(S-1,20,0.5),3)
binom.test(S-1,20,alternative="greater")

#approximate
z <- (p-0.5-.5*(1/n))/(0.5*sqrt(1/n))
cbind(z,2*(1-pnorm(z)))

#statistical power - signed rank test
p0 <- 0.5
p <- c(0.5,0.6,0.7,0.8,0.9)
for(n in c(20,30,40,50)){
c <- p0+1.645*0.5*sqrt(1/n)
z <- (c-p)/(0.5*sqrt(1/n))
print(round(c,3))
print(round(1-pnorm(z),2))
}

#Wilcox signed test
wilcox.test(c(17,3,4,5),c(10,8,6,1),paired=T,alternative="greater",
exact=T)

#smoking intervention trial data
wilcox.test(before,after,paired=T,alternative="greater",exact=T)
n <- 20
ex <- n*(n+1)/4
v <- n*(n+1)*(2*n+1)/24
W <- wilcox.test(before,after,paired=T,alternative="greater",exact=T)
$statistic
z <- (W-ex)/sqrt(v)
x2 <- z^2
pvalue <- (1-pchisq(x2,1))/2
cbind(n,ex,v,W,z,x2,pvalue)

#Kruskal-Wallis
x <- c(2.24,1.59,0.82,9.82,8.55,7.04,5.40,3.73,5.44,
6.58,7.43,8.14,8.76,2.88,4.51,3.05,1.69,0.57,
0.00,7.95,8.13,8.27,9.62,8.55,7.41,6.23,5.04,
3.84,2.76,4.47,11.91,11.52,11.08,10.50,9.75,
2.08,1.09,0.30,0.01,0.58,2.10,3.99,5.62,6.79)

n <- c(14,8,8,14)
id <- rep(1:4,n)
tapply(x,id,length)
tapply(x,id,mean)
tapply(x,id,var)
cbind(sum(n),mean(x),var(x))
summary(aov(x~factor(id)))

#analysis of variance
kruskal.test(x,id)
ranks <- rank(x)
tapply(ranks,id,length)
rbar <- tapply(ranks,id,mean)
rbar
```

```
tapply(ranks,id,var)
summary(aov(ranks~factor(id)))

#nonparametric
N <- length(ranks)
B <- sum(n*(rbar-(N+1)/2)^2)
T <- N*(N+1)*(N-1)/12
X2 <- (N-1)*B/T
pvalue <- 1-pchisq(X2,3)
cbind(T,B,X2,pvalue)

#Three-group regression
xx <- c(133,196,175,205,165,145,198,120,151,180,155,
210,191,146,164,170,157,162,160,135,148)
yy <- c(108,120,128,130,120,104,136,134,110,137,120,
154,125,146,130,148,126,136,127,120,128)

#nonparametric
x <- sort(xx)
m.x <- c(x[4],x[11],x[18])
y <- sort(yy)
m.y <- c(y[4],y[11],y[18])
rbind(m.x,m.y)
B <- (m.y[3]-m.y[1])/(m.x[3]-m.x[1])
A <- sum(m.y)/3-B*sum(m.x)/3
cbind(A,B)
#parametric
summary(lm(yy~xx))

#Tukey quick test
#example 1
A <- c(0.2,1.5,2.1,3.0,4.0,4.2,4.4,4.8,5.3,6.7)
B <- c(3.1,3.4,3.7,4.1,5.3,5.8,6.3,6.6,7.8,7.9)
#A contains the minimum and B contains the maximum
t1 <- sum(ifelse(A<min(B),1,0))
t2 <- sum(ifelse(B>max(A),1,0))
t <- t1+t2
cbind(t1,t2,t)

#example 2
B <- c(132,140,149,155,172,179,181,182,185,189,220,
220,225,247,254,255)
A <- c(176,177,177,180,192,202,209,210,211,223,227,
237,238,239,242,271,274,325)

ifelse(B<min(A),1,0)
t1 <- sum(ifelse(B<min(A),1,0))
ifelse(A<max(B),0,1)
t2 <- sum(ifelse(A>max(B),1,0))
t <- t1+t2
cbind(t1,t2,t)

#Friedman test
x <- c(2.3,4.3,5.3,1.6,1.9,6.9,3.5,1.2,0.2)
m <- matrix(x,3,3)
m
```

```
friedman.test(m)
M <- t(apply(m,1,rank))
M
rbar <- apply(M,2,mean)
s <- apply(M,2,sum)
rbind(s,rbar)
S <- 3*sum((rbar-mean(rbar))^2)
S

#age by parity - 6 by 6 table
x <- c(3.281,3.278,3.280,3.220,3.202,3.333,3.305,
3.354,3.360,3.345,3.332,3.315,3.291,3.371,3.397,
3.405,3.399,3.398,3.258,3.356,3.390,3.422,3.434,
3.443,3.225,3.322,3.374,3.412,3.435,3.467,3.202,
3.303,3.335,3.392,3.417,3.482)

m <- t(matrix(x,6,6))
m
M <- t(apply(m,1,rank))
M
friedman.test(m)

rbar <- apply(M,2,mean)
sums <- apply(M,2,sum)
rbind(sums,rbar)
M <- apply(m,1,rank)
R <- apply(M,1,mean)
R
r <- c <- 6
B <- r*sum((R-mean(R))^2)
T <- r*(c-1)*c*(c+1)/12
S <- r*(c-1)*B/T
cbind(B,T,S)

#k by 2 table
#binomial test
r <- 10
c <- 2
prop.test(c,r,correct=FALSE)

#binomial
c1 <- c(1,1,1,2,1,1,2,1,1,1)
c2 <- c(2,2,2,1,2,2,1,2,2,2)
p <- sum(ifelse(c1==2,1,0))/r
x2 <- ((p-.5)/sqrt(1/(4*r)))^2
x2
friedman.test(cbind(c1,c2))
```

#CHAPTER 16 - Rates

```
ci <- function(m,v) {
A <- m-1.960*sqrt(v)
B <- m+1.960*sqrt(v)
cbind(A,B)
}
```

```
#confidence intervals
d <- 234
pop <- 11139194
r <- d/pop
ci(log(r),1/d)
exp(ci(log(r),1/d))*100000

d0 <- 549
pop0 <- 63818574
r0 <- d0/pop0
rr <- r/r0
cbind(10^5*r,10^5*r0,rr)
v <- 1/d+1/d0
ci(log(rr),v)
exp(ci(log(rr),v))

#Mens Health Study data
#data: non-smokers
smk0 <- c(2,42,27,22,26,16,31,37,15,30,12,5,80,29,13,1,14)
c.smk0 <- c(0,0,0,1,0,1,1,1,1,1,0,1,1,1,1,1,1)

#data: non-smokers
d0 <- sum(c.smk0)
r0 <- d0/sum(smk0)
v0 <- 1/d0
cbind(r0,log(r0),v0)
#ci – log-rate
ci(log(r0),v0)
#ci – rate
exp(ci(log(r0),v0))
rev(1/exp(ci(log(r0),v0)))
#mean survival time
1/r0
#median
m0 <- log(2)/r0
m0
#ci – log-median
ci(log(m0),v0)
#ci – median
exp(ci(log(m0),v0))
rev(log(2)/exp(ci(log(r0),v0)))

#data: smokers
smk <- c(21,13,17,8,23,18)
c.smk <- c(0,1,1,1,1,1)

#smokers
d <- sum(c.smk)
r <- d/sum(smk)
v <- 1/d
cbind(r,log(r),v)
#mean survival time
1/r
#median
m <- log(2)/r
```

```
m
#ci – log-rate
ci(log(r),v)
#ci – rate
exp(ci(log(r),v))
rev(1/exp(ci(log(r),v)))
#ci – log-median
ci(log(m),v)
#ci – median
exp(ci(log(m),v))
rev(log(2)/exp(ci(log(r),v)))
```

#CHAPTER 17 - Nonparimetric survival analysis

```
library(survival)

#data
#survival probability distribution (no censoring)
t <- c(3,9,12,15,18,35,48,51,55,68)
n <- length(t)
cc <- rep(1,n)
f <- survfit(Surv(t,cc)~1)
summary(f)
plot(f,xlab="survival time",ylab="survival probability")

#survival probability distribution (censoring)
#data
t <- c(3,9,12,15,18,35,48,51,55,68)
n <- length(t)
cc <- c(1,0,1,1,0,1,0,1,1,1)
f <- survfit(Surv(t,cc)~1)
summary(f)
plot(f,xlab="survival time",ylab="survival probability")

#variance
#complete data
n <- c(10,9,8,7,6,5,4,3,2)
t <- c(6,3,3,3,17,13,3,4,13)

#censored data
n <- c(10,8,7,5,3,2)
t <- c(9,3,20,16,4,13)

#estimation
P <- cumprod(1-1/n)
a <- P*t
nn <- n*(n-1)
A <- rev(cumsum(rev(a)))
AA <- A^2/nn
d <- length(n)+1
v <- (d/(d-1))*sum(AA)
cbind(P,t,a,nn,A,AA)
tbar <- sum(c(1,P)*c(3,t))
cbind(tbar,v)

#complete: calculated directly from the data
s <- c(3,9,12,15,18,35,48,51,55,68)
```

```
mean(s)
var(s)/10

#Cumulative hazard function
t <- c(3,12,15,35,51,55,68)
n <- c(10,8,7,5,3,2,1)
p <- 1-1/n
H <- cumsum(-log(p))
round(cbind(t,n,p,H),3)
plot(t,H,type="s",ylim=c(0,2),xlim=c(0,68),xlab="Time",ylab=
"Cumulative hazard function")
title("cumulative hazard")

#log-rank analysis
#data - SFMHS - smokers (n = 8) and nonsmokers (n = 17)
smk <- c(2,8,13,24,25,38,43,48)
c.smk <- c(1,1,1,0,1,1,1,1)  #censored = 0
smk0 <- c(12,18,21,34,40,46,30,33,39,42,44,50,56,58,61,66,77)
c.smk0 <- c(1,1,1,1,1,1,0,1,1,1,1,0,1,1,1,0,1)
t <- c(smk,smk0)
cc <- c(c.smk,c.smk0)
status <- rep(0:1,c(length(smk),length(smk0)))

#log-rank analysis
f <- survfit(Surv(t,cc)~status)
summary(f)
f0 <- survdiff(Surv(t,cc)~status)
f0
cbind(f0$exp[1],f0$var[1,1],f0$chisq)
```

#CHAPTER 18 - Weibull survival function

```
library(survival)

#Weibull simulated data
t <- c(12.90,2.42,30.35,24.43,13.48,19.49,15.84,8.56,27.15,2.01,35.80,
49.89,8.66,17.35,3.24,7.84,9.41,3.91,16.59,37.48,24.06,8.08,14.91,
6.93,9.46,37.71,20.43,8.06,23.39,4.68,18.11,2.45,14.78,25.88,10.47,
9.38,9.62,24.17,20.85,6.26,22.09,32.77,39.17,43.04,6.69,4.92,8.54,
10.83,5.56,6.43)

#test - simulation
#G <- 1.5
#L <- 0.05
#iter <- 10000
#p <- runif(iter)
#t <- ((-log(p))^(1/G))/L

n <- length(t)
cc <- rep(1,n)
summary(t)
mean(t)
1/mean(t)

f <- survreg(Surv(t,cc)~1,dist="weibull")
summary(f)
```

```
s <- f$scale
lambda <- exp(-f$coefficients)
gamma <- 1/f$scale
w.log <- f$loglik[1]
v <- f$var
se <- sqrt(v[2,2])
x2 <- (log(gamma)/se)^2
pvalue <- 1-pchisq(x2,1)
cbind(gamma,se,lambda,x2,pvalue)

#exponential model
f <- survreg(Surv(t,cc)~1,dist="exponential")
summary(f)
e.log <- f$loglik[1]
lambda.0 <- 1/mean(t)
lambda.1 <- 1/exp(f$coefficients)
cbind(lambda.0,lambda.1)

#likelihood ratio comparison
x2 <- 2*(w.log-e.log)
pvalue <- 1-pchisq(x2,1)
cbind(w.log,e.log,x2,pvalue)
```

#CHAPTER 19 - Prediction performance

```
ci <- function(m,v) {
A <- m-1.960*sqrt(v)
B <- m+1.960*sqrt(v)
round(cbind(A,B),5)
}

#confidence intervals
ci(0.231,1/259+1/3595) #nri
ci(0.109,0.031^2) #nri - table
ci(0.0057,0.00213^2) #idi
ci(.150,0.0416^2)  #B - slope

#data
sbp <- c(108,117,108,96,91,131,117,130,145,98,118,
111,86,94,125,91,133,105,133,110,99,118,
99,87,115,95,124,118,121,123,121,78,107,
119,89,100,158,117,106,101,105,118,118,104,
119,135,92,111,99,117)
dbp <- c(67,88,68,59,56,76,65,87,90,62,80,
72,52,60,70,57,66,63,76,67,51,79,
66,51,70,58,75,75,71,75,68,52,68,
74,58,70,79,66,72,56,67,73,70,71,
69,83,52,55,57,62)

#least squares estimated line
summary(lm(sbp~dbp))

#perpendicular
D <- .5*(var(dbp)-var(sbp))/cov(sbp,dbp)
B <- -D+sqrt(1+D^2)
B0 <- -D-sqrt(1+D^2)
```

```
A <- mean(sbp)-B*mean(dbp)
round(cbind(D,A,B,B0,-1/B0),3)

#CHAPTER 20 - Attributable risk

ci <- function(m,v) {
A <- m-1.960*sqrt(v)
B <- m+1.960*sqrt(v)
round(cbind(A,B),5)
}

#usual estimation form a 2 by 2 table
a <- 159
b <- 1343
c <- 98
d <- 1553
matrix(c(a,c,b,d),2,2)
n <- a+b+c+d
P <- (a+c)/n
p <- c/(c+d)
rr <- (a/(a+b))/(c/(c+d))
e <- (a+b)/n
cbind(P,p,rr,e)

#attributable risk estimates
#version 1
(P-p)/P
#version 2
(a*d-b*c)/((a+c)*(c+d))
#version 3
e*(rr-1)/(e*(rr-1)+1)
#version 4
1-1/(e*rr+(1-e)*1.0)
#version 5
(a/(a+c))*((rr-1)/rr)

#confidence intervals - log(1-ar) and ar
ar <- (P-p)/P
V <- (b+ar*(a+d))/(n*c)
ci(log(1-ar),V)
rev(1-exp(ci(log(1-ar),V)))

#age-strata summary
d0 <- c(5,15,21,15)
P0 <- c(100,200,350,600)
d1 <- c(10,36,54,40)
P1 <- c(100,200,300,500)
D <- d1/sum(d1)
rr <- (d1/P1)/(d0/P0)
e <- sum(P1+d1)/(sum(P0+d0)+sum(P1+d1))
ar <- e*(rr-1)/(e*(rr-1)+1)
round(cbind(D,rr,ar),3)
arbar <- sum(D*ar)
arbar
```

```
#chd and smoking
chd <- c(59,37,100,61)
nochd <- c(271,325,1072,1228)
r <- (chd/(chd+nochd))/(chd[4]/(chd[4]+nochd[4]))
e <- (chd+nochd)/sum(chd+nochd)
round(cbind(e,r),3)
P <- sum(chd)/sum(chd+nochd)
p <- (chd[4]/(chd[4]+nochd[4]))
ar1 <- (P-p)/P
ar2 <- 1-1/sum(e*r)
cbind(P,p,ar1,ar2)

#data - chd and smoking at four level of cholesterol
a <- c(14,17,47,81)
b <- c(333,248,381,381)
c <- c(17,12,27,41)
d <- c(467,343,374,369)
n <- a+b+c+d
#estimates
D <- (a+c)/(sum(a+c))
p <- c/(c+d)
P <- (a+c)/n
ar <- (P-p)/P
v <- ((1-ar)^2*(b+ar*(a+d)))/(n*c)
round(cbind(D,p,P,ar,v),3)

#weighted average - estimated attributable risk
#version 1
w <- 1/v
V <- 1/sum(w)
arbar <- sum(w*ar)/sum(w)
arbar
ci(arbar,V)
#version 2
arbar0 <- sum(D*ar)
arbar0
ci(arbar0,V)

#simple misclassification model
A <- c(1.0,0.95,0.90,0.80)
a <- 159
b <- 1343
c <- 98
d <- 1553
n <- a+b+c+d
p <- (c*A)/(c*A+d)
P <- (a+c)/n
ar <- (P-p)/P
round(ar,3)
```

#CHAPTER 22 - ROC-curve

```
ci <- function(m,v) {
A <- m-1.960*sqrt(v)
```

```
B <- m+1.960*sqrt(v)
round(cbind(A,B),5)
}

#data: test (t) versus standard (s) - carotid disease
t <- c(20.0,30.0,40.0,45.0,50.0,50.0,55.0,55.0,57.5,
66.7,67.5,67.5,73.6,75.0,75.0,75.0,77.5,77.5,80.0,
80.0,80.0,80.0,82.5,82.5,85.0,87.5,87.5,92.5,85.0,85.0)

s <- c(39.5,55.9,59.5,60,63,64.9,66.8,67.2,68.0,70.6,
72.4,73.6,75.3,76.3,77.4,77.8,78.2,78.4,78.5,78.6,
80.8,85.6,86.3,87.4,88.5,88.6,88.7,91.5,87.4,88.3)

#probabilities
cbind(1-pnorm(20,18,3),1-pnorm(20,22.5,3))

#"optimum" roc-value - circle
ns <- 30
nt <- 30
tbar <- mean(t)
sbar <- mean(s)
c <- seq(20,150,0.05)
V <- ((ns-1)*var(s)+(nt-1)*var(t))/(ns+nt-2)
sen <- 1-pnorm(c,sbar,sqrt(V))
fpf <- 1-pnorm(c,tbar,sqrt(V))
plot(fpf,sen,type="l",xlab="false positive",ylab="sensitivity")
abline(0,1)
c0 <- (sbar+tbar)/2
points(cbind(1-pnorm(c0,tbar,sqrt(V)),1-pnorm(c0,sbar,sqrt(V))))
title("ROC curve")

#Parametric ROC analysis
ns <- length(s)
nt <- length(t)
tbar <- mean(t)
sbar <- mean(s)
V <- ((ns-1)*var(s)+(nt-1)*var(t))/(ns+nt-2)
R <- (tbar-sbar)/sqrt(2*V)
auc <- 1-pnorm(R)
z <- (tbar-sbar)/sqrt(V*(1/ns+1/nt))
pvalue <- pnorm(z)
round(cbind(tbar,sbar,V,R,auc,z,pvalue),3)

#95% confidence intervals for R and auc
v.r <- .5*(1/nt+1/ns)
ci(R,v.r) #confidence interval from R
1-pnorm(rev(ci(R,v.r))) #confidence interval from auc

#note:  t-test of difference of mean values tbar and sbar)
t.test(t,s,var.equal=TRUE)

#Nonparametric ROC analysis
#artificial data
x <- c(11.49,10.61,11.51,10.60,12.64,11.62,11.20,12.11,10.79,10.62,
10.70,11.05,8.12,9.97,12.31,10.97,10.96,9.46,10.67,10.50)
y <- c(0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1)
n <- length(y)
```

```
nt <- sum(y)
ns <- n-nt
U <- wilcox.test(x[y==0],x[y==1])$statistic
W <- U+ns*(ns+1)/2
P <- U/(ns*nt)
wbar <- sum(rank(x)[y==0])/ns
auc <- (wbar-.5*(ns+1))/nt
v <- (1/12)*(1/ns+1/nt)
z <- (auc-0.5)/sqrt(v)
pvalue <- 1-pnorm(z)
round(cbind(U,W,wbar,P,auc,z,pvalue),3)

#data from illustration
s <- c(2.8,5.2,5.6,6.5,8.8,9.9)
t <- c(3.4,5,6,6.1,7,7.3,7.7,8.4,9,9.2)
n <- length(t)+length(s)
nt <- length(t)
ns <- n-nt
cbind(n,ns,nt)
S <- sqrt(((nt-1)*var(t)+(ns-1)*var(s))/(nt+ns-2))
R <- (mean(s)-mean(t))/(S*sqrt(2))
auc <- 1-pnorm(R)
round(cbind(mean(t),mean(s),S,R,auc),3)

#roc plot - nonparametric
#Table 8.0
ns <- 6
nt <- 10
n <- nt+ns
id <- c(1,0,0,1,1,0,0,1,0,0,0,0,1,0,0,1)
X <- Y <- NULL
x <- y <- 0
for(i in 1:n) {
x <- ifelse(id[i]==1,x+1/ns,x)
X[i] <- 1-x
y <- ifelse(id[i]==0,y+1/nt,y)
Y[i] <- 1-y
}
round(cbind(id,X,Y),2)
plot(c(1,X),c(1,Y),type="s",xlim=c(0,1),ylim=c(0,1),
xlab="false positive fraction",ylab="true positive fraction")
points(c(1,X),c(1,Y),pch=20)
abline(0,1)
title("ROC PLOT")

#a few choices for plot symbols
plot(1:20,rep(5,20),pch=1:20,axes=FALSE,xlab="",ylab="")
text(1:20,rep(4.8,20),1:20)
text(10,5.3,"a few choices for plot symbols")
```

#CHAPTERS 23 - Genetics.r

```
ci <- function(m,v) {
A <- m-1.960*sqrt(v)
```

```
B <- m+1.960*sqrt(v)
round(cbind(A,B),5)
}

#mother/child pairs
#data
A <- 9
B <- 14
C <- 16
D <- 25
E <- 18
F <- 15
G <- 23

#genetic analysis
o <- c(A,B,C,D,E,F,G)
n <- sum(o)
N  <- 3*n-D
p <- (3*A+2*B+2*C+D+E+F)/N
q <- 1-p
cbind(n,N,p)
e <- n*c(p^3,p^2*q,p^2*q,p*q,p*q^2,p*q^2,q^3)
round(rbind(o,e/n,e),3)
x2 <- sum((o-e)^2/e)
pvalue <- 1-pchisq(x2,6)
cbind(x2,pvalue)
v <- p*(1-p)/N
ci(p,v)

#Synder's ratios
#data
x11 <- 194
x12  <- 19
x21 <- 117
x22 <- 28
x31 <- 0
x32 <- 26
n <- sum(x11+x12+x21+x22+x31+x32)
q <- sqrt((x12+x22+x32)/n)
p <- 1-q
cbind(q,p)

#calculation of Synder's ratios
s2 <- x12/(x11+x12)
s1 <- x22/(x21+x22)
c(s1,s2)
s2/s1^2

#chi-square analysis
e11 <- p^2*(1+2*q)
e12 <- p^2*q^2
e21 <- 2*p*q^2
e22 <- 2*p*q^3
E11 <- (x11+x12)*e11/(e11+e12)
E12 <- (x11+x12)*e12/(e11+e12)
```

```
E21 <- (x21+x22)*e21/(e21+e22)
E22 <- (x21+x22)*e22/(e21+e22)

o <- c(x11,x12,x21,x22)
e <- c(E11,E12,E21,E22)
cbind(round(e,1),o)
xx <- sum((o-e)^2/e)
pvalue <- 1-pchisq(xx,2)
cbind(xx,pvalue)

#Selection
#analysis
selection <- function(s1,s2,s3,p) {
q <- 1-p
for(ii in 1:50) {
D <- p^2*(1-s1)
H <- 2*p*q*(1-s2)
R <- q^2*(1-s3)
p <- (D+.5*H)/(D+H+R)
q <- 1-p
if(ii==1|ii==2|ii==3|ii==10|ii==30|ii==50) print(round(cbind(ii,D,
H,R,p,q),3))
}}

#balanced polymorphism - case 1
s1 <- 0.2
s2 <- 0
s3 <- 0.6
p <- 0.3
selection(s1,s2,s3,p)

#s=1.0 - aa fails to survive - case 2
s1 <- 0
s2 <- 0
s3 <- 1
p <- 0.3
selection(s1,s2,s3,p)

#s=0.5 -  probability - case 3
s1 <- 0
s2 <- 0
s3 <- 0.5
p <- 0.3
selection(s1,s2,s3,p)

#analysis
selection <- function(s1,s2,s3,p) {
q <- 1-p
for(ii in 1:50) {
D <- p^2*(1-s1)
H <- 2*p*q*(1-s2)
R <- q^2*(1-s3)
p <- (D+.5*H)/(D+H+R)
q <- 1-p
```

```
if(ii==1|ii==2|ii==3|ii==10|ii==30|ii==50) print(round(cbind(ii,D,
H,R,p,q),3))
}}

#Mr.A versus mr.B
a <- 100
b <- 10
pa <- 0.5
pb <- 0.1

for(i in 1:15) {
a0 <- a-a*pa+b*pb
b0 <- b+a*pa-b*pb
a.to.b <- a*pa
b.to.a <- b*pb
ratio <- a/b
print(round(cbind(i,a0,b0,a.to.b,b.to.a,ratio),2))
a <- a0
b <- b0
}

#admix model
m <- 0.1
p0 <- 0.1
P <- 0.8

#version 1
PP <- pp <- NULL
iter <- 10
for(k in 1:iter) {
p <- (1-m)*p0+m*P
p0 <- p
pp[k] <- p
}

#version 2
k <- 1:iter
p0 <- 0.1
PP <- (1-m)^k*p0+(1-(1-m)^k)*P
round(cbind(k,pp,PP),3)

#Wahlund model
p <- c(0.8,0.6,0.4,0.2)
p0 <- mean(p)
q <- 1-p
P <- p^2
PQ <- 2*p*q
Q <- q^2
grp <- cbind(mean(P),mean(PQ),mean(Q))
pop <- cbind(p0^2,2*p0*(1-p0),(1-p0)^2)
diff <- grp-pop
rbind(grp,pop,diff)
#partition
k <- length(p)
c(p0*(1-p0),sum(p*(1-p))/k,sum((p-p0)^2/k))
```

Index

2 × 2 table, 87, 89

acceleration, 360
actuarial assumption, 315
additive bivariate model, 120
additive linear regression model, 118
additive model, 119, 141, 150, 177, 178, 180, 181, 186
additive pairwise interaction, 196
additive proportional hazard model, 366
additive regression model, 114, 116, 122
additivity, 114, 162, 365
admixed, 458
age-adjusted rate ratio, 145
alternative distribution, 225
analysis of variance, 182
approximate average rate, 316
Arbuthnott, 291
ascertainment, 450
assortative mating, 482
attributable risk, 387
attributable risk percentage, 92
average survival probability, 341

Bailey method, 515
balanced polymorphism, 445, 448
bandwidth, 42, 43
Bartlett test of variance, 271, 276, 403
Bartlett test statistic, 272, 275
Bayes theorem, 109
Bernoulli variable, 51, 52, 54, 56
between group variability, 265, 266, 268
bilinear regression model, 113
binomial coefficient, 49, 50
binomial distribution, 52, 53, 55, 56, 60, 90, 170, 263, 289, 440
binomial mean value, 53
binomial probability distribution, 23, 24, 49, 51, 57, 67, 183, 326
binomial theorem, 498, 499
binomial variance, 51, 53
bivariate linear regression model, 112, 123, 124
Bonferroni inequality, 229
bootstrap estimate, 204, 207, 225
boxplot, 227

calibration, 250, 379
case-load estimate, 395
categorical variable, 186
censored, 311
central limit theorem, 5–7
chi-square based measures of association, 81
chi-square analysis, 134, 137
childhood leukemia mortality data, 39, 40
chi-square comparison, 105, 147
chi-square distribution, 9, 10, 12, 16, 82, 106, 108–110, 130, 135, 138, 142, 144
collinear, 123
comparison of the proportions, 89, 90
complete independence, 190, 193
computer software, 151
conditional independence, 191, 194, 195, 201
conditional likelihood estimation, 346
confidence band, 19, 20, 21, 23–27, 35, 53, 71, 72, 227
confounder bias, 121, 122
contingency coefficient, 70, 82
Cook distance, 255
correlation coefficient, 70
Cox, 346
Cramer coefficient, 82
critical value, 4, 5
crossover, 435
crude ratio, 160
cumulative distribution function, 230, 232, 240, 255, 419
cumulative hazard function, 333
cumulative normal probability, 5
cumulative probability distribution, 235, 324

degrees of freedom, 8, 10
deMoivre, 3
deMoivre-Laplace theorem, 53
direct adjustment, 38
dose-response relationship, 148

effective sample size, 288
Einstein, 52
Eisenhart, 300
exact confidence interval, 22
expected value, 47, 49, 51, 52

exponential cumulative probability function, 239, 248
exponential distribution, 242
exponential survival distribution, 238, 319, 320, 322, 353

f-ratio test statistic, 270
false positive, 108
family-wise error rate, 229
family laws, 434
Fisher, 3, 9, 71, 98
Fisher exact test, 101
Fisher information function, 508
Friedman rank test, 302
Friedman test statistic, 303

Galton, 69
gamma coefficient, 77
Gauss, 3
Gauss-Laplacian distribution, 3
genetic selection, 442
geometric mean, 283, 286
geometric probability distribution, 57, 58
goodness-of-fit, 11, 64, 65, 253, 257
Graunt, 291, 316, 322
Greenwood, 330
Greenwood estimated variance, 331
Greenwood variance, 330

Halley, 316
Hardy-Weinberg equilibrium, 434, 465
harmonic mean value, 212
harmonic series, 481
hazard rate, 339
hazard ratio, 339, 340, 343
heterogeneity, 264, 265
Hodgkins disease mortality, 95
Hodgkins disease mortality rate, 44
homogeneity, 145, 146, 148, 162, 169, 264, 265, 267
hypergeometric probability distribution, 98, 99, 100–102, 338

identically distributed binary variable, 102, 105
infant mortality rate, 38
interaction, 111, 119, 120, 150, 165, 188, 189, 192, 349, 365
inverse function, 233

joint independence, 191, 194, 197

Kaplan-Meier estimate, 324, 326
kappa statistic, 216
keno, 100
Knox analysis table, 400
Knox method, 402
Kolmogorov, 241, 242
Kolmogorov-Smirnov test, 243
Kruskal-Wallis analysis, 295–297, 302, 303

least squares estimation, 464
left-censored data, 283
Levene test of variance, 273, 276, 387
life table, 317
likelihood statistic, 170, 172, 173, 356, 371, 372
likelihood value, 171, 501
limit of detection, 284
linear bivariate regression model, 111
linear relationship, 70
log-cumulative hazard function, 362
log-hazard ratio, 345
log-likelihood value, 151, 171–174, 176, 189
log-normal distribution, 278, 281
log-rank test, 335, 339
log-rate, 265
logarithmic transformation, 71
logistic model, 257
logistic regression model, 371
loglinear Poisson model, 141, 187
loglinear model, 148

Maclaurin series, 497
Mann-Whitney, 128, 132, 133
Mann-Whitney estimation, 81, 130
Mann-Whitney test, 127, 129, 131, 417
Mantel-Haenszel odds ratio, 36
Mantel statistic, 404, 405
matched pair, 156, 182–184, 287
maximum likelihood estimate, 505, 510
Mcnemar test, 184
mean fitness, 448
mean polish, 179, 181
mean survival time, 318, 321, 327, 331
measures of association, 92
median value, 321
median polish, 188
median survival time, 334, 335, 360
median value, 23, 24, 283
Mendel, 450, 465
Mendelian inheritance pattern, 441
method of moments estimation, 51, 104
Miller, 271
model prediction, 371
mortality rate, 33, 34, 39, 309
mortality rate ratio, 46
mother-child assortment, 439
mother-child pair, 440
multiple correlation coefficient, 117
multivariable Weibull model, 363

nested, 189
net reclassification index (nri), 375
Newman, 19
noninformative, 330
noninformative censoring, 311, 312

nonparametric confidence interval, 207
nonparametric measure of association, 77
normal distribution, 4, 6, 7, 9, 16, 17, 42, 53, 55, 114, 129, 170, 278
normal distribution approximation, 3, 57, 60, 96, 98, 134
null distribution, 225
null hypothesis, 222, 300

odds ratio, 35–37, 39, 82, 92, 94–96, 190–192, 202
one-way analysis of variance, 266, 295, 276, 294
ordinal data, 300
ordinal variable, 126, 129
ordinary least squares estimate, 31, 383, 386

partial correlation coefficient, 117
Pascal, 50
Pascal triangle, 50
Pearson, 3, 9
Pearson chi-square statistic, 102, 106, 126, 170
Pearson product-moment correlation coefficient, 103
percentage decrease, 209, 210
performance table, 407
perinatal mortality, 158, 161
perinatal mortality ratio, 163
perpendicular least squares, 75, 76, 382, 386
phi-coefficient, 82
Poisson, 58
Poisson additive model, 146, 396
Poisson distribution, 63, 64, 296
Poisson loglinear regression, 390
Poisson probability distribution, 47, 58, 60–62, 142, 263, 264, 313, 402
population stratification, 458
positive assortative, 483
power residue method, 245
power of the sign test, 291
product-limit estimated survival probability, 324, 326, 330, 339
product-moment correlation coefficient, 69, 466
proportional hazard analysis, 345–346
proportional reduction in error criterion, 83
pseudo-random number, 245, 246, 249, 396

quantile, 282
quantile-quantile plot, 235, 238
quasi-independence, 156
Quetelet, 3
quick test, 300

random genetic drift, 473, 474
randomization response survey technique, 56
randomization test, 220
randomized response interview technique, 516
randomized test statistic, 225
rate ratio, 34, 35, 143, 144–147, 161, 162
ratio, 29–31
ratio estimate, 33
ratio of rates, 46
reclassification technique, 372
recombination, 435
regression diagnostics, 250
regression model, 297
relative hazard ratio, 92, 94, 95, 345, 389, 390
replicate estimation, 215
replicate sample, 205
residual value, 113–115, 179–182, 253
ridit value, 81, 132
right censored, 328
risk ratio, 148
robust, 271, 297
Rutherford, 263

sample covariance, 490
sample mean value, 3
sample variance, 3, 260
sampling with replacement, 204
saturated model, 175
Scott, 289
seed-value, 245
segregation probability, 450, 453
segregation ratio, 450
sensitivity, 108, 380, 407
shift model, 409
Siegel-Tukey nonparametric evaluation, 274
signed rank test, 291
significance probability, 15
sign test, 289, 306
simple linear regression model, 31
simulated data, 235, 248
smoothed curve, 46
smoothing technique, 40
Somer coefficient, 80, 81
spatial pattern, 62
Spearman rank correlation coefficient, 73, 74
specificity, 108, 407
standard error, 20
standard mortality ratio, 39, 40
standard normal distribution, 4, 5, 9
statistical transformation, 71
Stein, 20
stochastically independent, 88
structural zero, 106, 154
Student (Gosset), 352
Student t-test statistic, 76
survival probability, 319, 324
Synder ratio, 441

t-distribution, 7
Taylor series, 496, 500, 516
test for linear trend, 134
test of variance, 262, 268
  Bayes, 109
three-group regression analysis, 297
three-way table, 189

tricubic relationship, 44
truncated data, 450, 451
truncated Poisson distribution, 65, 66
Tshuprow coefficient, 82
Tukey, 179, 299, 300
two-sample comparison, 98
two-sample rank test, 417
two-sample test statistic, 139

uniform probability distribution, 314
US presidents, 67

Wahlund model, 471, 472
Wald test, 511
Weibull, 352
Weibull parametric survival model, 352, 360, 366, 471, 472
Weibull two-sample model, 359
weighted average, 29, 30, 35, 39, 41, 42, 44, 48, 49, 136, 160, 163, 164, 259, 390, 394, 395
Wilcoxon, 417
Wilcoxon rank sum test, 126–128, 131, 132, 291, 293
Wilcoxon two-sample nonparametric analysis, 297
Wilcoxon two-sample rank test, 243, 274, 297
within group variability, 266, 268

Yule coefficient, 82, 83