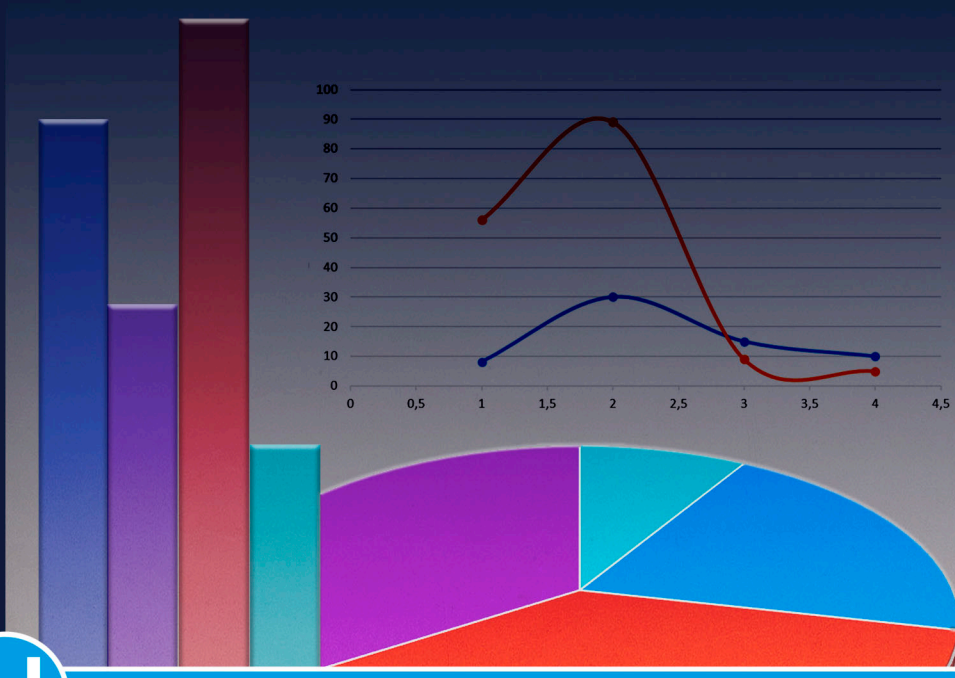
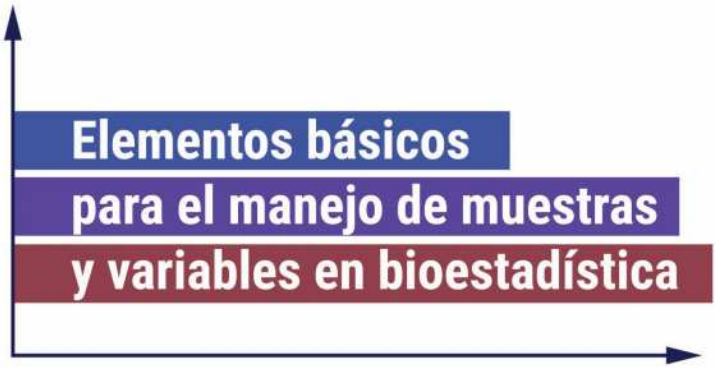
EDITORIAL CIENCIAS MÉDICAS

Elementos básicos para el manejo de muestras y variables en bioestadística

Dr. Jesús Romero Madero

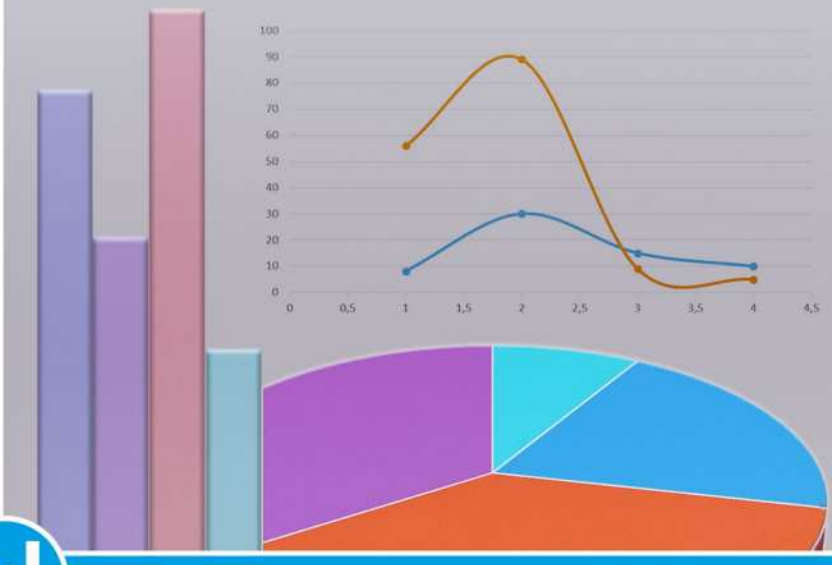


SALUD PÚBLICA • BIOESTADÍSTICA


Elementos básicos
para el manejo de muestras
y variables en bioestadística

Elementos básicos para el manejo de muestras y variables en bioestadística

Dr. Jesús Romero Madero



SALUD PÚBLICA • BIOESTADÍSTICA

La Habana • 2021

Catalogación Editorial Ciencias Médicas

Romero Madero, Jesús
Elementos básicos para el manejo de muestras y variables en bioestadística/
Jesús Romero Madero. —La Habana: Editorial Ciencias Médicas, 2021.
809 p.: il., tab. — (Colección Salud Pública. Serie Bioestadística)
-
-
Bioestadística, Estadísticas de Salud, Estadísticas Sanitarias, Análisis Multivariante

WA 900

Cómo citar:
Romero Madero J. Elementos básicos para el manejo de muestras y variables en bioestadística [Internet]. La Habana: Editorial Ciencias Médicas; 2021. Disponible en: http://www.bvscuba.sld.cu/libro/elementos-basicos-para-el-manejo-de-muestras-y-variables-en-bioestadistica/

Edición: Ing. José Quesada Pantoja
Diseño interior y de cubierta: DI. Meylín Sisniega Lorigados
Ilustraciones: DCV. Amanda Rodríguez Sánchez y DI. Meylín Sisniega Lorigados
Emplane: María Pacheco Gola

© Jesús Romero Madero, 2021
© Sobre la presente edición:
Editorial Ciencias Médicas, 2021

ISBN 978-959-313-953-3 (PDF)
ISBN 978-959-313-954-0 (EPUB)

Editorial Ciencias Médicas
Centro Nacional de Información de Ciencias Médicas
Calle 23, No. 654, entre D y E, El Vedado
La Habana, CP 10400, Cuba
Teléfono: (53) 7836-1893
ecimed@infomed.sld.cu
www.ecimed.sld.cu

Autor

Jesús Romero Madero

Máster en Salud de los Trabajadores
Especialista de II Grado en Gastroenterología
Profesor Auxiliar
Facultad de Ciencias Médicas de Artemisa

Dedicatoria

A mis hijos Jesús Junior y Jimmy Anyelo, que son mi máxima fuente de inspiración y ojalá este libro les motive e ilumine al conducirse por el camino del conocimiento y el buen actuar en sus vidas.

A mis padres Pastora Nélida y Jesús, quienes me dieron la vida y con ello tener el privilegio de poder ayudar a los demás.

A mi esposa Sara, por su apoyo ante los momentos de desaciertos para la realización de este proyecto.

A la memoria de mis abuelos, quienes me cuidaron y orientaron a tomar la senda del aprendizaje.

A mis colegas, estudiantes universitarios e investigadores jóvenes en las ciencias médicas, que necesitan contar con una herramienta más para el procesamiento estadístico de la información en sus proyectos investigativos, confiaron en mí para su obtención y han esperado pacientemente la culminación de este proyecto.

Agradecimientos

A nuestra Revolución, por haberme dado la posibilidad de formarme como profesional en el área de la salud pública y tener la oportunidad de poder trasmitir todo lo aprendido a quien lo necesite dentro y fuera de nuestras aulas.

A los profesores que he tenido desde mi infancia, quienes me han permitido crecer bebiendo en la fuente de su sabiduría, lo que ha sido vital en mi desempeño. Especialmente agradezco para llevar a efecto este libro al profesor Mario Soler Saavedra, investigador, ejemplo de humanismo y sabiduría de la Universidad de Ciencias Médicas de La Habana que me acercó inicialmente al maravilloso mundo de la investigación en bioestadística, al profesor e investigador Luis Carlos Silva, experto en temas de epidemiología y bioestadística de la Escuela Nacional de Salud Pública, al doctor Maizel Monzón, Especialista en Bioestadística del Centro Nacional de Ensayos Clínicos y al profesor de Bioestadística Elio López León, de la Facultad de Ciencias Médicas de Artemisa, quienes contribuyeron con gran parte de la bibliografía revisada y empleada en la confección de este texto y me permitieron las consultas personales de algunos de los contenidos presentados en esta obra.

A los estudiantes, doctores residentes y profesores de la Facultad de Ciencias Médicas de Artemisa, en su sede de San Antonio de los Baños, por mostrarme la necesidad que tenían de contar con un texto que les facilitara el modo de realizar el procesamiento estadístico de la información recogida en sus investigaciones biomédicas y proyectos de grado científico.

Al personal de informática del Hospital Universitario General Docente Iván Portuondo Echevarría de la provincia Artemisa por su ayuda profesional.

Al grupo de trabajo del Centro Editorial Ciencias Médicas por darme la posibilidad de que fuese evaluado este proyecto, sus indicaciones oportunas y precisas, así como por mostrar su interés, dedicación y consagración para que se llevara a efecto, lo que es muy digno de destacar.

Prefacio

En la prédica de nuestro héroe nacional José Julián Martí Pérez, se resalta la premisa de que: "[…] todo hombre cuando viene al mundo tiene el derecho a que se le eduque y se le enseñe". En justo tributo por ese hecho, este hombre debería trasmitir lo aprendido a todo nuevo ser que lo necesite. Al poner a consideración de los lectores este libro he pretendido ser consecuente con esta premisa.

Antes de dirigir las primeras palabras al lector, quiero expresar el sentimiento de agradecimiento eterno hacia los profesores, los profetas eternos para estos tiempos, con quienes hemos convivido y aún lo hacemos hoy día en las aulas a la luz de su sabiduría y ejemplo, con los cuales trabajar, siempre es un verdadero honor, privilegio y compromiso para mí.

La motivación que he tenido para tratar de llevar a efecto este libro se ha reforzado en los últimos años a partir de la detección personal de las necesidades de los estudiantes universitarios y profesionales jóvenes en general de poder contar con alguna nueva herramienta académica que favorezca una mejor comprensión y aplicación más eficaz de la bioestadística en sus proyectos investigativos; aunque mis primeras inquietudes al respecto se retraen en el tiempo desde la etapa de mis estudios de pregrado en la carrera de Medicina, con el despertar de los primeros acordes de la asignatura Bioestadística, pasando por los compaces y melodías recibidos de las investigaciones en que participé durante los estudios de pregrado y posgrado, contando con la guía oportuna de mis tutores y más recientemente desde mi posición de tutor para proyectos de tesis para la terminación de residencias en las ciencias médicas y poder alcanzar el respectivo grado científico de los estudiantes en Cuba o en el extranjero.

Para la confección de esta obra fue necesario rescatar y engarzar varios factores académicos, administrativos y logísticos en general, sin dejar de mencionar el sacrificio personal y de quienes colaboraron desde la etapa de planificación, desarrollo y culminación para presentar de esta primera edición que se pone a consideración del lector.

En este gran esfuerzo se ha consultado un gran volumen de bibliografías nacionales e internacionales de autores clásicos muy reconocidos en el terreno de la bioestadística; así como ha resultado imprescindible contar con el apoyo de algunos expertos y colegas nacionales e internacionales para este propósito, en conjunción con los apuntes y resúmenes realizados durante mis estudios de pregrado y posgrado.

El objetivo fundamental que subraya este libro es intentar brindarle al lector la posibilidad de tener un acercamiento resumido al infinito universo de la bioestadística, en la vertiente descriptiva y la inferencial de forma conjunta, mediante la consulta de algunos de sus aspectos básicos: teóricos conceptuales y algorítmicos, que constituyen verdaderas claves, fundamentalmente enfocado hacia el manejo de muestras y variables con los correspondientes contrastes de hipótesis en la inferencia bioestadística; lo que se encuentra en el epicentro de esta disciplina aplicada en las investigaciones científicas, al mismo tiempo que el lector pueda apoderarse y aplicar los conocimientos derivados de sus contenidos, como una forma legítima de alcanzar un primer escalón en la subida al Olimpo del procesamiento estadístico eficaz de la información.

La línea de continuidad y consecutividad articulada de los resúmenes referentes a los contenidos básicos de la bioestadística, giran principalmente en torno al manejo de muestras y variables relacionados con las pruebas de significación o de contraste de hipótesis en la inferencia bioestadística abordadas en las dos secciones del libro, lo que puede constituir un elemento docente-didáctico que facilite la comprensión integrada de estos aspectos, tratando de proveerle al joven investigador la idea de que la bioestadística es una disciplina que funciona como un todo armónico, en la que se pone de manifiesto el empleo del método científico; resaltando que para su enseñanza, comprensión, aprendizaje y aplicación se requiere contar con una visión integral e integrada que conjugue al par alumno-profesor y absorber el recorrido histórico de los contenidos de esta disciplina hasta la actualidad.

Este libro está pensado y orientado al servicio especialmente de los estudiantes universitarios, profesionales e investigadores jóvenes, pues se pretende poder agilizar y complementar el recuerdo de los contenidos básicos de la Bioestadística impartida en las clases de la universidad en Cuba.

En la primera sección se abordan de modo particular los aspectos conceptuales básicos necesarios y en la segunda sección se presentan propuestas algorítmicas esquematizadas y no esquematizadas que permitan la mejor comprensión, por parte del estudiante universitario, del manejo de las muestras y variables en su relación con los respectivos contrastes de hipótesis de la inferencia bioestadística, en estrecho acuerdo con los objetivos investigativos respectivos, como parte del procesamiento estadístico de la información.

Como tendencia se ha tratado de abarcar la mayor cantidad posible de los contenidos resumidos en las distintas temáticas seleccionadas de consulta con expertos en estos temas, prevaleciendo una intención didáctica y dialéctica; aunque ello llevó implícito cierta desproporción en la extensión de los contenidos abordados en algunos capítulos, fundamentalmente en la primera sección. En ese mismo sentido, con el ánimo de incrementar las posibilidades de adquisición y comprensión en algunos de estos contenidos, se emplearon variados tipos de gráficos, tabulaciones y esquemas bajo la influencia de la formación docente-metodológica; aunque esto pudiera en ocasiones percibirse como redundante por el lector, en cuyo caso, sería involuntaria con la intención noble de facilitar el aprendizaje y la aplicación de la bioestadística.

Complementariamente se anexan aspectos tales como tablas para representar la distribución teórica-esperada de probabilidades para diferentes estadígrafos o estadísticos de prueba con los respectivos valores críticos a emplear en los distintos tipos de contrastes de hipótesis mediante técnicas paramétricas y no paramétricas; así como un grupo de formularios aplicados en estos contrastes de hipótesis y un resumen de términos estadísticos ordenados alfabéticamente, que se utilizan en este material o la bibliografía requerida para su consulta, de ser necesario.

En la primera sección se articulan aspectos propios de la estadística descriptiva con los de la estadística inferencial, soportando los referentes a la metodología de la investigación y la epidemiología para los tipos de diseños de estudios que incluyen los que presentan un enfoque de riesgo y que resultan necesarios involucrar en el procesamiento estadístico de la información, fundamentalmente en el manejo de muestras y variables con sus respectivos contrastes de hipótesis en la inferencia bioestadística, siendo consecuentes con el acercamiento a los objetivos y aspiraciones planteados para este texto.

De los aspectos de la estadística descriptiva se abordan los que con más frecuencia se emplean en los procesamientos de los datos, tales como: las clasificaciones de las variables, los tipos de medidas de resumen de las variables, el modo de la operacionalización de las variables, las distintas distribuciones de variables aleatorias, tipos de coeficientes, de estadígrafos y de representaciones gráficas para los datos de las variables; así como los tipos de muestras y de muestreos; todo lo cual converge en una caracterización general más precisa del fenómeno o problema objeto de estudio, lo que a su vez puede servir de premisa y admite hilvanarse con la vertiente o perspectiva inferencial.

Respecto a los aspectos de la estadística inferencial se abordan también los que con más frecuencia se emplean en las investigaciones para el procesamiento estadístico de la información como los referentes a las pruebas de significación o de hipótesis por

técnicas paramétricas y no paramétricas, y las estimaciones por intervalo de confianza de ciertos parámetros para estudios epidemiológicos analíticos observacionales y experimentales, incluyendo los referentes para el enfoque de riesgo, siempre partiendo de algoritmos propuestos para el manejo de muestras y variables y viceversa, con el asesoramiento de expertos, utilizando los contrastes de hipótesis respectivos y relacionados con las opciones de técnicas bioestadísticas investigativas para el procesamiento estadístico de la información que permiten estimar: diferencias, asociaciones, correlaciones, predicciones y construcción de modelos de pronóstico para el comportamiento de las variables dependientes a partir de las independientes en términos probabilísticos haciendo uso del enfoque frecuentista, por ser el más estudiado y a su vez más generalizada su aplicación en las investigaciones universitarias en Cuba hasta este momento; sin dejar de mencionarse y exponerse algunos ejemplos de opciones en los contrastes de hipótesis mediante el enfoque bayesiano, con las respectivas ventajas de este último en el ejercicio diario del profesional de la salud e investigador en general.

Con toda intencionalidad se ha querido mostrar al investigador joven, mediante la articulación de los contenidos de este material, que el procesamiento estadístico de la información debe ser visto e interpretado integradamente, lo que se evidencia cuando en la primera sección se aborda la explicación de los aspectos teóricos conceptuales, básicos y esenciales de las pruebas de significación o de hipótesis, y quedan insertadas gráficamente estas entre las diferentes opciones de técnicas estadísticas de la inferencia bioestadística para el procesamiento de la información con la respectiva combinación del manejo algorítmico de muestras y variables y viceversa en la segunda sección, donde se representan estos algoritmos en forma esquematizada y no esquematizada, con potencial para convertirse en una aplicación digital o *software* que sintetice estas opciones para las situaciones investigativas más frecuentes a confrontar en particular; todo está en íntima interacción con diferentes elementos de la bioestadística descriptiva e inferencial y requiere del engarce con los conocimientos de epidemiología, metodología de la investigación, cibernética-matemática y las ciencias informáticas.

De modo que los aspectos de los contenidos desarrollados en este texto pueden servir de base para comprender lo que ya es imprescindible emplear en el momento actual de la bioestadística descriptiva e inferencial en las diferentes investigaciones científicas, o sea, las herramientas estadísticas desarrolladas para los procesadores estadísticos en las computadoras en relación con las distintas opciones del manejo de los datos procedentes de una o más variables (univariadas, bivariadas y multivariadas) en una muestra y de dos o más muestras relacionadas o no; pero al mismo tiempo, los resultados estimados con estas herramientas requieren de la revisión permanente de los aspectos de las distintas secciones de este texto por su base teórica, dadas

las bondades que ofrecen estas técnicas estadísticas para un análisis más profundo, multilateral, multifactorial e integrado de los resultados, sin olvidar la consulta bibliográfica permanente por parte del investigador joven.

Realizar un resumen de los contenidos básicos más importantes de la bioestadística relacionados con el manejo de muestras y variables y sus respectivas pruebas de significación en la inferencia bioestadística en tan solo un texto y cuidando mantener la estructura exigida por las editoriales para la confección de libros de corte científico (aún teniendo la fortuna de poder contar con el criterio autorizado de algunos expertos), es una tarea muy difícil, sobre todo si se quiere ser bien explícito y poder "llegarle al estudiante", al tiempo que se corre el riesgo de dejar escapar involuntariamente algunos detalles informativos relacionados con este tema, en los cuales quizás necesite profundizar el lector; en todo caso, se sugiere la consulta de una clásica, variada, amplia y actualizada bibliografía expuesta en la parte final de este texto y quizás, a partir de la misma, realizar vínculos o consultas con otros autores nacionales e internacionales que le permitan enriquecer la información necesaria para el análisis de los respectivos contenidos presentes en esta edición, partiendo siempre del principio de que se dejan las "puertas abiertas" a cualquier aclaración o sugerencia que pueda tributar en sentido general o particular, en una mejor comprensión de los aspectos abordados en este texto para ediciones futuras.

Los aspectos a exponer en este texto giran en torno a los elementos básicos conceptuales y algorítmicos útiles para el correcto manejo de las muestras y las variables en su relación con los respectivos tipos de pruebas de significación en la inferencia bioestadística; con lo cual el dominio de esos elementos puede significar una pieza medular para la realización del procesamiento estadístico eficaz de la información en lo referente fundamentalmente a la inferencia, en franco acercamiento a las principales y más frecuentes situaciones investigativas en que se puede llevar a efecto en el contexto de las investigaciones científicas modernas.

De esta manera se aspira a poder contribuir, mediante los aspectos desarrollados en esta obra, a la mejor formación de los estudiantes universitarios, profesionales e investigadores jóvenes complementando la información brindada en los textos utilizados para la docencia médica por la asignatura Bioestadística del sistema nacional de salud en Cuba o en otras sedes universitarias; así como despertar la sed insaciable de conocimientos que ha de caracterizar al hombre moderno, lo que queda en total consonancia con las palabras proféticas pronunciadas en diferentes momentos en los años finales del siglo xx por el Comandante en Jefe Fidel Castro Ruz y que tienen total vigencia, al expresar que: "[…] en el futuro nuestra patria ha de convertirse en un país de hombres de ciencia, en hombres de pensamiento".

Me sentiría muy reconfortado, junto a todo el personal involucrado que ha aportado, apoyado y trabajado en la conformación de este libro, si el esfuerzo realizado para ponerlo en manos de los estudiantes universitarios e investigadores jóvenes, se viera coronado con un uso provechoso por el lector, que le permita crecer científicamente y que quizás en los próximos tiempos le ayude a presentar la generación de nuevos conocimientos que contribuyan a mejorar la calidad de vida del hombre; siendo así, este libro podría ocupar al menos un espacio en el caudal de conocimientos del lector para este tema, en pleno siglo XXI.

El autor

Introducción

Desde los comienzos de la civilización, el hombre ha utilizado formas sencillas de la estadística como las "representaciones gráficas y símbolos" para contar el número de personas, animales y objetos encontrados en cuevas, pieles, rocas, maderas y otros materiales. Hacia el año 3000 a.C. los babilonios usaban tablillas de arcilla para recopilar datos sobre la producción agrícola y el comercio; los egipcios analizaban datos sobre la población y las rentas antes de la construcción de las pirámides; así mismo en China hacia el año 2000 a.C. se llevaban registros poblacionales y los griegos realizaban censos hacia el año 594 a.C. Siguiendo en el tiempo, aunque el Imperio Romano fue el primer gobierno en recopilar gran cantidad de datos sobre la población, superficie y renta de los territorios que dominaba y durante la Edad Media se realizaron algunos censos con cierto rigor en Europa fue en el año 1662 en que apareció en Inglaterra el primer estudio estadístico resaltable sobre la población titulado: "Observations of the London Bills of Mortality". El surgimiento de tres escuelas simultáneas en Europa en el siglo XVII provocó que la estadística floreciera como ciencia:

- La administrativa (de origen alemán): dirigida a proporcionar información al Estado; sus exponentes principales fueron Vito de Seckendorf, Germann Coring y Godofredo de Achenwalt.
- La probabilística (de origen italiano): devenida francesa, que consideraba problemas relacionados con el azar; sus exponentes principales fueron Blaise Pascal, Pierre de Fermat, Pierre Simon (marqués de Laplace), Simeón Denis Poisson, los hermanos Jean, Jacques y Daniel Bernoulli y además el alemán Carl Friedrich Gaus.
- La demográfica (de origen inglés): consideraba problemas actuariales, relacionados con situaciones financieras y las rentas de la población; sus exponentes principales fueron Petty, Halley, Davenant y Graun.

En el inicio del siglo XIX surge una segunda escuela inglesa que se ocupaba de los problemas de la agricultura y la biometría; contaba como exponentes a Sir Francis Galton, Carl Pearson, Ronald Fisher y William S. Gosset, quienes establecieron los fundamentos principales de la ciencia que se conoce en la actualidad como estadística.

La oposición a la utilización de procedimientos numéricos para el análisis de datos durante mucho tiempo fue la "posición oficial" entre muchos conocedores y desconocedores de los aspectos de la estadística, ya que se argumentaba que cada persona era un ser único y no podían hacerse conclusiones generalizables. Es con la aparición en el año 1937 del libro *Principles of the Medical Statistics* de Sir Austin Bradford Hill, que la utilización de la estadística empieza a marchar paralelamente con el desarrollo de la medicina y otras ciencias afines.

En la actualidad la estadística, vocablo derivado del griego statera (balanza), del latín status (estado, posición) o del alemán staat (estado, situación), constituye un campo del estudio relacionado con la recopilación, organización y resumen de los datos en su aspecto descriptivo, así como con la obtención de predicciones acerca de un conjunto de datos cuando solo se observa una parte de ellos en su aspecto inferencial.

La bioestadística es la rama de la estadística aplicada que estudia la utilización de métodos estadísticos en problemas médicos y biológicos. La mayoría de los profesionales de las áreas biomédicas que atienden el comportamiento del estado de salud en las poblaciones requieren conocer los principios que guían la aplicación de los métodos estadísticos a los temas propios en esta área del conocimiento.

Considerando el protagonismo moderno de la bioestadística en las investigaciones científicas se hace notar que puede guardar estrecha relación con las ciencias sociales, las de la economía, las de la información, la gran mayoría de las biológicas y la metodología de la investigación, entre otras.

El desarrollo y nivel de aplicación de la bioestadística en el campo de la investigación en todas las ciencias en general y sobre todo en las ciencias de la salud, ha experimentado en los últimos años un incremento considerable. Este auge ha estado asociado al sucedido en el área de las ciencias informáticas, lo que ha llevado a una sociedad donde la computadora personal se ha convertido en un instrumento de uso habitual, lo que hace posible la realización de pruebas estadísticas que de forma rutinaria hubieran requerido un gran esfuerzo, así como el manejo de grandes cantidades de datos (minería de datos) que hasta hace algún tiempo eran impensables por un investigador.

Paralelamente a este fenómeno se expresa otro factor que contribuye a un mayor desarrollo y aplicación de la bioestadística en tiempos actuales y es el relacionado con el entorno profesional, donde el desempeño de la profesión, tradicionalmente limitado a la justa aplicación de los conocimientos adquiridos en los estudios de pregrado, exige ahora que la investigación forme parte inseparable de la labor profesional cotidiana. Este requerimiento ha demandado que todos los profesionales necesiten la formación, aprendizaje y aplicación de técnicas básicas bioestadísticas para hacer frente al necesario análisis de datos en las investigaciones. Este cambio en los requisitos para la labor profesional representa un nuevo reto para los actuales y futuros profesionales y por ello casi todas las carreras universitarias de una forma u otra incluyen la bioestadística en sus planes de estudio de pregrado y posgrado.

El conocimiento de los conceptos y principios básicos de la bioestadística es clave para realizar el diseño de las futuras investigaciones y poder utilizar *a priori* los elementos indispensables de esta para la planificación, el manejo e interpretación de los datos a recoger, relacionados con los sucesos observados y cuyo propósito final sea condensar la información para que pueda ser trasmitida y entendida por otros, así como poder extrapolar sus conclusiones a las poblaciones de las que fueron tomadas las medidas experimentales.

La aplicación práctica de los principios y los aspectos metodológicos de la bioestadística constituyen un soporte imprescindible para la obtención, procesamiento y análisis de la información en los diferentes diseños de estudios investigativos. Al mismo tiempo facilita la evaluación de la significación de los resultados y la conformación de nuevos modelos predictivos de ciertos eventos en la naturaleza y la sociedad para su mejor comprensión.

El vertiginoso desarrollo alcanzado en la aplicación de las diferentes técnicas bioestadísticas en las investigaciones científicas ha conducido a una ampliación de las posibilidades con que cuenta el investigador para el procesamiento de la información obtenida, su análisis y discusión de los resultados relacionados con diversos problemas de investigación, así mismo las opciones contenidas en cada una de estas técnicas forman parte de las herramientas aliadas del investigador que le permiten visualizar nuevas aristas de los fenómenos de la naturaleza y la sociedad con incidencia en la calidad de vida del ser humano en sus componentes biológico, sicológico y social de forma integrada.

La generación de nuevas hipótesis por el investigador, en el estudio de antiguos y nuevos problemas sociales que enfrenta el ser humano, la mayoría de las ocasiones derivan de nuevos resultados obtenidos e interpretaciones en las investigaciones científicas, en las cuales están estampadas habitualmente algunas aplicaciones novedosas de técnicas bioestadísticas en el procesamiento de la información recogida, tanto en la vertiente descriptiva como en la inferencial o en ambas, en medio de diseños de estudio cada vez más complejos. Esto puede conducir a una mayor aproximación en la comprensión de los elementos involucrados en esos problemas, todo lo que descansa y lleva el sello determinante en la mirada oportuna y experiencia acumulada del investigador contando con la asesoría de expertos, en total armonía con el nivel de conocimientos generalizados existentes para el tema en cuestión.

La manera primaria de introducirse en el modo operante del investigador es reconocer el camino que recorre desde la identificación del problema a investigar a partir de su observación activa, la descripción, la generación de hipótesis, el trazado de sus objetivos, la planificación en el empleo del material con que cuenta y los métodos a desarrollar para la consecuente obtención de sus resultados, que deben ser evaluados como parte del procesamiento estadístico de la información, mediante las respectivas pruebas de significación o probabilidad estadística en relación con los objetivos trazados, para arribar a conclusiones que permitan responder con la subsiguiente y esperada creación

de nuevos conocimientos de valor científico y social al problema identificado e hipótesis generadas a partir de este.

Posteriormente el investigador debe aplicar esos conocimientos generados según el problema abordado y comprueba el impacto de su puesta en práctica para lo que nuevamente requiere de la aplicación de diferentes opciones de técnicas bioestadísticas que brinden la oportunidad de evaluar la significación y comprobar la validez de los nuevos resultados con el respectivo nivel o grado de confianza que se les debe considerar en términos de probabilidades.

Si se pone en la mirada concentrada la aplicación de la metodología de la investigación y su relación con la bioestadística, la lectura que se puede realizar es que la primera lleva generalmente como primer paso a la observación y descripción de los fenómenos que rodean al hombre, donde el investigador apela a la experiencia acumulada (lo que se sabe) y trata de identificar problemas con los que convive; así como trata de establecer teóricamente posibles factores, situaciones o nexos que puedan explicar, influir o ser relevantes en estos hechos. De ello brotan hipótesis con posibles explicaciones relacionadas con ese problema, al tiempo que el investigador intenta la experimentación sobre variados aspectos que pueden influir o relacionarse con el problema detectado.

Los pasos que comienza a dar apuntan a recoger todos los datos de interés de ciertos rasgos, características o aspectos relevantes (caracterizados como variables) involucrados en esos fenómenos, seleccionando preferentemente de forma aleatoria a los individuos de la población en las muestras que la representen para obtener estos datos, que los agrupa, clasifica y almacena en bases de datos para realizar su descripción detallada en un primer intento de aproximación a la comprensión del fenómeno en la población de estudio y se vale de diferentes tipos de descriptores en esta etapa investigativa denominada como *exploratoria-descriptiva*. De ello deviene la necesidad de relacionar estos datos, compararlos, asociarlos, correlacionarlos, predecirlos, etcétera, empleando los parámetros poblacionales (estimados a partir de los muestrales), así como diferentes tipos de estadígrafos para contrastar las hipótesis estadísticas (aspecto analítico), lo que permite obtener resultados que requieren una evaluación de su significación y sobre lo que se facilita poder arribar a conclusiones extrapolables a nivel poblacional que ayuden a responder a los objetivos investigativos trazados previamente desde el nivel muestral, como "representante" de la población de estudio (etapa inferencial).

La aplicación de las diferentes opciones de las técnicas bioestadísticas por parte del investigador para la generación de nuevos conocimientos, pasa por diferentes etapas en los diferentes diseños investigativos. Estas técnicas, acciones y procedimientos se pueden aplicar desde la etapa de recolección de la información para la puesta al día respecto al fenómeno estudiado, pasando por la selección de sus muestras, las respectivas variables, los estimadores-parámetros y estadígrafos, hasta las más rigurosas aplicadas en la descripción e inferencia que garanticen la confiabilidad de los resultados obtenidos en términos probabilísticos, todo esto puede ser utilizado para contribuir a dar respuesta

al problema de investigación con la consiguiente aplicación práctica y evaluación de su impacto a escala social.

Cabe señalar que la evaluación de la significación de los resultados obtenidos, como parte de la inferencia bioestadística, es un momento trascendental del procesamiento estadístico de la información, que se efectúa a partir de los datos recogidos de los estimadores respectivos de las variables seleccionadas en las muestras de estudio y con apego a los objetivos del estudio. Esta evaluación es imprescindible para determinar el verdadero valor, peso, dimensión, cuantía o, en fin, hasta donde puede ser "confiable" un determinado resultado o efecto probabilísticamente hablando e intentar lograr una homogeneidad de criterios en la interpretación de los resultados por los investigadores, lo que equivale a tratar de "hablar un lenguaje común", al manifestarse sobre ciertos resultados obtenidos.

Desde la perspectiva investigativa descriptiva, si se cuenta con todos los descriptores necesarios y se emplean con el rigor científico esperado, se puede obtener una información muy valiosa que "abre las puertas de par en par" a la perspectiva inferencial, la que expresa la arista medular para realizar un eficaz procesamiento estadístico de la información en la mayoría de los diseños de investigación modernos; toda vez que incluye las distintas y más variadas opciones de las novedosas técnicas bioestadísticas de ese procesamiento mediante las pruebas de significación o para el contraste de hipótesis con que se cuenta en el trabajo con muestras y variables y viceversa para ese objetivo.

La ejecución del procesamiento estadístico de la información desde la combinación de estas perspectivas puede ser facilitado si se está proveído de la tutela, guía o punto referencial (de ser posible) de algún algoritmo presentado como diagrama de flujo, donde queden involucrados y relacionados de forma concatenada los elementos fundamentales y más frecuentemente empleados que se requieren para ese procesamiento en el manejo de muestras y variables según los respectivos contrastes de hipótesis, así como en la evaluación de los resultados obtenidos.

Como parte de la evaluación estadística de los resultados obtenidos de las investigaciones científicas un lugar preponderante lo ocupa la determinación de la significación estadística. En el momento actual coexisten con sus respectivas ventajas y desventajas dos tipos de enfoques, paradigmas o modelos principales: el frecuentista o clásico (que fue el primero en aparecer y el más ampliamente empleado) y el bayesiano (más moderno), siendo ambos son "atraídos por un centro de gravedad" que constituyen las pruebas de significación, las que utilizan el contraste de hipótesis estadísticas respectivas y son aplicables a la mayoría de las situaciones investigativas o problemas de estudio específicos.

Por medio del modelo frecuentista se realiza un contraste de hipótesis estadísticas de nulidad o nula (H_0) y alternativa (H_1), mediante el empleo de los parámetros poblacionales estimados a partir de las muestras, los que son combinados en formulaciones matemáticas (para el caso específico en cuestión) que constituyen o devienen en los

estadísticos de prueba o de contraste experimentales observados-calculados. El conjunto de los posibles valores numéricos ordenados a tomar por el estadístico de contraste o de prueba y la probabilidad o frecuencia que cada uno de sus valores representa forman pares ordenados para un sistema gráfico de coordenadas x, y, respectivamente, en el que al unir los puntos que representan esos pares se puede describir gráficamente la curva de la distribución de probabilidades para el estadígrafo o estadístico de contraste o de prueba específico, que representa una variable aleatoria conocida y cuyos valores teóricos se representan tabuladamente, para su consulta respectiva a cada contraste de hipótesis en el anexo 1.

Al realizar el contraste de hipótesis el valor experimental observado-calculado del estadístico de contraste para las muestras de estudio se compara con el valor crítico o teórico-esperado-tabulado de referencia, según el nivel de confianza seleccionado por el investigador de ese estadístico. Esa comparación se efectúa ubicando gráficamente (siguiendo los ejes x, y) el valor calculado-observado del estadístico experimental respecto a la región crítica de rechazo de la hipótesis nula (H_0), delimitada según la distribución teórica-esperada del estadístico de contraste hacia los extremos en el eje x por su valor correspondiente al valor α preseleccionado por el investigador, lo que facilita el análisis acerca de la posible significación estadística del resultado observado, por lo que se debe tomar de forma complementaria en el análisis, el valor de probabilidad estimada bajo la hipótesis nula o donde se cumpliría esta, según el resultado observado del estadístico de contraste, denominado como p-valor, valor p o probabilidad crítica, que expresa un índice de la fuerza de la evidencia observada con los datos muestrales y se expresa como el resultado numérico probabilístico obtenido bajo H_0, en relación con el estadístico de contraste experimental observado-calculado.

La comparación numérica con la representación gráfica agregada del p-valor versus el nivel de significación α, que expresa la probabilidad de equivocarnos al rechazar H0 al ser esta cierta o de cometer un error tipo 1, facilita la toma de decisión respecto al rechace o no de H_0 y la consiguiente aceptación de la hipótesis alternativa o no, respectivamente. La regla de decisión para el rechace o no de H0 expresa que si el p-valor es mayor que el valor α prefijado no existen suficientes evidencias para el rechace de H_0.

El modelo bayesiano de forma general también emplea el contraste de hipótesis estadísticas, aunque no se busca el rechazo o no de la hipótesis nula (H) ni de escoger entre esta y la hipótesis complementaría (F); más bien identifica en que medida una hipótesis es más razonable que la otra. Este método toma en cuenta de forma directa en su análisis la información histórica previa del valor de los parámetros estimados e involucrados en el procesamiento estadístico de la información (a diferencia del frecuentista), establece las probabilidades a priori para estos y las combina con la verosimilitud estimada de esos parámetros estimados a partir de los datos en las muestras estudiadas y como resultado del procesamiento estima el valor del factor de Bayes a favor y en contra de la hipótesis nula, así como las probabilidades *a posteriori* de cumplirse la hipótesis nula en

relación con el tipo de contraste seleccionado, lo que es expresado mediante intervalos de probabilidades, lo que constituye un enfoque más cercano a la labor cotidiana del personal médico-investigador.

Para el eficaz procesamiento estadístico de la información en la inferencia bioestadística en los distintos tipos de diseños de estudio, al menos es necesario apoyarse en tres pilares principales:

- Conocer los aspectos conceptuales básicos de los contenidos de la bioestadística, que resultan necesarios para el manejo de muestras y variables según los correspondientes tipos de contrastes de hipótesis, utilizando las propuestas algorítmicas que combinan esos elementos y que permiten representar el despliegue de las diferentes opciones de las técnicas bioestadísticas incluidas en los contrastes en los distintos tipos de diseños de estudio con el empleo de los paradigmas, modelos o enfoques frecuentista y bayesiano.
- Conformar las bases de datos que permitan recoger, organizar y contener toda la información necesaria para cumplir los objetivos investigativos trazados, haciendo hincapié en su correspondencia con las necesidades aportadas por la operacionalización de las variables y las muestras requeridas a utilizar, al tiempo que su modo de confección permita el despliegue de las distintas opciones de técnicas bioestadísticas del procesamiento de la información contenidas en las distintas pruebas de contrastes de hipótesis o de significación empleadas para la evaluación de la significación de los resultados obtenidos para un cierto nivel de confianza, en estrecha relación con las situaciones investigativas derivadas del problema e hipótesis de investigación planteadas.
- Aplicar las distintas herramientas u opciones de técnicas bioestadísticas contenidas en los distintos programas o procesadores estadísticos empleados para la inferencia bioestadística, con la ayuda de bases de datos conformadas según el número y tipo de muestras y variables requeridos en los distintos contrastes de hipótesis a seleccionar para las situaciones investigativas respectivas y poder interpretar la salida de los resultados que brindan en relación con las hipótesis relacionadas con el problema de investigación.

Para tener éxito en la obtención e interpretación consecuente de los resultados que pueden brindar las distintas, modernas y potentes herramientas contenidas en los diversos procesadores estadísticos se hace necesario, en primer lugar, el dominio de los principios y toda la base teórica conceptual involucrada, así como el adecuado manejo de muestras y variables según los contrastes de hipótesis en la inferencia bioestadística a seleccionar mediante guías algorítmicas prácticas por parte del investigador, sobre la que se levantan estos instrumentos o herramientas informáticas que ponen a su servicio los profesionales que se dedican a la programación de estos sistemas automatizados de la información.

Aunque en la actualidad todos los aspectos relacionados con las formulaciones matemáticas empleadas en el procesamiento de la información estadística, desde el cálculo (estimación) del tamaño muestral para cada diseño de estudio, pasando por los de los parámetros y estadísticos de prueba según su distribución de probabilidades, hasta llegar incluso a la formulación de funciones matemáticas que expresan un modelo multivariado, quedan incluidos en toda la parte de la programación de los procesadores estadísticos y sus cálculos solo están "en manos de potentes computadoras" (invisibles para el investigador), es muy oportuno tener el conocimiento de la procedencia de las herramientas estadísticas informáticas para esos paquetes estadísticos, lo que puede incrementar la posibilidad de emplear eficazmente estos instrumentos en la investigación científica, solo así se puede tener una imagen integral e histórica de la bioestadística, que permita conectar el conocimiento referente a de dónde vienen, dónde se encuentran y hacia dónde se proyectan las opciones de las técnicas y tendencias en el procesamiento estadístico de la información con que se cuenta para su mejor utilización.

Un aspecto muy importante al enfrentarse a un problema investigativo, sus hipótesis y los objetivos trazados para su comprobación, es la orientación acerca del material a seleccionar y el método a seguir para el procesamiento de la información que se pretende recoger respecto al problema. Aunque no existe una "receta única" para esa orientación, el elemento primario a considerar puede ser la consulta de los objetivos de la investigación y en base a esto planificar y ejecutar desde la perspectiva bioestadística descriptiva o la inferencial el procesamiento estadístico de la información.

Cada objetivo trazado en la investigación desde la perspectiva analítica-inferencial se puede intentar hacerlo corresponder con las posibles y respectivas pruebas de significación del amplio espectro de opciones de técnicas estadísticas incluidas que aportan los paquetes estadísticos computacionales, mediante el contraste de hipótesis estadísticas. Una probable manera de facilitar la selección de cada tipo de prueba de significación es mediante el establecimiento de un vínculo de las pruebas con los verbos empleados en los enunciados de los objetivos analíticos que se trace el investigador a partir de su problema de investigación y para lo cual es requisito indispensable el adecuado manejo de muestras y variables, en lo que hay implícito gran cantidad de elementos básicos conceptuales y algorítmicos, todo esto puede ser la clave para dirigir la mirada en los diferentes diseños de estudio al realizar el procesamiento estadístico siguiendo los principios de la metodología de la investigación y bajo un formato algorítmico que constituya un elemento didáctico de incalculable valor para el estudiante universitario y el investigador joven.

El dominio del primero de los pilares antes expuestos puede constituir una pieza básica para intentar la realización del procesamiento estadístico eficaz de la información en lo referente fundamentalmente a la inferencia bioestadística, en franco acercamiento a las principales y más frecuentes situaciones investigativas en que se puede llevar a efecto, en el contexto de las investigaciones científicas modernas y como tal expresa el objetivo fundamental de este libro.

Contenido

SECCIÓN I. CONCEPTOS BÁSICOS EN BIOESTADÍSTICA RELACIONADOS CON EL MANEJO DE MUESTRAS Y VARIABLES/ 1

Capítulo 1. Aspectos de metodología de la investigación involucrados en el manejo de muestras y variables/ 2

Formas de abordar el problema de una investigación según el alcance de sus resultados/ 2
Tipos de diseño de estudio a emplear en la investigación epidemiológica/ 3
- Planteamiento del problema/ 7
- Fijación de los objetivos/ 7
- Formulación de las hipótesis de trabajo/ 7
- Definición de la unidad de observación y de la unidad de medida/ 8
- Determinación de la población y de la muestra/ 8
- Recolección de la información/ 9
- Crítica, clasificación y ordenamiento de la información/ 9
- Presentación de la información/ 9
- Análisis de los resultados obtenidos/ 10
- Publicación de los resultados obtenidos/ 10

Momentos de la investigación para el control o disminución de sesgos/ 10
Etapas en el procesamiento estadístico de la información/ 12
Estadística descriptiva e inferencial en el procesamiento de la información en las investigaciones biomédicas/ 12
- Estadística descriptiva/ 12
- Estadística inferencial/ 12

Elementos de metodología de la investigación y de la bioestadística empleados en la técnica del metaanálisis/ 13
- Formulación del problema/ 20
- Búsqueda de la literatura/ 21
- Codificación de las características y resultados de los estudios/ 21
- Análisis estadístico e interpretación/ 22
- Publicación del metaanálisis / 22

Capítulo 2. Aspectos relacionados con el empleo de las muestras/ 29

- Representatividad de una muestra/ 29

Conceptos fundamentales del muestreo/ 31
- Tipos de métodos de muestreo a emplear/ 33

Formulario empleado en el cálculo del tamaño de muestra para la estimación de parámetros por intervalos de confianza y para el contraste de hipótesis en estudios epidemiológicos/ 39
- Estimación de parámetros por intervalos de confianza/ 39

Estimación de parámetros del contraste de hipótesis/ 44
Procedimiento para la selección de la muestra empleando algunos tipos de muestreo/ 56
Muestreo simple aleatorio/ 56
Muestreo sistemático (Silva, 2000: 126-131)/ 56
Muestreo aleatorio estratificado
(Silva, 2000: 126-131)/ 58
Muestreo por conglomerados monoetápico (Silva, 2000: 157-159)/ 59
Muestreo por conglomerados bietápico (Silva, 2000: 177-180)/ 59
Muestreo por conglomerados monoetápico estratificado (Silva, 2000: 181-190)/ 60
Muestreo por conglomerados bietápico estratificado (Silva, 2000: 192-194) / 61
Clasificación de la muestra en bioestadística/ 62

Capítulo 3. Aspectos relacionados con el empleo de las variables/ 63

Clasificación de las variables de las muestras/ 63
Descripción de los tipos de las medidas de resumen para variables cuantitativas, numéricas, paramétricas, métricas (continuas y discretas)/ 66
Descripción de los tipos de medidas de resumen para variables cualitativas, categóricas o no paramétricas (nominales y ordinales)/ 79
Medidas de frecuencia/ 79
Medidas de asociación/ 82
Medidas de impacto/ 91
Descripción de las posibles representaciones gráficas según el tipo de variables de estudio/ 94
Gráfico de barras simples/ 97
Gráfico de barras múltiples/ 98
Gráfico de barras compuestas o proporcionales/ 98
Gráfico de Pareto/ 99
Gráfico de pastel, de sectores o circular/ 99
Gráfico aritmético simple/ 100
Histograma/ 100
Polígono de frecuencias/ 101
Gráfico de frecuencias acumuladas u ojivas/ 102
Gráfico de tronco y hojas/ 103
Gráfico de caja y bigote/ 104
Diagrama de dispersión/ 105
Gráfico de series de tiempo/ 106
Tablas de contingencia, de doble entrada o cruzadas de filas por columnas/ 106
Tipos de tablas de contingencia (cruzadas) para determinar la asociación de 2 variables cualitativas/ 110
Tablas 2x2 simples/ 110
Tablas 2x2 estratificadas/ 114
Tablas de contingencia 2xN simples/ 116
Tablas de contingencia 2xN estratificadas/ 117
Tablas MxN/ 119
Eficacia de un procedimiento diagnóstico (prueba diagnóstica)/ 120
Medidas o índices a estimar que permiten evaluar la eficacia de una prueba diagnóstica/ 121
Concepto y el vínculo de estadígrafos muestrales y poblacionales/ 125
Estimación por intervalo de confianza
de los parámetros (poblacionales)
más frecuentemente empleados/ 129
Estadígrafos que permiten determinar relaciones o asociaciones entre variables/ 152

Coeficiente de correlación lineal simple muestral (momento producto de Pearson)/ 153
Coeficiente de concordancia de rangos de Kendall/ 154
Coeficiente de correlación punto biserial/ 156
Coeficiente de determinación/ 157
Coeficiente de correlación por rangos de Spearman poblacional estimado o poblacional muestral/ 157
Coeficiente de correlación lineal múltiple/ 158
Coeficiente de determinación múltiple o correlación parcial/ 158
Coeficiente de regresión lineal simple muestral/ 158
Coeficiente de regresión lineal múltiple/ 159
Análisis de riesgo en estudios analíticos observacionales/ 162
Prueba de la probabilidad exacta o prueba exacta de Fisher/ 171
Estimación puntual y por intervalo de confianza de las medidas de frecuencia, asociación e impacto en los estudios epidemiológicos analíticos observacionales empleando tablas de contingencia/ 176
Estudio de corte o transversal/ 176
Resultados en estudios transversales/ 182
Estudio de caso-control no apareado/ 183
Resultados en estudios de caso-control no apareado (sin *matching* o no apareamiento)/ 187
Estudio de cohorte con datos de incidencia acumulada/ 188
Resultados en estudios de cohorte, datos de la incidencia acumulada/ 194
Estudio de cohorte con datos de tasas de incidencia/ 194
Resultados en estudios de cohorte, datos por tasas de incidencia/ 199
Para estudio de cohorte, datos de densidad de incidencia/ 200
Resultados en estudios de cohorte, datos de densidad de incidencia/ 201
Para estudio de cohorte, datos de mortalidad estandarizada/ 201
Resultados en estudios de cohortes, datos de mortalidad estandarizada/ 202
Para estudio de caso-control apareado (caso 1-control 1)/ 203
Estudios caso-control pareados (caso 1-control 1) / 204
Resultados en Estudios caso-control pareados (con apareamiento o *matching*), 1 caso para 1 control/ 206
Estudios caso-control pareados, K controles por caso (1< K< 5)/ 206
Resultados en Estudios caso-control pareados (con apareamiento o *matching*), relación caso: 1, control >1/ 207
Formulario resumen de las medidas de frecuencia, asociación e impacto empleadas en estudios epidemiológicos analíticos observacionales / 208
Medidas de frecuencia/ 208
Medidas de asociación/ 209
Medidas de impacto/ 215
Teoría de probabilidades y las distribuciones de probabilidades de variables aleatorias/ 217
Distribuciones discretas/ 222
Distribuciones continuas/ 233

Capítulo 4. Algunos aspectos conceptuales de la inferencia bioestadística/ 263

Contraste de hipótesis estadísticas como parte del procesamiento estadístico de la información y bases para su interpretación/ 263
Pasos generales a seguir en el contraste de hipótesis/ 267
Resumen de algunas pruebas de significación empleadas en el enfoque de riesgo/ 446

Para tablas 2x2 simples/ 446
Para tablas 2x2 estratificadas/ 457
Para tablas 2xn simples/ 460
Para tablas 2xn estratificadas/ 462
Para tablas MxN (M, N ≥ 2) / 463
Usos de los estadísticos de prueba en contrastes de hipótesis mediante pruebas paramétricas/ 464
Estadístico de prueba Z ~ N (0, 1) (distribución normal estándar)/ 464
Estadístico de prueba t-student (t)/ 464
Estadístico de prueba chi-cuadrado (χ^2) / 465
Estadístico de prueba F de Snedecor-Fisher/ 466
Opciones de técnicas bioestadísticas a realizar en el contraste de hipótesis estadísticas/ 469
Por el enfoque o paradigma frecuentista/ 469
Por el enfoque o paradigma bayesiano/ 472
Métodos bivariados de análisis de regresión lineal simple y de la correlación lineal simple/ 472
Correlación entre variables/ 484
Premisas y aspectos necesarios para realizar un análisis de varianza (ANOVA) de una vía o de un factor/ 487
Nociones sobre el empleo del ANOVA en diseños de estudio experimentales/ 502
Métodos de análisis multivariados/ 510
Métodos de dependencia/ 511
Métodos de interdependencia/ 513
Método multivariado de regresión logística/ 515
Interpretación de los estadígrafos resultantes a mostrar en una regresión logística/ 520
Bases para el ajuste del modelo/ 522
Relevancia del cociente o razón de verosimilitudes (Colectivos de autores, 2014)/ 523
Método multivariado de análisis de sobrevida o supervivencia/ 534
Concepto de curvas de sobrevida/ 534
Construcción de curvas de sobrevida o de supervivencia/ 535
Comparación de curvas de sobrevida empleando el método no paramétrico del *log-rank* o prueba generalizada de Savage/ 538
Comparación de curvas de sobrevida (método paramétrico)/ 540
Método multivariado de regresión de Cox/ 542
Interpretación de los estadígrafos resultantes a mostrar en una regresión de Cox/ 544
Aplicación práctica al estimar un modelo de predicción por medio de la regresión de Cox/ 546
Método multivariado regresión lineal múltiple o de regresión múltiple/ 548
Primer paso del modelo de regresión múltiple/ 552
Segundo paso del modelo de regresión múltiple/ 554
Tercer paso del modelo de regresión múltiple/ 564
Cuarto paso del modelo de regresión lineal múltiple/ 565
Predicción con el modelo de regresión lineal múltiple/ 566
Ventajas y desventajas de la regresión lineal múltiple/ 567
Relación entre la regresión lineal múltiple y la correlación lineal múltiple/ 569
Método multivariado de análisis discriminante/ 574
Pasos a tener en cuenta al realizar el análisis discriminante/ 575
Algunas claves a tener en cuenta al realizar el análisis discriminante/ 576

Aspectos teóricos-matemáticos involucrados en el análisis discriminante/ 577
Método multivariado de análisis multivariante de la varianza (MANOVA)/ 596
Evaluación del valor teórico dependiente del MANOVA/ 600
Análisis múltiple de covarianza (MANCOVA)/ 601
Validación de los resultados obtenidos en el MANOVA/ 602
Método multivariado de análisis factorial/ 602
Validación del análisis factorial/ 608
Consideraciones finales/ 608

Capítulo 5. Empleo del método o paradigma bayesiano en la inferencia bioestadística

Diferencias entre el método frecuentista (clásico) y el bayesiano/ 612
Aplicación del paradigma bayesiano en el análisis de un método diagnóstico/ 625
Aplicación bayesiana en la inferencia bioestadística en una población con el empleo del odds ratio/ 626
Aplicación bayesiana en la inferencia bioestadística en una población, para la estimación de una proporción/ 627
Estimación de una proporción/ 627
Valoración de hipótesis sobre una proporción/ 629
Aplicación bayesiana en la inferencia bioestadística en dos poblaciones/ 632
Estimación de efectos basados en dos proporciones/ 632
Valoración de hipótesis sobre una diferencia de proporciones/ 633
Probar que la diferencia de proporciones está dentro de un intervalo/ 634
Aplicación bayesiana en la inferencia bioestadística en una población, empleo de una media/ 635
Estimación de una media/ 636
Valoración de hipótesis sobre el empleo de una media/ 637
Aplicación bayesiana en la inferencia bioestadística en dos poblaciones para la stimación de una diferencia de medias/ 639
Estimación de la diferencia de dos medias/ 639
Valoración de una hipótesis sobre la diferencia de dos medias/ 640
Aplicación bayesiana en la inferencia bioestadística en la valoración de pruebas convencionales/ 642
Prueba de comparación de medias/ 642
Prueba chi-cuadrado/ 643

SECCIÓN II. PROPUESTAS ALGORÍTMICAS PARA EL MANEJO DE MUESTRAS Y VARIABLES EN LA INFERENCIA BIOESTADÍSTICA/ 644

Capítulo 6. Propuestas algorítmicas esquematizadas para el manejo de muestras y variables / 645

Capítulo 7. Propuesta algorítmica no esquematizada para el manejo de muestras y variables con un enfoque frecuentista / 712

ANEXOS/ 712

Anexo 1 Aspectos de metodología de la investigación involucrados en el manejo de muestras y variables/ 712

Anexo 2. Formularios para el manejo de muestras y variables relacionados con los contrastes de hipótesis de la inferencia estadística y algunos aspectos afines/ 776

Estadísticos de prueba (una población con una variable)/ 784
Métodos paramétricos/ 784
Métodos no paramétricos/ 784
Intervalos de confianza (una población, una variable)/ 785
Estadísticos de prueba (una población con dos variables)/ 786
Métodos paramétricos/ 786
Métodos no paramétricos/ 787
Estadísticos de prueba (dos poblaciones con una variable)/ 788
Métodos paramétricos/ 788
Métodos no paramétricos/ 789
Estadísticos de prueba (más de dos poblaciones y una variable)/ 790
Métodos paramétricos/ 790
Métodos no paramétricos/ 792
Determinación del tamaño de muestra/ 793
Propiedades relacionadas con las probabilidades/ 795
Distribuciones de probabilidad/ 796

Anexo 3. Glosario de términos/ 797

Bibliografía/ 811

Sección I

Conceptos básicos en bioestadística relacionados con el manejo de muestras y variables

La sección incluye cinco capítulos, que abordan los aspectos esenciales de los contenidos básicos de la bioestadística, tanto en la vertiente descriptiva como en la inferencial, los que son muy importantes tomar en cuenta para poder encausar el procesamiento estadístico de la información recogida en las investigaciones científicas, mediante el correcto manejo de muestras y variables y viceversa en el aspecto inferencial, lo que está en estrecha relación con los principios de la metodología de la investigación y la epidemiología. Estos aspectos constituyen el escalón primario para el estudio de la bioestadística asociada con otras ramas de las ciencias en el largo camino del proceso investigativo.

Capítulo 1

Aspectos de metodología de la investigación involucrados en el manejo de muestras y variables

Formas de abordar el problema de una investigación según el alcance de sus resultados

Al analizar el problema de una investigación las formas de abordarlo de acuerdo con el alcance que pueden tener sus resultados son:

- Forma descriptiva: a partir de la observación de un fenómeno e identificación de un determinado problema de investigación se realiza la recolección, organización, resumen y presentación de toda la información estadística empleando unas muestras con las respectivas variables que permitan la caracterización del problema.
- Forma exploratoria: a partir de la observación de un fenómeno e identificación de un determinado problema de investigación se realiza la recolección, organización, resumen y presentación de toda la información estadística que permita familiarizar al investigador con el problema a investigar.
- Forma analítica-explicativa: a partir de la observación de un fenómeno e identificación de un determinado problema de investigación se determina cuales factores están relacionados con este, es decir, se analizan relaciones de causa-efecto.
- Forma experimental: a partir de la observación de un fenómeno e identificación de un determinado problema de investigación se introduce y manipula un factor causal para la determinación de un efecto correspondiente, con empleo de grupo control obtenido de forma aleatorizada, lo que es de amplia aplicación en estudios de clínica médica (ensayos clínicos).
- Forma *cuasi* experimental: a partir de la observación de un fenómeno e identificación de un determinado problema de investigación se introduce y manipula un factor causal para la determinación de un efecto correspondiente, pero con el empleo de grupo control obtenido de forma no aleatorizada, lo que es de amplia aplicación en estudios de clínica médica.

– Forma de evaluación: a partir de la observación de un fenómeno e identificación de un determinado problema de investigación se realiza s evaluación en este determinando la eficiencia, eficacia y efectividad de algún factor involucrado en el problema investigativo.

Los tipos de investigaciones epidemiológicas según los objetivos que persiguen se pueden clasificar en:
– Investigaciones descriptivas (para algunos autores incluye las denominadas investigaciones exploratorias).
– Investigaciones analíticas.
– Investigaciones de intervención.

La investigación según la aplicabilidad de sus resultados se puede clasificar en:
– Investigación fundamental.
– Investigación aplicada.
– Investigación de desarrollo.

Se presenta de forma resumida los diferentes niveles de actuación en la investigación epidemiológica según sus objetivos (Tabla 1.1).

Tabla 1.1. Niveles de actuación de la investigación epidemiológica

Nivel de actuación	Objetivos
Nivel descriptivo	Describir, en forma cuantitativa y cualitativa, la distribución de un problema de salud en relación con variables de tiempo, lugar y persona
Nivel de conocimiento etiológico	Inferir la participación causal o etiológica de las variables condicionantes en el problema de salud en estudio
Nivel de intervención	Evaluar el impacto probable de las medidas tomadas en relación con el problema de salud

Tipos de diseño de estudio a emplear en la investigación epidemiológica

De manera general los diseños de estudio en la investigación epidemiológica pueden ser clasificados en:
1. Según la posición que adopta el investigador:
   a) Descriptivo: el investigador solo realiza la caracterización del fenómeno sin su aspecto analítico, es decir, resume, organiza y presenta gráficamente la información contenida en una matriz de datos sin realizar análisis comparativo-asociativo con pruebas de hipótesis, ni tampoco la modificación de las condiciones del fenómeno observado y suelen clasificarse en:

- Estudios transversales: el estudio de uno o más eventos epidemiológicos, en el que se examinan la relación entre estos eventos (enfermedad) y una serie de variables, en una población y momento determinados. También se les conoce como estudios de corte o estudios de prevalencia, por ser esta la medida de frecuencia por excelencia de este tipo de investigación
- Estudios longitudinales: puede definirse como la revisión de uno o más eventos epidemiológicos durante un intervalo de tiempo suficientemente largo en virtud de las características del evento. Estos estudios suelen diseñarse en forma prospectiva (hacia delante), retrospectiva (hacia atrás) o ambispectiva (en ambas direcciones).

b) Analítico-observacional: el investigador describe el fenómeno, realiza el análisis e interpreta los resultados a partir de comparaciones o asociaciones significativas de lo observado con empleo de la estimación de parámetros, contraste de hipótesis y elaboración de modelos estadísticos-matemáticos (aspecto inductivo-inferencial), pero no modifica las condiciones del fenómeno observado). Sus tipos pueden ser:
   - De corte o transversal: se selecciona a individuos de la población sin tener en cuenta su estado respecto a la enfermedad o exposición, se observa cuáles de ellos están expuestos y cuales no o cuales presentan la enfermedad o no y se compara el grupo expuesto con el no expuesto para cierta enfermedad simultáneamente.
   - De caso y control: se selecciona la población en función de su estado respecto a la enfermedad, se escoge a un grupo de enfermos (casos) y un grupo de no enfermos y se comparan los antecedentes de exposición de cada grupo, es decir, el comportamiento de riesgos en grupos casos (expuestos y no expuestos) y grupo control (expuesto y no expuesto) de forma retrospectiva para un periodo puntual.
   - De cohorte: se selecciona a un grupo de individuos de la población sin la enfermedad de interés y se dividen en función de su exposición o no, es decir, un grupo de expuestos y un grupo de no expuestos y se observa como se desarrolla la enfermedad en ambos grupos durante un periodo de tiempo continuo. Los datos de las variables pueden presentarse como incidencia acumulada, densidad de incidencia y mortalidad estandarizada.

c) Analítico experimental: el investigador tiene un enfoque inductivo-deductivo y modifica las condiciones del fenómeno observado de forma aleatoria, empleando grupo control *versus* grupos casos obtenidos de forma aleatoria. Sus tipos son: ensayos clínicos controlados, evaluación de nueva tecnología e intervención en el terreno social entre los mas frecuentes.

d) Analítico cuasi experimental: el investigador tiene un enfoque inductivo-deductivo y realiza manipulaciones del fenómeno observado de forma no aleatoria, carece de grupos de comparación (verdaderos controles) al no emplear muestra obtenida de forma aleatoria.

2. Según el periodo y secuencia del estudio pueden ser los diseños de estudio:
   a) Longitudinal: toda la información necesaria para la investigación es recogida a lo largo de un periodo prolongado de tiempo, desde el pasado o hacia el futuro.
   b) Transversal: toda la información necesaria para la investigación es recogida al unísono, en un momento puntual del tiempo.
3. Según el tiempo de ocurrencia de los hechos y el registro de la información pueden ser los diseños de estudio:
   a) Retrospectivo: se recoge toda la información necesaria para realizar la investigación de un periodo pasado respecto al comienzo del estudio, es decir, desde el presente hacia el pasado.
   b) Prospectivo: se recoge toda la información necesaria para la investigación a medida que va pasando el tiempo, es decir, desde el presente al futuro.
   c) Combinado: se obtiene la información necesaria para la investigación desde el comienzo del estudio en lo adelante (hacia el futuro, componente prospectivo), pero se emplea información recogida previamente a la vez (componente retrospectivo) relacionada con el mismo problema de investigación.

En la tabla 1.2 se presenta la relación del tipo de diseño de estudio, su subtipo y los objetivos respectivos.

Tabla 1.2. Resumen del tipo de diseño de estudio, su subtipo y objetivos

Tipo de diseño	Subtipo del diseño de estudio	Objetivos
Experimentales: se realiza la manipulación por el investigador del factor en estudio con empleo de la aleatorización	Laboratorio	– Probar hipótesis etiológicas, estimar comportamientos agudos y efectos biológicos – Determinar eficacia de intervenciones para modificar factores de riesgo en una población
	Ensayos clínicos	– Probar hipótesis etiológicas y estimar efectos en la salud a largo plazo – Probar la eficacia de intervenciones para modificar el estado de salud – Sugerir factibilidad de intervenciones poblacionales
	Intervenciones comunitarias	– Identificar personas o grupos con alto riesgo – Probar eficacia y efectividad de intervenciones clínicas para modificar el estado de salud de determinadas poblaciones – Sugerir programas y políticas de salud pública

Tipo de diseño	Subtipo del diseño de estudio	Objetivos
Cuasi experimentales: se realiza la manipulación por el investigador del factor en estudio sin empleo de la aleatorización	Ensayos clínicos y de laboratorio	– Los mismos que en los estudios experimentales
	Programas y políticas de salud	– Evaluar el alcance de los objetivos de la salud pública – Determinar problemas no anticipados o consecuencias de implementar y las razones para el éxito o fracaso de una intervención – Comparar costos y beneficios de una intervención – Sugerir cambios en las actuales políticas y programas de salud
Observacionales: no se realiza la manipulación por el investigador del factor en estudio	Descriptivos	– Estimar la frecuencia de ciertas enfermedades o características, tendencias temporales e identificar individuos enfermos – Generar nuevas hipótesis y sugerir la racionalidad de nuevos estudios
	Analíticos	– Probar hipótesis etiológicas específicas y estimar efectos crónicos en la salud – Generar nuevas hipótesis etiológicas y sugerir mecanismos de causalidad – Generar hipótesis y sugerir su potencial para prevenir enfermedades

El método estadístico parte de la observación de un fenómeno y como no puede siempre mantener las mismas condiciones predeterminadas o a voluntad del investigador, deja que actúen libremente, pero se registran las diferentes observaciones y se analizan sus variaciones.

Para el planeamiento de una investigación, por norma general, se siguen las siguientes etapas en las cuales de forma general está la utilización del método estadístico:

1. Planteamiento del problema.
2. Fijación de los objetivos.
3. Formulación de las hipótesis de trabajo.
4. Definición de la unidad de observación y de la unidad de medida.

5. Determinación de la población y de la muestra.
6. Recolección de la información.
7. Crítica, clasificación y ordenamiento de toda la información.
8. Presentación de la información recolectada y ordenada.
9. Análisis de los resultados obtenidos.
10. Publicación de las conclusiones.

Las ideas más importantes resumidas para cada uno de estos aspectos de una investigación se presentan a continuación y es más evidente la utilización del método estadístico en las etapas (3 a la 10) referentes desde el planteamiento de hipótesis estadísticas hasta emitir conclusiones en términos probabilísticos según los resultados obtenidos.

Planteamiento del problema

Al abordar una investigación se debe tener bien definido qué se va a investigar y por qué se pretende estudiar algo, es decir, se debe establecer una delimitación clara, concreta e inteligible sobre los fenómenos que se pretenden estudiar, para lo cual se deben tener en cuenta, entre otras cosas, la revisión bibliográfica del tema, para ver su accesibilidad y consultar los resultados obtenidos por investigaciones similares, someter las proposiciones básicas a un análisis lógico, es decir, se debe hacer una ubicación histórica y teórica del problema.

Fijación de los objetivos

Luego de tener claro lo que se pretende investigar, se debe presupuestar hasta dónde se quiere llegar; en otras palabras, fijar cuáles son las metas y objetivos. Estos deben plantearse de tal forma que no haya lugar a confusiones o ambigüedades y debe, además, establecerse diferenciación entre lo de corto, mediano y largo plazo, así como entre los objetivos generales y los específicos.

Formulación de las hipótesis de trabajo

Una hipótesis es, ante todo, una explicación provisional de los hechos objeto de estudio y su formulación depende del conocimiento que el investigador posea sobre la población investigada. Una hipótesis estadística debe ser susceptible de docimar, esto significa que se deba probar para su aceptación o rechazo.

Una hipótesis que se formula acerca de un parámetro (media, proporción, varianza, entre otras), con el propósito de rechazarla, se llama *hipótesis de nulidad* y se representa por H_0 y a su hipótesis contraria se le llama *hipótesis alternativa* y se representa por H_1.

Definición de la unidad de observación y de la unidad de medida

La unidad de observación, entendida como cada uno de los elementos constituyentes de la población estudiada, debe definirse previamente, resaltando todas sus características, pues, al fin de cuentas, es a ellas a las que se les hará la medición.

La unidad de observación puede estar constituida por uno o varios individuos u objetos y denominarse respectivamente simple o compleja.

El criterio sobre la unidad de medición debe ser previamente definido y unificado por todo el equipo de investigación. Si se trata de medidas de longitud, volumen, peso, entre otras, debe establecerse bajo qué unidad se tomarán las observaciones ya sea en metros, pulgadas, libras, kilogramos, entre otras.

Asociado a la unidad de medida, deben establecerse los criterios sobre las condiciones en las cuales se ha de efectuar la toma de la información.

Determinación de la población y de la muestra

Estadísticamente la población se define como un conjunto de individuos o de objetos que poseen una o varias características comunes. No se refiere esta definición únicamente a los seres vivientes, una población puede estar constituida por los habitantes de un país o por los peces de un estanque, así como por los establecimientos comerciales de un barrio o las unidades de vivienda de una ciudad.

Existen, desde el punto de vista de su manejabilidad, poblaciones finitas e infinitas. Aquí el término infinito no es tomado con el rigor semántico de la palabra; por ejemplo, los peces dentro de un estanque son un conjunto finito, sin embargo, en términos estadísticos puede ser considerado como infinito.

La muestra es un subconjunto de la población a la cual se le efectúa la medición para estudiar las propiedades del conjunto del cual es obtenida.

En la práctica estudiar todos y cada uno de los elementos que conforman la población no es aconsejable, ya sea por la poca disponibilidad de recursos, por la homogeneidad de sus elementos, porque a veces es necesario destruir lo que se está midiendo o por ser demasiado grande el número de sus componentes o no se pueden controlar; por eso se recurre al análisis de los elementos de una muestra para hacer inferencias respecto al total de la población. Existen diversos métodos para calcular el tamaño de la muestra y también para tomar los elementos que la conforman, pero no es el objetivo de este texto estudiarlos, pero se debe tener en cuenta que la muestra debe ser representativa de la población y sus elementos escogidos al azar para asegurar la objetividad de la investigación.

Recolección de la información

Una de las etapas más importantes de la investigación es la recolección de la información, en la que se ha de partir, a menos que se tenga experiencia con muestras análogas, de una o varias muestras piloto en las cuales se ponen a prueba los cuestionarios y se obtiene una aproximación de la variabilidad de la población para calcular el tamaño exacto de la muestra que conduzca a una estimación de los parámetros con la precisión establecida.

El establecimiento de las fuentes y cauces de información, así como la cantidad y complejidad de las preguntas, de acuerdo con los objetivos de la investigación, son decisiones que se han de tomar teniendo en cuenta la disponibilidad de los recursos financieros, humanos y de tiempo, y las limitaciones que se tengan en la zona geográfica, el grado de desarrollo, la ausencia de técnica, entre otros.

Es importante descubrir dónde está la información y cómo y a qué "costo" se puede conseguir, es determinar si la encuesta se debe aplicar por teléfono, por correo o si se necesitan agentes directos que recojan la información, establecer su número óptimo y preparar su entrenamiento adecuado.

Crítica, clasificación y ordenamiento de la información

Después de haber reunido toda la información pertinente se necesita la depuración de los datos recogidos. Para hacer la crítica de una información es fundamental el conocimiento de la población por parte de quien depura para poder detectar falsedades en las respuestas, incomprensión a las preguntas, respuestas al margen, así como todas las posibles causas de nulidad de una pregunta o nulidad de todo un cuestionario.

Separado el material de "desecho" con la información depurada se procede a establecer las clasificaciones respectivas y con la ayuda de hojas de trabajo, en las que se establecen los cruces necesarios entre las preguntas, se ordenan las respuestas y se preparan los modelos de tabulación de las diferentes variables que intervienen en la investigación.

El avance tecnológico y la popularización de los medios digitales como las computadores y teléfonos móviles actuales hacen que estas tareas, manualmente difíciles, puedan ser realizadas en corto tiempo.

Presentación de la información

Una información estadística adquiere más claridad cuando se presenta en la forma adecuada. Los cuadros, tablas y gráficos facilitan el análisis, pero se debe tener cuidado con las variables que se van a presentar y la forma de hacerlo. No es aconsejable saturar un informe con tablas y gráficos redundantes que, antes que claridad, crean confusión.

Una tabla es un resumen de información respecto a una o más variables, que ofrece claridad al lector sobre lo que se pretende describir, por lo que para su fácil interpretación una tabla debe tener por lo menos un título adecuado el cual debe ser claro y conciso, la tabla propiamente dicha con los correspondientes subtítulos internos y la cuantificación de los diferentes ítems de las variables y las notas al pie que muestren claridad sobre situaciones especiales de la tabla u otorguen los créditos a la fuente de la información. De la misma forma los gráficos y cuadros pueden contribuir a describir aspectos relevantes de la información recogida y complementar la tabulación realizada. La elección de determinada tabla o gráfico para mostrar los resultados debe hacerse no solo en función de las variables que relaciona, sino del lector a quien va dirigido el informe.

Análisis de los resultados obtenidos

La técnica estadística ofrece métodos y procedimientos objetivos que convierten las especulaciones de primera mano en aseveraciones cuya confiabilidad puede ser evaluada y ofrecer una premisa medible en la toma de una decisión. Es el análisis donde se cristaliza la investigación. Esta es la fase de la determinación de los parámetros y estadísticos muestrales para las estimaciones e inferencias respecto a la población, el ajuste de modelos y las pruebas de las hipótesis planteadas para establecer y redactar las conclusiones definitivas.

Publicación de los resultados obtenidos

Toda conclusión a la que se arribe y responda a los objetivos previamente trazados a partir del problema identificado es digna de ser comunicada a un auditorio. Puede haber otros estudiosos del mismo problema a quienes se les puede aportar información, conocimientos y otros puntos de vista acerca de este, lo que enriquece el cúmulo de conocimientos en ese entonces.

El fluido intercambio entre los investigadores en cuanto a las conclusiones arribadas para el problema de investigación analizado permite también la identificación de nuevas aristas del problema de investigación u otro relacionado que puede constituir un nuevo motivo para reemplear el método estadístico con lo cual se cierra así el ciclo concerniente a este.

Momentos de la investigación para el control o disminución de sesgos

Un aspecto muy importante a tomar en cuenta al realizar el diseño de investigación son los sesgos, que constituyen errores cometidos por el investigador desde la planificación

hasta la culminación de la investigación, lo que puede dar al traste con los resultados investigativos y esto compromete el valor científico del estudio. Es preciso saber, por tanto, los principales momentos o áreas en que se puedan cometer:

- Selección y obtención de la muestra.
- Procedimiento para la recogida de la información (datos).
- Selección y operacionalización de las variables.
- Medición de las variables.
- Control de factores de confusión.

Los sesgos más frecuentemente señalados en los diseños de estudio son:

- Sesgo de selección: error debido a diferencias sistemáticas entre las características de los sujetos seleccionados para el estudio y las de los que no fueron seleccionados, es decir, cuando la muestra estudiada no es adecuada para lograr la representatividad de la población de origen. Ocurre cuando hay un error sistemático en los procedimientos utilizados para seleccionar los sujetos del estudio. Este sesgo conduce a una estimación del efecto distinta del obtenible para la población total.
- Sesgo de información o de observación: es el resultado de diferencias sistemáticas en la forma en que son obtenidos los datos, al medir la exposición o la ocurrencia de la enfermedad, dando lugar a una deficiente calidad (precisión) de la información entre los grupos que se comparan. Esta clasificación incorrecta puede presentarse en clasificación incorrecta no diferencial o aleatoria y diferencial.
- Sesgo de confusión: situación en la que la medición del efecto de una exposición sobre un riesgo se altera debido a la asociación de esa exposición con otro factor que influye sobre la evolución del resultado estudiado.

Los sesgos, aunque generalmente no pueden ser totalmente eliminados en las investigaciones, pueden controlarse realizando un diseño orientado a este fin o mediante la realización de un análisis estadístico adecuado.

El valor obtenido de cualquier medición en una variable consta de dos partes: el valor verdadero y el error de medida, es decir:

Valor obtenido = Valor verdadero + Error de medida

Donde:

- Error de medida se compone de dos partes: una sistemática (sesgo) y la otra aleatoria (al azar).
- El sesgo puede alterar la validez de los resultados, mientras que el error aleatorio, que por definición sucede por igual en todos los grupos y subgrupos, no afecta en principio la validez, pero puede disminuir la posibilidad de encontrar una verdadera asociación.

Etapas en el procesamiento estadístico de la información

Durante la planificación de un diseño de estudio a la hora de realizar el procesamiento estadístico de la información se debe tomar en cuenta sus etapas:

- Organización de la información (sus datos):
  - Clasificar los datos procedentes de las variables del estudio a partir de escalas para datos cuantitativos (valores continuos y discretos) y cualitativos (atributos o categorías: nominal y ordinal).
  - Distribuir las frecuencias para datos cuantitativos (simples o agrupados) y cualitativos.
- Resumen o condensación de la información (sus datos): emplear medidas descriptivas para variables cuantitativas (de tendencia central, dispersión, posición, forma de distribución) y cualitativas (de frecuencia, asociación, impacto).
- Presentación de la información (sus datos): emplear tablas o gráficos según el tipo de variable y sus medidas descriptivas respectivas para ilustrar y facilitar el entendimiento de toda la información previamente resumida.

Estadística descriptiva e inferencial en el procesamiento de la información en las investigaciones biomédicas

Estadística descriptiva

- Organización de la información: clasificar los datos, distribuirlos según sus frecuencias y operacionalizar las variables.
- Resumen o condensación de la información: empleo de las respectivas medidas para cada variable, de modo que faciliten su operacionalización.
- Presentación de la información (de los datos) de las variables empleando las diferentes medidas respectivas de cada una, de forma tabulada o graficada.

Estadística inferencial

Selección de una técnica de muestreo de preferencia probabilística que facilite emplear una muestra lo más representativa posible de la población para:

- Estimación puntual y por intervalo de confianza de medidas respectivas de variables poblacionales (parámetros) a partir de los estimadores muestrales para un determinado nivel de confianza.

– Estimación puntual y por intervalos de confianza de diferencias y cocientes de parámetros (poblacionales) a partir de estimadores (muestrales) para un determinado nivel de confianza.
– Comprobar el comportamiento de ciertos parámetros estimados (poblacionales) por medio del análisis de sus diferencias, cocientes, asociaciones y en la construcción de modelos probabilísticos a partir de los estimadores (muestrales) respectivos con empleo de contrastes de hipótesis o pruebas de hipótesis que permitan evaluar la significación estadística o probabilidades de sus resultados para cierto nivel de confianza.

De modo general la estadística descriptiva brinda las herramientas necesarias para caracterizar a un grupo de elementos o unidades de análisis, pero sin pretender generalizar esos resultados a un grupo mayor, objetivo que sí persigue la estadística inferencial.

Elementos de metodología de la investigación y de la bioestadística empleados en la técnica del metaanálisis

En el siglo xx se vio el paso desde el descubrimiento de las penicilinas hasta la manipulación genómica. El avance de los conocimientos biomédicos está dirigido a mejorar la calidad de vida de los individuos y esto se logra a través de una mejor asistencia clínica y el uso adecuado de las herramientas diagnósticas y terapéuticas, en tanto debe existir un nexo entre la producción de conocimiento y su aplicación práctica. Para ello existe una cadena de eventos que inician con la aparición de nuevos conocimientos, su transcripción al papel (artículos, guías, entre otros), su transmisión y llegada a los profesionales de la salud. Este último, en principio, depende de todos los anteriores, a los que se suman la interpretación y aplicabilidad que se les pueda dar posteriormente.

Muchas veces los avances de la ciencia no son tenidos en cuenta en la práctica clínica hasta años después. La aplicabilidad de los conocimientos está directamente relacionada con las creencias y experiencias de los que van a utilizarlos. Asociado a estas barreras se encuentra el desarrollo exponencial que sufre la información. El gran número de publicaciones científicas existentes actualmente pone de manifiesto la necesidad de gestionar de manera eficiente el conocimiento científico.

Una forma de resumir y poder analizar esta información es agruparla a través de la técnica que se conoce como metaanálisis, el cual constituye un mecanismo para la integración de datos, mediante el análisis cuantitativo de los resultados de distintos estudios relacionados con una pregunta específica. Según la propuesta de la Scottish Intercollegiate Guidelines Network (SIGN) y la Agència d'Avaluació de Tecnologia Mèdica (AATM) el metaanálisis es considerado dentro de la evidencia más alta (evidencia 1) y su sola presencia genera grados de recomendación A.

El metaanálisis consiste en realizar una sistemática identificación, valoración, síntesis y, si es pertinente, la agregación estadística de todos los estudios sobre un mismo tema, siguiendo un método explícito y predeterminado.

El método del metaanálisis puede ser utilizado para cambiar los resultados de muchos estudios de alcance reducido, para llegar a una conclusión autorizada, en lugar de aguardar años, dicho esto a modo de ejemplo, hasta que terminen ensayos clínicos para comprobar si un nuevo tratamiento es eficaz o no.

El metaanálisis se ha utilizado ampliamente en sicología y en ciencias de la educación y actualmente tiene una gran aplicación en diferentes ramas de la biomedicina. Es de carácter retrospectivo y no debe emplearse en forma indiscriminada. Se diferencian de los artículos de revisión en que generalmente enfocan un solo aspecto clínico, por ejemplo, una causa, la exactitud de pruebas utilizadas para el diagnóstico, un tratamiento. Con este método se pueden combinar los resultados cuantitativos de varios estudios para evaluar un efecto que será de forma más precisa que en cada estudio individual, pero manteniendo siempre su validez clínica y estadística. Ofrece ventajas sobre los enfoques tradicionales de revisión, pues impone a quien los lleva a cabo un estudio sistemático con criterios explícitos, facilita la investigación cuantitativa y con ello la síntesis de los resultados de estudios individuales.

Los metaanálisis representan un instrumento para compendiar grandes números de estudios relevantes. Este tipo de artículo se acepta en la actualidad como una disciplina científica y su elaboración como un artículo de investigación. Para llevar a cabo esos estudios se exige un diseño cuidadoso, con un protocolo igual al de cualquier trabajo de investigación. Su validez está dada, siempre y cuando se consignen claramente las cuestiones que se planteen, los criterios para incluirlos y que sean reproductibles los estudios que se impusieron. La idea de fondo para el metaanálisis es entregar un conocimiento sintetizado, reduciendo la subjetividad, por lo que ha adquirido una importancia creciente en la toma de decisiones y conductas en medicina clínica, aunque paralelamente cuando se está familiarizado con esta metodología, también se debe conocer sus limitaciones y no confiar ciegamente en sus resultados. Pese a que la información entregada por esta herramienta epidemiológica parece ser más precisa y consistente, debe ser complementaria al juicio clínico en el proceso de toma de decisiones médicas y como todo trabajo de investigación debe ser leído críticamente.

Básicamente el metaanálisis es una revisión sistemática en la cual se combinan matemáticamente los resultados de varios estudios para contestar una misma pregunta. El Centro Cochrane define una revisión sistemática como la agrupación de estudios que tienen por objetivo reunir toda la evidencia empírica que cumple unos criterios de elegibilidad previamente establecidos, para responder una pregunta específica de investigación.

Para la realización de este tipo de revisiones se emplea una metodología sistemática que pretende reducir los sesgos dándole validez interna, de forma tal que ofrezca información adecuada para las tomas de decisión.

Existe una serie de criterios que una revisión sistemática debe cumplir: existencia de objetivos claros, criterios (de inclusión y exclusión) preestablecidos de elegibilidad, planteamiento de la metodología y que esta sea reproducible, búsqueda sistematizada que permita la identificación de los estudios que cumplan los criterios planteados, validación de los resultados de los estudios que cumplan los criterios y presentación de una síntesis de los estudios que se incluyeron, tanto de las características como de los resultados.

Un metaanálisis permite, esencialmente, fusionar estadísticamente los resultados de diversos estudios independientes, pero en cierta medida combinables entre sí con el objeto de verificar si existe un efecto que pueda ser evaluado estadísticamente. La primera referencia que se tiene de un estudio de este tipo es un artículo de Pearson en 1904 que resumió los resultados de 11 estudios que se referían a los efectos de la vacuna contra la fiebre tifoidea.

El término metaanálisis fue inicialmente aplicado en el ámbito de las ciencias sociales y la sicología y solo a partir de la década de los años ochenta del siglo xx empezó a usarse en medicina, para generalizarse en la década de los años noventa del mismo siglo en las publicaciones médicas. Glass tiene el mérito de haber acuñado la palabra metaanálisis en un artículo publicado en 1976, definiéndolo como el análisis estadístico de una colección de resultados con el propósito de integrar sus hallazgos.

Matemáticamente la agrupación de los resultados es una estrategia que inicia Pearson (1904) en el siglo xx, sin embargo, es al final de la década de los años setenta del siglo xx que Glass, estadounidense, presidente de la Asociación Americana de Investigación en Educación y Cochrane, epidemiólogo inglés, impulsan su uso, con gran impacto en el área biomédica a partir de la década de los años ochenta en pleno siglo xx.

La revisión sistemática y el metaanálisis son términos complementarios, pero no sinónimos. Generalmente las revisiones sistemáticas incluyen un metaanálisis; pero no necesariamente. Un metaanálisis es una técnica estadística que permite estimar cuantitativamente el beneficio del total de estudios técnicamente adecuados recopilados en una revisión sistemática. En términos simples la revisión sistemática es la metodología y el metaanálisis es un conjunto de herramientas estadísticas. Un metaanálisis puede efectuarse después de una revisión sistemática o bien sin esta, simplemente combinando los resultados de varios estudios, lo que no es muy aconsejable porque probablemente se omitiría información de importancia, incurriendo en un sesgo de publicación. Por lo tanto, la calidad de un metaanálisis (análisis cuantitativo) depende de la calidad de la revisión sistemática que lo precede (análisis cualitativo).

No está de más recordar que los estudios susceptibles de ser incluidos en un metaanálisis deben establecer una comparación del efecto entre al menos dos grupos. La principal dificultad de combinar los resultados de múltiples estudios radica en la naturaleza diversa de estos, tanto en diseño como en la metodología utilizada; los más útiles son los ensayos clínicos controlados y aleatorizados, otros estudios clínicos son menos estrictamente controlados y, por lo tanto, tienen un mayor sesgo, como los es-

tudios de cohorte, los de casos y control o de corte, entendiendo como sesgo un error sistemático que lleva a una distorsión del parámetro que refleja el efecto investigado alejándolo del valor verdadero. Además, existe diversidad en los tamaños muestrales y poblaciones estudiadas, lo que puede causar que la combinación de esos estudios no entregue un efecto consistente y por el contrario este sea heterogéneo o diverso. Todo esto debe ser tomado en cuenta al evaluar la calidad y credibilidad de un metaanálisis.

Las etapas en que puede dividirse un metaanálisis de forma genérica son:

- Extracción de los datos de cada estudio individual y obtención del punto estimado o medida resumen del efecto.
- Decidir si es apropiado combinar los resultados de los estudios y presentar el efecto final, esto involucra dar mayor importancia a los estudios que aportan mayor información. Habitualmente estos son los estudios con mayor cantidad de participantes (mayor tamaño muestral).

El metaanálisis puede emplear como sustrato los datos agregados de cada estudio o los datos individuales (*pooled analyses*) de cada integrante del estudio.

En el caso de contar con los datos individuales se obtendrá un metaanálisis de mayor calidad, lo que permite una mayor fiabilidad del resultado final y a su vez analizar los subgrupos, que al emplear los datos agregados no es posible. Sin embargo, hay que tener en cuenta que los estudios individuales son más caros, requieren de mayor tiempo, así como de la colaboración entre los investigadores de los diferentes estudios.

Otro tipo de metaanálisis es el acumulativo, que es aquel que luego de haberse realizado y por la posterior aparición de nuevos estudios se vuelve a realizar acumulando los estudios nuevos a los ya existentes. El metaanálisis de redes se emplea para el cálculo del efecto directo a través del análisis de trabajos de investigación que contengan las intervenciones de estudio evaluadas entre si y de cada una de estas con placebo.

Existen múltiples limitaciones a tener en cuenta a la hora de evaluar un metaanálisis:

- La calidad de los estudios que lo incluyen: estos pueden ser ensayos clínicos, estudios observacionales, estudios diagnósticos, entre otros. Al momento de evaluar los resultados finales de los metaanálisis este hecho tiene que ser tenido en cuenta. Los sesgos propios de los diferentes estudios que lo incluyen, influyen en la calidad de los resultados que ofrecen y, por lo tanto, estudios sesgados ofrecen resultados sesgados, por lo que es importante, además, evaluar el tamaño muestral, las variables y el análisis estadístico de los datos, al analizar los resultados.
- El sesgo propio de publicación como consecuencia de lo cual muchos estudios no son publicados, porque los resultados van en contra de los intereses de los promotores del mismo o no son significativos.
- El sesgo de elección de los artículos que se incluirán en el metaanálisis.
- La variabilidad o heterogeneidad de los estudios, ya sea en características del contexto (por ejemplo, el sociodemográfico), de la metodología (por ejemplo, el grupo

control), medición de puntos de final (*end points*) diferentes, diferencias en la magnitud de los resultados, entre otros.
– Dificultad en algunas interpretaciones de los resultados obtenidos.

El objetivo principal que engloba la realización de un metaanálisis es resolver si existe un efecto, si este efecto es positivo o negativo (favorable o desfavorable) y cuantificarlo en una cifra. El efecto se puede cuantificar como una medida de asociación binaria, si se trata de datos categóricos, o bien una diferencia estandarizada de medias en caso de datos continuos.

Para corroborar la existencia de un efecto, obviamente deben seleccionarse los trabajos en que exista una comparación entre dos grupos. Idealmente estos estudios deben ser exclusivamente ensayos clínicos aleatorizados para disminuir el riesgo de la aparición de sesgos. Cuando se realiza un metaanálisis es de gran utilidad examinar los resultados de cada ensayo incluido y los globales en una representación gráfica. Esto se efectúa mediante un gráfico denominado *forest plot,* que permite una interpretación sencilla y rápida en este asunto, todo lo cual se asocia al resultado de pruebas de significación estadísticas.

Los objetivos específicos que generalmente llevan al empleo de un metaanálisis son:
– La necesidad de guía para la toma de decisiones a través de revisión sistemática con resúmenes matemáticos de sus resultados.
– Dar una solución a la incertidumbre que no ha podido ser resuelta por la existencia de resultados dispares en los estudios existentes.
– Estimar el tamaño del efecto (al sumar el "n" poblacional).
– Evaluar la heterogeneidad de los estudios.
– Evaluar, si es posible, los subgrupos.
– Aportar información para la realización de evaluaciones de tecnologías empleadas en salud.

Existen múltiples guías que se pueden emplear a modo de recordatorio de cómo leer un metaanálisis. En 1999 se publica la Declaración Quality of Reporting of Metanalysis (QUORUM), que ofrece las primeras normas de publicación con el objetivo de mejorar la calidad de revisiones sistemáticas y metaanálisis. Estas se han ido actualizando hasta llegar al 2010 donde se publica la Declaración Preferred Reporting Items for Systematic reviews and MetAnalyses (PRISMA), una actualización y ampliación de QUOROM consistente en una lista de comprobación estructurada de 27 ítems, en la que a diferencia de la anterior se detalla y justifica cada ítem que debe ser tenido en cuenta para la elaboración de estos estudios de metaanálisis. PRISMA sigue la estructura básica de un artículo científico, detallando los contenidos específicos que deben reportarse en cada sección; también propone la utilización de un diagrama de flujo para ilustrar el proceso de la revisión.

Cabe destacar que, además de la ampliación en el número de ítems como resultado de recientes avances metodológicos, PRISMA tiene una aplicabilidad mayor que QUOROM, ya que no se limita solo a los metaanálisis de ensayos clínicos aleatorizados, sino que también es útil para las revisiones de otro tipo de estudios (Tabla 1.3).

Tabla 1.3. Declaración PRISMA. Lista de comprobación de los ítems para incluir en la publicación de una revisión sistemática con metaanálisis o sin este

Sección/tema	Ítem	Recomendaciones
Título		
Título	1	Identificar la publicación como una revisión sistemática, metaanálisis o ambos
Resumen		
Resumen estructurado	2	Facilitar un resumen estructurado que incluya, según corresponda, antecedentes, objetivos, fuente de los datos, criterios de elegibilidad de los estudios, participantes e intervenciones, evaluación crítica de los estudios y métodos de síntesis, resultados, limitaciones, conclusiones e implicaciones de los hallazgos principales, y número de registro de la revisión sistemática
Introducción		
Justificación	3	Describir la justificación de la revisión en el contexto de lo que ya se conoce sobre el tema
Objetivos	4	Plantear de forma explícita las preguntas que se desea contestar detallando los participantes, las intervenciones, las comparaciones, los resultados y el diseño de los estudios (PICOS)
Métodos		
Protocolo y registro	5	Indicar sí existe un protocolo de revisión al que se pueda acceder y dónde (por ejemplo, dirección web), y si está disponible, la información sobre el registro incluyendo su número de identificación
Criterios de elegibilidad	6	Especificar las características de los estudios (por ejemplo, PICOS, duración del seguimiento) y de las publicaciones (por ejemplo, años abarcados, idiomas o estado de publicación) utilizadas como criterios de elegibilidad, y su justificación
Fuentes de los datos	7	Describir todas las fuentes de los datos (por ejemplo, bases de datos bibliográficas y periodos de la búsqueda, contacto con los autores para identificar estudios adicionales) utilizadas en la búsqueda, y la fecha de la última búsqueda realizada
Búsqueda	8	Presentar la estrategia completa de búsqueda electrónica en, al menos, una base de datos, incluyendo los límites utilizados, de tal forma que pueda ser reproducida.

Sección/tema	Ítem	Recomendaciones
Selección de los estudios	9	Detallar el proceso de selección de los estudios (por ejemplo, estudios cribados, evaluados para su elegibilidad, incluidos en la revisión sistemática y, cuando sea pertinente, incluidos en el metaanálisis)
Proceso de extracción de datos	10	Describir los métodos para la extracción de los datos de las publicaciones (por ejemplo, uso de formularios pilotado, extracción de datos por duplicado y de forma independiente) y cualquier proceso para obtener y confirmar datos por parte de los investigadores
Lista de los datos	11	Listar y definir ludas las variables para las que se buscaron datos (por ejemplo, PICOS, fuente de financiación), y cualquier asunción y simplificación que se haya hecho
Riesgo de sesgo en los estudios individuales	12	Describir los métodos utilizados para evaluar el riesgo de sesgo de los estudios individuales (especificar si se realizó sobre el estudio o en los resultados), y cómo esta información se ha utilizado en la síntesis de datos.
Medidas de resumen	13	Especificar las principales medidas de resumen (por ejemplo, razón de riesgos o diferencia de medias)
Síntesis de resultados	14	Describir los métodos para manejar los datos y combinar los resultados de los estudios, cuando esto fue posible, incluyendo medidas de consistencia (por ejemplo, I2) para cada metaanálisis
Riesgo de sesgo entre estudios	15	Especificar cualquier evaluación del riesgo de sesgo que pueda afectar la evidencia acumulativa (por ejemplo, sesgo de publicación o de comunicación selectiva).
Análisis adicionales	16	Describir los métodos de análisis adicionales (por ejemplo, análisis de sensibilidad o de subgrupos, metarregresión), en el caso de que se hicieran, indicando cuáles habían sido preespecificados
Resultados		
Selección de los estudios	17	Facilitar el número de estudios cribados, evaluados para su elegibilidad e incluidos en la revisión, detallando las razones de las exclusiones en cada etapa, idealmente mediante un diagrama de flujo
Características de los estudios	18	Para cada estudio presentar las características para las que se extrajeron datos (por ejemplo, tamaño, PICOS y duración del seguimiento), y proporcionar las citas bibliográficas
Riesgo de sesgo de los estudios	19	Presentar datos sobre el riesgo de sesgo de cada estudio y, si está disponible, cualquier evaluación del sesgo en los resultados (véase el ítem 12).
Resultados de los estudios individuales	20	Para todos los resultados considerados (beneficios o daños), presentar: a) el dato resumen para cada grupo de estudio, y b) la estimación del efecto con su intervalo de confianza, idealmente de forma gráfica mediante un diagrama de bosque (*forestplot*)

Sección/tema	Ítem	Recomendaciones
Síntesis de los resultados	21	Presentar los resultados de todos los metaanálisis realizados, incluyendo los intervalos de confianza y las medidas de consistencia
Riesgo de sesgo entre estudios	22	Presentar los resultados de cualquier evaluación del riesgo de sesgo entre estudios (véase el ítem 15).
Análisis adicionales	23	Facilitar los resultados de cualquier análisis adicional, en el caso de que se hayan realizado (por ejemplo, análisis de sensibilidad o de subgrupos, metarregresión (véase el ítem 16)
Discusión		
Resumen de la evidencia	24	Resumir los hallazgos principales, incluyendo la fuerza de la evidencia para cada resultado principal; considerar su relevancia para los distintos usuarios clave (por ejemplo, proveedores de cuidados, usuarios y decisores en salud)
Limitaciones	25	Discutir las limitaciones de los estudios y de los resultados (por ejemplo, riesgo de sesgo), y de la revisión (por ejemplo, obtención incompleta de los estudios identificados, comunicación selectiva)
Conclusiones	26	Proporcionar una interpretación general de los resultados en el contexto de otras evidencias, así como las implicaciones para la futura investigación
Financiación	**27**	
Financiación	28	Describir las fuentes de financiación de la revisión sistemática y otro tipo de apoyos recibidos (por ejemplo, aporte de datos), así como el posible rol de los financiadores en la revisión sistemática

Los metaanálisis, al ser revisiones sistemáticas, objetivas y científicas, precisamente se desarrollan a través de una serie de fases claramente establecidas y que son muy similares a las propias de cualquier investigación empírica:

- Formulación del problema.
- Búsqueda de la literatura.
- Codificación de los estudios.
- Análisis estadístico e interpretación.
- Publicación del metaanálisis.

En cada una de estas fases el revisor debe detallar con claridad todas las decisiones adoptadas y los procedimientos aplicados, facilitando de esta forma la posibilidad de que otros puedan replicar el metaanálisis.

Formulación del problema

Al igual que en cualquier investigación empírica, el primer paso en un metaanálisis consiste en delimitar el problema objeto de la revisión, definiendo con precisión las va-

riables y conceptos implicados. El problema de investigación va a determinar tanto los objetivos del metaanálisis como las características que han de tener los estudios que se incluyan en la revisión. Es fundamental que el metaanálisis tome en consideración la forma en que se ha investigado previamente el fenómeno de interés, la validez de las conclusiones de tal investigación, las hipótesis con mayor apoyo empírico y las cuestiones para las que todavía no hay una respuesta clara y consistente a lo largo de los estudios.

Búsqueda de la literatura

Una vez formulado el problema el siguiente paso de un metaanálisis consiste en definir los criterios de selección que han de cumplir los estudios empíricos y desarrollar una búsqueda de la literatura lo más exhaustiva posible. Entre los criterios de selección figuran el periodo temporal de realización de los estudios, el diseño de investigación que deben tener, el lenguaje en el que deben estar escritos y otras características más específicas determinadas por los objetivos de cada metaanálisis. En cuanto a los procedimientos de búsqueda, deben combinar la consulta de bases de datos bibliográficas especializadas, por ejemplo, Medline, Cochrane Central y PEDro, con otras estrategias de búsqueda como la consulta de las referencias bibliográficas de trabajos relevantes y otros metaanálisis en el área de estudio. Así mismo es muy importante recurrir a otros procedimientos menos sistemáticos que permitan la localización de estudios no publicados (tesis doctorales, comunicaciones presentadas a congresos e informes técnicos internos de centros de investigación), por ejemplo, a través del contacto con autores o grupos de investigación especialistas en el tema.

Codificación de las características y resultados de los estudios

Los estudios que se integran en un metaanálisis, aún evaluando el mismo fenómeno, no suelen presentar idénticos resultados. Los resultados van a variar de estudio a estudio por múltiples factores aleatorios, pero también por determinadas características diferenciales de los estudios que se pueden identificar y codificar. Tales características moderadoras de los resultados pueden explicar buena parte de su heterogeneidad. Además de la codificación y del registro de las características moderadoras de los estudios, una cuestión esencial en metaanálisis es la forma en que se miden y registran los resultados, habitualmente a través de índices cuantitativos del tamaño del efecto como el *odds ratio*, la diferencia media tipificada, la diferencia de riesgos, la razón de riesgos o el coeficiente de correlación lineal de Pearson. La ventaja de cualquiera de estos índices es que permiten transformar a una misma medida resumen de variable los resultados procedentes de diferentes escalas de medida, por lo que se puede combinar y comparar a través de los estudios. Tanto las características moderadoras de los estudios como el cálculo de los tamaños del efecto se registran en un protocolo o

cuestionario de codificación, cuyas instrucciones para su correcta implementación se deben recoger en un manual de codificación. Las instrucciones deben detallarse de la forma más clara, precisa y objetiva, con el propósito de que cualquier otro investigador que decidiera aplicar el manual alcance similares valoraciones.

Análisis estadístico e interpretación

Una vez que se dispone de los datos fundamentales del metaanálisis a saber, las características y resultados de los estudios integrados, el siguiente paso es el de analizar e interpretar estos datos guiados por los objetivos e hipótesis de la revisión. Las técnicas de análisis habitualmente aplicadas responden a tres estrategias generales:

- Promediar los resultados o tamaños del efecto a través de los estudios.
- Evaluar la heterogeneidad de los resultados.
- Explicar tal heterogeneidad en función de las características diferenciales de los estudios integrados para el caso de que se confirme que los resultados son heterogéneos.

Los resultados del tamaño del efecto estudiado, según el estimador seleccionado para el tipo de variable a estudiar en los estudios seleccionados a comparar, se representan gráficamente mediante el *forest plot*, que en el caso de ensayos clínicos dicho tamaño del efecto se obtiene mediante la comparación de grupos tratados *versus* grupo control o con placebo y el procedimiento consiste en el análisis de la significación estadística de las diferencias en los tamaños de efecto, lo que además conlleva su presentación mediante valores puntuales promedios o ponderados e intervalos de confianza respectivos, según sea el caso. Conjuntamente se debe evaluar la heterogeneidad de los tamaños del efecto a través de la prueba que emplea el estadístico Q de Cochrane, cuyo resultado de no ser significativa, permite indicar si existe cierto grado de representatividad sobre el conjunto de estudios integrados en el metaanálisis del tamaño del efecto promedio estimado.

Respecto a la interpretación de todos estos análisis es muy importante que, junto con la significación estadística, los metaanálisis también informen de la significación clínica de los tamaños del efecto medio obtenido.

Publicación del metaanálisis

Una vez concluida la fase de análisis de datos e interpretación de los resultados, solo queda emprender la redacción formal del estudio metaanálisis para su posterior publicación. La estructura que se debe seguir es similar a la de cualquier investigación científica con introducción, método, resultados y discusión.

De forma general para llevar a cabo el metaanálisis se requiere emplear una estrategia por pasos como se presenta a continuación.

Estrategia de ingreso de estudios al metaanálisis

Todo estudio, independientemente del método que emplee, debe informar al lector sobre el problema que pretende resolver, la justificación del mismo y la pregunta a contestar. Muchas veces esta no es formulada explícitamente planteándose exclusivamente los objetivos de la investigación a realizar. La comprensión del problema permite delimitar el tema y evaluar si los objetivos del estudio están en relación con este. Lo cual es importante al evaluar los resultados finales obtenidos.

Una vez delimitado el tema y los objetivos debe estar detallada la información de la metodología empleada para alcanzarlos, pues debe permitir la replicación del trabajo realizado. Los autores deben haber detallado si se realizó un protocolo previo a la realización del estudio, si se realizaron modificaciones al mismo y las razones que la justifiquen. Debe constar, además, el tipo de estudio que se realizó, criterios de inclusión y de exclusión, participantes, las intervenciones y los resultados (*outcomes*) primarios y secundarios a evaluar.

En la metodología a emplear en el metaanálisis se debe especificar las fuentes de búsqueda incluidas en el trabajo. Se debe emplear una amplia gama de herramientas para obtener la mayor información posible. Se debe plantear, además, la consulta a investigadores sobre la existencia de estudios que por sus resultados contrarios a lo buscado o sin significación no se hayan publicado.

Otro tema que debe estar especificado está relacionado con la recolección de datos. En este ítem se debe haber considerado identificar qué tipos de datos se extraen, cómo se manejan, la medida del efecto del tratamiento, así como el manejo de los datos que faltan. También hay que tener en cuenta la especificación de cómo se evaluó el sesgo en los estudios incluidos, la heterogeneidad entre los estudios que ingresen al metaanálisis y el sesgo de publicación. Se debe exponer, si es necesario, como se realiza la síntesis de los datos, el análisis por subgrupos (si se realiza) y la metodología empleada para el análisis de sensibilidad.

Estrategia de análisis de los resultados obtenidos

Para esto se debe tener en cuenta:

– La forma de determinar el *tamaño del efecto* a evaluar: hay que considerar que el mismo varía en relación al tamaño de la muestra y al número de eventos existentes. La varianza es el elemento estadístico que toma en cuenta ambos componentes lo que permite ponderar adecuadamente el peso de cada estudio que ingresa al metaanálisis. En función del tipo de variable, si esta es cualitativa se puede emplear el *odds ratio*, el riesgo relativo o la diferencia de riesgos. En el caso de ser cuantitativa, si los estudios presentan igual medida para el efecto se puede emplear una media ponderada; en cambio, si se emplean diferentes medidas se emplea la media estandarizada que equivale a la media dividida por la desviación estándar o por el

coeficiente de correlación lineal de Pearson. Estos resultados pueden presentarse a través de un gráfico (*forest plot*) que los permite visualizar en forma resumida. En el eje de las "x" (abscisas) se representa la medida del efecto y en el eje de la "y" (ordenadas) se colocan los diferentes estudios empleando algún criterio de ordenación (por ejemplo, la fecha). También se traza una línea vertical que se ubica en el punto en el que no existe diferencia entre los eventos estudiados. Cada estudio debe contar con una estimación puntual, así como con un intervalo de confianza dado. A los lados del estudio se coloca el nombre del primer autor y el año del estudio y se identifica el "n" de eventos existentes para cada situación estudiada. Otro elemento presente en este tipo de gráficas es la medida del resultado y su intervalo de confianza para cada estudio.

- Tipo de modelo que se empleó para la medición del efecto, como se evaluó la heterogeneidad entre los estudios, el sesgo de publicación y el análisis de sensibilidad: en relación a los modelos que se pueden emplear para analizar estas variables en general son dos: el modelo de efectos fijos y el de efectos aleatorios. La diferencia está en la concepción de la población de estudios de partida:
  - En el *modelo de efectos fijos* se asume que la diferencia entre los tamaños del efecto de los estudios individuales se debe solo al hecho de que los estudios utilizan muestras de sujetos diferentes, o sea, la variabilidad es por error de muestreo aleatorio exclusivamente. Por lo tanto, cada estudio tendrá una varianza intraestudio diferente al de otro. En este caso se pueden emplear diferentes modelos como son el de Wolf, el de Mantel-Haenzel y el de Peto, estos últimos son más específicos.
  - En el *modelo de efectos aleatorios* no solo existe la variabilidad intraestudio (de muestreo) sino que, además, existe variabilidad entre los estudios. Es por esto que la varianza entre estudios es un valor constante para todos los estudios incluidos en el metaanálisis. El impacto de uno u otro modelo se observa en el grado de generalización de los resultados obtenidos en el metaanálisis. En el caso del modelo fijo solo permite generalizar los resultados a aquellos estudios que tengan características similares a los que se incluyeron en el metaanálisis. En cambio, con el modelo de efectos aleatorios se puede generalizar a una población mayor de estudios y es más representativo, sin embargo, existe un problema conceptual en el que no está claro el rango de diferencia admitido entre los estudios que participan en el metaanálisis como aceptable.
- Para evaluar si existe *heterogeneidad entre los diferentes estudios* que ingresaron al metaanálisis se debe plantear como hipótesis nula que no existe diferencia entre los ellos, es decir, que son homogéneos sus resultados. Para este cálculo se emplea el estadístico Q de Cochrane que se distribuye según una distribución chi-cuadrado de Pearson con k – 1 grados de libertad y permite determinar si existe o no heterogeneidad. Si no se rechaza la hipótesis de homogeneidad significa que el tamaño del efecto medio calculado en el metaanálisis puede usarse como resumen numérico

apropiado, en cambio, si se rechaza, es que existen variables de confusión que están influyendo en los resultados las que deben ser analizadas. Otro estadístico empleado es el índice de inconsistencia (I^2), que es menos influenciado por el número de estudios que se incluyen en el metaanálisis como ocurre con el Q de Cochrane, con lo que permite complementar la información que aporta el estadístico Q, al expresar el grado de heterogeneidad que existe entre los estudios; este estadístico se interpreta siguiendo una clasificación normativa en términos porcentuales como 25 % (heterogeneidad baja), 50 % (heterogeneidad moderada) y 75 % (heterogeneidad alta). La "heterogeneidad" de los estudios utilizados no es más que la variabilidad que existe entre los estimadores que se han obtenido en cada uno de esos estudios; mientras menos heterogéneos sean los resultados, más probable es que las diferencias entre los estudios se deban al azar y mientras mayor sea la heterogeneidad es más probable es que los resultados se deban a universos distintos.

- Para evaluar la existencia de *sesgo de publicación* se emplea un gráfico conocido como *gráfico de embudo o funnel plots*. Se utiliza para ver si existe heterogeneidad sistemática entre los estudios ingresados al metaanálisis. Un buen modelo es el que presenta una imagen de embudo simétrico. Cuando este no es así es que existe una relación entre el tamaño del estudio y el efecto del tratamiento. Se plantea que los estudios pequeños generalmente son de menor calidad y presentan sesgos por lo que influyen en el resultado final. La forma de construir este gráfico es ubicando en el eje vertical alguna de las medidas evaluadas que estiman el efecto del tratamiento; en este caso a menor intervalo de confianza el estudio queda más arriba y es más preciso; en el eje horizontal se mide el efecto del tratamiento a escala. Una vez establecidos los ejes se coloca la estimación puntual de cada estudio colocando una línea vertical donde se encuentra la estimación combinada del metaanálisis. A mayor número de estudios ingresados más se aprecia la forma de un embudo invertido. Están más dispersos los estudios con menor cantidad de individuos y eventos por lo que son menos precisos. Cuando el embudo es asimétrico puede deberse a que existen sesgos de publicación u otros, lo que puede hacer cuestionar la validez del estudio.
- Para el *análisis de sensibilidad en el metaanálisis*, por medio del cual se evalúa la robustez del estudio, se modifican las condiciones de realización del metaanálisis y se evalúa nuevamente el resultado. Básicamente se retira un estudio y se realizan nuevamente los cálculos del metaanálisis. Se reintegra y retira otro realizando un nuevo metaanálisis, tantas veces como estudios se incluyan en el metaanálisis. Con esto se evidencia si alguno de los estudios incluidos modifica por si solo el resultado final. En el caso que los resultados de los diferentes metaanálisis tengan igual dirección, magnitud y significación estadísticas se dice que el metaanálisis tiene resultados robustos. Una apreciación a tener en cuenta es que cuando se retiran estudios de buena calidad y precisos, la magnitud del resultado final se amplía, lo que permite identificar este tipo de estudios.

Todo lo anterior se refiere a los metaanálisis que incluyen exclusivamente ensayos clínicos aleatorizados, pero también existen metaanálisis de estudios observacionales. En este caso se debe tomar en cuenta que los pacientes sometidos al factor o tratamiento en investigación pueden diferir en otros aspectos que pueden influenciar el efecto resultante (confusión), de manera que llegan a ser más importantes que el sesgo de publicación. Aún si dentro del análisis se ajusta por las variables que diferencian a los grupos en estudio puede persistir una confusión residual. La gran diversidad de diseños en los estudios observacionales y en las poblaciones estudiadas hace particularmente problemática la interpretación de una medida resumen a partir de la combinación de sus resultados. Por esta razón no es de extrañar que exista un claro predominio de metaanálisis de ensayos clínicos en la literatura biomédica. No obstante, el número de metaanálisis de estudios observacionales publicados ha aumentado progresivamente, pues son una herramienta de gran ayuda en la toma de decisiones, en especial en el área de la salud pública y la política sanitaria. Por esta razón y dados los retos particulares que presentan este tipo de estudios a causa del riesgo inherente de sesgos y las diferencias en sus diseños, se hace necesario disponer de una herramienta que ayude a mejorar la calidad de los informes publicados de este tipo de metaanálisis.

En se creó el grupo The Metanalysis of Observational Studies in Epidemiology (MOOSE), formado por 27 expertos procedentes de diferentes áreas (clínicos, investigadores clínicos y sociales, estadísticos, epidemiólogos y editores), cuyo objetivo era confeccionar una lista de comprobación que ayudara a mejorar la calidad de la presentación de los metaanálisis de estudios observacionales. El grupo de trabajo celebró un seminario en 1997 que dio como resultado la propuesta MOOSE, que consiste en una lista de comprobación donde se establece un total de 35 recomendaciones estructuradas en seis grandes apartados:

- *Título:* debe identificar el estudio como metaanálisis (o revisión sistemática).
- *Resumen:* utilizar el formato estructurado de la revista.
- *Introducción:* debe incluir:
  - Definición del problema.
  - Hipótesis planteada.
  - Descripción de las medidas de resultado.
  - Tipo de exposición o intervención estudiada.
  - Tipo de diseños de estudio utilizados.
  - Población de estudio.
- *Estrategia de búsqueda:* debe describir:
  - Calificación de quién realizó la búsqueda (documentalistas e investigadores).
  - Estrategia de búsqueda, incluyendo el período de tiempo y palabras clave utilizadas.
  - Esfuerzos realizados para incluir todos los estudios disponibles, incluyendo contacto con los autores.

- Bases de datos y registros consultados.
- *Software* utilizado para la búsqueda, su nombre y versión, incluyendo funcionalidades especiales (por ejemplo, explosión).
- Búsqueda manual (por ejemplo, lista de referencias de los estudios recuperados).
- Listado de las citaciones localizadas y de las excluidas, con su justificación.
- Método empleado para los estudios publicados en otros idiomas distintos del inglés.
- Método empleado para los estudios no publicados y los resúmenes.
- Descripción de los contactos con autores.

– *Selección de estudios:* debe incluir:
- Descripción de la relevancia y justificación de los estudios seleccionados para demostrar la hipótesis del estudio.
- Justificación de la selección y sistema de codificación de los datos (por ejemplo, según criterios clínicos establecidos o por conveniencia).
- Documentación acerca de cómo los datos fueron clasificados y codificados (por ejemplo, múltiples evaluadores, cegamiento y fiabilidad interobservador).
- Valoración de los factores de confusión (por ejemplo, comparabilidad entre casos y controles).
- Valoración de la calidad de los estudios, incluyendo el cegamiento de los evaluadores, estratificación o regresión para posibles factores predictivos de los resultados.
- Valoración de la heterogeneidad.
- Descripción suficientemente detallada de los métodos estadísticos (por ejemplo, una descripción completa del modelo de efectos fijos o aleatorios, justificación acerca de si el modelo seleccionado tiene en cuenta los factores predictores de los resultados, modelos de dosis-respuesta, o metaanálisis acumulativo) para poder ser reproducidos.

– *Resultados:* debe presentar:
- Resumen gráfico de los resultados (estimadores) individuales y globales.
- Tabla descriptiva con información de todos los estudios incluidos.
- Resultados de los análisis de sensibilidad o de subgrupo.
- Medida de la incertidumbre estadística de los resultados.

– *Discusión:* debe incluir:
- Valoración cuantitativa de los sesgos (por ejemplo, sesgo de publicación).
- Justificación de las exclusiones (por ejemplo, exclusión de estudios en idioma distinto al inglés).
- Valoración de la calidad de los estudios incluidos.

– *Conclusiones:* debe incluir:
- Posibles explicaciones alternativas para los resultados observados.

- Generalización de las conclusiones (por ejemplo, a partir de los datos disponibles y en el contexto de lo que se sabe sobre el tema).
- Orientaciones para la investigación futura.
- Fuentes de financiamiento.

Los metaanálisis tienen sus propios problemas metodológicos que en gran medida ya se han descrito, así como la forma de enfrentarlos. Una lectura crítica supone tener presentes estas limitaciones, tratar de identificarlas en las publicaciones y asegurarse que los autores lidiaron adecuadamente con ellas:

– Heterogeneidad: es por esto que se critica a los metaanálisis satirizándoselos porque se comparan peras con manzanas.
– Sesgo de publicación: dada su naturaleza es muy difícil impedir el sesgo de publicación o, por lo menos, hacerse una idea de su magnitud.
– Sesgo impulsado por agenda: ocurre cuando los autores tienen una agenda económica, social o política. Aunque ocurre más en metaanálisis realizados en el ámbito de las ciencias sociales, la interacción cada vez más estrecha entre investigadores y la industria farmacéutica hace que el conflicto económico surja en el ámbito de la investigación médica.

Existen dos aspectos claves para la correcta interpretación de los resultados de un metaanálisis:

– Número inadecuado de estudios: no está determinado el número mínimo de estudios a partir del cual se puede realizar un metaanálisis, porque es muy dependiente de la pregunta formulada. Parece lógico que los que están realizados con un mayor número de estudios tengan más confiabilidad y obtengan resultados más consistentes, independientemente de la dispersión.
– Número de pacientes requeridos en los estudios: se ha sugerido el cálculo del tamaño óptimo de información u *optimal information size* (OIS). El tamaño óptimo de información corresponde al número de pacientes que se habrían requerido al calcular el tamaño muestral tradicional para un estudio único, con potencia suficiente. Otra manera de calcular el tamaño óptimo de información es consultar un nomograma para encontrar el número de pacientes o, más precisamente, el número de eventos.

Una vez se cuenta con los resultados y su análisis se debe leer la discusión que los autores realizan. También evaluar las limitaciones que plantearon al realizar el trabajo, así como las que no plantearon, pero que existieron y son identificadas al leer el trabajo.

Al leer las conclusiones realizadas por los autores se debe recordar los objetivos que tenía el estudio ya que estas deben ser la respuesta a lo planteado. Finalmente, y una vez que se tiene opinión en relación con el estudio, se debe determinar es viable de aplicar a la población con la que se trabaja habitualmente o al paciente en particular sobre el que se quiere intervenir.

Capítulo 2

Aspectos relacionados con el empleo de las muestras

Representatividad de una muestra

En las situaciones en las que se está interesado en describir determinadas características de una población o en extraer conclusiones, la mayor parte de las veces resulta inviable llevar a cabo el estudio de cada uno de sus individuos o elementos. Es por este motivo que se recurre al estudio de una parte de la población (*muestra*) a partir de la cual se extrapolan los resultados a nivel población, es decir, por inferencia estadística.

Una muestra para ser lo más representativa posible de su población de origen se requiere:

- Que las propiedades que caracterizan a la población o universo de estudio se distribuyan de forma análoga en su subconjunto, o sea, en la muestra seleccionada.
- Que todos los elementos de la población o del universo de estudio tengan la misma probabilidad de formar parte de la muestra o, si no la misma, que sea conocida esa probabilidad para todos los elementos de la población.

Las bases para declarar una muestra como *buena*, lo que denota su representatividad, es su capacidad de satisfacer el propósito para el cual fue diseñada, es decir, que reproduzca como un modelo a pequeña escala las características más importantes de la población. Para esto se debe utilizar como apoyo los factores más importantes que describen la *bondad de una muestra*:

- Tamaño muestral: se refiere a la utilización de la respectiva fórmula en su estimación para cada tipo de diseño de estudio y estimadores a emplear en el mismo.
- Condiciones de selección: se refiere al empleo de métodos o técnicas de selección de las unidades de análisis, es decir, de tipos de muestreo probabilísticos y no probabilísticos).
- Contenido de la investigación: se refiere a que la muestra sea portadora de toda la información necesaria para aquellas variables realmente relevantes en el estudio.

Es necesario recordar que:

- La bioestadística solo permite hacer inferencias a la población real de la cual fue extraída la muestra específica.
- El concepto universo es de existencia teórica y se refiere en general a todas las poblaciones a las cuales se pueden hacer extensivos los resultados muestrales, apoyado en los conocimientos de la ciencia particular de que se trate y se obvia la ubicación espacio-temporal del grupo de sujetos estudiados.
- La población es un conjunto de elementos (personas, animales, objetos, entre otros) o unidades de análisis que al menos comparten una característica que interesa conocer o estudiar o que son susceptibles de verse afectados por el fenómeno que se quiere investigar, circunscrito a una región geográfica bien determinada y en un periodo o momento del tiempo dado; mientras que la muestra no es más que una parte cualquiera de la población, un conjunto cualesquiera no vacío, o sea, que teóricamente un solo elemento de la población puede ser considerado como una muestra (aunque en la práctica esto ocurra con poca frecuencia). Los conceptos de muestra y población son relativos, lo que en un estudio puede ser una población en otro puede ser una muestra.
- Unidad de análisis: es cada uno de los elementos o sujetos de la población sobre los cuales se recoge la información y su naturaleza está determinada por los objetivos que se persiguen.
- Unidad de muestreo: es la que se utiliza para realizar la selección muestral, es aquella donde ir a seleccionar las unidades de análisis y en los casos más simples puede coincidir con la unidad de análisis, pero en general no es así.
- Unidad de observación: es aquella de donde se obtiene la información, lo cual no siempre se logra a partir de la propia unidad de análisis.
- Marco muestral: es la lista de todas las unidades de muestreo susceptibles de ser seleccionadas para integrar la muestra.
- Error de muestreo: es la diferencia en que se incurre, al menos en teoría, al inferir la situación de la población (para algún parámetro) por medio de lo observado en la muestra (a partir del estimador correspondiente). La magnitud de este error muestral está estrechamente vinculada de forma inversa con el tamaño muestral, es decir, para estimar un menor error muestral se requiere de un mayor tamaño muestral y viceversa.

Los aspectos relacionados con las propiedades de la distribución teórica de las medias en el muestreo son:

- La media de la distribución de las medias en el muestreo es igual a la media poblacional.
- La forma de distribución de las medias en el muestreo es aproximadamente el de la distribución normal, suponiendo que n sea suficiente grande (teorema central del límite).

– La desviación estándar de la distribución de las medias en el muestreo es: $\sigma / \sqrt{n}$ y se conoce como error estándar de la media.

Conceptos fundamentales del muestreo

En estadística se conoce como muestreo a la técnica para la selección de una muestra a partir de una población.

Aunque nunca se puede estar enteramente seguro de que el resultado sea una muestra representativa, sí se puede actuar de manera que esta condición se alcance con una probabilidad alta. Al conjunto de muestras que se pueden obtener de la población se denomina *espacio muestral*. La variable que asocia a cada muestra su probabilidad de extracción sigue la llamada *distribución muestral*.

Los objetivos generales de la técnica de muestreo son:

– Obtener una muestra que sea tan representativa de la población o el universo de estudio como sea posible.
– Plantear una forma de recogida de datos que se ajuste a los recursos socioeconómicos y temporales con que se cuente.

Las ventajas del uso del muestreo son:

– Reducción de los costos.
– Mayor rapidez en la obtención de resultados.
– Mayor alcance o esfera de acción.
– Mayor precisión en los datos.

La desventaja del uso del muestreo es la presencia de errores de apreciación (error del muestreo), al esperarse que las muestras sean un modelo reducido o a pequeña escala de la población, al inferir la situación de la población por medio de lo observado en la muestra.

Para diseñar la muestra se debe tomar en cuenta:

– Definir la población o universo de estudio.
– Obtener, de ser necesario y posible, el marco muestral pertinente, un listado de los elementos que componen el universo y que permitirán su identificación y es la base de los muestreos probabilísticos.
– Estimar la variabilidad poblacional de la característica en estudio o la frecuencia (probabilidad) de aparición o presentación del fenómeno que se estudia.
– Escoger la técnica de muestreo más adecuada: muestreo probabilístico o no.
– Tomar decisiones sobre el tamaño y el margen de error de la muestra: emplear una base estadística para la confiabilidad y la precisión deseada de los resultados.
– Seleccionar el tipo de diseño de estudio a realizar.
– Escoger la opción de técnica estadística a emplear en la inferencia para el problema de estudio en cuestión.

Entre los aspectos conceptuales imprescindibles al abordar la técnica del muestreo se debe tomar en cuenta que toda investigación epidemiológica exige la selección de una muestra de la población que se estudia y necesariamente siempre hay que tomar una decisión sobre:

- El tamaño que ha de tener la muestra.
- El procedimiento para seleccionar el número de unidades de análisis (elementos o sujetos que componen la población y que pueden ser, en su momento, objeto de interrogatorio o medición), esto se corresponde con el tipo de muestreo (abordado a continuación en tipos de muestreos).
- La realización de estimaciones de parámetros por su valor puntual y sus respectivos márgenes de incertidumbre (por intervalo de confianza) si el problema que se encara es "de estimación", es decir, si es de naturaleza básicamente descriptivo-inferencial.

Con respecto al elemento *tamaño de la muestra* se deben tomar en cuenta estas recomendaciones:

- La determinación *a priori* de los valores que intervienen en las fórmulas del tamaño muestral es una decisión esencialmente subjetiva. En concreto, por ejemplo, no es fácil establecer cuál es el error máximo aceptable para la estimación de un parámetro. Esto puede conducir a que dos investigadores independientes, aplicando la misma fórmula, obtengan tamaños de muestra muy diferentes por el hecho de que tienen visiones distintas (defendibles ambas) del problema.
- Para el cálculo de un tamaño de muestra se debe conocer, en la fase de diseño, el valor de un parámetro que solo se puede estimar una vez extraída la muestra, lo que constituye una contradicción habitual. Para resolver este problema hay dos opciones:
  - Estimar el parámetro a partir de una muestra piloto.
  - Obtener un valor aceptable a partir de las referencias que se tengan de trabajos o experiencias similares ya realizadas.
- En la práctica, las encuestas están generalmente diseñadas para estimar un conjunto grande de parámetros y no uno solo. Sin embargo, es necesario determinar un tamaño de muestra único. Esta dificultad suele solventarse adoptando uno de los parámetros como "el más importante", lo que implica otra decisión subjetiva en el cálculo.
- El error muestral en que se incurre al estimar un parámetro depende del tipo de diseño establecido para seleccionar los elementos que integran la muestra. Por lo general, a tamaños de muestra iguales, el error es mayor si se utiliza un diseño complejo, por ejemplo, muestreo estratificado bietápico, que bajo el muestreo simple aleatorio. Sin embargo, las fórmulas que permiten obtener el tamaño de muestra en función de la precisión siempre asumen que se va a realizar un mues-

treo simple aleatorio, lo que conduce a tamaños de muestra más pequeños que los que realmente son necesarios para garantizar el grado de precisión exigido si el diseño es otro. Una forma de "corregir" el tamaño de muestra en función del diseño muestral es multiplicar el tamaño obtenido con muestreo simple aleatorio por el denominado efecto de diseño. Naturalmente, no es fácil hallar una estimación adecuada de este valor, pero en la práctica suele asumirse algún valor entre 1,5 y 3 para el efecto de diseño. Un valor igual a 2, por ejemplo, significa que para obtener la misma precisión hay que estudiar el doble de individuos que con el muestreo simple aleatorio.

– El tamaño muestral está determinado por varios factores, entre los que se pueden mencionar:
  - La precisión de la estimación, que se mide a través del error muestral. La magnitud del máximo margen de error admisible debe ser fijada por el investigador, pues es él quien mejor identifica las consecuencias prácticas de un error dado. En general se cumple que, al aumentar la precisión, es decir, al disminuir el grado de error admisible, aumenta el tamaño de muestra necesario, algo coherente con lo que indica el sentido común. Cuando el error se expresa como un porcentaje del valor que puede tener el parámetro, por ejemplo, un 5 % o un 10 %, en ese caso se habla de precisión relativa.
  - El nivel de confianza, que generalmente se fija en 95 %. El efecto de este valor sobre el tamaño de muestra es directo, es decir, para una precisión dada, el tamaño de muestra aumenta al aumentar el nivel de confianza.

Tipos de métodos de muestreo a emplear

Existen dos métodos para seleccionar muestras de poblaciones: el muestreo no aleatorio o de juicio y el muestreo aleatorio, que incorpora el azar como recurso en el proceso de selección. Cuando este último cumple con la condición de que todos los elementos de la población tienen alguna oportunidad de ser escogidos en la muestra y si la probabilidad correspondiente a cada sujeto de la población es conocida de antemano, recibe el nombre de *muestreo probabilístico*.

La razón fundamental para el empleo del muestreo es de naturaleza económica, por lo general resulta literalmente imposible estudiar a todos los individuos que integran una población (hacer un censo). El imperativo de hallar recursos metodológicos que sean práctica y económicamente viables, pero a la vez, científicamente válidos, es de muy larga data. Sin embargo, no fue hasta la década de los años treinta del siglo xx, con un célebre trabajo presentado por el polaco Jerzy Neyman ante la Royal Statistical Society, cuando se comenzó a desarrollar una teoría orgánica sobre las técnicas que permiten superar de manera científica esa dificultad.

Un paso de medular relevancia en el progreso histórico de consolidación de la teoría del muestreo fue dado con la introducción del azar en el procedimiento de selección muestral. El azar puede intervenir de muchos modos en el proceso, pero se considera que se ha conferido un carácter estadísticamente riguroso al método de selección solo cuando se cumplen dos condiciones: que el azar opere de manera tal que el procedimiento otorgue a todos y cada uno de los elementos de la población cierta probabilidad conocida de antemano de integrar la muestra, y que tal probabilidad no sea nula para elemento alguno. En tal caso este muestreo es aleatorio y se dice que se ha seguido un método probabilístico de muestreo.

Los procedimientos probabilísticos satisfacen la exigencia intuitiva de eliminar o mitigar, al menos, la carga subjetiva que puede influir en la elección de los elementos que se van a examinar, interrogar o medir y, por ende, en las conclusiones. Por otra parte, solo cuando el método es probabilístico se puede medir el grado de precisión con que se realizan las estimaciones, que es una de las tareas fundamentales del muestreo.

Según el modo de obtener la muestra, si se permite la valoración del error de muestreo o no, se clasifican los métodos de muestreo en:

– Muestreo probabilístico: todos los elementos elegibles de la población tienen la misma probabilidad de formar parte de la muestra o, si no la misma, por lo menos que sea conocida esa probabilidad para todos los elementos de la población y permiten la valoración del error de muestreo. El calificativo probabilístico se refiere a que la selección de las unidades muestrales se realiza utilizando un esquema basado en las probabilidades (medida de las posibilidades) que tienen los sujetos de la población de formar parte de la muestra. Los tipos pueden ser:
  - Muestreo aleatorio simple: los elementos de la muestra son elegidos al azar de entre todos los de la población y requiere la identificación y listado de todos los elementos de la población y pueden ser con reposición (se reincorpora la unidad de observación extraída a la población antes de volver a seleccionar otra unidad) y sin reposición.
  - Muestreo sistemático (en fases): se realiza eligiendo cada k-ésimo caso de alguna lista o grupo, a partir de la previa división del tamaño poblacional entre el muestral y se establece la amplitud del intervalo de muestreo (k) de donde se extrae cada elemento.
  - Muestreo aleatorio estratificado: los elementos de la muestra son elegidos al azar por muestreo simple aleatorio a partir de estratos establecidos en la población y respetando la proporción de cada estrato en la población.
  - Muestreo aleatorio por conglomerados: se trata de una forma de muestreo simple aleatorio en que la unidad de muestreo no son los elementos de la población, sino agrupaciones de estos que de forma natural existan en aquélla (conglomerados). Antes de seleccionar la muestra, la población se divide en unidades de

primera etapa, cada una de las cuales contiene un conjunto de varias unidades de análisis. En cualquier caso, las unidades de muestreo de primera etapa deben cubrir por entero a la población sin intersecarse, o sea, todo miembro de la población o unidad de análisis pertenece a una y solo una unidad de primera etapa. Si se admiten en la muestra a todas las unidades de análisis que integran las unidades de primera etapa elegidas, se ha optado por un muestreo por conglomerados monoetápico. Si luego de seleccionadas las unidades de primera etapa que aportan elementos a la muestra, dentro de cada una se eligen algunas unidades de análisis, entonces se está ante un muestreo por conglomerados bietápico.

- Muestreo por conglomerados monoetápico estratificado: se trata de aquel en que se divide a la población en cierto número de partes o conglomerados, que se consideran como unidades de primera etapa e incluir en la muestra a todas las unidades de análisis pertenecientes a los conglomerados que se elijan. Sin embargo, en esta variante las unidades de primera etapa tienen que estar inicialmente conformadas en estratos.
- Muestreo por conglomerados bietápico estratificado: se utiliza cuando se quiere aplicar un muestreo bietápico, pero habiendo separado antes las unidades de primera etapa según los estratos.

– Muestreo no probabilístico: todos los elementos elegibles de la población no tienen la misma probabilidad de formar parte de la muestra o no es conocida esa probabilidad para todos los elementos de la población y no permiten la valoración del error de muestreo. Los tipos pueden ser:
  - Fortuita, incidental o de voluntarios: se realiza a partir de unidades de observación (sujetos) disponibles de una población que puedan portar o se relacionan con la información conveniente y necesaria respecto a un efecto determinado que desee estudiar el investigador, los cuales llegan a él de forma fortuita, casual o voluntaria.
  - Por selección de experto, por juicios, opinática o por conveniencia: se realiza a partir de unidades de observación (sujetos) de una población que están relacionados específicamente con cierto suceso o que portan información que influye, incide o está relacionada con cierto efecto según el juicio u opinión del encuestador o investigador. Es usada por expertos con la intención de seleccionar especímenes típicos o representativos de un fenómeno en particular
  - Por cuotas: se realiza dividiendo la población en estratos homogéneos o subpoblaciones, para asegurar que haya proporciones representativas de los diversos estratos de la muestra y se eligen los sujetos por conveniencia o juicios.
  - De bola de nieve o en cascada: se realiza a partir de unidades de observación (sujetos) de una población donde un contacto deriva a otro contacto y así suce-

sivamente, es decir, que el encuestado deriva al encuestador a otros posibles casos. Se utiliza frecuentemente cuando los casos resultan muy difíciles de contactar.

- De poblaciones móviles: es denominado también de captura-marcaje-recaptura de insectos, aves, peces, entre otros, es muy utilizado en algunas ramas como la zoología para estudiar el comportamiento migratorio de poblaciones y la estimación de su tamaño poblacional.

Otros tipos de muestreos no probabilísticos o generalmente bajo la óptica de conveniencia para los objetivos del investigador, que pueden encontrarse en diferentes textos aparecen con las denominaciones de muestreo de casos extremos o atípicos, muestreo de máxima variación, muestreo de unidades homogéneas, muestreo de casos típicos, muestreo de casos críticos**,** muestreo de casos para confirmar o refutar, muestreo de casos políticamente importantes, entre otros y no tienen de gran aplicación en las ciencias biomédicas, por lo que no son objetivo de este texto.

Apuntes finales para recordar sobre la teoría del muestreo:

– El muestreo constituye un recurso de la investigación científica cuya función básica consiste en determinar que parte de una realidad objeto de estudio debe ser analizada, con el objetivo de hacer inferencias al todo de la cual procede. A esa parte analizada se le llama *muestra* y al todo de la cual procede *población* o *universo*. Se debe distinguir entre población objeto, la que realmente se desea estudiar, y la población muestreada, la que realmente se estudió, no siempre ambas coinciden como es lo deseado, por lo que se debe ser cuidadoso a la hora de emitir conclusiones, estas deben recaer sobre la población muestreada.
– A partir de los resultados obtenidos al estudiar los elementos de una muestra es posible realizar estimaciones de los parámetros poblacionales, esta es la razón de ser de la muestra. Este procedimiento de estimación se realiza de manera diferente en dependencia del método de selección empleado e inevitablemente está sujeto a un margen de error, que es llamado *error de muestreo* y puede ser cuantificado siempre y cuando se emplee un método probabilístico de selección. A su vez la magnitud del error depende del tamaño muestral y de la variabilidad de los datos, que es menor en la medida que aumenta el tamaño de la muestra.
– Se dice que un método de selección de muestras es probabilístico cuando otorga una probabilidad conocida, no nula, a cada uno de los elementos de la población de ser incluidos en la muestra. A su vez, este es equiprobabilístico si la probabilidad de inclusión es la misma para todos los elementos de dicha población. Siempre que se emplean métodos equiprobabilísticos la media y la proporción muestrales son estimadores insesgados de los respectivos parámetros poblacionales.

- El muestreo simple aleatorio es un método que descansa exclusivamente en el azar. Además de ser equiprobabilístico se distingue por otorgar igual probabilidad de ser seleccionada a todas y cada una de las muestras de tamaño n susceptibles de ser formadas de esa población de N elementos. Es el método que se supone se emplee para realizar los procedimientos de inferencia estadística.
- El muestreo sistemático constituye una buena opción cuando existe algún tipo de ordenación de la población objeto, aunque no se disponga de una lista. Cuando el criterio de ordenación de los elementos es tal que los elementos más parecidos tienden a estar más cercanos, el muestreo sistemático suele ser más preciso que el aleatorio simple, ya que recorre la población de un modo más uniforme. Por otro lado, es a menudo más fácil no cometer errores con un muestreo sistemático que con el muestreo simple aleatorio.
- El muestreo aleatorio estratificado implica la división de la población en k subconjuntos o estratos mutuamente excluyentes y exhaustivos, dentro de los cuales se seleccionan las unidades de análisis que conforman la muestra mediante un procedimiento que involucra el azar. Este produce muestras equiprobabilísticas solamente cuando se realiza la asignación proporcional, es decir, el tamaño de muestra de cada estrato es proporcional al tamaño del estrato correspondiente respecto a la población total, de manera que:

  $$\frac{n_i}{N_i} = \frac{n}{N} \quad \text{para todo i}$$

  .
- El muestreo por conglomerados suele ser más económico que el muestreo simple aleatorio cuando se desea estudiar una población grande, pues permite controlar la dispersión geográfica de la muestra al seleccionar las unidades de análisis solamente de algunas partes o conglomerados de la población. La selección puede ser realizada en una o varias etapas, lo que le da nombre a los tipos específicos de este método, el más empleado es el muestreo por conglomerado monoetápico, este produce muestras equiprobabilísticas, pero el tamaño muestral es aleatorio, ya que depende del tamaño de los conglomerados seleccionados. No obstante, este sesgo en la práctica puede ser despreciado y emplear la media y proporción muestrales como estimadores de los parámetros correspondientes.
- Determinar el tamaño de muestra mínimo necesario para realizar una investigación constituye uno de los problemas más complejos de la teoría muestral. Es de vital importancia para el proceso inferencial, de él depende la precisión de las estimaciones, además de ser un factor clave para el contraste de hipótesis. Este procedimiento exige que el investigador piense en términos para los cuales generalmente no está preparado, por su parte, el uso de las fórmulas requiere de una serie de elementos

con un gran margen de subjetividad. Pero, lo más importante, ninguna fórmula sustituye el buen juicio y el sentido común de un investigador bien preparado, además, al final lo que cuenta son los recursos disponibles.

- En la práctica, dada la complejidad de muchos fenómenos a estudiar, se requiere la selección de muestras complejas, donde solo en contadas ocasiones se puede emplear los esquemas básicos de muestreo probabilísticos por sí mismo.
- No siempre una muestra probabilística es mejor que una no probabilística, además de que en muchas ocasiones la única forma de realizar el estudio es a través de muestras no probabilísticas.
- Es justo aclarar que muestra no probabilística no es sinónimo de muestra no representativa. La ventaja de la muestra probabilística (aleatoria) está en que permite evaluar a *posteriori* los errores o desviaciones que pueden ocurrir o presentarse al estimar las características poblacionales a través de la muestra, aunque el criterio del investigador siempre debe estar presente en la selección de la muestra o al determinar el alcance en tiempo y espacio de sus resultados.
- Tanto las muestras probabilísticas como las no probabilísticas pretenden ser representativas de la población objeto de estudio, pero su equivalencia es un error que se comete sistemáticamente.
- No se puede asegurar jamás que una muestra es representativa, pero se pueden comparar algunas características generales conocidas de la población con las de la muestra y en caso de similitud se puede afirmar que la muestra la representa adecuadamente.
- Por encima de todos los criterios técnicos para llevar a cabo un diseño de estudio con un tamaño muestral seleccionado y el empleo de un esquema muestral aleatorio más o menos complicado está su factibilidad, dadas las condiciones materiales existentes y el nivel de conocimiento científico del problema en cuestión.
- El investigador debe siempre estar en condiciones de fundamentar sus criterios de selección del tipo de muestreo a emplear, sin olvidar que el principio esencial es el volumen, la calidad y la confiabilidad de la información que se pueda conseguir de cada unidad muestral.
- La estrategia de combinar varias opciones metodológicas que incluyan el muestreo, el diseño y el análisis, se conoce como *triangulación*, tal vez debido al hecho alegórico de que el triángulo es la más sólida de todas las formas geométricas planas.

En la figura 2.1 se presenta a modo de resumen un esquema integrador de los aspectos necesarios a tomar en cuenta en la técnica del muestreo para la inferencia estadística.

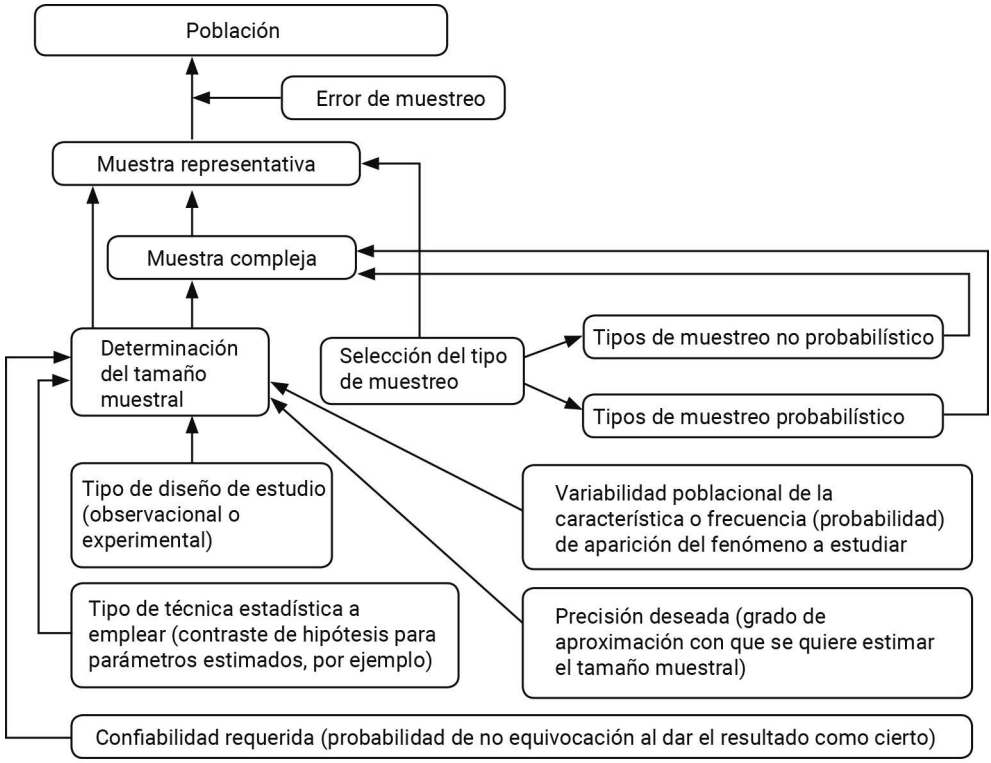


Fig. 2.1. Esquema para integrar los aspectos básicos necesarios a tomar en cuenta en la técnica de muestreo para la inferencia estadística.

Formulario empleado en el cálculo del tamaño de muestra para la estimación de parámetros por intervalos de confianza y para el contraste de hipótesis en estudios epidemiológicos

Estimación de parámetros por intervalos de confianza

En este caso se puede calcular el tamaño de muestra mínimo para estimar un parámetro θ mediante un intervalo de confianza, para el de nivel de confianza $(1 - \alpha)$ % con una cierta precisión (absoluta o relativa). También se calcula la precisión del intervalo de confianza para un tamaño dado n.

Nota aclaratoria: en todas las opciones que se presentan a continuación se representa a Z_{α} como el percentil de la distribución normal estándar N(0, 1), que deja a la izquierda una cola de probabilidad α y $(1 - \alpha) \cdot 100$, que expresada en por ciento es el nivel de confianza:

– Una media (Machin, 1997: 133-135):
  - El tamaño de muestra (n) se expresa como:

    $n = \left(z_{1-\alpha/2} \frac{\sigma}{e} \right)^2$, si la población es infinita;

    $n_F = \frac{Nn}{N+n}$, si la población es finita.

    (los tamaños resultantes se multiplican por el efecto de diseño.)
    donde:
    σ: desviación estándar esperada en la población;
    e: precisión absoluta de un intervalo de confianza para la media;
    N: tamaño de la población;
    Nn: es la multiplicación del tamaño de la población seleccionada N por el tamaño muestral n a seleccionar en la misma asumiéndola como infinita;
    deff: es el efecto de diseño (valores a seleccionar entre 1,5 y 3).
  - La precisión absoluta (e) se calcula como:

    $e = z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$, si la población es infinita;

    $e_F = e\sqrt{1-f}$, si la población es finita.

    (los valores resultantes se multiplican por el efecto de diseño).
    donde:
    n: tamaño de la muestra;

    $f = \frac{n}{N}$.

– Una proporción (Machin, 1997: 132-133):
  - Tamaño de muestra:

    $n = \left(\frac{Z_{1-\alpha/2}}{e} \right)^2 P(1-P)$, si la población es infinita,

    $n_F = \frac{Nn}{N+n}$, si la población es finita.

    (los tamaños resultantes se multiplican por el efecto de diseño).
    donde:
    P: proporción esperada en la población;
    e: precisión absoluta de un intervalo de confianza para la proporción;
    deff: efecto de diseño (seleccionar valores entre 1,5 y 3);
    N: tamaño de la población;

Nn: es la multiplicación del tamaño de la población seleccionada N por el tamaño muestral n a seleccionar en la misma asumiéndola como infinita.

- Precisión absoluta:

$e = Z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}}$, si la población es infinita;

$e_F = e\sqrt{1-f}$, si la población es finita.

(Los valores resultantes se multiplican por el efecto de diseño y por 100).

Donde:

n: tamaño de la muestra;

$f = \frac{n}{N}$.

– *Odds ratio* (Machin, 1997: 134):

- Tamaños de muestra:

$$n_1 = \frac{Z_{1-\alpha/2}\left[\frac{1}{P_1(1-P_1)} + \frac{1}{\varphi P_0(1-P_0)}\right]}{\left[\ln(1-\varepsilon)\right]^2} \text{ (casos)}$$

$n_0 = \phi n_1$ (controles)

donde:

P_1: proporción de casos expuestos;

P_0: proporción de controles expuestos;

ε: es la precisión relativa de un intervalo de confianza para la *odds ratio*;

ϕ: es el número de controles por caso.

P_1, P_0 y OR se relacionan de este modo:

$$P_1 = \frac{OR\,P_0}{(1-P_0) + ORP_0}$$

$$P_0 = \frac{P_1}{OR(1-P_1) + P_1}$$

$$OR = \frac{P_1(1-P_0)}{P_0(1-P_1)}$$

OR: *odds ratio*.

- Precisión relativa:

$$\varepsilon = \left[1 - \exp\left(Z_{1-\alpha/2}\sqrt{Var(\ln OR)}\right)\right]\cdot 100$$

donde:

$$Var(\ln OR) = \frac{1}{n_1}\left[\frac{1}{P_1(1-P_1)} + \frac{1}{\varphi P_0(1-P_0)}\right]\cdot 100$$

n_1: tamaño de la muestra de casos.

– Riesgo relativo (Lwanga, 1991: 52):
  - Tamaños de muestra:

$$n_1 = \frac{Z^2_{1-\alpha/2}\left[\frac{1-P_1}{P_1} + \frac{1-P_0}{\varphi P_0}\right]}{\left[\ln(1-\varepsilon)\right]^2} \text{ (expuestos)}$$

$n_0 = \phi n_1$ (no expuestos)
donde:
P_1: riesgo en expuestos;
P_0: riesgo en no expuestos;
RR: riesgo relativo;
P_1, P_0 y RR se relacionan de este modo:
$P_1 = P0RR$

$$P_0 = \frac{P_1}{RR}$$

$$RR = \frac{P_1}{P_0}$$

ε: precisión relativa de un intervalo de confianza para el riesgo relativo;
ϕ: razón entre los tamaños de no expuestos y expuestos.

  - Precisión relativa:

$$\varepsilon = \left[1 - \exp\left(Z_{1-\alpha/2}\sqrt{\text{Var}(\ln RR)}\right)\right] \cdot 100$$

donde:

$$\text{Var}(\ln RR) = \frac{1}{n_1}\left[\frac{1-P_1}{P_1} + \frac{1-P_0}{\varphi P_0}\right]$$

$n_1 = \frac{n}{1+\varphi}$: tamaño de la muestra de expuestos;

n: tamaño total de muestra.

– Concordancia (Cantor, 1996):
  - Tamaño de muestra:

$$n = \left(\frac{Z_{1-\alpha/2}}{e}\right)^2 Q$$

Donde:
e: precisión absoluta de un intervalo de confianza para el coeficiente kappa;

$$Q = \frac{A + B - C}{(1-P_e)^2}$$

$A = P_{11}[1 - (P_{1.} + P_1)(1 - k)]^2 + P_{22}[1 - (P_{2.} + P_2)(1 - k)]^2$
$B = (1 - k)^2[P_{12}(P_1 + P_2)^2 + P_{21}(P_2 + P_1)^2]$

$C = [k - P_e(1 - k)]^2$

k: concordancia esperada (coeficiente kappa);

$P_{1.} = P_1$: proporción de clasificaciones positivas por el primer observador;

$P_1 = P_2$: proporción de clasificaciones positivas por el segundo observador;

$P_{2.} = 1 - P_1$ y $P_2 = 1 - P_1$

$P_e = P_1.P_1 + P_2.P_{.2}$

$P_0 = k(1 - P_e) + P_e$

$$P_{22} = \frac{P_0 - P_{1.} + P_{.2}}{2}$$

$P_{11} = P_0 - P_{22}$, $P_{22} = P_{1.} - P_{11}$ y $P_{21} = P_{.1} - P_{11}$

- Precisión absoluta:

$$e = Z_{1-\alpha/2}\sqrt{\frac{Q}{n}}$$

donde:

n: tamaño de muestra.

– Sensibilidad y especificidad (Obuchowski, 1998):

Sensibilidad:

- Tamaños de muestra:

n_E y $n_{NE} = \phi n_E$, si se conoce *a priori* la condición del enfermo;

$n = \frac{n_E}{P}$, si no se conoce *a priori* la condición del enfermo;

$$n_E = \left(\frac{Z_{1-\alpha/2}}{e}\right)^2 \theta_s(1-\theta_s)$$ (fórmula para una proporción, población infinita);

donde:

θ_s: sensibilidad esperada;

n_E: número de enfermos;

n_{NE}: número de no enfermos;

n: tamaño total de muestra;

ϕ: razón entre los tamaños de no enfermos y enfermos;

P: prevalencia de enfermedad en la población;

e: precisión absoluta de un intervalo de confianza para θ_s.

- Precisión absoluta:

$$e = Z_{1-\alpha/2}\sqrt{\frac{\theta_s(1-\theta_s)}{n_E}}$$ (fórmula para una proporción, población infinita)

donde:

$n_E = \frac{n}{1+\varphi}$, si se conoce *a priori* la condición de enfermo;

$n_E = n_P$, si no se conoce *a priori* la condición de enfermo;

n: tamaño total de muestra.

Especificidad:

- Tamaños de muestra:

  n_{NE} y $n_E = \frac{n_{NE}}{\varphi}$, si se conoce *a priori* la condición del enfermo;

  $n = \frac{n_{NE}}{1-P}$, si no se conoce *a priori* la condición del enfermo;

  $n_{NE} = \left(\frac{Z_{1-\alpha/2}}{e}\right)^2 \theta_E(1-\theta_E)$ (fórmula para una proporción, población infinita)

  donde:

  θ_E: especificidad esperada;

  e: precisión absoluta de un intervalo de confianza para θ_E;

- Precisión absoluta:

  $e = Z_{1-\alpha/2}\sqrt{\frac{\theta_E(1-\theta_E)}{n_{NE}}}$ (fórmula para una proporción, población infinita)

  donde:

  $n_{NE} = \frac{n\varphi}{1+\varphi}$, si se conoce *a priori* la condición del enfermo;

  $n_{NE} = n(1 - P)$, si no se conoce *a priori* la condición del enfermo;

  n: tamaño total de muestra.

Sensibilidad y especificidad:

- Tamaños de muestra:

  n_E y n_{NE}, si se conoce *a priori* la condición del enfermo;

  $n = \text{máx}\left\{\frac{n_E}{P}, \frac{n_{NE}}{1-P}\right\}$, si no se conoce *a priori* la condición del enfermo;

  n_E: se calcula con la fórmula para la sensibilidad;

  n_{NE}: se calcula con la fórmula para la especificidad.

  La precisión se calcula por separado para sensibilidad y especificidad.

Estimación de parámetros del contraste de hipótesis

En este caso se puede calcular el tamaño de muestra mínimo para aplicar una prueba de hipótesis bilateral (salvo que se indique unilateral) con nivel de confianza $(1 - \alpha)$% y potencia $(1 - \beta)$%. También se calcula la potencia de la prueba para un tamaño de muestra dado n.

Nota aclaratoria: En todas las opciones que se presentan a continuación se representará a:

Z_α: percentil de la distribución normal estándar N(0, 1), que deja a la izquierda una cola de probabilidad α;

$1 - \alpha$%: nivel de confianza;

$1 - \beta$%: potencia;

ϕ: función de distribución normal estándar.

– Comparación de medias, grupos independientes (Schouten, 1999):

  • Tamaños de muestra:

$$n_1 = \left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{d} \right)^2 \frac{(\tau + \phi)\sigma_1^2}{\phi} + \frac{(\tau + \phi^3) Z^2_{1-\alpha/2}}{2\phi(\tau + \phi)^2}$$, si las varianzas son distintas;

$$n_1 = \left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{\Delta} \right)^2 \frac{1 + \phi}{\phi} + \frac{Z^2_{1-\alpha/2}}{2(\tau + \phi)^2}$$, si las varianzas son iguales;

$n_2 = \phi n_1$

donde:

σ_i^2: varianza esperada en la población i, i = 1, 2;

$$\tau = \frac{\sigma_2^2}{\sigma_1^2};$$

d: diferencia esperada entre las dos medias;

ϕ: razón entre los dos tamaños muestrales;

$\Delta = \frac{d}{\sigma}$: diferencia estandarizada de medias;

σ: varianza común esperada en las dos poblaciones.

  • Potencia:

$1 - \beta = \Phi (Z_1 - \beta)$

$$Z_{1-\beta} = |d| \sqrt{\left(n_1 - \frac{(\tau^2 + \phi^3) Z^2_{1-\alpha/2}}{2\phi(\tau + \phi)^2} \right)^2 \frac{\phi}{(\tau + \phi)\sigma_1^2}} - Z_{1-\alpha/2}$$, si las varianzas son distintas;

$$Z_{1-\beta} = |\Delta| \sqrt{\left(n_1 - \frac{Z^2_{1-\alpha/2}}{2(1+\phi)} \right) \frac{\phi}{1+\phi}} - Z_{1-\alpha/2}$$, si las varianzas son iguales;

donde:

$$n_1 = \frac{n}{1+\phi};$$

n: tamaño total de la muestra.

– Comparación de medias, grupos emparejados (Machin, 1997: 73):

  • Número de pares:

$$n = \left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{\Delta} \right)^2 + \frac{Z^2_{1-\alpha/2}}{2}$$

donde:

$\Delta = \frac{d}{\sigma_d}$: diferencia estandarizada de medias;

d: diferencia esperada entre las dos medias;

$\sigma_d = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$: desviación estándar de las diferencias;

σ_i: desviación estándar esperada en la población i, i = 1, 2;

ρ: coeficiente de correlación entre ambas poblaciones.

- Potencia:

$1 - \beta = \Phi (Z_1 - \beta)$

$$Z_{1-\beta} = \left[|\Delta| \sqrt{\left(n - \frac{Z^2_{1-\alpha/2}}{2} \right)} \right] - Z_{1-\alpha/2}$$

donde:

n: número de pares de la muestra.

– Comparación de proporciones, grupos independientes (Machín, 1997: 19-20):

- Tamaños de muestra para aplicar el test chi-cuadrado (χ^2):

$$n_1 = \frac{\left(Z_{1-\alpha/2} \sqrt{(1+\phi)\bar{P}(1-\bar{P})} + Z_{1-\beta} \sqrt{\phi P_1(1-P_1) + P_2(1-P_2)} \right)^2}{\phi (P_1 - P_2)^2}$$

$n_2 = \phi n_1$

- Tamaños de muestra para aplicar el test chi-cuadrado (χ^2) con la corrección por continuidad de Yates (χ_c^2) o la prueba exacta de Fisher:

$$m_1 = \frac{n_1}{4} \left[1 + \sqrt{1 + \frac{2(1+\phi)}{\phi n_1 |P_1 - P_2|}} \right]^2$$

$m_2 = \phi m_1$

donde:

Pi: proporción esperada en la población i, i = 1, 2;

ϕ: razón entre los dos tamaños muestrales;

$$\bar{P} = \frac{P_1 + \phi P_2}{1 + \phi}$$

- Potencia del test chi-cuadrado (χ^2):

$1 - \beta = \Phi (Z_1 - \beta)$

$$z_{1-\beta} = \frac{|P_1 - P_2| \sqrt{\phi n_1} - z_{1-\alpha/2} \sqrt{(1+\phi)\bar{P}(1-\bar{P})}}{\sqrt{\phi P n_1 (1-P_1) + P_2(1-P_2)}}$$

- Potencia del test chi-cuadrado (χ^2) con la corrección por continuidad de Yates (χ_c^2) o del test exacto de Fisher:

$1 - \beta = \Phi (Z_1 - \beta)$

$$z_{1-\beta} = \frac{|P_1 - P_2|\sqrt{\phi\left(m_1 - \lambda + \frac{\lambda_2}{4m_1}\right)} - z_{1-\alpha/2}\sqrt{(1+\phi)\overline{P}(1-\overline{P})}}{\sqrt{\phi P_1(1-P_1) + P_2(1-P_2)}}$$

donde:

$$\lambda = \frac{1+\phi}{\phi|P_1 - P_2|};$$

$$n_1 = \frac{n}{1+\phi};$$

$$m_1 = \frac{m}{1+\phi}, \text{ donde n o m es el tamaño total de la muestra.}$$

– Comparación de proporciones, grupos emparejados (Connor, 1987):
  - Número de pares para aplicar el test de McNemar:

$$n = \frac{\left(z_{1-\alpha/2}\sqrt{P_d} + z_{1-\beta}\sqrt{Pd - (P_1 - P_2)^2}\right)^2}{(P_1 - P_2)^2}$$

donde:

P_i: proporción esperada en el grupo i, i = 1, 2...;

$P_d = P_1(1 - P_2) + P_2(1 - P_1)$, proporción de pares discordantes.

  - Potencia:

$1 - \beta = \Phi (Z_1 - \beta)$

$$z_{1-\beta} = \frac{|P_1 - P_2|\sqrt{n} - z_{1-\alpha/2}\sqrt{P_d}}{\sqrt{P_d - (P_1 - P_2)^2}}$$

donde:

n: número de pares en la muestra.

– Estudios de casos y controles, grupos independientes (Machin, 1997: 19-20):
  - Tamaños de muestra para aplicar el test chi-cuadrado (χ^2):

$$n_1 = \frac{\left(Z_{1-\alpha/2}\sqrt{(1+\phi)\overline{P}(1-\overline{P})} + Z_{1-\beta}\sqrt{\phi P_1(1-P_1) + P_2(1-P_2)}\right)^2}{\phi(P_1 - P_2)^2}$$

$n_2 = \phi n_1$

  - Tamaños de muestra para aplicar el test χ^2 con la corrección por continuidad de Yates (χ_c^2) o el test exacto de Fisher:

$$m_1 = \frac{n_1}{4}\left[1 + \sqrt{1 + \frac{2(1+\phi)}{\phi n_1|P_1 - P_2|}}\right]^2$$

$m_2 = \phi m_1$

donde:

n_1 o m_1: número de casos en la muestra;

n_2 o m_2: número de controles en la muestra;

ϕ: número de controles por caso;

P_1: proporción de casos expuestos;

P_2: proporción de controles expuestos;

P_1 y P_2: se relacionan con OR de este modo:

$$P_1 = \frac{ORP_2 + \phi P_2}{(1-P_2) + ORP_2}$$

$$P_2 = \frac{P_1}{OR(1-P_1) + P_1}$$

– Estudios de casos y controles, grupos emparejados (Connor, 1987):

  • Número de pares para aplicar el test de McNemar:

$$n = \frac{\left(Z_{1-\alpha/2}\sqrt{P_d} + Z_{1-\beta}\sqrt{Pd-(P_1-P_2)^2}\right)^2}{(P_1-P_2)^2}$$

donde:

P_1: proporción de casos expuestos;

P_2: proporción de controles expuestos;

P_1 y P_2: se relacionan con OR de este modo:

$$P_1 = \frac{ORP_2}{(1-P_2) + ORP_2}$$

$$P_2 = \frac{P_1}{OR(1-P_1) + P_1}$$

– Estudios de cohorte (Machin, 1997: 19-20):

  • Tamaños de muestra para aplicar el test chi-cuadrado (χ^2):

$$n_1 = \frac{\left(Z_{1-\alpha/2}\sqrt{(1+\phi)\bar{P}(1-\bar{P})} + Z_{1-\beta}\sqrt{\phi P_1(1-P_1) + P_2(1-P_2)}\right)^2}{\phi(P_1-P_2)^2}$$

$n_2 = \phi n1$

  • Tamaños de muestra para aplicar el test chi-cuadrado (χ^2) con la corrección por continuidad de Yates (X_c^2) o el test exacto de Fisher:

$$m_1 = \frac{n_1}{4}\left[1 + \sqrt{1 + \frac{2(1+\phi)}{\phi n_1 |P_1 - P_2|}}\right]^2$$

$m_2 = \phi m_1$

donde:

n_1 o m_1: tamaño de la muestra de expuestos;

n_2 o m_2: tamaño de la muestra de no expuestos;

P_1: riesgo en expuestos;

P_2: riesgo en no expuestos;

P_1 y P_2: se relacionan con RR de este modo:

$P_1 = P_2 RR$

$$P_2 = \frac{P_1}{RR}$$

ϕ: razón entre el tamaño muestral de no expuestos y el de expuestos.

– Equivalencia de medias (Machin, 1997: 102):

• Tamaños de muestra para un contraste unilateral con varianzas iguales:

$$n_1 = \frac{\sigma^2(1+\phi)\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{\phi(|d| - \varepsilon)^2}$$

$n_2 = \phi n_1$

donde:

d: diferencia esperada entre las dos medias;

σ: varianza común esperada en las dos poblaciones;

ε: diferencia máxima que indica equivalencia entre las medias;

ϕ: razón entre los dos tamaños muestrales.

• Potencia:

$1 - \beta = \Phi(z_1 - \beta)$

donde:

$$z_{1-\beta} = \frac{|d - \varepsilon|\sqrt{n}}{\sigma}\sqrt{\frac{\phi n_1}{1+\phi}} - z_{1-\alpha}$$

$$n_1 = \frac{\phi n}{1+\phi}$$

n: tamaño total de la muestra.

– Equivalencia de proporciones (Machin, 1997: 101-102):

• Tamaños de muestra para un contraste unilateral:

$$n_1 = \frac{\left(Z_{1-\alpha}\sqrt{\phi\bar{P}_1(1-\bar{P}_1) + \phi\bar{P}_2(1-\bar{P}_2)} + Z_{1-\beta}\sqrt{\phi P_1(1-P_1) + P_2(1-P_2)}\right)^2}{\phi(|P_1 - P_2| - \varepsilon)^2}$$

$n_2 = \phi n_1$

donde:

P_i: proporción esperada en la población i, i = 1, 2;

ε: diferencia máxima que indica equivalencia entre las proporciones;

ϕ: razón entre los dos tamaños muestrales;

$$\bar{P}_1 = 2u\cos(w) - \frac{b}{3a}$$

$$\bar{P}_2 = \bar{P}_1 - \varepsilon$$

$$w = \frac{1}{3}\left[\pi + \arccos\left(\frac{v}{u^3}\right)\right]$$

$$v = \left(\frac{b}{3a}\right)^3 - \frac{bc}{6a^2} + \frac{d}{2a}$$

$$u = \text{sig}(v)\sqrt{\left(\frac{b}{3a}\right)^2 - \frac{c}{3a}}$$

$$a = \frac{1+\phi}{\phi}$$

$$b = -\frac{1}{\phi}\left[1+\phi+\phi P_1 + P_2 + \varepsilon(1+2\phi)\right]$$

$$c = -\frac{1}{\phi}\left[\phi\varepsilon^2 + \varepsilon(2\phi P_1 + 1 + \phi) + \phi P_1 + P_2\right]$$

$$d = -P_1\varepsilon(1+\varepsilon)$$

- Potencia:

  $1 - \beta = \Phi(Z_1 - \beta)$

  donde:

$$z_{1-\beta} = \frac{|P_1 - P_2| - \varepsilon\sqrt{n}_1 - z_{1-\alpha}\sqrt{\phi\bar{P}_1(1-\bar{P}_1)(1-P_2)}}{\sqrt{\phi P_1(1-P_1) + P_2(1-P_2)}}$$

$$n_1 = \frac{n}{1+\phi}$$

  n: tamaño total de la muestra.

– Pruebas diagnósticas, grupos independientes (Machin, 1997: 19-20):

  En este caso se calcula el tamaño de muestra necesario para comparar la sensibilidad o la especificidad de dos pruebas diagnósticas en grupos independientes de igual tamaño mediante el test chi-cuadrado ($\chi^{2)}$.

  Sensibilidad:

  - Número de enfermos para aplicar el test chi-cuadrado (χ^2):

$$n_E = \frac{\left(z_{1-\alpha/2}\sqrt{2\bar{\theta}(1-\bar{\theta})} + z_{1-\beta}\sqrt{\theta_1(1-\theta_1) + \theta_2(1-\theta_2)}\right)^2}{(\theta_1 - \theta_2)^2}$$

- Número de enfermos para aplicar el test chi-cuadrado (χ^2) con la corrección por continuidad de Yates o el test exacto de Fisher:

$$m_E = \frac{n_E}{4}\left[1+\sqrt{1+\frac{4}{n_E|\theta_1-\theta_2|}}\right]^2$$

Si se conoce *a priori* la condición de enfermo se calculan los tamaños de enfermos y no enfermos:

n_E y $n_{NE} = \phi n_E$

m_E y $m_{NE} = \phi mE$

Si no se conoce *a priori* la condición de enfermo se calcula el tamaño total de muestra:

$$n = \frac{n_E}{P} \quad o \quad m = \frac{m_E}{P}$$

donde:

θ_i: sensibilidad esperada para la prueba i, i = 1, 2;

$$\bar{\theta} = \frac{\theta_1 - \theta_2}{2};$$

ϕ: razón entre el tamaño muestral de no enfermos y el de enfermos;

P: es la prevalencia de enfermedad en la población.

- Potencia:

$1 - \beta = \Phi(z_1 - \beta)$

Para el test chi-cuadrado (χ^2) sin corrección:

$$z_{1-\beta} = \frac{|\theta_1-\theta_2|\sqrt{n_E}}{\sigma}\sqrt{\frac{\phi n_1}{1+\phi}} - z_{1-\alpha}$$

Para el test chi-cuadrado (χ^2) con la corrección por continuidad de Yates (X_c^2) o el test exacto de Fisher:

$$z_{1-\beta} = \frac{|\theta_1-\theta_2|\sqrt{m_E - \lambda + \frac{\lambda^2}{4m_E}} - z_{1-\alpha/2}\sqrt{2\bar{\theta}(1-\bar{\theta})}}{\sqrt{\theta_1(1-\theta_1)+\theta_2(1-\theta_2)}}$$

donde:

$n_E = \frac{n}{1+\phi}$ o $m_E = \frac{m}{1+\phi}$, si se conoce a priori la condición del enfermo;

$n_E = nP$ o $m_E = mP$, si no se conoce *a priori* la condición del enfermo;

n: tamaño total de la muestra de cada grupo (los dos iguales);

$$n_E = \frac{2}{|\theta_1-\theta_2|}$$

Especificidad:

Las fórmulas para calcular el tamaño de muestra son las mismas que para la sensibilidad, con la diferencia de que se obtienen los no enfermos n_{NE} y m_{NE}, respectivamente, para el test chi-cuadrado (χ^2) con o sin corrección por continuidad. También coinciden las fórmulas para la potencia, sustituyendo n_E y m_E por n_{NE} y m_{NE}, respectivamente.

- Si se conoce *a priori* la condición de enfermo se calculan los tamaños de no enfermos y enfermos:

$$n_{NE\ y}\ n_E = \frac{n_{NE}}{\phi}$$

$$m_{NE\ y}\ m_E = \frac{m_{NE}}{\phi}$$

- Si no se conoce *a priori* la condición de enfermo se calcula el tamaño total de muestra:

$$n = \frac{n_{NE}}{1-P} \quad \text{o} \quad m = \frac{m_{NE}}{1-P}$$

- Potencia: se calcula primero el tamaño de muestra de no enfermos a partir del tamaño total de muestra:

$$n_{NE} = \frac{\phi n}{1+\phi} \quad \text{o} \quad m_{NE} = \frac{\phi m}{1+\phi}$$, si se conoce *a priori* la condición del enfermo;

$n_{NE} = n(1 - P)$ o $m_{NE} = m(1 - P)$, si no se conoce *a priori* la condición del enfermo.

Sensibilidad y especificidad:

- Si se conoce *a priori* la condición del enfermo se calcula el tamaño de enfermos para sensibilidad y el de no enfermos para especificidad.
- Si no se conoce *a priori* la condición del enfermo se calcula el tamaño total de muestra:

$$n = \text{máx}\left\{\frac{n_E}{P}, \frac{n_{NE}}{1-P}\right\}$$

- Potencia: se calcula por separado para sensibilidad y especificidad.

– Pruebas diagnósticas, grupos emparejados (Connor, 1987):

Sensibilidad:

Se aplican las fórmulas empleadas anteriormente en la comparación de proporciones, grupos emparejados por el test de McNemar para obtener el número de enfermos, n_E, donde:

P_1: sensibilidad esperada para la prueba 1;

P_2: sensibilidad esperada para la prueba 2.

- Si se conoce *a priori* la condición del enfermo se calculan los tamaños de enfermos y no enfermos:

  n_E y $n_{NE} = \phi n_E$
- Si no se conoce *a priori* la condición del enfermo se calcula el tamaño total de muestra:

$$n = \frac{n_{NE}}{P}$$

- Potencia: se calcula primero el tamaño de muestra de enfermos a partir del tamaño total de muestra:

$$n_E = \frac{n}{1+\phi}$$, si se conoce *a priori* la condición del enfermo;

$n_E = nP$, si no se conoce *a priori* la condición del enfermo;

donde:
ϕ: razón entre el tamaño muestral de no enfermos y el de enfermos;
P: prevalencia de enfermedad en la población.

Especificidad:
Se aplican las fórmulas empleadas anteriormente en la comparación de proporciones, grupos emparejados por el test de McNemar para obtener el número de no enfermos, n_{NE}, donde:
P_1: es la especificidad esperada para la prueba 1;
P_2: es la especificidad esperada para la prueba 2.

- Si se conoce *a priori* la condición del enfermo se calculan los tamaños de no enfermos y enfermos:

$$n_{NE} \text{ y } n_E = \frac{n_{NE}}{\phi}$$

- Si no se conoce *a priori* la condición del enfermo se calcula el tamaño total de muestra:

$$n = \frac{n_{NE}}{1-P}$$

- Potencia: se calcula primero el tamaño de muestra de no enfermos a partir del tamaño total de muestra:

$$n_{NE} = \frac{\phi n}{1+\phi}$$, si se conoce *a priori* la condición del enfermo;

$n_{NE} = n(1 - P)$, si no se conoce *a priori* la condición del enfermo;
donde:
ϕ: razón entre el tamaño muestral de no enfermos y el de enfermos;
P: prevalencia de enfermedad en la población.

Sensibilidad y especificidad:

- Si se conoce *a priori* la condición de enfermo se calcula el tamaño de enfermos para sensibilidad y el de no enfermos para especificidad; en caso contrario, si no se conoce *a priori* la condición de enfermo, se calcula el tamaño total de muestra como sigue:

$$n = \text{máx}\left\{\frac{n_E}{P}, \frac{n_{NE}}{1-P}\right\}$$

- Potencia: se calcula por separado para sensibilidad y especificidad.

– Comparación de dos curvas de supervivencia (Schoenfeld, 1983): en este caso se calcula el tamaño de muestra necesario para comparar dos curvas de supervivencia mediante el test *log-rank*, asumiendo un modelo de riesgos proporcionales para los tiempos de supervivencia.

- Número de eventos necesarios:

$$E = \left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \frac{1}{\phi}\left(\frac{1+\phi}{\ln\Delta}\right)^2$$

- Tamaño total de muestra (corregido por las pérdidas de seguimiento):

$$n_c = \frac{n}{1-P_c} \quad \text{con } n = \frac{E}{P_e}$$

- Tamaño de muestra de cada grupo:

$$n_1 = \frac{n_c}{1+\phi} \quad \text{y } n_2 = \frac{n_c}{1+\phi}$$

donde:

ϕ: razón entre los dos tamaños muestrales (n_2/n_1);

$\Delta = \frac{\ln P_2}{\ln P_1}$: hazard ratio;

P_i: es la probabilidad de supervivencia en el grupo i, i = 1, 2;

Pc: es la proporción de pérdidas de seguimiento;

$Pe = \frac{(1-P_1)+\phi(1-P_2)}{1+\phi}$: probabilidad de ocurrencia de un evento durante el periodo de estudio.

– Comparación de tres o más curvas de supervivencia (Ahnn,1995):

Esta opción calcula el tamaño de muestra necesario para comparar k curvas de supervivencia mediante el test log-rank, asumiendo un modelo de riesgos proporcionales para los tiempos de supervivencia y que todos los grupos son de igual tamaño.

- Número de eventos necesarios:

$$E = \frac{\delta k^2}{(k-1)\sum_{j=2}^{k}(\ln\Delta_j)^2 - 2\sum_{j=2q>j}^{k}\sum \ln\Delta_j \ln\Delta_q}$$

- Tamaño total de muestra (corregido por las pérdidas de seguimiento):

$$n_c = \frac{n}{1-P_c} \quad \text{con } n = \frac{E}{P_e}$$

- Tamaño de muestra de cada grupo:

$$n_j = \frac{n_c}{k}, \quad \Delta j = 1,\ldots, k$$

donde:

k: número de grupos que se comparan;

$$\Delta_j = \frac{\ln P_j}{\ln P_1}, \quad j = 2,\ldots, k;$$

P_j: probabilidad de supervivencia en el grupo j, j = 1, 2,..., k;

δ: parámetro de no centralización de la distribución del estadístico bajo la hipótesis alternativa: chi-cuadrado no central con k – 1 grados de libertad. Depende del nivel de confianza y de la potencia del test (Narula y Desu, 1981);

Pc: proporción de pérdidas de seguimiento;

$P_e = \frac{1}{k}\sum_{j=1}^{k}(1-P_j)$: probabilidad de ocurrencia de un evento durante el periodo de estudio.

– Coeficiente de correlación (Machin,1997: 168-9):

- El tamaño de muestra se calcula siguiendo un procedimiento iterativo:

Paso 1:

$$U_1 = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right); \; n_1 = \frac{(z_{1-\alpha} + z_{1-\beta})^2}{U_1^2} + 3$$

Paso i (i > 1):

$$U_i = U_1 + \frac{\rho}{2(n_{i-1}-1)}; \; n_i = \frac{(z_{1-\alpha} + z_{1-\beta})^2}{U_i^2} + 3$$

Si $|n_i - n_{i-1}| < 1$ se para el proceso y se toma $n = n_i$.

- Potencia:

$1 - \beta = \Phi (Z_1 - \beta)$

$$z_{1-\beta} = \left(U\sqrt{n-3}\right) - z_{1-\alpha}$$

donde:

ρ: coeficiente de correlación a detectar;

$$U = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) + \frac{\rho}{2(n-1)}$$

Procedimiento para la selección de la muestra empleando algunos tipos de muestreo

La selección de la muestra tiene sus particularidades en el procedimiento según el tipo de muestreo que se emplee. A continuación se presentan las recomendaciones para algunos tipos de muestreos.

Muestreo simple aleatorio

- Algoritmo de selección sin reemplazamiento:
  - Se asigna un número de 1 a N a cada elemento de la población.
  - Paso 1: se selecciona un número aleatorio entre 1 y N.
  - Paso k + 1: se selecciona un número aleatorio entre los N − k números no seleccionados en los pasos anteriores, k = 1,..., n − 1.
- Probabilidad de selección (%):

$$f = \frac{n}{N} \cdot 100$$

donde:
N: tamaño de la población;
n: es el tamaño de la muestra.

Muestreo sistemático (Silva, 2000: 126-131)

- Muestreo sistemático regular:
  - Se asigna un número de 1 a N a cada elemento de la población.
  - Se calcula el intervalo de selección, $k = \frac{N}{n}$ parte entera.
  - Se selecciona el arranque aleatorio r entre 1 y k.
  - Los números seleccionados son los de la secuencia $\{r + j_{k'j} = 0, 1,..., s - 1\}$, donde s es el menor entero tal que $r + s_k > N$.
- Probabilidad de selección (%):

$$f = \frac{n}{N} \cdot 100$$

donde:
n: tamaño de muestra resultante.
- Muestreo sistemático en fases:
  - Se asigna un número de 1 a N a cada elemento de la población.

- Se calcula h:

$$h = \begin{cases} \dfrac{f}{100} \text{ si } f < 51 \\ 1 - \dfrac{f}{100} \text{ si } f > 50 \end{cases}$$

- En la tabla 2.1 se buscan los valores de k_1 y k_2 que corresponden a h.
- Se selecciona un número aleatorio entre 1 y k_1 (r_1) y, si $k_2 \neq 0$, un número aleatorio entre 1 y $|k_2|$ (r_2).
- Se calcula la secuencia de números: $S_1 = \{r_1 + jk_1, j = 0, 1, \ldots, s - 1\}$, donde s es el menor entero tal que $r_1 + s_{k1} > N$.
- Si $k_2 = 0$ se acaba el proceso y los números seleccionados son los de S1.
- Si $k_2 > 0$ a la secuencia S_1 se adicionan algunos valores tomados de entre los no seleccionados. Para esto se consideran los N − s números resultantes de eliminar la secuencia S_1 de los números 1 a N. En esa nueva lista se seleccionan los números S_2 que ocupan las posiciones $r_2 + jk_2$, con j = 0, 1,..., t − 1, donde t es el menor entero tal que $r_2 + t_{k2} > N - s$. Los números seleccionados son la unión de las dos secuencias S_1 y S_2.
- Si $k_2 < 0$, de la secuencia S1 se eliminan los números que ocupan las posiciones $r_2 + j|k_2|$, con j = 0, 1,..., g − 1, donde g es el menor entero tal que $r_2 + g|k_2| > s$. Los números seleccionados son los que quedan en S1 luego de eliminar esos valores.
- Nota: Si f > 50, los números seleccionados corresponden al complementario de los números resultantes del proceso anterior respecto del conjunto de los N primeros números naturales.
  donde:
  N: tamaño de la población;
  n: tamaño de la muestra;
  f: probabilidad de selección o fracción de muestreo.

Tabla 2.1. Intervalos de selección (k_1, k_2) para aplicar el muestreo sistemático en fases con probabilidad de selección h

h	k_1	k_2	h	k_1	k_2
.01	50	−2	.26	6	9
.02	50	0	.27	4	38
.03	33	0	.28	4	25
.04	25	0	.29	4	19
.05	20	0	.30	3	−10
.06	16	−25	.31	3	−14
.07	14	−50	.32	3	−25

Tabla 2.1. Continuación

h	k_1	k_2	h	k_1	k_2
.08	12	−25	.33	3	0
.09	10	−10	.34	3	60
.10	10	0	.35	3	40
.11	11	48	.36	3	25
.12	8	−25	.37	3	18
.13	7	−11	.38	3	14
.14	7	−50	.39	3	12
.15	6	−10	.40	2	−5
.16	6	−25	.41	3	9
.17	7	32	.42	2	−6
.18	5	−10	.43	2	−7
.19	5	−20	.44	2	−8
.20	5	0	.45	2	−10
.21	9	9	.46	2	−12
.22	5	40	.47	2	−17
.23	5	27	.48	2	−25
.24	4	−25	.49	2	−50
.25	4	0	.50	2	0

Muestreo aleatorio estratificado (Silva, 2000: 126-131)

- Tamaño total de la población:

$$N = \sum_{h=1}^{L} N_h$$

- Algoritmo de selección: en el estrato h (h = 1,..., L) se selecciona una muestra simple aleatoria de tamaño n_h.
- Probabilidad de selección (%) en el estrato h:

$$f_h = \frac{n_h}{N_h} \cdot 100 \text{ , } h = 1,..., L$$

Donde:

L: número de estratos;

N_h: tamaño poblacional del estrato h, h = 1,..., L;

nh: tamaño de muestra del estrato h, h =1,..., L. Si el reparto es proporcional:

$$n_h = n\frac{N_h}{N} \cdot 100 \text{ , } h = 1,..., L, \text{ siendo n el tamaño total de la muestra.}$$

Muestreo por conglomerados monoetápico (Silva, 2000: 157-159)

– Tamaño total de la población:

$$N = \sum_{k=1}^{M} N_k$$

– Algoritmo de selección: se selecciona una muestra simple aleatoria de m conglomerados.

– Probabilidad de selección (%):

$$f = \frac{m}{M} \cdot 100$$

Donde:

N_k: tamaño poblacional del conglomerado k, k = 1,..., M;

M: número de conglomerados en la población;

n: tamaño de muestra previsto;

m: número de conglomerados a seleccionar:

$$m = n\frac{M}{N}$$

Muestreo por conglomerados bietápico (Silva, 2000: 177-180)

1. Tamaño total de la población:

$$N = \sum_{k=1}^{M} N_k$$

2. Algoritmo de selección en la primera etapa. Hay dos opciones:
   a) Se seleccionan m conglomerados con probabilidades proporcionales a sus tamaños N_k, k = 1,..., M:
      – Se calculan los tamaños acumulados de los conglomerados:

$$C_k = \sum_{i=1}^{k} M_i \text{ , } k = 1,..., M$$

      – Se calcula el intervalo de selección $I = \left(\frac{N}{m}\right)$ parte entera.
      – Se selecciona el arranque aleatorio r entre 1 e I.
      – Se calculan los números $z_j = r + j_I$, j = 0, 1,..., m – 1.
      – Para cada z_j se selecciona el primer conglomerado k que cumpla $C_k \geq z_j$.
   b) Se selecciona una muestra simple aleatoria de m conglomerados.
3. Algoritmo de selección en la segunda etapa: en cada conglomerado de la muestra, se selecciona una muestra simple aleatoria de tamaño u_k.

4. Probabilidad de selección (%) en el conglomerado k:

   a) $f_k = \frac{m}{N} u_k \cdot 100$, k = 1,..., M, si los conglomerados se seleccionan con probabilidades proporcionales al tamaño

   b) $f_k = \frac{mu_k}{MN_k} \cdot 100$, k = 1,..., M, si se selecciona una muestra simple aleatoria de conglomerados.

   donde:

   N_k: tamaño poblacional del conglomerado k, k = 1, ..., M;

   M: número de conglomerados en la población;

   m: número de conglomerados a seleccionar;

   u_k: tamaño de muestra a seleccionar en el conglomerado k, k = 1,..., M.

Muestreo por conglomerados monoetápico estratificado (Silva, 2000: 181-190)

– Tamaño de la población de cada estrato:

$$N_h = \sum_{k=1}^{M_h} N_{hk}$$

– Tamaño total de la población:

$$N = \sum_{h=1}^{L} N_h$$

– Algoritmo de selección: en el estrato h (h = 1,..., L) se selecciona una muestra simple aleatoria de m_h conglomerados.

– Probabilidad de selección (%) en el estrato h:

$$f_h = \frac{m_h}{M_h} \cdot 100, \; h = 1,..., L$$

donde:

L: número de estratos;

M_h: número de conglomerados en el estrato h, h = 1, ..., L;

N_{hk}: tamaño poblacional del conglomerado k del estrato h, k = 1,..., M_h, h = 1,..., L;

m_h: número de conglomerados a seleccionar en el estrato h, h = 1,..., L.

- Si el reparto es proporcional a partir del número total de conglomerados en la muestra, m:

$$m_h = m\frac{N_h}{N}, \; h = 1,..., L$$

- Si m_h se calcula a partir del tamaño de muestra del estrato h, n_h:

$$m_h = n_h \frac{M_h}{N_h}, \; h = 1,..., L$$

Muestreo por conglomerados bietápico estratificado (Silva, 2000: 192-194)

1. Tamaño de la población de cada estrato:

$$N_h = \sum_{k=1}^{M_h} N_{hk}$$

2. -Tamaño total de la población:

$$N = \sum_{h=1}^{L} N_h$$

3. Algoritmo de selección en el muestreo equiprobabilístico:
   a) Primera etapa: en el estrato h (h = 1,..., L) se seleccionan mh conglomerados con probabilidades proporcionales a sus tamaños N_{hk}, k = 1,..., M_h (véase "Muestreo por conglomerados bietápico"). El reparto por estratos del número total de conglomerados en la muestra m es proporcional:

$$m_h = m\frac{N_h}{N}\text{ , h = 1,..., L}$$

   b) Segunda etapa: en cada conglomerado de la muestra del estrato h (h = 1,..., L) se selecciona una muestra simple aleatoria de tamaño u.
4. Algoritmo general de selección:
   a) Primera etapa. Hay dos opciones:
      – En el estrato h (h = 1,..., L) se seleccionan mh conglomerados con probabilidades proporcionales a sus tamaños N_{hk}, k = 1, ..., M_h (véase "Muestreo por conglomerados bietápico").
      – En el estrato h (h = 1,..., L) se selecciona una muestra simple aleatoria de m_h conglomerados.
   b) Segunda etapa: en cada conglomerado de la muestra del estrato h (h = 1,..., L), se selecciona una muestra simple aleatoria de tamaño u_{hk}.
5. Probabilidad de selección (%) en el conglomerado k del estrato h:

   a) $f_k = \frac{m_h}{N_h} u_{kh} \cdot 100$, k = 1,..., M_h, h = 1,..., L, si los conglomerados se seleccionan con probabilidades proporcionales al tamaño,

   b) $f_k = \frac{m_h}{M_h}\frac{u_{kh}}{N_{hk}} \cdot 100$, k = 1,..., M_h, h = 1,..., L, si se selecciona una muestra simple aleatoria de conglomerados.

   donde:

   L: número de estratos,

   M_h: número de conglomerados en el estrato h, h = 1, ..., L;

   N_{hk}: tamaño poblacional del conglomerado k del estrato h, k = 1, ..., M_h, h = 1,..., L;

m: número total de conglomerados a seleccionar;
m_h: número de conglomerados a seleccionar en el estrato h, h = 1,..., L;
u: tamaño de muestra a seleccionar en cada conglomerado, igual para todos;
u_{hk}: l tamaño de muestra a seleccionar en el conglomerado k del estrato h, k = 1,..., M_h, h = 1,..., L.

- Se calculan: $w_i = \frac{1}{s_i}$, $w = \sum_{i=1}^{k} w_i$ y $w^* = \frac{w_i}{w}$
- Se selecciona un número aleatorio r entre 0 y 1.
- Si el sujeto s + 1 se asigna al grupo 1.
- Si el sujeto s + 1 se asigna al grupo j (j = 2,..., k).
  donde:
  k: número de tratamientos;
  n: número de sujetos.

Clasificación de la muestra en bioestadística

- Según su número: una, dos o más de dos.
- Según el tipo de relación de las mismas:
  - Pareadas, emparejadas, dependientes y relacionadas.
  - No pareadas, no emparejadas, independientes y no relacionadas.
- Según el método de muestreo empleado para su obtención:
  - Probabilística.
  - No probabilística.
- Según la presencia de los rasgos o características más abarcadores posibles de la población en la misma o no:
  - Representativa.
  - No representativa.

Capítulo 3

Aspectos relacionados con el empleo de las variables

Clasificación de las variables de las muestras

Una variable es cualquier característica, rasgo o atributo de la población que puede asumir diferentes comportamientos, valores o grados de intensidad en diferentes momentos de un experimento, para los diferentes elementos, individuos o unidades de análisis que la conforman.

Una variable aleatoria es toda función que atribuye un único número real a cada suceso elemental del espacio muestral de un experimento aleatorio. Es la función cuyos valores no pueden ser predichos exactamente antes de realizar el experimento y no siempre toman el mismo valor aunque las condiciones del experimento se mantengan constantes. También puede ser denominada como la función en que sus valores se originan como resultado de factores aleatorios (al azar), que no pueden predecirse con exactitud y anticipación, es decir, es cualquier característica medible que toma diferentes valores con probabilidades determinadas. Los valores resultantes de los procedimientos de medición se denominan observaciones o medidas.

Las variables aleatorias pueden ser clasificadas como:

- Variable aleatoria discreta: es aquella que solo puede tomar un número finito o numerable de valores; por ejemplo: el sexo de un recién nacido.
- Variable aleatoria continua: es la que puede tomar un número infinito no numerable de valores, por ejemplo: la talla y en general variables que expresan tiempo, medidas de longitud, peso, entre otras.

Las variables en general se pueden clasificar de esta manera:

1. Por su expresión o naturaleza:
   a) Cuantitativa-numérica-paramétrica-métrica (miden cantidades, magnitudes). Las variables cuantitativas son aquellas en las que las diferencias existentes entre los diferentes elementos de la población son medibles numéricamente.
   b) Cualitativa-categórica-no paramétrica-no métrica (miden cualidades, atributos). Las variables cualitativas son aquellas en las que las diferencias entre un elemento y otro son atributos o cualidades no medibles en términos numéricos.

c) Cualitativa-cuantitativa, combinada o compleja (miden aspectos cualitativos y cuantitativos al mismo tiempo). Las variables que combinan aspectos cuantitativos y cualitativos complejizan el análisis de los datos; pero pueden facilitar la caracterización de un fenómeno de estudio dado, por ejemplo, la variable: presencia de evento, si o no (cualitativa) para un periodo de tiempo dado (cuantitativa) presenta un tipo combinado de otras dos variables.

2. Por la forma de su expresión:
   a) La variable cuantitativa-numérica-paramétrica-métrica puede ser:
      – Continua: hay continuidad entre sus valores enteros, incluye valores decimales o fraccionarios.
      – Discreta: no hay continuidad entre dos valores enteros.
   b) La variable cualitativa-categórica-no paramétrica-no métrica puede ser:
      – Nominal: tiene cualidades excluyentes sin orden o graduación de las mismas y puede ser dicotómica (dos nominaciones o categorías) o multinominal (más de dos nominaciones o categorías).
      – Ordinal: tiene cualidades no excluyentes con orden o graduación de las mismas, por lo general presenta más de dos cualidades de ordenamiento.

        Para recordar:
        - Variable cuantitativa continua permite: medir.
        - Variable cuantitativa discreta permite: contar.
        - Variable cualitativa nominal permite: clasificar.
        - Variable cualitativa ordinal permite: jerarquizar.

        Aunque en el caso de las variables cualitativas las mediciones no se llevan a cabo en el sentido usual de la palabra, se puede contar el número de personas, lugares o cosas pertenecientes a varias categorías, a estos conteos se les denomina como frecuencia (absoluta, relativa y acumulada) y entonces pueden ser asumidos los valores variables de frecuencia correspondientes de cada categoría en las muestras a estudiar, como una variable cuantitativa discreta.
3. Por su causalidad o la función que desempeñan en el problema de investigación (papel o posición que se le asigna) las variables se clasifican en:
   a) Variable dependiente: efecto, de respuesta, de salida, explicada, riesgo de…, pronóstico de…
   b) Variable independiente: causal-factor de riesgo, factor pronóstico, explicativa-covariable-confusora, de entrada, predictora. Condiciones para que una variable independiente sea un potencial confusor son:
      – Debe ser un factor de riesgo independiente para el resultado o efecto de interés.
      – Debe estar asociado con la exposición o factor de riesgo de interés (otra variable independiente).
      – No debe estar en la ruta causal entre exposición o factor de riesgo y resultado o evento de interés.

4. Por la forma de medición de las escalas de clasificación a emplear en las variables, se clasifican en:
   a) Variable con escala cualitativa: las variables se miden a partir de sus categorías (nominales) o el orden de una categoría (ordinales).
   b) Variable con escala cuantitativa: las variables se miden por intervalos o por proporción o razón.

      Recordar que:

      - Solo las variables cuantitativas pueden tener continuidad en sus valores, aunque no todas las variables numéricas son continuas. Las variables cualitativas se pueden expresar en valores numéricos como discretas.
      - Podemos asignar valores numéricos arbitrarios a las mediciones de una característica cualitativa; pero no se deben efectuar operaciones algebraicas con estos valores.
      - Una medición se define como la asignación de números a objetos o eventos de acuerdo con un conjunto de reglas.

Una escala de medición no es más que el conjunto de clases o categorías que se definen para clasificar la información en función de determinada variable.

Las diversas escalas de medición son consecuencia de que la medición puede llevarse a cabo según diferentes conjuntos de reglas.

La escala de medición para la clasificación de la información recogida debe cumplir con dos requisitos o condiciones para que sea considerada como buena:

– Que sea exhaustiva: si el número de clases o categorías que la constituyen garantizan que todos los elementos que integran la población pueden ser clasificados.
– Que las categorías sean mutuamente excluyentes, de forma tal que no quede duda de en qué clase debe ser incluido cada uno de los elementos que conforman la muestra o la población.

Los tipos de escalas de medición son:

– Escala nominal: es la más baja, consiste en designar o "nombrar" las observaciones clasificándolas en varias categorías mutuamente excluyentes y colectivamente exhaustivas.
– Escala ordinal: consiste en clasificar las observaciones por grados, fuerza o intensidad de acuerdo con algún criterio. En ocasiones se puede hacer su equivalencia o conversión a una escala nominal.
– Escala de intervalos: es más especializada que la nominal o la ordinal en el sentido de que con ella, no solo es posible ordenar las mediciones (por clases); sino que también se conoce la distancia entre dos mediciones cualesquiera.
– Escala de razones o de proporciones: es el nivel más alto de medición, se caracteriza por el hecho de que puede determinarse tanto la igualdad de las razones como la de los intervalos. Los datos pertenecen a una escala de intervalo donde existe un valor cero verdadero.

La fuerza de estas escalas aumenta en sentido ascendente, la escala nominal es la más débil y la de razones la más fuerte con vistas a realizar un procesamiento estadístico de los datos (Fig. 3.1).

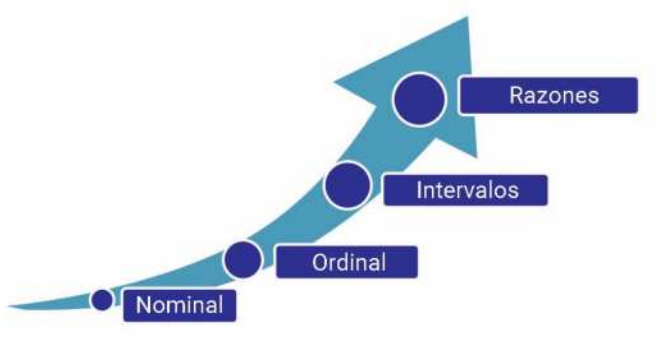


Fig. 3.1. Sentido de la fuerza de las escalas de medición para las variables en el procesamiento estadístico de la información.

Los aspectos a tomar en cuenta para la operacionalización de las variables en los diferentes diseños de estudio son:

1. Nombre de la variable.
2. Clasificación de la variable (por su expresión y causalidad).
3. Determinar el tipo de medida de resumen a emplear para cada variable y conformar la escala de medición o clasificación según el tipo de variable.
4. Descripción del aspecto según el cual se construye la escala de medición o clasificación.
5. Criterio de evaluación (indicador) empleado en cada tipo de variable.

Nota: Los aspectos 3, 4 y 5 son los que propiamente determinan la operacionalización de la variable.

Descripción de los tipos de las medidas de resumen para variables cuantitativas, numéricas, paramétricas, métricas (continuas y discretas)

Los tipos de medidas para variables cuantitativas son:

1. De tendencia central o de centralización:
   a) Media aritmética: indica el valor promedio de la variable en la muestra.
      – Para datos simples:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

donde:

$X_{i:}$ valores de la variable en cada observación;

n: total de observaciones.

– Para datos simples y empleando valores de frecuencias absoluta:

$$\overline{X} = \frac{\sum_{i=1}^{n} f_i X_i}{n}$$

donde:

f_i: frecuencias absolutas;

X_i: categorías de la variable;

n: total de observaciones.

– Para datos agrupados por intervalos de clase:

$$\overline{X} = \frac{\sum_{i=1}^{n} MC_i \cdot f_i}{n}$$

donde:

MC_i: marcas de clase de los intervalos;

f_i: frecuencias absolutas de los intervalos;

n: total de observaciones.

b) Media aritmética ponderada:

$$(\overline{X}) = \frac{\sum_{i=1}^{n} W_i \cdot X_i}{\sum_{i=1}^{n} W_i}$$

donde:

W_i (*weight*): peso que se otorga a cada valor i de la variable X en la observación.

X_i: valor de la variable en cada observación.

n: total de observaciones.

En función del tipo de problema de análisis se pueden emplear otros tipos de medias denominadas: medias generalizadas, que se aplican a unas observaciones: x1… xn.

c) La media geométrica es la media de los logaritmos de los valores de la variable X y se expresa como:

$$\overline{x}_g = \sqrt[n]{x_1 x_2 \ldots x_n}$$

$$\overline{x}_g = \sqrt[n]{x_1 x_2 \ldots x_n} = \sqrt[n]{\prod_{1}^{n} X_i}$$

Donde Π (letra griega pi mayúscula) es el multiplicatorio y significa que se debe multiplicar uno por otro, todos los valores de la variable X. La raíz no es una raíz

cuadrada, sino es una raíz del total de valores de X (según), siendo n el tamaño de muestra, por lo que el término incluido en la raíz es elevado a la 1/n para despejar la raíz, quedando:

$$\bar{x}_g = \left(\prod_1^n X_i \right)^{1/n}$$

Si los datos están agrupados en una tabla entonces:

$$\bar{x}_g = \sqrt[n]{x_1^{n_1} x_2^{n_2} \ldots x_k^{n_k}}$$

Otra forma de calcular la media geométrica es:

- Se calculan los logaritmos neperianos de los datos de la variable X ($_i$log X).
- Se calcula la media de esos logaritmos neperianos: (Σlog X_i)/n.
- Se eleva el número e a la cantidad calculada en el paso anterior.

$$\log \bar{x}_g = \frac{\log x_1 + \ldots + \log x_n}{n} = e^{\left(\sum \log x_i / n\right)}$$

d) Media armónica: se calcula dividiendo el número de observaciones entre la sumatoria del inverso de cada valor (x_i) de la variable (de cada observación):

$$\bar{x}_a = \frac{n}{\sum \left(\frac{1}{x_i} \right)}$$

e) Media cuadrática: es la raíz cuadrada de la media aritmética de los cuadrados y se expresa como:

$$\bar{x}_c = \sqrt{\frac{x_1^2 + \ldots x_n^2}{n}}$$

f) Mediana: es el valor central que toma la variable. Tiene unas propiedades de las que carece la media, por lo que es un buen complemento informativo e incluso en algunos casos puede ser una medida más útil. Estas propiedades son:
   - Es más robusta que la media frente a la presencia de anomalías o valores atípicos extremos.
   - La mediana deja un 50 % de las observaciones por encima y otro 50 % por debajo y esto le da unas ventajas que la media no tiene.
   - La mediana para datos simples y muestras impares se obtiene al ordenar numéricamente los valores de las observaciones n, la que queda en la posición del centro es la medida buscada, es decir, su posición se calcula: med = (n + 1)/2.
   - La mediana para datos simples y muestras pares se obtiene al ordenar numéricamente los valores de las observaciones n, es la media aritmética de los valores del centro de la serie de observaciones n. Se puede aplicar la fórmula anterior (n + 1)/2 y el resultado expresa la posición entre los dos valores que habría que hallar la media aritmética, el cual resulta en la mediana.

– Para datos agrupados en intervalos de clases la mediana se calcula como:

$$med = LRI + \left[\frac{\frac{n}{2} + FAA}{f}\right] \cdot c$$

donde:

LRI: límite real inferior del intervalo que contiene la mediana.

n: total de observaciones de la serie de datos.

FAA: frecuencia absoluta acumulada del intervalo de confianza que antecede al que contiene la mediana.

f: la frecuencia absoluta del intervalo de confianza que contiene la mediana.

C: amplitud del intervalo de confianza que contiene la mediana.

g) Moda: valor, clase o categoría que aparece con más frecuencia en una serie de datos; o sea, es el que más se repite). Tiene como características:
   – No siempre existe.
   – Puede no ser única.
   – En ocasiones puede usarse para datos cualitativos.

   Moda (para datos agrupados) = $L_{Mo} + [(f_{Mo} - f_A)/(2f_{Mo} - f_p - f_A)]$ a

   donde:

   L_{Mo}: límite real inferior de la clase modal;

   f_{Mo}: frecuencia absoluta de la clase modal;

   f_A: frecuencia absoluta de la clase anterior a la clase modal;

   f_p: frecuencia absoluta de la clase posterior a la clase modal;

   a: amplitud de la clase modal.

2. De dispersión o variabilidad (absoluta o relativa).
   a) De variabilidad absoluta: rango, recorrido o amplitud, es la diferencia entre el valor máximo y mínimo de la distribución de la variable y se calcula por:

      Rango = Valor máximo de X – Valor mínimo de X

   b) Desviación media absoluta: es la sumatoria de las diferencias entre cada valor de x y su media y se calcula por:

$$DMA = \sum_{i=1}^{n} (X_i - \overline{X})$$

   c) Desviación media promedio o desviación promedio: es la sumatoria de las diferencias entre cada valor de x y su media dividida entre el número de observaciones o el tamaño de muestra y se calcula:
      – Desviación media promedio para datos simples, no agrupados en clases:

$$DM = \frac{\sum_{i=1}^{n} (X_i - \overline{X})}{n}$$

donde:

X_i: observaciones de la serie (i = 1,..., n);

$\overline{X}$: media aritmética de la serie;

n: total de observaciones.

– Desviación media promedio para datos agrupados en intervalos de clases:

$$DM = \frac{\sum_{i=1}^{n} \left| MC_i - \overline{X} \right| \cdot f_i}{n}$$

donde:

MC_i: marca de clase de cada intervalo de confianza;

f_i: frecuencia absoluta de cada intervalo de confianza;

$\overline{X}$: media aritmética de la serie;

n: total de observaciones.

d) Desviación estándar o típica (s): es la raíz cuadrada de la sumatoria de las diferencias entre cada valor de x y su media al cuadrado dividida entre el número de muestra y se calcula por:

– Desviación estándar para datos simples:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}}$$

donde:

X_i: son las observaciones de la serie (i = 1,..., n);

$\overline{X}$: media aritmética de la serie;

n: total de observaciones.

– Desviación estándar para datos agrupados en intervalos de clases:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (MC_i - \overline{X})^2 f_i}{n-1}}$$

donde:

MC_i: marca de clase de cada intervalo de confianza;

f_i: frecuencia absoluta de cada intervalo de confianza;

$\overline{X}$: media aritmética de la serie;

n: total de observaciones.

– Desviación estándar o típica poblacional (σ):

$$\sigma = \sqrt{\frac{\sum (Xi - \mu)^2}{N}}$$

donde:

μ: media poblacional;

N: total de observaciones.

e) Varianza (s^2): es la desviación estándar al cuadrado y se calcula por:

- Varianza para datos simples, no agrupados en intervalos de clases:

$$S^2 = \left[\frac{\sum (X_i - \overline{X})^2}{n-1} \right]$$

donde:

X_i: son las observaciones de la serie (i = 1,..., n);

$\overline{X}$: media aritmética de la serie·

n: total de observaciones·

- Varianza para datos agrupados en intervalos de clases:

$$S^2 = \left[\frac{\sum (MC_i - \overline{X})^2 f_i}{n-1} \right]$$

donde:

MC_i: marca de clase de cada intervalo de confianza;

f_i: frecuencia absoluta de cada intervalo de confianza;

$\overline{X}$: media aritmética de la serie·

n: total de observaciones·

- Varianza poblacional (σ^2):

$$\sigma^2 = \left[\frac{\sum (X_i - \mu)^2}{N} \right]$$

donde:

μ: media poblacional;

N: total de observaciones.

f) Coeficiente de variación (CV): es una medida de dispersión relativa que expresa la relación a través de un cociente de la desviación estándar entre la media expresada en porcentaje, es decir, la desviación típica como porcentaje de la media, y se calcula por:

$$CV = \frac{SD}{\overline{X}} \cdot 100$$

donde:

SD: desviación estándar;

$\overline{X}$: media aritmética de la serie·

h) Error estándar de la media (EE) o error típico: mide el grado de dispersión de las medias de todas las posibles muestras que pudieran extraerse de la población y se calcula por:

$$EE = \frac{S}{\sqrt{n}}$$

Donde:

S: desviación estándar de todas las posibles muestras extraídas de la población;
n: número de sujetos extraídos de la población en estudio.

i) Rango intercuartílico (RCI): es la distancia entre los valores de la distribución de una variable determinada (en el eje X) que representan el 75 y 25 percentil, es decir el valor obtenido de la diferencia entre el valor del tercer y primer cuartil y se calcula por:

$RCI = Q_3 - Q_1 = P_{75} - P_{25}$

Su ventaja frente al rango es que este último suele ver bastante afectado por la presencia de cualquier valor anómalo o atípico (*outliers*) muy alto o bajo, por lo que suele ser preferible utilizarlo como medida de dispersión de los datos).

j) Estandarización de la variable X por Z: es otra forma de medir la dispersión sin depender de las unidades de medida, es decir, su transformación a:

$$z = \frac{X - \overline{X}}{S}$$

Su valor resultante expresa cada desvío ($X - \overline{X}$) como razón respecto a la desviación estándar (S) y se puede emplear para comparar distribuciones.

3. Posicional o de posición:

a) Percentiles: representa la posición que corresponde a cada 1 % de la distribución de los valores de la variable X, es decir, existen 100 por cada 100 posibles posiciones de la distribución de X:

– Para una serie de datos simples, la posición del percentil en cuestión (P_i), se calcula por:

$$P_i = \frac{n+1}{100} p_i$$

donde:

n: total de observaciones;

p_i: es el percentil deseado (i = 1, 2,..., 99).

– Para una serie de datos agrupados, la posición del percentil en cuestión (P_i), se calcula por:

$$P_i = LRI + \left(\frac{\frac{np_i}{100} - FAA}{f} \right) c$$

donde:

p_i: es el percentil deseado (i = 1, 2,..., 99);

LRI: límite real inferior de la clase o intervalo que contiene al percentil;
n: total de observaciones;
FAA: frecuencia acumulada absoluta de la clase que antecede a la del percentil;
f: frecuencia absoluta de la clase que contiene al percentil;
c: amplitud del intervalo de clase que contiene al percentil.

b) Deciles: representa la posición que corresponde a cada 10 % de la distribución de los valores de la variable x, es decir, existen 10 por cada 100 posibles posiciones de la distribución de x:
   - Para una serie de datos simples, la posición del decil en cuestión (D_I), se calcula por:

$$D_I = \frac{n+1}{100} d_i$$

donde:
n: total de observaciones;
d_i: decil deseado (i = 1, 2, 3, 4, 5, 6, 7, 8, 9).

   - Para una serie de datos agrupados, la posición del decil en cuestión (DI) se calcula por:

$$D_i = LRI + \left(\frac{\frac{nd_i}{10} - FAA}{f} \right) c$$

donde:
d_i: decil deseado (i = 1, 2,..., 9);
LRI: límite real inferior de la clase o intervalo que contiene al decil;
n: total de observaciones;
FAA: frecuencia acumulada absoluta de la clase que antecede a la del decil;
f: frecuencia absoluta de la clase que contiene al decil;
c: amplitud del intervalo de clase que contiene al decil.

c) Cuartiles: representa la posición que corresponde a cada 25 % de la distribución de los valores de la variable X, es decir, existen 4 por cada 100 posibles posiciones de la distribución de X.
   - Para una serie de datos simples, la posición del cuartil en cuestión (Q_i), se calcula por:

$$Q_i = \frac{n+1}{4} q_i$$

donde:
n: número de observaciones;
q_i: es el cuartil deseado.

– Para una serie de datos agrupados, la posición del cuartil en cuestión (Q_i), se calcula por:

$$Q_i = LRI + \left(\frac{\frac{nq_i}{4} - FAA}{f} \right) c$$

donde:

q_i: es el cuartil deseado (i = 1, 2, 3);
LRI: límite real inferior del intervalo que contiene al cuartil;
n: total de observaciones;
FAA: frecuencia acumulada absoluta de la clase que antecede a la del cuartil;
f: frecuencia absoluta de la clase que contiene al cuartil;
c: amplitud del intervalo de clase que contiene al cuartil.

Cuando las distribuciones son simétricas, coinciden la media, la mediana, la moda, el segundo cuartil, el quinto decil y el percentil cincuenta.

4. De la forma de distribución de los datos: estas medidas permiten sospechar a partir de las representaciones gráficas de la distribución de los datos como es el comportamiento de la distribución sus frecuencias, o sea, si es simétrica o no simétrica, respecto a un valor central; así como si hay unimodalidad o multimodalidad respecto al número de agrupaciones de los datos en su distribución (moda) ya sea una o más de una respectivamente.
   a) Medidas de asimetría: la expresión matemática de la asimetría es complicada y habitualmente se recurre al ordenador para calcularla, su expresión matemática puede ser:

$$A = \frac{n\sum \left(\frac{XI - X}{S} \right)^3}{(n-1)(n-2)}$$

   – Otra forma de plantearla sería:

$$A = \frac{m_3}{m_2^{3/2}}$$

donde:

$m_k = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^k$ es el momento central de orden k, k = 2, 3, 4.

   – Cuando hay asimetría positiva la cola de la derecha es más prolongada y su coeficiente de asimetría será positivo (A > 0); mientras que si la asimetría es negativa la cola de la izquierda es más larga y el coeficiente será negativo (A < 0) (Fig. 3.2).

También se definen toda una familia de estadísticos que ayudan a interpretar la asimetría, denominados momentos o índices de asimetría el más empleado es:

– Momento central de tercer orden, que se expresa como:

$$m_p = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^p$$

donde:

p: es el número de orden;

$m_p = 0$: la distribución es simétrica, no hay asimetría;

$m_p > 0$: hay asimetría derecha o positiva en la distribución de frecuencias de los datos;

$m_p < 0$: hay asimetría izquierda o negativa en la distribución de frecuencias de los datos.

– Coeficiente momento de asimetría (momento adimensional de orden 3 respecto a la media aritmética):

• Para datos simples:

$$Ma_3 = \frac{M_3}{S^3} = \frac{\sum_{x=1}^{j}\frac{\left(X_j - \overline{X}\right)3}{n}}{\left(\sqrt{M_2}\right)^3} = \frac{\sum_{x=1}^{j}\frac{\left(X_j - \overline{X}\right)3}{n}}{\sqrt{\left(\sum_{x=1}^{j}\frac{\left(X_j - \overline{X}\right)2}{n}\right)^3}}$$

• Para datos agrupados:

$$Ma_3 = \frac{M_3}{S^3} = \frac{\sum f_j\frac{\left(X_j - \overline{X}\right)3}{n}}{\left(\sqrt{M_2}\right)^3} = \frac{\sum f_j\frac{\left(X_j - \overline{X}\right)3}{n}}{\sqrt{\left(\sum f_j\frac{\left(M_j - \overline{X}\right)2}{n}\right)^3}}$$

donde:

M_J: Marca de clase;

f_j: frecuencia de clase;

$Ma_3 = 0$: distribución es simétrica, no hay asimetría;

$Ma_3 > 0$: hay asimetría derecha o positiva en la distribución de frecuencias de los datos.

$Ma_3 < 0$: hay asimetría izquierda o negativa en la distribución de frecuencias de los datos (Fig. 3.2).

– Coeficiente de asimetría (CA) de Pearson (1 y 2):

$$CA_1 = \frac{(\overline{X} - \text{moda})}{S} \quad \text{y} \quad CA_1 = \frac{3(\overline{X} - \text{mediana})}{S}$$

– Coeficiente cuartílico de asimetría (CCA):

$$CAA = \frac{Q_3 - 2Q_2 + Q_1}{(Q_3 - Q_1)}$$

– Coeficiente percentílico de asimetría (CPA):

$$CPA = \frac{P_{90} - 2P_{50} + P_{10}}{(P_{90} - P_{10})}$$

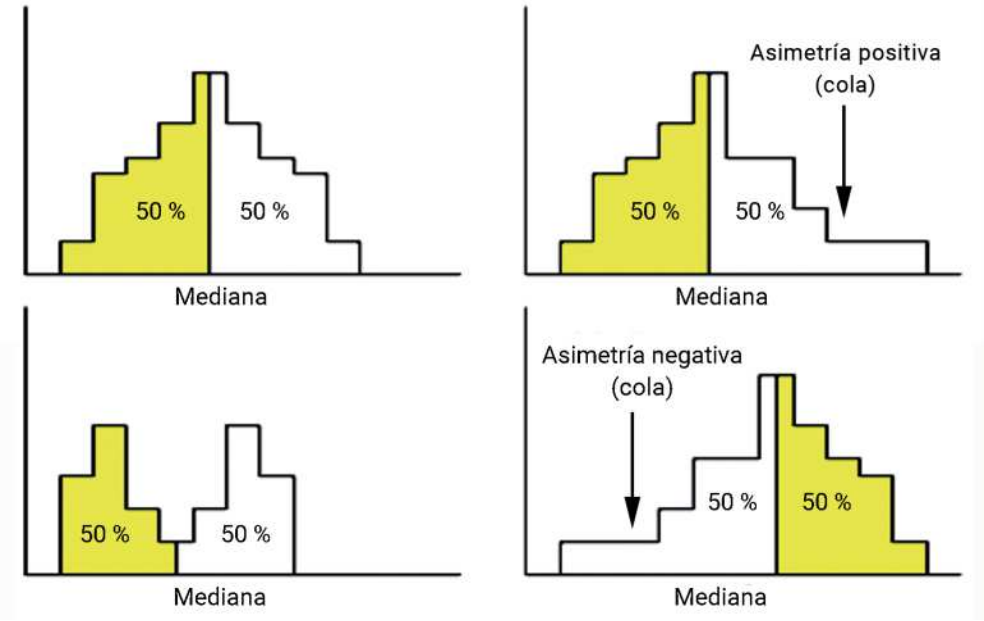


Fig. 3.2. Representación gráfica para los tipos de simetría en la distribución de una variable.

5. Otros estadísticos empleados en la forma de distribución de los datos:
   a) Índice basado en los tres cuartiles o cociente de asimetría de Yule-Bowley:
      – Si $Q_3 - Q_2 = Q_2 - Q_1$, la distribución de frecuencia de los datos es simétrica.
      – Si $Q_3 - Q_2 > Q_2 - Q_1$, la distribución de frecuencia de los datos es asimétrica positiva.
      – Si $Q_3 - Q_2 < Q_2 - Q_1$, la distribución de frecuencia de los datos es asimétrica negativa.
   b) Medidas de la curtosis: para medir el grado de esbeltez, agudeza, aplastamiento o apuntamiento de la distribución, es decir, la curtosis (K) se emplean:

$$K = \frac{m_4}{(m_2)^2} - 3$$

donde:

$m_k = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^k$ es el momento central de orden k, k = 2, 3, 4.

– Estadístico de apuntamiento (coeficiente de aplastamiento de Fisher = curtosis):

$$\gamma_2 = \frac{m_4}{\sigma^4} - 3$$

donde:

m_4: momento empírico de cuarto orden.

Según el resultado de γ_2 para la curtosis las distribuciones de frecuencia se clasifican en:

- Leptocúrtica: cuando $\gamma_2 > 0$, o sea, si la distribución de frecuencias es más apuntada que la normal.
- Mesocúrtica: cuando $\gamma_2 = 0$, es decir, cuando la distribución de frecuencias es tan apuntada como la normal.
- Platicúrtica: cuando $\gamma_2 < 0$, o sea, si la distribución de frecuencias es menos apuntada que la normal.

Los valores de los datos de la variable cuantitativa se pueden representar por medio de histogramas en relación con las medidas de curtosis (Fig. 3.3).

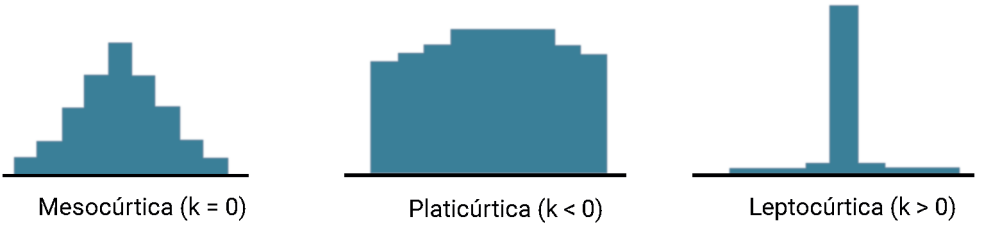


Fig. 3.3. Representación gráfica de los tipos de distribuciones de los valores de una variable por medio de histogramas para analizar la curtosis.

– Coeficiente momento de curtosis (momento adimensional de orden 4 respecto a la media aritmética):

$$Ma_4 = \frac{M_4}{(M_2)^2} = \frac{\sum_{x=1}^{j} \frac{(X_j - \overline{X})4}{n}}{\sum_{x=1}^{j} \frac{(X_j - \overline{X})2}{n}}$$

– Coeficiente de curtosis percentílico:

$$K = \frac{Q}{(P_{90} - P_{10})}$$

donde:

$$Q = \frac{Q_3 - Q_1}{2}$$

Los valores de los datos de la variable cuantitativa se pueden representar por medio de curvas de distribución en relación con las medidas de curtosis (Fig. 3.4).

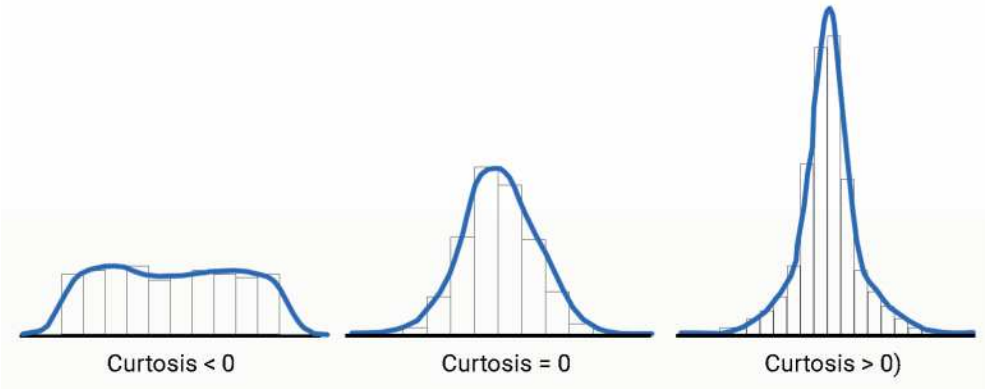


Fig. 3.4. Representación gráfica de los tipos de distribuciones de los valores de una variable cuantitativa por medio de curvas e histogramas para analizar la curtosis.

La curtosis es una medida de las llamadas "de forma" que cuantifica lo esbelta o aplanada que resulta una distribución de probabilidad (versión poblacional) o su equivalente cuando se refiere a un conjunto de datos (versión muestral). La determinación del valor de curtosis y el coeficiente de curtosis forman parte de las señas de identidad de una distribución de probabilidad y tiene interés estudiarlas en el marco de la caracterización de los modelos teóricos. Se toma como referencia de la curtosis el valor que corresponde a la distribución normal (=3) y el coeficiente de curtosis para la misma (= 0). Si una distribución tiene una curtosis mayor que la normal hay que interpretarlo como que su parte central es más picuda (con más "apuntamiento") que una normal con su misma desviación tipo, y si el valor es menor es más plana, lo que se traduce en que sus colas son más "pesadas'', es decir, que es más probable encontrar valores alejados de la media.

A modo de resumen graficado se relacionan los tipos de medidas empleados en la descripción de las variables cuantitativas con la respectiva área del gráfico para la distribución de frecuencias de una variable (Fig. 3.5).

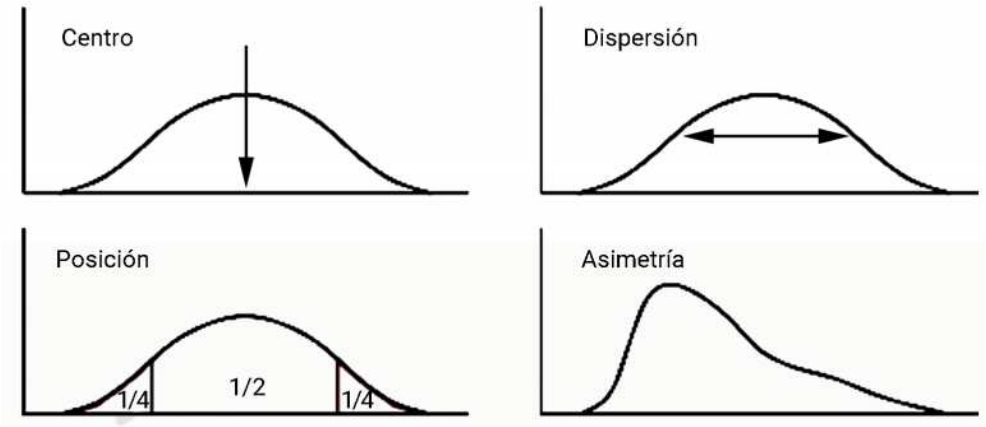


Fig. 3.5. Relación entre los tipos de medidas a emplear para describir una variable cuantitativa y la región de la curva para la distribución de frecuencias donde se expresan.

Descripción de los tipos de medidas de resumen para variables cualitativas, categóricas o no paramétricas (nominales y ordinales)

Las distintas medidas de resumen para variables cualitativas requieren para su expresión y cálculo auxiliarse de la representación graficada mediante tablas de contingencia fundamentalmente, de las que se emplean las denominaciones de las celdas; para lo cual es preciso consultar los acápites "Descripción de las posibles representaciones gráficas según el tipo de variables de estudio" y "Tipos de tablas de contingencia (cruzadas) para determinar la asociación de dos variables cualitativas".

Los tipos de medidas para variables cualitativas son las medidas de frecuencia, las medidas de asociación y las medidas de impacto.

Medidas de frecuencia

- Frecuencia absoluta: número absoluto de casos o unidades de estudio que presentan una modalidad perteneciente a una clase
- Frecuencia relativa: expresa la frecuencia o el número de observaciones en que ocurre un evento en relación con el total de observaciones o la población total en el cuál este puede ocurrir; es decir, el número de casos o con una condición del total de individuos probables de tenerla y se calcula: P = (A/A+B) y puede ser expresada en por ciento. Es el cociente entre las frecuencias absolutas de una clase(n_i) y el número total de observaciones para todas las clases (n), es decir:

$$f_i = \frac{n_i}{n}$$

- Frecuencia absoluta acumulada: es el número de elementos de la población cuya modalidad es inferior o equivalente a la modalidad de la clase o la empleada como referencia, es decir:

$$N_i = n_1 + n_2 + \ldots + n_i = \sum_{j=1}^{i} n_j$$

- Frecuencia relativa acumulada: representa el tanto por ciento de los elementos de la población que presentan una modalidad inferior o igual a la modalidad de la clase o la empleada como referencia, es decir:

$$f_i = \frac{N_i}{n} = \frac{n_1 + \ldots + n_i}{n} = f_1 + \ldots + f_i = \sum_{j=1}^{i} f_j$$

- Tasa: expresa la velocidad en que ocurre un evento en un periodo de tiempo establecido, forma especial de proporción estimada para un número elevado de la población

(10n) y se calcula: P= (A/A+B) · 100 (expresada en por ciento). Estos son algunos ejemplos de tasas y sus respectivas fórmulas de estimación (empleando las letras de cada celda en tablas de contingencia 2x2 simple):

- Tasa de mortalidad: número de fallecidos para una población determinada (10n) para un periodo de tiempo dado por 100 (expresada en por ciento).
- Tasa de natalidad: número de nacidos vivos para una población determinada (10n) para un periodo de tiempo dado por 100 (expresada en por ciento).
- Tasa de prevalencia (total, poblacional o global) (TPrev): número de casos con cierta condición, enfermedad o riesgo para una población determinada (10n) en un periodo de tiempo dado por 100 (expresada en por ciento) en tablas cruzadas:

$$TPrev = \frac{(a-c)}{n} = \frac{(a+c)}{(a+b+c)} \cdot 100$$

Otros conceptos:

– Prevalencia puntual (PrevP): es la probabilidad de que un individuo en una población presente una determinada característica, por ejemplo, enfermedad, en el tiempo t. Se estima por:

$$Prev\ P = \frac{C_t}{N_t}$$

donde:

C_t: número de casos prevalentes;

N_t: población encuestada.

– Prevalencia lápsica (PL): es la probabilidad de que un individuo de una población sea un caso en cualquier momento del periodo (t0, t). Se estima por:

$$PL_{(t0,\,t)} = \frac{C_{(t0,\,t)}}{N}$$

o también se puede presentar como:

$PL_{(to,\,t)}$ = C0 + I (to, t)/N, donde C (to, t) incluye los casos prevalentes C0 y los casos incidentes I(to, t).

– Tasa de enfermar los expuestos al factor (TEExp): expresa la proporción de enfermos del total de expuestos por 100. En tablas cruzadas o de contingencia, siguiendo la nomenclatura de sus celdas se expresa como:

$$TEExp = \frac{a}{a+b} \cdot 100$$

– Tasa de enfermar los no expuestos al factor (TENoExp): expresa la proporción de enfermos del total de no expuestos al factor por 100:

$$TENoExp = \frac{c}{c+d} \cdot 100$$

– Tasa de prevalencia relativa (TPrevR): es la razón de la prevalencia para cierta condición, enfermedad o riesgo relacionada con cierto factor, entre la prevalencia para una condición diferente para el mismo factor o con la misma condición, pero expresión diferente del factor en un momento dado por 100:

$$TPrevR = \frac{\left(\frac{a}{a+c}\right)}{\left(\frac{c}{a+c}\right)} \cdot 100$$

– Tasa de incidencia (TI): número de casos nuevos aparecidos con cierta condición en la población durante un periodo de tiempo dado por 100:

$$TI = \frac{(a+c)}{n} \cdot 100 = \frac{(A+C)}{(A+B+C)} \cdot 100$$

- Incidencia acumulada (riesgo acumulado o cumulative risk) (IA): IA = A+C
- Tasa de incidencia acumulada para la cohorte expuesta (TIA):

$$TIA = \frac{A}{(A+B)} \cdot 100$$

- Tasa de incidencia acumulada para la cohorte no expuesta (TIACNoExp):

$$TIACNoExp = \frac{C}{(C+D)} \cdot 100$$

- Tasa de incidencia acumulada para toda la cohorte (TIAC):

$$TIAC = \left[\frac{A}{(A+B)} + \frac{C}{(C+D)}\right] \cdot 100$$

– Tasa de incidencia para expuestos (TIExp): es el número de nuevos enfermos con exposición previa entre el total de personas expuestas por tiempo de seguimiento por 100:

$$TIExp = \frac{A}{PTSE} \cdot 100$$

– Tasa de incidencia para no expuestos (TINoExp): es el número de nuevos enfermos sin exposición previa entre el total de personas no expuestas por tiempo de seguimiento por 100:

$$TINoExp = \frac{C}{PTSNE} \cdot 100$$

– Densidad de incidencia en expuestos (DIExp): es el número de nuevos enfermos con previa exposición entre todo el tiempo por personas de seguimiento de los expuestos:

$$DIExp = \frac{A}{T1} = \frac{A}{TPSE}$$

– Densidad de incidencia en no expuestos (DINoExp): es el número de nuevos enfermos sin previa exposición entre todo el tiempo x personas de seguimiento de los no expuestos:

$$DIExp = \frac{C}{T2} = \frac{C}{TPSNE}$$

– Densidad de incidencia total (*incidence rate*) (IR): es el total de nuevos enfermos en la población entre todo el tiempo x personas de seguimiento de la cohorte:

$$IR = \frac{A + C}{T1 + T2}$$

Medidas de asociación

Incluye el enfoque para la estimación de indicadores de riesgo:

– Razón (R): expresa la relación a través de un cociente de dos posibilidades o riesgo de ocurrir un evento o suceso o también puede ser de "chances" u oportunidades para la ocurrencia de un evento, es decir, las probabilidades para que ocurra un suceso (A) entre las probabilidades de otro suceso o efecto, opuesto o no, (B) y se calcula:

$$R = \frac{A}{B}$$

– Índice (Ind): es una razón expresada en por ciento y se calcula:

$$Ind = \frac{A}{B} \cdot 100$$

A continuación se presentan ejemplos de razón y sus fórmulas respectivas de estimación, empleando las letras convencionales para denominar las celdas en las tablas de contingencia 2x2 simples:

1. Riesgo relativo o razón de riesgo (RR): razón de prevalencia de enfermedad o prevalencia relativa de enfermedad (para estudio transversal y de caso-control no pareado) o razón de incidencia de enfermedad (para estudio de cohorte por incidencia acumulada). Expresa el cociente entre las proporciones: efecto positivo con factor causal positivo (a) total con factor positivo (a + b) entre efecto positivo con factor causal negativo (c) / total con agente causal negativo (c + d).

   Es una medida relativa de la intensidad o fuerza, así como de la dirección de la asociación entre el factor de riesgo y la enfermedad u otra probable consecuencia asociada al factor, las veces más probables que se disminuye la tasa en expuestos si se eliminara el factor, por tablas de contingencia:

   $RR = \frac{a / m_1}{c / m_2} = \frac{a / (a+b)}{c / (c+d)} = \frac{a(c+d)}{c(a+b)}$ es estimado como:

$$RR = \frac{a/c}{b/d} \text{ o } RR = \frac{ad}{bc} \text{ para caso-control}$$

$$RR = \frac{\text{Riesgo de enfermar en expuestos}}{\text{Riesgo de enfermar en no expuestos}}$$

$$RR = \frac{\text{Riesgo específico de enfermar los expuestos al factor}}{\text{Riesgo específico de enfermar los no expuestos al factor}}$$

a) En estudios de corte o transversal la razón de riesgo (RR) es estimado como razón de prevalencias según exposición:

$$RR = \frac{\text{Prevalencia en expuestos}}{\text{Prevalencia en no expuestos}} = \frac{a/m_1}{c/m_2}$$

b) En estudios de casos y controles no pareados RR es estimado a partir de la OR de exposición:

$$RR = \frac{\text{Oportunidad de estar enfermo según exposición (Odds 1 = casos enfermos)}}{\text{Oportunidad de no estar enfermo según exposición (Odds 2 = controles sanos)}} = \frac{a/c}{b/d}$$

c) En estudios de cohorte RR es estimado como razón de incidencias acumuladas según exposición:

$$RR = \frac{\text{Incidencia en expuestos}}{\text{Incidencia en no expuestos}} = \frac{A/(A+B)}{C/(C+D)}$$

d) En estudios observacionales con estratificación RR se emplea de un factor de estratificación:

$$RR(\text{ajustado})_{\text{Mantel Haenzsel}} = \frac{\sum a(c+d)/n}{\sum c(a+b)/n} = \frac{a_{e1}(c+d)_{e1}/n_{e1} + a_{e2}(c+d)_{e2}/n_{e2}}{c_{e1}(a+b)_{e1}/n_{e1} + c_{e2}(a+b)_{e2}/n_{e2}}$$

donde:

e: estrato;

1, 2: número de la tabla para cada estrato.

Regla de interpretación:

- Si RR > 1: asociación positiva, el factor investigado constituye un factor causal, de riesgo o de pronóstico a tomar en cuenta y significa las veces mayor de tener un efecto (riesgo, enfermedad y pronóstico) en la presencia del factor causal o también las veces más probable de tener el efecto para una condición o categoría del factor respecto a su otra condición o categoría del mismo.
- Si RR < 1: asociación negativa, el factor investigado no es causal, de riesgo o de pronóstico; tiene un comportamiento protector el factor causal o exposición para el respectivo efecto (riesgo, enfermedad o pronóstico.)
- Si RR = 1: los riesgos son iguales, esté o no el factor, por ende no hay asociación entre factor causal investigado y el efecto.

2. Riesgo atribuible absoluto, diferencia de riesgo o exceso de riesgo (RA): expresa la diferencia de las proporciones: efecto positivo con factor causal positivo (a) / total

con factor positivo (a + b) menos efecto positivo con factor causal negativo (c) / total con agente causal negativo (c + d). Es una medida del impacto que provoca el factor causal para cierto efecto, las unidades que se reduce la tasa en los expuestos si se eliminara el factor de riesgo, por tablas de contingencia:

$$RA = \frac{a}{m_1} - \frac{c}{m_2} = \frac{A}{A+B} - \frac{C}{C+D} = RR - 1$$

RA = tasa de incidencia en expuestos – tasa de incidencia en no expuestos
RA = número de enfermos de total de expuestos – número de enfermos del total de no expuestos
RA = riesgo de enfermar en expuestos – riesgo de enfermar en no expuestos

Regla de interpretación:

- Si RA > 0: el factor investigado constituye un factor causal o de riesgo para enfermar (efecto) a tomar en cuenta y significa el valor ganado en tasa debido al factor, en cuantas unidades se reduciría la tasa en los expuestos si se eliminara el factor de riesgo.
- Si RA = 0: el factor investigado no tiene influencia o impacto en el riesgo, al no existir diferencias entre tasas.
- Si RA < 0: el factor investigado constituye un factor de protección para enfermar (efecto) a tomar en cuenta y significa el valor disminuido en tasa debido al factor, en cuantas unidades se aumentaría la tasa (de protección) en los expuestos si se mantuviera el factor protector.

3. Riesgo atribuible absoluto porcentual (%RAA) o porcentaje de exceso de riesgo (%ER): es la expresión porcentual de la diferencia de riesgos entre los expuestos y los no expuestos. Por tablas de contingencia,

$$\%RAA = \left(\frac{A}{A+B} - \frac{C}{C+D} \right) \cdot 100$$

4. Riesgo atribuible en expuestos, proporcional en expuestos (RAE) o fracción etiológica atribuible en expuestos (FEE): es el exceso de mortalidad o morbilidad por la presencia del factor de riesgo en los expuestos de forma proporcional respecto al total de expuestos; las unidades que se reduciría la tasa en expuestos de forma proporcional de no existir el factor de riesgo. Por tablas de contingencia:

$$RAE = \frac{\frac{a}{a+b} - \frac{c}{c+d}}{\frac{a}{a+b}}$$

$$RAE = \frac{\text{Tasa de los expuestos} - \text{Tasa no expuestos}}{\text{Tasa en expuestos}}$$

a) En estudios de cohortes es:

$$\text{RAE o FEE} = \frac{RR-1}{RR} \text{ o } \text{RAE o FEE} = \frac{RR-1}{IR}$$

b) En casos y controles es:

$$\text{RAE o FEE} = \frac{OR-1}{OR}$$

5. Riesgo atribuible en expuestos porcentual, riesgo atribuible proporcional en expuestos porcentual (%RAE) o fracción etiológica atribuible en expuestos porcentual (FEE%): es la expresión porcentual del riesgo atribuible proporcional en los expuestos; las unidades que se reduciría la tasa en expuestos de forma proporcional de no existir el factor de riesgo expresada en porcentaje. Por tablas de contingencia:

$$\%\text{RAE o FEE\%} = \left(\frac{\frac{a}{a+b} - \frac{c}{c+d}}{\frac{a}{a+b}} \right) \cdot 100$$

$$\%\text{RAE o FEE\%} = \left(\frac{\text{Tasa de los expuestos} - \text{Tasa no expuestos}}{\text{Tasa en expuestos}} \right) \cdot 100$$

a) En estudios de cohortes es:

$$\%\text{RAE o FEE\%} = \left(\frac{RR-1}{RR} \right) \cdot 100 \text{ o } \%\text{RAE o FEE\%} = \left(\frac{RR-1}{IR} \right) \cdot 100$$

b) En casos y controles es:

$$\%\text{RAE o FEE\%} = \left(\frac{OR-1}{OR} \right) \cdot 100$$

6. Riesgo absoluto poblacional (RAbsP), diferencia de riesgo poblacional (DRP), exceso de riesgo poblacional (ERP): es el exceso de mortalidad o morbilidad por la presencia del factor de riesgo en toda la población, las unidades que se reduciría la tasa poblacional de no existir o si se eliminara el factor de riesgo. Por tablas de contingencia:

$$\text{RAbsP} = \frac{a+c}{n} - \frac{c}{c+d}$$

7. Riesgo absoluto poblacional porcentual (RAbsP%), porcentaje de diferencia de riesgo poblacional (%DRP), exceso de riesgo poblacional porcentual (%ERP): es el exceso de mortalidad o morbilidad por la presencia del factor de riesgo en toda la población, las unidades que se reduciría la tasa poblacional de no existir o si se eliminara el factor de riesgo. Por tablas de contingencia:

$$\text{RAbsP\%} = \left(\frac{a+c}{n} - \frac{c}{c+d} \right) \cdot 100$$

8. Riesgo atribuible poblacional, riesgo atribuible proporcional poblacional (RAP) o fracción etiológica atribuible poblacional (FEP): es el exceso de mortalidad o morbili-

dad por la presencia del factor de riesgo de forma proporcional en toda la población; las unidades que se reduciría la tasa poblacional de forma proporcional de no existir o se eliminara el factor de riesgo. Por tablas de contingencia:

$$\text{RAP o FEP} = \frac{\frac{a+c}{n} - \frac{c}{c+d}}{\frac{a+c}{n}}$$

RAP o FEP = Tasa total (total de enfermos para población) – Tasa (enfermos no expuestos).

9. Riesgo atribuible poblacional porcentual, riesgo atribuible proporcional poblacional porcentual (%RAP) o fracción etiológica atribuible poblacional porcentual (FEP%): es la expresión porcentual del riesgo atribuible proporcional en la población, es decir, el exceso de mortalidad o morbilidad por la presencia del factor de riesgo de forma proporcional en toda la población expresada en por ciento; las unidades que se reduciría la tasa en la población de forma proporcional de no existir o se eliminara el factor de riesgo en la población, expresada en porcentaje. Por tablas de contingencia:

$$\%\text{RAP o FEP\%} = \left(\frac{\frac{a+c}{n} - \frac{c}{c+d}}{\frac{a+c}{n}} \right) \cdot 100$$

$$\%\text{RAP o FEP\%} = \left(\frac{\text{Tasa total} - \text{Tasa de los no expuestos}}{\text{Tasa total}} \right) \cdot 100$$

10. Riesgo específico de enfermar los expuestos (REEEexp): expresa la probabilidad de enfermar los expuestos al factor de riesgo. Por tablas de contingencia:

$$\text{REEEexp} = \frac{a}{a+b}$$

11. Riesgo específico de enfermar los no expuestos (REENoExp): expresa la probabilidad de enfermar los no expuestos al factor de riesgo. Por tablas de contingencia:

$$\text{REENoExp} = \frac{c}{c+d}$$

12. Riesgo total (RT): expresa la sumatoria de la probabilidad de enfermar los expuestos y los no expuestos para una población determinada. Se calcula en tablas de contingencia:

$$\text{RT} = \frac{\frac{a}{a+b} + \frac{c}{c+d}}{n} = \frac{\frac{a}{a+b} + \frac{c}{c+d}}{a+b+c+d}$$

13. Razón de riesgo o *hazards ratio* (HR) o razón de tasas instantáneas para un evento entre dos grupos:

$$HR = \frac{\dfrac{\text{Sujetos del grupo 1 con el evento}}{\text{Total de sujetos con posibilidad del evento en grupo 1} \cdot \text{Tiempo de seguimiento}}}{\dfrac{\text{Sujetos del grupo 2 con el evento}}{\text{Total de sujetos con posibilidad del evento en grupo 2} \cdot \text{Tiempo de seguimiento}}}$$

Ejemplo para estudios observacionales con enfoque de riesgo:

$$HR = \frac{\dfrac{\text{Expuestos que enferman}}{\text{Total de expuestos} \cdot \text{Tiempo de seguimiento}}}{\dfrac{\text{No expuestos que enferman}}{\text{Total de no expuestos} \cdot \text{Tiempo de seguimiento}}}$$

Nota: El tiempo de seguimiento a tomar en cuenta para el cálculo de HR es el mismo en el denominador y el denominador.

14. Razón de disparidad, de desventaja, de oportunidades, de momios (RM) o de productos cruzados (odds ratio) (OR): expresa el cociente del número de sujetos con cierto efecto con y sin la presencia de un factor de riesgo; entre el número de sujetos sin el mismo efecto anterior con y sin el factor de riesgo presente o la razón (relación) de estar enfermo con y sin exposición a cierto factor de riesgo entre no estar enfermo con y sin exposición a dicho factor. Indica las veces que es mayor o menor la oportunidad de suceder un evento para cierto factor de riesgo presente respecto a la oportunidad de no suceder dicho evento por la influencia del mismo factor. Se calcula para tablas de contingencia:

$$OR == \frac{a \cdot d}{b \cdot c}$$

que se obtiene a partir de la siguiente expresión o formulación matemática desarrollada:

$$RT = \left(\frac{\dfrac{a}{a+c}}{\dfrac{c}{c+d}} \right) / \left(\frac{\dfrac{b}{b+d}}{\dfrac{d}{b+d}} \right) = \frac{\dfrac{a}{c}}{\dfrac{b}{d}} = \frac{ad}{bc}$$

En estudios observacionales epidemiológicos se puede expresar como OR de exposición y como OR de enfermedad, donde se varían en sus formulaciones los órdenes de los valores de las distintas celdas para las tablas de contingencia 2x2 simples, pero al final en la expresión matemática de la OR se conduce al mismo planteamiento en la formulación de esta ecuación:

a) Para la OR de exposición:
   - Oportunidad o *chance* (*odds*) de haber tenido la exposición para la aparición de la enfermedad (Odds casos o enfermos u Odds 1): expresa una razón de proporciones de enfermos según la presencia o no de exposición:

$$\text{(Odds casos o enfermos u Odds 1)} = \frac{\text{Enfermos o casos expuestos en relación con el total de enfermos o casos}}{\text{Enfermos o casos no expuestos en relación con el total de enfermos o casos}}$$

Por tablas de contingencia:

$$\text{(Odds casos o enfermos u Odds 1)} = \frac{a/n_1}{c/n_1} = \frac{a/(a+c)}{c/(a+c)}$$

– Oportunidad o chance (Odds) de haber tenido la exposición para no tener la enfermedad (Odds de controles o sanos) u (Odds 2): expresa una razón de proporciones de controles o sanos según la presencia o no de exposición:

$$\text{(Odds sanos o controles u Odds 2)} = \frac{\text{Expuestos sanos o controles en relación con el total de controles o sanos}}{\text{No expuestos o controles en relación con el total de controles o sanos}}$$

Por tablas de contingencia:

$$\text{(Odds sanos o controles u Odds 2)} = \frac{b/n_2}{d/n_2} = \frac{b/(b+d)}{d/(b+d)}$$

La OR de exposición (OR Exp = Odds 1/ Odds 2) también se expresa como:

$$OR\ Exp = \frac{\text{Oportunidades de exposición para enfermar}}{\text{Oportunidades de exposición para no enfermar}}$$

Por tablas de contingencia 2x2 se expresa como:

$$\text{OR Exp} = \frac{(a/n_1)/(c/n_1)}{(b/n_2)/(d/n_2)} = \frac{(a/a+c)/(c/a+c)}{(b/b+d)/(d/b+d)} = \frac{(a/c)}{(b/d)} = \frac{a \cdot d}{b \cdot c}$$

b) Para la OR de enfermedad es necesario tener en cuenta:

– Oportunidad o chance (Odds) de tener la enfermedad previa exposición (Odds de expuestos u Odds 1 en estudios de cohorte): expresa una razón de proporciones entre enfermos y sanos con previa exposición.

$$\text{(Odds expuestos u Odds 1)} = \frac{\text{Enfermos expuestos en relación con el total de expuestos}}{\text{No enfermos expuestos en relación con el total de expuestos}}$$

Por tablas de contingencia:

$$\text{(Odds expuestos u Odds 1)} = \frac{A/(A+B)}{B(A+B)}$$

– Oportunidad o chance (Odds) de tener la enfermedad sin previa exposición (Odds no expuestos u Odds 2 para estudios de cohorte): expresa una razón de proporciones entre enfermos y sanos sin previa exposición:

$$\text{(Odds no expuestos u Odds 2)} = \frac{\text{Enfermos no expuestos en relación con el total de no expuestos}}{\text{Sanos no expuestos en relación con el total de no expuestos}}$$

Por tablas de contingencia:

$$\text{(Odds no expuestos u Odds 2)} = \frac{C/(C+D)}{D(C+D)}$$

La OR de enfermedad (OR Enf = Odds 1 / Odds 2) también se expresa como:

$$\text{OR enfermedad} = \frac{\text{Oportunidades de enfermar para expuestos}}{\text{Oportunidades de enfermar para no expuestos}}$$

Por tablas de contingencia:

$$\text{OR de enfermedad} = \frac{\left(\frac{A}{A+B}\right)/\left(\frac{B}{A+B}\right)}{\left(\frac{C}{C+D}\right)/\left(\frac{D}{C+D}\right)} = \frac{a/b}{c/b} = \frac{a \cdot d}{b \cdot c}$$

c) En estudios observacionales con estratificación (empleo de un factor de estratificación):

$$\text{OR o RM(ajustada)}_{\text{Mantel Haenzer}} = \frac{\sum(ad/n)}{\sum(bc/n)} = \frac{(ad_{e1}/n_{e1}) + (ad_{e2}/n_{e2})}{(bc_{e1}/n_{e1}) + (bc_{e2}/n_{e2})}$$

donde:

e: estrato;

1, 2: número de la tabla de cada estrato.

d) En estudios observacionales empleando el método o paradigma bayesiano:

$$\text{OR a posteriori (puntual)} = OR_{pt} = \exp(\mu_{pt})$$

donde:

$$\mu_{pt} = \frac{\frac{\mu_{pr}}{\sigma^2_{pr}} + \frac{\mu_{dt}}{\sigma^2_{dt}}}{\frac{1}{\sigma^2_{pr}} + \frac{1}{\sigma^2_{dt}}};$$

$\mu_{pr} = \ln(OR_{pr})$;

$\mu_{dt} = \ln(OR_{dt})$;

$$\sigma_{pr} = \frac{\ln(Sup_{pr}) - \ln(Inf_{pr})}{3{,}92};$$

$$\sigma_{dt} = \frac{\ln(Sup_{dt}) - \ln(Inf_{dt})}{3{,}92};$$

OR_{pr}: OR *a priori*, que se da como referencia por estudios realizados previamente;
OR_{dt}: OR de los datos o empírica, calculada para de los datos muestrales;
Sup_{pr}, Inf_{pr}: valores superior e inferior correspondientes al intervalo de confianza de OR_{pr}.
Sup_{dr}, Inf_{dt}: valores superior e inferior correspondientes al intervalo de confianza de OR_{dt}.

Regla de interpretación para OR:

- Si OR > 1: asociación positiva, la exposición o presencia del probable factor causal (de riesgo, de pronóstico, variable independiente) está asociado, guarda relación o puede comportarse como una causa o factor de riesgo, de pronóstico para el efecto relacionado (riesgo, enfermedad, pronóstico, variable dependiente).
- Si OR < 1: asociación negativa, la exposición o presencia del probable factor causa l (variable independiente) está asociado, guarda relación o puede comportarse como una causa o factor protector para el efecto relacionado (variable dependiente).
- Si OR = 1: no asociación, la exposición o presencia del probable factor causal (variable independiente) no está asociado, no guarda relación o no puede comportarse como una causa o factor para el efecto relacionado (variable dependiente).

15. Razón de prevalencias o de prevalencias según exposición (RPEnf) (similar al RR): expresa la relación entre el número de enfermos (nuevos y existentes) expuestos del total de expuestos entre el número de enfermos (nuevos y existentes) no expuestos del total de no expuestos. Por tablas de contingencia:

$$RPEnf = \frac{a/(a+b)}{c/(c+d)}$$

$$RPEnf = \frac{\text{Riesgo específico de enfermar los expuestos al factor}}{\text{Riesgo específico de enfermar los no expuestos al factor}}$$

16. Razón de incidencia (RI): expresa la relación de los nuevos enfermos para un total de personas expuestas por un tiempo de seguimiento entre los nuevos enfermos para un total de personas no expuestas por el mismo tiempo de seguimiento. Por tablas de contingencia:

$$RI = \frac{A/PTSE}{C/PTSNE}$$

17. Razón de incidencia acumulada según exposición (RPEnf) (similar al RR): Expresa la incidencia acumulada, de enfermos según el total de expuestos entre la incidencia acumulada, de enfermos según el total de no expuestos durante todo el periodo de seguimiento. Por tablas de contingencia:

$$RPEnf = \frac{a/(a+b)}{c/(c+d)}$$

$$RPEnf = \frac{\text{Riesgo específico de nuevos enfermos en expuestos}}{\text{Riesgo específico de nuevos enfermos en no expuestos}}$$

18. Razón de densidad de incidencia (RDI): expresa la relación de los nuevos enfermos para un tiempo por persona de seguimiento del total de expuestos entre los nuevos

enfermos para un tiempo por persona de seguimiento del total de los no expuestos. Por tablas de contingencia

$$RDI = \frac{A/TPSE}{C/TPSNE}$$

19. Razón de prevalencia para exposición (RPExp), oportunidad o chance (odds) de tener la enfermedad previa exposición: expresa la prevalencia (enfermos) expuestos del total de expuestos entre la no prevalencia (sanos) expuestos del total de sanos. Por tablas de contingencia:

$$RPExp = \frac{a/(a+c)}{b/(b+d)}$$

$$RPExp = \frac{\text{Riesgo específico de enfermar en expuestos}}{\text{Riesgo específico de no enfermar en expuestos}}$$

20. Razón de incidencia acumulada o de incidencias acumuladas según exposición (RIAcu):

$$RIAcu = \frac{A/(A+B)}{C/(C+D)}$$

21. Razón de incidencia o de tasas de incidencia (RI):

$$RI = \frac{A/PTE}{C/PTNE}$$

22. Razón de densidades de incidencias (RDI) (*rate ratio*):

$$RI = \frac{A/T_1}{C/T_2}$$

23. Razón de mortalidad estandarizada (RME):

$$RI = \frac{(\sum A)/(\sum C \cdot B)}{1000} \cdot 100 = \frac{(\text{Total de A})/(\text{Total de C} \cdot B)}{1000}$$

En resumen, podemos emplear sus formulaciones auxiliándonos de las respectivas tablas de contingencia con la nomenclatura establecida de forma estandarizada para las letras correspondientes a las celdas, según los distintos tipos de estudios observacionales epidemiológicos.

Medidas de impacto

Se incluye el enfoque para la estimación de indicadores de riesgo.

Las medidas de impacto expresan una estimación de los riesgos o probabilidad de consecuencias negativas (de impacto negativo) y positivas (de impacto positivo) en la población debido a la exposición a cierto factor causal(de riesgo o protección).

De impacto negativo

Se sospechan cuando IR, RR, OR (según el tipo de estudio) es mayor de 1.

Nota: no existe unificación de criterios entre todos los autores al incluir los tipos de riesgos abordados anteriormente como medidas de asociación, también como de impacto negativo y solo expresan como medidas de impacto negativo a las fracciones atribuibles en expuestos y en población.

Entre las medidas de impacto negativo se pueden mencionar:

- Riesgo relativo o razón de riesgo (RR), razón de riesgo de incidencias (IR) y razón de riesgo de prevalencias (RP).
- Riesgo absoluto (RA), diferencia de riesgo (DR) y exceso de riesgo(ER).
- Riesgo absoluto porcentual (%RA) o diferencia de riesgo porcentual (DRP%) y exceso de riesgo porcentual (ER%).
- Riesgo atribuible en expuestos, riesgo atribuible proporcional en expuestos (RAE) o fracción etiológica atribuible en expuestos (FEAE).
- Riesgo atribuible en expuestos porcentual, riesgo atribuible proporcional en expuestos porcentual (%RAE) o fracción etiológica atribuible en expuestos porcentual (FEAE%).
- Riesgo absoluto poblacional (RAbsP), diferencia de riesgo poblacional(DRP) y exceso de riesgo poblacional (ERP).
- Riesgo absoluto poblacional porcentual (RAbsP%), porcentaje diferencia de riesgo poblacional (%DRP) y exceso de riesgo poblacional porcentual (%ERP).
- Riesgo atribuible poblacional, riesgo atribuible proporcional poblacional (RAP) o fracción etiológica atribuible poblacional (FEAP).
- Riesgo atribuible poblacional porcentual, riesgo atribuible proporcional poblacional porcentual (%RAP) o fracción etiológica atribuible poblacional porcentual (FEAP%).
- Riesgo total (RT).

De impacto positivo

Se sospechan cuando IR, RR, OR (según el tipo de estudio) es menor de 1.

Entre las medidas de impacto positivo, de prevención o de protección para expuestos se pueden mencionar:

- Fracción de prevención para expuestos porcentual o porcentaje de protección puntual por la exposición (FPE%) o porcentaje (puntual) de protección de expuestos en la población:
  $FPE\% = PF_e = (1 - IR) \cdot 100$ (a partir de la densidad de incidencia en estudio de cohorte).
  $FPE\% = PF_e = (1 - RR) \cdot 100$ (a partir de la incidencia acumulada en estudio de cohorte).
  $FPE\% = PF_e = (1 - OR) \cdot 100$ (a partir del OR de exposición en estudio de caso-control).
- Fracción de prevención máxima en expuestos porcentual o porcentaje de protección máximo posible ($\%FPE_{Máx}$) o porcentaje máximo posible de protección por la exposición al factor en los expuestos:

$\%FPE_{Máx}$ = (1 – valor IC inferior FPE o PFe a partir del IR) · 100 (estudio de cohorte por densidad de incidencia).

$\%FPE_{Máx}$ = (1 – valor IC linferior FPE o PFe a partir del RR) · 100 (para estudio de cohorte por incidencia acumulada).

$\%FPE_{Máx}$ = (1 – valor IC inferior FPE o PFe a partir del OR) · 100 (para estudio de caso-control).

– Fracción de prevención mínima en expuestos porcentual o % de protección mínimo posible ($\%FPE_{Mín}$) o porcentaje mínimo posible de protección por la exposición al factor en los expuestos:

$\%FPE_{Mín}$ = (1 – valor IC superior FPE o PFe a partir del IR) · 100 (para estudio de cohorte por densidad de incidencia o tasas de incidencia).

$\%FPE_{Mín}$ = (1 – valor IC superior FPE o PFe a partir del RR) · 100 (para estudio de cohorte por incidencia acumulada).

$\%FPE_{Mín}$ = (1 – valor IC superior FPE o PFe a partir del OR) · 100 (para estudio de caso-control).

Entre las medidas de impacto positivo, de prevención o de protección para la población se puede mencionar:

– Fracción de prevención para la población porcentual (FPP%) o porcentaje de protección puntual para la población por la exposición al factor protector:

$$FPP = PF_p = PF_e \frac{T_1}{T} \cdot 100$$ (a partir de la densidad de incidencia en estudios de cohorte).

$$FPP = PF_p = PF_e \frac{n_1}{n} \cdot 100$$ (a partir de la incidencia acumulada en estudios de cohorte).

$$FPP = PF_p = \frac{p_1 PF_e}{p_1 PF_e + OR} \cdot 100$$ (a partir del OR de exposición en estudio de caso-control).

Aunque el índice de posición (IP) no es un tipo de medida o no constituye un estimador propiamente dicho de las variables cualitativas, permite describir el comportamiento de las variables ordinales respecto a las muestras de donde proceden. Representa una medida de gran utilidad para cuantificar la posición global de una muestra con respecto a una variable categórica medida en escala ordinal (variable ordinal), sin necesidad de tener en cuenta el número de clases que la componen. Por ejemplo, si se tiene una muestra de n individuos que se evalúan a través de una escala ordinal de k clases ($A_1, A_2, ..., A_k$), el Índice de Posición permite obtener un valor resumen (situado entre 0 y 1) de las respuestas de todos los individuos de la muestra.

Este Índice toma el valor 0 cuando toda la muestra está ubicada en el extremo inferior de la escala ordinal, o sea, como A1 y el valor 1 cuando se encuentra ubicada en el extremo superior A_k. Además, si la distribución de los individuos es simétrica respecto a la clase o clases centrales de la escala (en función de si k es impar o par, respectiva-

mente), entonces el índice toma el valor de 0,5. En términos de inferencia estadística. Su estimación puntual es:

$$IP = \frac{M-1}{k-1};$$

mientras que por intervalo de confianza (IC):

$$\text{IC inferior} = IP + z_{1-\alpha/2}\sqrt{Var(IP)}$$

$$\text{IC superior} = IP + z_{1-\alpha/2}\sqrt{Var(IP)}$$

donde:

k: número de categorías de la escala;

$$M = \sum_{i=1}^{k} ip_i ;$$

$p_i = \frac{x_i}{n}$: proporción de sujetos que eligen la categoría i con i = 1,..., k;

xi: de sujetos que eligen la categoría i, con i = 1,..., k;

$n = \sum_{i=1}^{k} x_i$: tamaño de la muestra;

$$Var(IP) = \frac{1}{n(k-1)^2}\left\{\sum_{i=1}^{k} i^2 p_i(1-p_i) - 2\sum_{i=1}^{k}\sum_{j>1} ijp_ip_j\right\};$$

$z_{1-\alpha/2}$: percentil de la distribución normal estándar N(0,1), que deja a la izquierda una cola de probabilidad $1-\alpha/2$;

$1-\alpha$: nivel de confianza.

Descripción de las posibles representaciones gráficas según el tipo de variables de estudio

El aspecto descriptivo de la bioestadística en el trabajo con las variables de las muestras o la caracterización de las muestras con el auxilio de sus variables comprende generalmente la realización de la distribución de frecuencia para sus datos con su respectiva representación tabular y gráfica según los tipos de variables a utilizar, conjuntamente con el empleo de las medidas de resumen correspondiente en esas variables.

Aunque la disposición de la tabla de frecuencia o cuadro estadístico varía de acuerdo con los datos que intenta resumir, hay algunos elementos comunes que deben tenerse en cuenta como:

- Identificación y título de la tabla o cuadro (presentación).
- Cuerpo de la tabla.

– Fuente.
– Notas explicativas de ser necesarias.

La identificación no es más que otorgar un orden consecutivo ascendente a las tablas o cuadros dentro del informe que se elabora. El título debe reunir los requisitos de ser completo y conciso, la condición de completo exige que el título responda a: ¿de qué trata, qué se presenta?, ¿cómo se midieron los datos, a través de que variables?, ¿dónde y cuándo se realizó el estudio? La condición de conciso, consiste en escribir justamente lo necesario para que sea completo.

El cuerpo de la tabla siempre está formado por filas (disposición horizontal) y columnas (disposición vertical). La primera columna suele identificarse como columna matriz y se utiliza para consignar las categorías o clases de la variable del estudio. La segunda columna es para mostrar la frecuencia absoluta, es decir, el conteo que se hizo para cada categoría de la variable. La tercera columna expresa la frecuencia relativa pudiendo ser expresada en por ciento y la cuarta columna para la frecuencia acumulada puede ser presentada también en por ciento. La primera fila (o las primeras filas si fuera necesario utilizar más de una) se utiliza para el encabezamiento de cada columna. La columna y fila últimas de las tablas se dedican a los totales. La fuente se refiere al documento de donde se extrajo la información presentada. Resulta válido y oportuno aclarar que en las tablas o cuadros solo se consignan fuentes secundarias. Las notas explicativas se utilizan cuando se considere que es oportuno aclarar algo en la tabla.

El tipo de tabla o cuadro descrito se conoce como tabla de distribución de frecuencia variable simple o también como tabla unidimensional, pues solo involucra a una variable con las frecuencia de sus datos, aunque también se pueden emplear en el trabajo con variables en las muestras las llamadas tablas bidimensionales, es decir, donde se relacionan dos variables y que suelen denominarse también como tablas de relación de variables y que se ejemplifican en las tablas de contingencia cruzadas o de doble entrada.

La distribución de frecuencias de los datos de una variable consiste en resumir y presentar en una tabla los diferentes valores que tomó la variable de interés en el estudio, a partir de una escala seleccionada y haciendo el conteo de los valores o de las categorías (según sea la variable) que cumpla con la escala definida.

Respecto a una escala de clasificación se debe recordar que se conceptualiza como el conjunto de clases o categorías que se definen para clasificar la información en función de determinada variable. Su construcción por clases o intervalos, la que se relaciona con una distribución de frecuencias para una variable cuantitativa, comienza por identificar en el conjunto de valores, cuál es el mayor valor y cuál es el menor valor.

Seguidamente se restan ambos valores para obtener el recorrido de los valores de la variable y este a su vez se divide entre el número de intervalos o clases seleccionados por el investigador, según su experiencia y propósitos del estudio, para obtener la amplitud de la clase. Siempre que el resultado del cálculo de la amplitud tenga parte fraccionaria, hay que aproximarla al número entero inmediato superior.

En una escala con pocas clases se concentra la información y en una escala con muchas clases la información se disgrega. Los intervalos de clase se construyen partiendo, primeramente, del valor menor del conjunto de valores y sumándole la amplitud calculada previamente, para obtener la primera clase, que va del valor menor del conjunto a la suma de este a la amplitud, para obtener la segunda clase se debe sumar la amplitud al valor superior de la clase anterior. Para obtener las siguientes clases se repite sumar la amplitud al valor superior de la anterior clase y sucesivamente se repite el procedimiento hasta la última clase determinada que incluya al valor superior del conjunto de datos.

Al valor inicial de la clase se le conoce como límite inferior de la clase y al valor final como límite superior de la clase, mientras que el valor medio de cada clase se denomina *marca de clase*. Otro elemento a describir son los valores fronteras de clase, que son los valores de la variable cuantitativa continua que pueden estar ubicados entre los límites de las clases consecutivas para delimitar las frecuencias acumuladas hasta una determinada clase.

Por último se coloca el signo positivo o mayor que delante del valor inferior de cada clase a partir de la segunda, lo que garantiza que se cumpla con el requisito de ser mutuamente excluyente, es decir, que cada valor sea incluido o clasificado en una sola clase y solo una. Antes de utilizar la escala se debe asegurar también que se cumpla con el requisito de ser exhaustiva, lo que se puede comprobar observando si admite valores desde el valor mínimo y al máximo del conjunto de valores, o sea, que están incluidos todos los valores a utilizar en las clases establecidas.

Una vez obtenida la escala a emplear y garantizando que cumple con los requisitos exigidos de ser exhaustiva y mutuamente excluyente, se pasa a confeccionar la distribución de frecuencia para los datos de la variable de estudio y efectuar su representación tabular.

Se debe recordar que una tabla o cuadro estadístico es un recurso que emplea la estadística para representar de forma resumida, organizada y entendible determinada información (Tabla 3.1).

Tabla 3.1. Ejemplo de tabla general para representar la distribución de frecuencia de los datos de las variables según el tipo de escala a emplear

Tipo de escala para variable	Frecuencia absoluta	Frecuencia relativa	% frecuencia relativa	Frecuencia acumulada	% frecuencia acumulada
Totales					

Otra manera complementaria y no sustitutiva a las tablas mencionadas de presentar la información estadística, es permitir al lector obtener de una forma más rápida y explícita una impresión del conjunto del material presentado en los gráficos estadísticos. El gráfico es la representación de un suceso estadístico por medio de figuras geométricas

(puntos, líneas, rectángulos, círculos) o figuras. Unas veces su finalidad es simplemente mostrar a otras personas la evolución de determinado suceso, pues mientras que la interpretación de un cuadro estadístico requiere ciertos conocimientos, cualquiera puede comprender fácilmente que una línea ascendente traduce un aumento del suceso estudiado y una línea descendente significa una disminución. Otras veces la finalidad de un gráfico es ayudar al análisis de la información aclarando ciertas relaciones poco aparentes en el material tabulado.

Los componentes de un gráfico son:

- Identificación y título (presentación).
- Gráfico propiamente dicho.
- Fuente.
- Notas explicativas (de ser necesarias).
- Leyenda (siempre necesaria).

En cuanto a la identificación y título se mantienen los mismos requisitos que para las tablas o cuadros estadísticos. En relación con el gráfico propiamente dicho, en su construcción hay reglas generales y, además, aspectos específicos para cada tipo de gráfico. Para la fuente y notas explicativas es similar a lo exigido para las tablas. La leyenda si es obligatoria y su fin es identificar los elementos del gráfico con su correspondiente origen.

Gráfico de barras simples

Para representar los datos de una variable cualitativa o cuantitativa discreta y la información se dispone en frecuencias absolutas o relativas o en medidas de resumen a través de un sistema de coordenadas. Se colocan las categorías de la variable cualitativa o valores de la cuantitativa discreta en el eje x (abscisas) y las frecuencias en el eje y (ordenadas). Se puede emplear para representaciones de series cronológicas de pocos datos (Fig. 3.6).

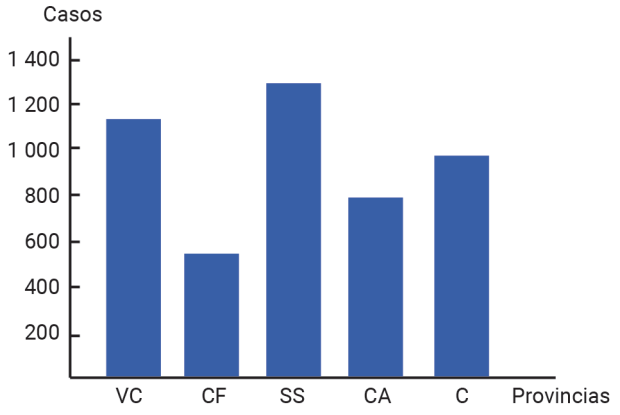


Fig. 3.6. Ejemplo de gráfico de barras simples sin ordenar.

Gráfico de barras múltiples

Para representar los datos de dos variables, que pueden ser cualitativas o cuantitativas discretas ambas o una cualitativa y la otra cuantitativa discreta, la información se dispone en frecuencias absolutas o relativas o en medidas de resumen. Los datos se representan mediante barras agrupadas. Sobre el eje de las x (abscisas) se colocan las categorías representativas de una de las variables y sobre cada una de esas categorías se levantan tantas barras como categorías tenga la otra (Fig. 3.7).

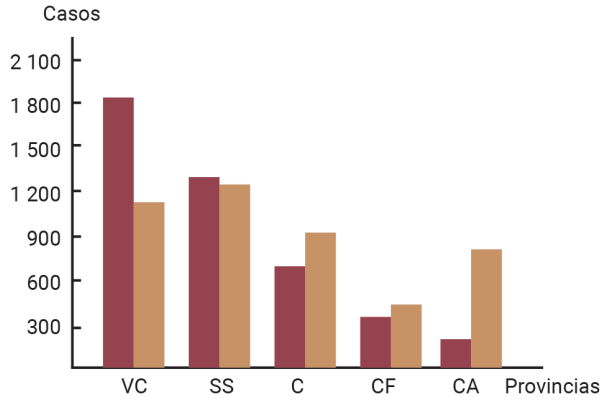


Fig. 3.7. Ejemplo de gráfico de barras múltiples.

Gráfico de barras compuestas o proporcionales

Para representar los datos de dos variables, que pueden ser ambas cualitativas o cuantitativas discretas, o una cualitativa y la otra cuantitativa discreta, de esta forma se dispone de la información en frecuencias relativas y la información perteneciente a una variable se representa en su totalidad en una sola barra (Fig. 3.8).

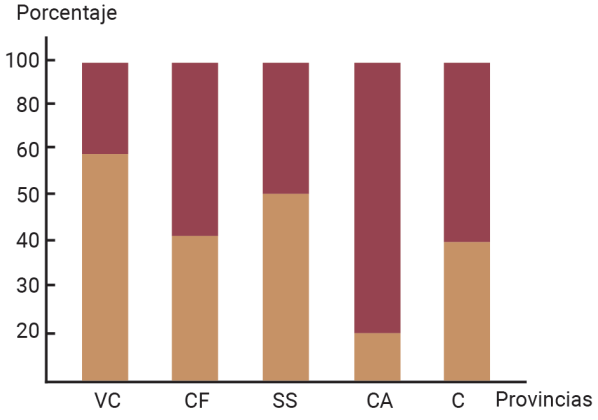


Fig. 3.8. Ejemplo de gráfico de barras compuestas o proporcionales.

Gráfico de Pareto

Se representa por barras que parten desde el eje horizontal o las abscisas relacionadas con datos categorizados para variables cualitativas, donde dichas barras se ordenan de acuerdo con las frecuencias absolutas o relativas correspondientes con cada una. La barra más alta se coloca a la izquierda y las más pequeñas a la derecha. Al ordenar las barras por frecuencias, este gráfico enfoca la atención en las categorías más importantes (Fig. 3.9.).

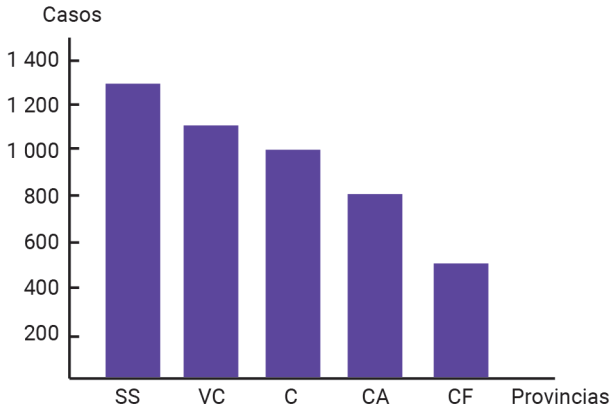


Fig. 3.9. Ejemplo de gráfico de Pareto.

Gráfico de pastel, de sectores o circular

Para representar los datos de una variable cualitativa o cuantitativa discreta, la información se dispone en porcentaje. Básicamente es un círculo dividido en sectores que representan las categorías de la variable y brinda la misma información que el gráfico de barras simples (Fig. 3.10).

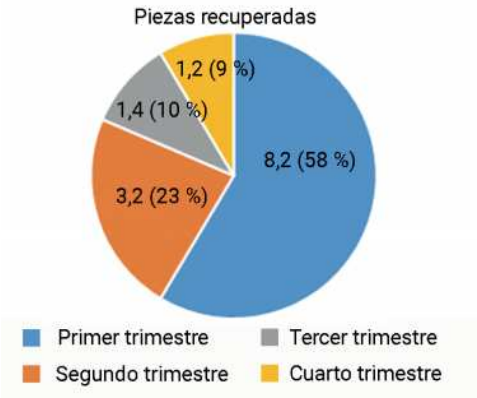


Fig. 3.10. Ejemplo de gráfico de pastel, de sectores o circular.

Gráfico aritmético simple

Para representar los datos de una variable cualitativa o cuantitativa discreta a partir de una representación de puntos que se unen entre sí por segmentos, quedando en el eje de las ordenadas las frecuencias, donde se toma la escala aritmética y en el eje de las abscisas una serie de valores cronológicos. Si para la representación fuera necesario usar la escala logarítmica se llama *gráfico semilogarítmico*. En este tipo de gráfico se admite el uso de interrupciones en los ejes cuando se desea reflejar la tendencia del fenómeno (Fig. 3.11).

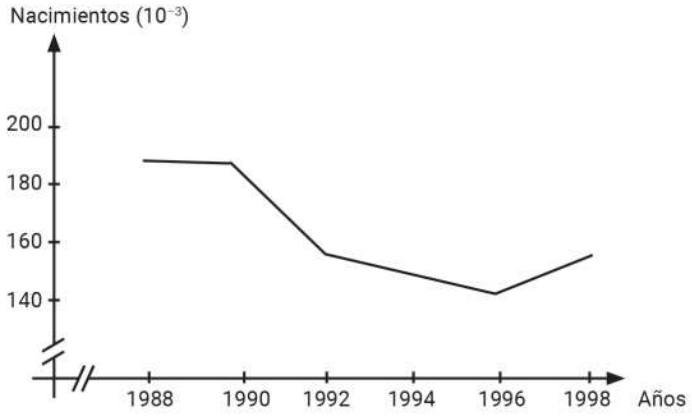


Fig. 3.11. Ejemplo de gráfico aritmético simple.

Histograma

Tiene como base en su construcción una tabla de distribución de frecuencia, o sea, se considera en la práctica una versión gráfica de dicha tabla referida a una variable cuantitativa continua (aunque se generaliza a las discretas) y es muy importante para describir la naturaleza de dicha distribución. Dicha variable cuantitativa emplea una escala por clases o intervalos, la cual ubica sus datos por las marcas de las clases o intervalos en la abscisa, mientras que en la ordenada la información se dispone en frecuencias absolutas o relativas, así como, en medidas de resumen. Cada una de las clases se corresponde con solo una barra vertical. Dichas barras verticales son adyacentes (sin espacio ni huecos entre las mismas) y alcanzan la altura tope que se corresponde con la frecuencia absoluta o relativa correspondiente, ubicada en el eje vertical de la ordenada. Específicamente, el centro de cada barra se corresponde con las marcas de la respectiva clase, el ancho con el tamaño del intervalo o clase y la longitud se corresponde con la frecuencia del valor de la clase.

Una regla sencilla para tomar como referencia a la hora de decidir el número de intervalos o clases a seleccionar a partir del número total de los datos de la variable

cuantitativa en estudio para construir un histograma se presenta en la tabla 3.2 y se representa en la figura 3.12.

Tabla 3.2. Relación entre el número de datos de la variable y el de intervalos a seleccionar en la confección de histogramas

Número de datos	Número de intervalos
<20	(utilizar diagrama de puntos)
20-50	7
>50-75	10
>75-100	12
Más de 100	15

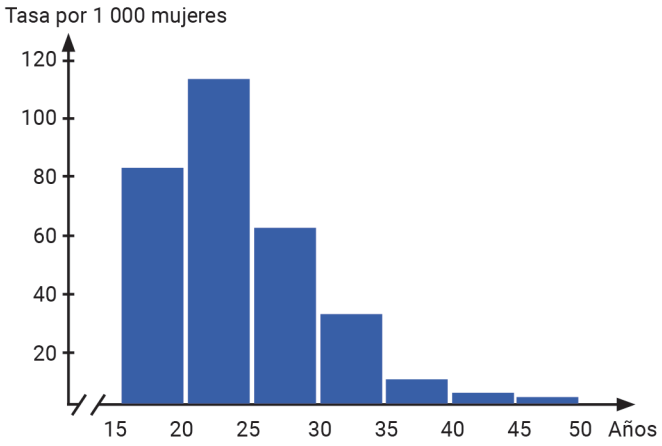


Fig. 3.12. Ejemplo de gráfico tipo histograma.

Polígono de frecuencias

Tiene como base una o varias tablas de distribuciones de frecuencias, esto último con el objetivo de poder compararlas y representar varias clasificaciones (con varias líneas en un mismo gráfico) a diferencia del histograma, que utiliza una sola distribución de frecuencias. En este gráfico se pueden emplear dos o más variables, donde al menos una tiene que ser cuantitativa continua (aunque se generaliza a las discretas igual que el histograma), la cual ubica sus datos por las marcas de las clases o intervalos en la abscisa, mientras que en la ordenada la información se dispone en frecuencias absolutas o relativas, así como, en medidas de resumen.

La otra variable cualitativa o cuantitativa por clases generalmente permite estratificar o formar grupos de diferentes distribuciones de frecuencia para la cuantitativa antes mencionada. Su representación no se hace por rectángulos, sino a través de una línea quebrada para dar idea de continuidad, pudiendo esta comenzar en contacto con la abscisa desde el valor inferior del primer intervalo o clase y terminar en contacto con la misma abscisa en el valor superior correspondiente al último intervalo o clase para la variable cuantitativa empleada. En este gráfico, los puntos se inscriben en la marca de clase, y esto lo diferencia del gráfico aritmético simple, donde el punto se coloca en los valores de los datos. No se puede usar, al igual que el histograma, cuando hay clases abiertas (Fig. 3.13.).

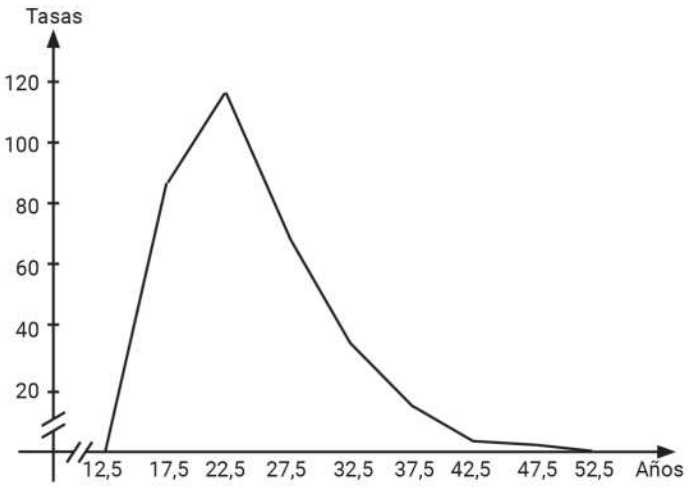


Fig. 3.13. Ejemplo de gráfico tipo polígono de frecuencias.

Gráfico de frecuencias acumuladas u ojivas

Tiene como base una o varias distribuciones de frecuencias tomando en cuenta específicamente las acumuladas o acumulativas, esto último para poder compararlas y poder representar varias de ellas como ojivas (con varias líneas unidas por puntos en un mismo gráfico). En este gráfico se pueden emplear dos o más variables, donde al menos una tiene que ser cuantitativa continua (aunque se generaliza a las discretas igual que el histograma), la cual ubica sus datos por las fronteras de las clases o intervalos en la abscisa, mientras que en la ordenada la información se dispone en frecuencias acumuladas o acumulativas.

La otra variable cualitativa o cuantitativa por clases generalmente permite estratificar o formar grupos de diferentes distribuciones de frecuencia para la cuantitativa antes

mencionada. La ojiva utiliza fronteras de clase en vez de marcas de clase a lo largo de la escala horizontal a diferencia del polígono de frecuencias y el histograma. Su representación comienza con el valor frontera inferior de la primera clase y termina con el valor frontera superior de la última clase de la variable cuantitativa seleccionada de la cual se cuenta con su distribución de frecuencias acumuladas. Las ojivas son útiles para determinar el número de valores (equivalentes con un porcentaje) que se encuentran por debajo de un valor específico de la variable cuantitativa en estudio (Fig. 3.14).

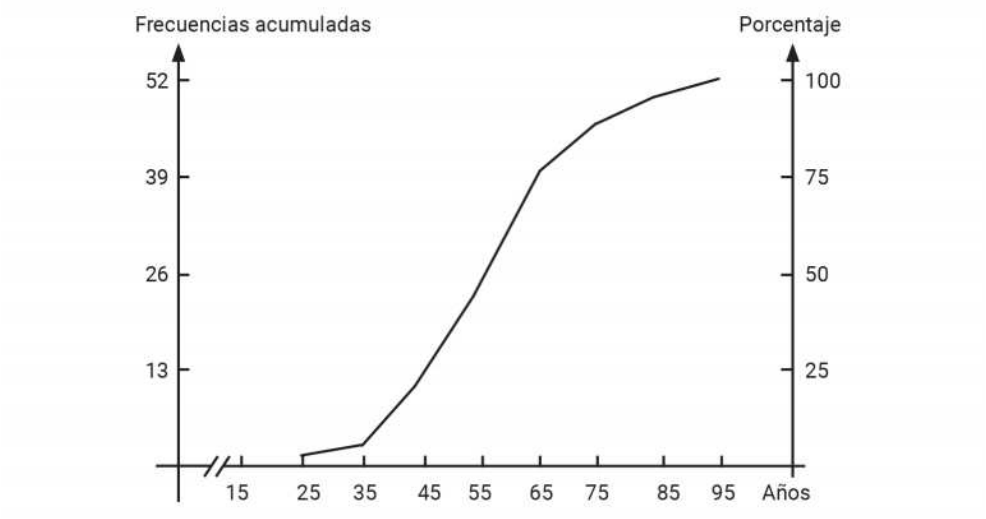


Fig. 3.14. Ejemplo de gráfico de frecuencias acumuladas u ojivas.

Gráfico de tronco y hojas

Utiliza en su conformación como base una distribución de frecuencias para una variable cuantitativa en que cada valor de sus datos se separa en dos partes: el tallo, que es el dígito ubicado en el extremo izquierdo, y la hoja, que se corresponde con el dígito del extremo derecho. Esta última tiene valores ordenados ascendentemente de izquierda a derecha, lo cual guarda similitud con el histograma, solo que con inversión de las variables para cada eje (los valores de la variable cuantitativa en las ordenadas y las frecuencias respectivas en las abscisas), donde "las barras" para este caso serían sustituidas por los propios valores.

Este gráfico aventaja al histograma en que es muy fácil de construir y no se pierden los datos originales. Una gran ventaja de la gráfica de tallo y hojas radica en que permite ver la distribución de los datos y, al mismo tiempo, retener toda la información de la lista original: en caso de ser necesario, se podría reconstruir la lista de valores original. Otra ventaja es que la construcción de una gráfica de tallo y hojas implica

una forma fácil y rápida de ordenar datos (acomodarlos en orden), y algunos procedimientos estadísticos requieren de un ordenamiento (como el cálculo de la mediana o de los percentiles).

Uno de los lineamientos para la construcción de histogramas es que se incluyan entre 5 y 20 clases, lo cual se aplica a la gráfica de tallo y hojas por las mismas razones. Por lo general, se obtienen mejores gráficas de tallo y hojas al redondear primero los valores de los datos originales. Además, este tipo de gráficas puede expandirse para incluir más renglones y condensarse para disminuir el número de renglones (Fig. 3.15).

14	8	9	9	9			
15	0	3	3	4	4	4	4
15	6	7	7	8	8	8	
16	1	2	3	3	4	4	
16	7	8	8	9			
17	0	2					

Fig. 3.15. Ejemplo de gráfico de tronco y hojas.

Gráfico de caja y bigote

Para su conformación se parte también de la utilización de una distribución de frecuencias para los datos de una variable cuantitativa, y es un gráfico que contribuye a dar una clara imagen del comportamiento de dichos datos, trabaja con valores máximo, mínimo (que delimitan los bigotes o líneas que llegan desde el tercer y primer cuartil respectivamente) y cuartiles. Entre el primer y tercer cuartil se construye una caja (rectángulo) y a la altura del valor de la mediana se traza una línea horizontal que llega a los bordes. Las regiones más estrechas entre los cuartiles indican mayor concentración de datos.

Se emplea en estudios bivariados para representar las diferencias de valores en la variable cuantitativa a partir de nominaciones en la cualitativa independiente; es decir, son muy apropiados para mostrar el comportamiento en la posición y distribución de los datos cuando interesa presentarlos estratificados por alguna variable cualitativa. Además, facilitan la comparación entre grupos y permiten una rápida identificación de valores atípicos en su distribución de frecuencias.

Lo que se debe realizar ante la existencia de valores atípicos (extremos) en los datos de una variable cuantitativa es no ignorarlos como si no existieran ni eliminarlos sin intentar averiguar por qué se han producido. Si está claro que la causa es un error, se elimina el valor; en cambio, si no es un error, habrá que valorar la conveniencia de incluirlos en el estudio según sea la razón por la que se han producido y la frecuencia con que se esperan valores similares (Fig. 3.16).

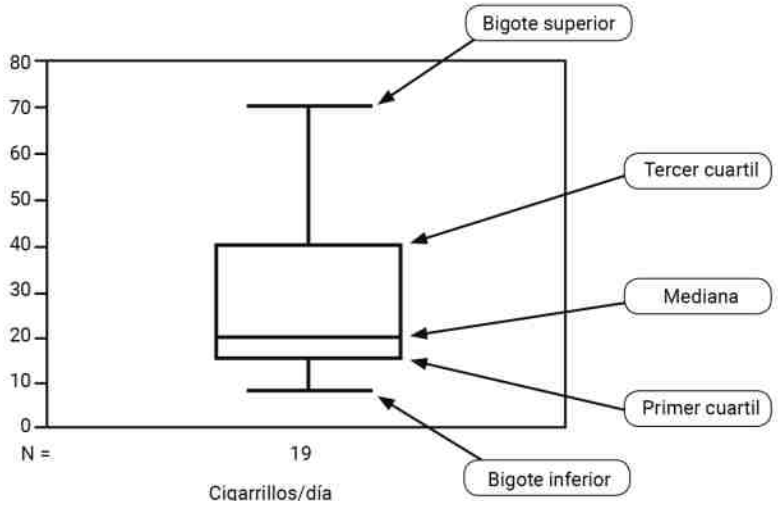


Fig. 3.16. Ejemplo de gráfico de caja y bigote.

Diagrama de dispersión

Constituye una gráfica de datos apareados (x, y) procedentes de dos variables cuantitativas, con un eje x horizontal y un eje y vertical. Los datos se aparean de tal forma que cada valor de un conjunto de datos corresponde a un valor de un segundo conjunto de datos. Se utiliza cuando se quiere estudiar o descubrir si existe relación o asociación entre dos variables cuantitativas.

Se presenta como el resultado de la proyección en el plano de un grupo de puntos en relación a un sistema de ejes coordenados donde los valores de los datos organizados de una de las variables, la que se considera independiente, se ubican en el eje de las abscisas y los de la independiente en el eje de las ordenadas. El patrón de los puntos graficados suele ser útil para determinar si existe alguna relación entre las dos variables (Fig. 3.17).

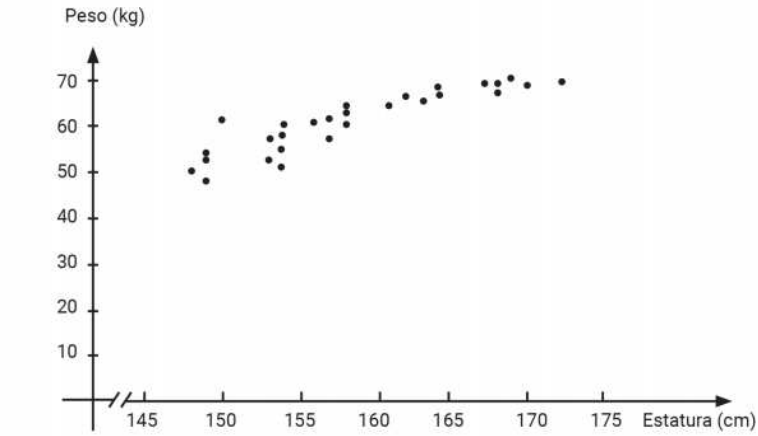


Fig. 3.17. Ejemplo de gráfico de diagrama de dispersión.

Gráfico de series de tiempo

Para la conformación de gráficos de series de tiempo se requiere contar con el registro de los valores de una variable cuantitativa en diferentes momentos, durante un periodo de tiempo determinado. Se emplea en su construcción un sistema de ejes de coordenadas donde en la ordenada o eje vertical se sitúan los valores de la variable cuantitativa de estudio y en el eje horizontal o de las abscisas la variable cuantitativa continua cronológica (tiempo). Los datos de cada serie de tiempo permiten evaluar la evolución y posibles tendencias en el tiempo y en un área determinada, de ciertos fenómenos representados o relacionados con la variable seleccionada y con esto inducen a proyectar la necesidad de haber tomado o no ciertas medidas para poder influir en el devenir de los resultados aportados para el fenómeno de estudio (Fig. 3.18).

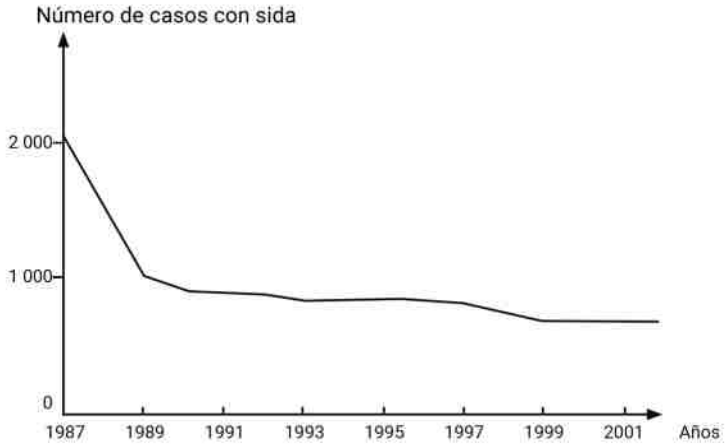


Fig. 3.18. Ejemplo de gráfico de series de tiempo.

Tablas de contingencia, de doble entrada o cruzadas de filas por columnas

Se construyen mediante el empleo de un número igual o mayor de dos filas por dos columnas en que se establece un cruzamiento de la información de las dos variables cualitativas y se exponen los valores de frecuencia correspondientes a la coincidencia de dos categorías, una de cada variable en las respectivas celdas ubicadas para este propósito en el interior de la tabla. Estructuralmente en la primera columna se declara el nombre de la primera variable y en las filas que esta contiene las categorías para la escala empleada respectiva, mientras que en la segunda columna se declara el nombre de la segunda variable de cuya columna se derivan dos o más columnas para representar

las categorías de esa variable según la escala empleada. La información contenida para ambas variables es empleada para el cálculo de las medidas de frecuencia y asociación entre las variables y evaluar la significación estadística de dicha relación (Tabla 3.3).

Tabla 3.3. Ejemplo de tabla de contingencia, de doble entrada o cruzada 2x2 columnas simple

Variable independiente (causa)	**Variable dependiente (efecto)**		
	Tipo 1	Tipo 2	Totales
Categoría 1	a	b	$a + b = m_1$
Categoría 2	c	d	$c + d = m_2$
Totales	$a + c = n_1$	$b + d = n_2$	$n_1 + n_2 = n$ $m_1 + m_{2 =} n$

A continuación se presenta un resumen tabular que relaciona el tipo de escala para variable, el número y tipo de las mismas y la representación gráfica a utilizar (Tabla 3.4).

Tabla 3.4. Relación entre el tipo de escala para el trabajo con variables, el número y tipo de variables y la representación gráfica a utilizar

Tipo de escala para variable	**Número y tipo de variables**	**Tipo de gráfico a emplear**
Cualitativa o cuantitativa por clases	Una variable con categorías o clases en abscisa y datos de frecuencia absoluta o relativa en ordenada.	De barras simples, de pastel, sectores o circular, de Pareto
	– Dos cualitativas en abscisa y datos de frecuencia absoluta o relativa en ordenada – Una cualitativa y otra cuantitativa discreta por clases en abscisa y datos de frecuencia absoluta o relativa en ordenada. – Dos cuantitativas discretas por clases en abscisa y datos de frecuencia absoluta o relativa en ordenada.	De barras múltiples, de barras compuestas o proporcionales
Cualitativa categorizada o cuantitativa con cronológica combinadas	Una cuantitativa continua (tiempo) en la abscisa combinada con una cualitativa (para una categoría seleccionada) o cuantitativa con datos en escala aritmética en la ordenada	Aritmético simple

Tipo de escala para variable	Número y tipo de variables	Tipo de gráfico a emplear
Cualitativa, cuantitativa o ambas combinadas.	Una cuantitativa continua con datos por marca de clase en la abscisa y datos de frecuencia absoluta o relativa en ordenada	Histograma y polígono de frecuencias.
	Una cuantitativa continua con datos de marca de clase combinada con una cualitativa que estratifica distribuciones de frecuencias en la abscisa y datos de frecuencia absoluta o relativa en ordenada	Polígono de frecuencias (más de una distribución)
	Una cuantitativa continua con valores de frontera de clases en abscisa combinada o no con cualitativa que estratifica distribuciones de frecuencias y datos de frecuencia acumulada en la ordenada	Polígono de frecuencias acumuladas u ojiva
Cuantitativa	Una cuantitativa discreta (dígitos en decenas o centenas, que representa el tallo) en ordenada y datos de frecuencia absoluta (dígitos en unidades que representa las hojas) en abscisa	De tallo o tronco y hojas
	Una cuantitativa discreta con todos sus valores originales en ordenada y un dato único correspondiente al total de frecuencias para la variable cuantitativa discreta estudiada en abscisa	De caja y bigote
	Dos cuantitativas con sus valores originales en cada eje (abscisa y ordenada) respectivamente	Diagrama de dispersión
	Dos cuantitativas discretas, una con datos cronológicos (tiempo) en la abscisa y la otra con datos de frecuencia absoluta o relativa en la ordenada	Series de tiempo

A continuación se presenta una propuesta de algoritmo para facilitar la selección de las representaciones gráficas que con mayor frecuencia se emplean en el trabajo con variables durante el procesamiento estadístico descriptivo de la información (Fig. 3.19).

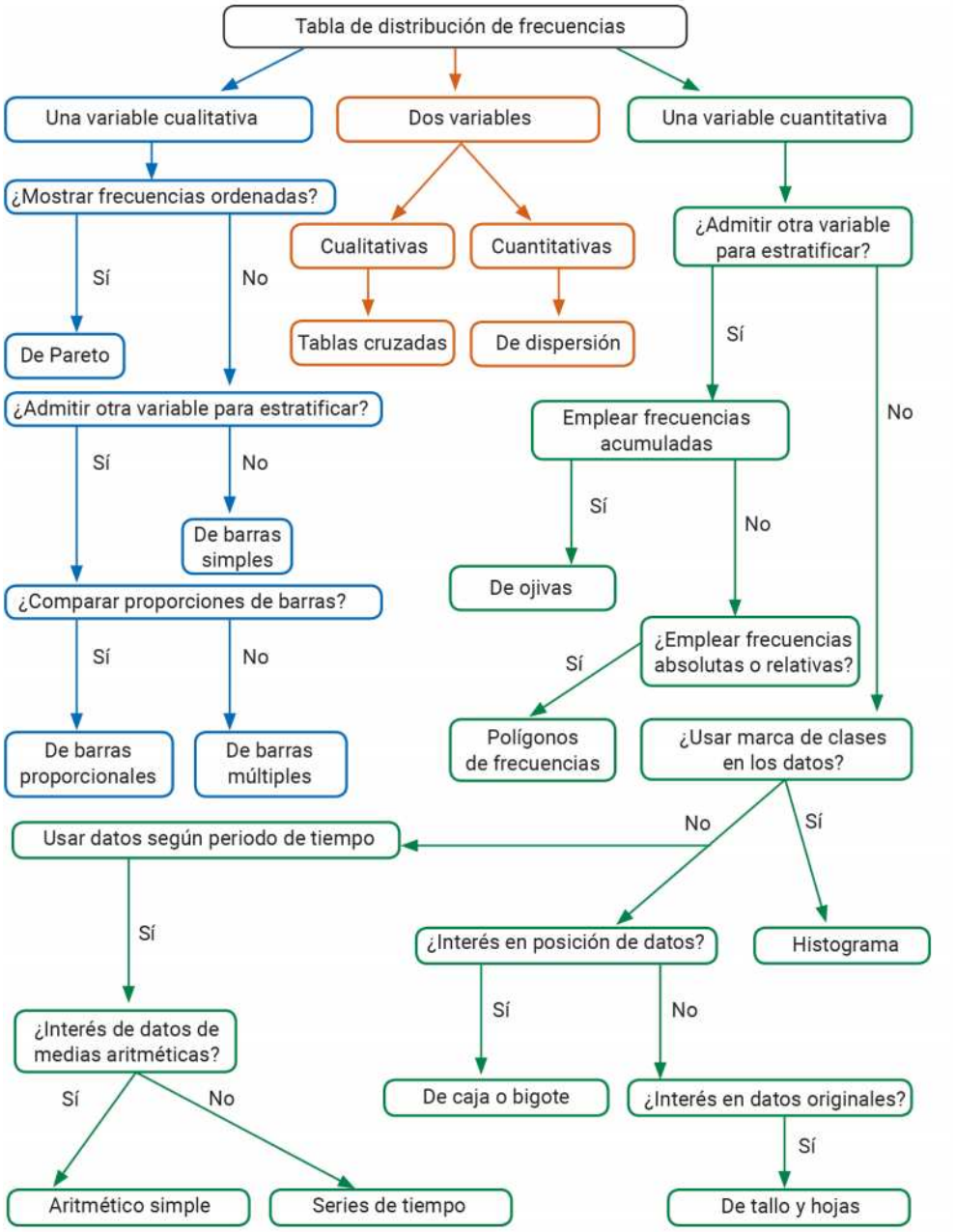


Fig. 3.19. Algoritmo propuesto para la selección de la representación gráfica en el trabajo con variables de estadística descriptiva.

Tipos de tablas de contingencia (cruzadas) para determinar la asociación de 2 variables cualitativas

Tablas 2x2 simples

En los estudios epidemiológicos observacionales que toman en cuenta el enfoque de riesgo se analiza la relación entre la exposición a cierto factor de riesgo y la aparición de un determinado efecto o consecuencia. Las tablas de contingencia representan la manera graficada de la cual se puede auxiliar el investigador para la exposición de los datos de dos variables: una independiente y otra dependiente que se desea evaluar su asociación o no. A continuación se presentan algunos ejemplos.

Tablas 2x2 simples (estudio transversal)

En la tabla 3.5 se representa una tabla de contingencia 2x2 simple (el número 2 expresa el número de filas y columnas) de un estudio transversal empleando la prevalencia, para determinar la probable asociación a modo de ejemplo entre la variable independiente: presencia de exposición (factor de riesgo positivo o no) y la dependiente: presencia de enfermedad (enfermo o sano).

Tabla 3.5. De contingencia, cruzada, tetracórica o de doble entrada para el análisis de prevalencia en enfermos y sanos para exposición

Variable independiente (exposición)	**Variable dependiente (enfermedad)**		
	Enfermos	**Sanos**	
Factor de riesgo positivo	a	b	m_1
Factor de riesgo negativo	c	d	m_2 o m_0
	n_1	n_2 o n_0	n

Donde:
Celda a: sujetos enfermos con antecedente de exposición.
Celda b: sujetos sanos con antecedente de exposición.
Celda c: sujetos enfermos sin antecedente de exposición.
Celda d: sujetos sanos sin antecedente de exposición.
Celda n_1: total de enfermos.
Celda n_2 o n_0: total de sanos.
Celda m_1: total de sujetos expuestos.
Celda m_2 o m_0: total de sujetos no expuestos.
Celda n: total de sujetos.

Tablas 2x2 simples (estudio caso-control no apareado)

En la tabla 3.6 se representa una tabla de contingencia 2x2 simple de un estudio caso-control no pareado para determinar la probable asociación a modo de ejemplo entre la variable independiente: presencia de exposición (con factor de riesgo o no) y la dependiente: presencia de enfermedad (caso o control).

Tabla 3.6. De contingencia, cruzada, tetracórica o de doble entrada para el análisis de prevalencia en caso y control para exposición

Variable independiente (exposición)	Variable dependiente (enfermedad)		Total
	Presente (caso)	Ausente (control)	
Presente	a	b	F1 = m_1 = (a + b)
Ausente	c	d	F2 = m_0 = (c + d)
Total	C1 = n_1 = (a + c)	C2 = n_0 = (b + d)	n = n_1 + n_0 n = (a + c) + (b + d)

Nota: se comparan incidencias acumuladas según exposición o no. Se realizan análisis verticales de los datos preferentemente.
Donde:
Celda a: sujetos enfermos (casos) con antecedente de exposición.
Celda b: sujetos sanos (controles) con antecedente de exposición.
Celda c: sujetos enfermos (casos) sin antecedente de exposición.
Celda d: sujetos sanos (controles) sin antecedente de exposición.
Celda C1: total de casos.
Celda C2: total de controles.
Celda F1: total de sujetos expuestos.
Celda F2: total de sujetos no expuestos.
Celda n: total de sujetos.

Tablas 2x2 simples (estudio de casos y controles pareados)

En la tabla 3.7 se representa una tabla de contingencia 2x2 simple de un estudio caso-control pareado para determinar la probable asociación a modo de ejemplo entre la variable independiente: presencia de exposición (expuesto o no) y la dependiente: presencia de enfermedad (caso o control; relación 1 caso: 1 control).

Tabla 3.7. De contingencia para el análisis de asociación entre variables cualitativas: presencia de exposición y presencia de enfermedad, con datos apareados en un estudio de caso-control (relación 1:1, caso-control)

		Controles		
		Expuestos	No expuestos	
Casos	Expuestos	(n_{11})(a)	(n_{10})(b)	Número de pares n_{11} + n_{10} (a+b)

Tabla 3.7. Continuación...

	Controles		
	Expuestos	**No expuestos**	
No expuestos	$(n_{01})(c)$	$(n_{00})(d)$	Número de pares $n_{01} + n_{00}$ (c+d)
	Número de pares $n_{11} + n_{01}$ (a+c)	Número de pares $n_{10} + n_{00}$ (b+d)	Total de pares $n = (n_{11} + n_{10}) + (n_{01} + n_{00})$ n = (c+d)+ (a+b

Donde:
Celda n_{11} (a): caso y control expuestos.
Celda n_{10} (b): controles no expuestos, casos sí.
Celda n_{01} (c): caso no expuesto, control sí.
Celda n_{00} (d): caso no expuesto, control no expuesto.
Celda $n_{11} + n_{01}$ (a+c): total de pares de controles expuestos.
Celda $n_{10} + n_{00}$ (b+d): total de pares de controles no expuestos.
Celda $n_{11} + n_{10}$ (a+b): total de pares de casos expuestos.
Celda $n_{01} + n_{00}$ (c+d): total de pares de casos no expuestos.
Celda n: $(n_{11} + n_{10}) + (n_{01} + n_{00})$ o (c + d) + (a + b): total de pares de la muestra.

Tablas 2x2 simples (estudio de cohorte)

En la tabla 3.8 se representa una tabla de contingencia 2x2 simple de un estudio de cohorte empleando la incidencia acumulada, para determinar la probable asociación a modo de ejemplo entre la variable independiente: presencia de exposición (expuesto o no), y la dependiente: presencia de enfermedad (enfermo o sano).

Tabla 3.8. De contingencia, cruzada, tetracórica o de doble entrada para el análisis de incidencia acumulada en estudios de cohorte

	Enfermos	**Sanos**	
Expuestos	A	B	A + B
No expuestos	C	D	C + D
Total	A + C	B + D	A + B + C + D

Nota: se comparan incidencias acumuladas según exposición o no. Se realizan análisis horizontales de los datos preferentemente.
Donde:
Celda A: expuestos que desarrollan la enfermedad (evento presente).
Celda B: expuestos que no desarrollan la enfermedad (evento ausente).
Celda C: no expuestos que desarrollan la enfermedad (evento positivo).
Celda D: no expuestos que no desarrollan la enfermedad (evento negativo).
Celda A + C: total de nuevos enfermos.
Celda B + D: total de los que permanecen sanos.
Celda A + B: total de sujetos expuestos.
Celda C + D: total de sujetos no expuestos.
Celda: A + B + C + D: total de sujetos de la cohorte.

En la tabla 3.9 se representa una tabla de contingencia 2x2 simple de un estudio de cohorte empleando las tasas de incidencia, para determinar la probable asociación a modo de ejemplo entre la variable independiente: presencia de exposición (expuesto o no) y la dependiente: presencia de enfermedad o efecto, donde se agrega la columna referente al total de personas por tiempo de seguimiento.

Tabla 3.9. De contingencia, cruzada, tetracórica o de doble entrada para el análisis de densidad de incidencia por tasas de incidencia en estudios de cohorte

Exposición a	Riesgo (enfermedad-defunción)	Total de personas por tiempo de seguimiento
Expuestos	A	PTSE o (T_1)
No expuestos	C	PTSNE (T_0)
Total	A + C	PTSE + PTSNE ($T_1 + T_0$)

Donde:
Celda A: nuevos enfermos o fallecidos con exposición previa.
Celda B: nuevos enfermos o fallecidos sin previa exposición.
Celda PTSE (T_1): número de personas expuestas para un determinado tiempo de seguimiento.
Celda PTSNE (T_0): número de personas no expuestas para un determinado tiempo de seguimiento.
Celda A + C: total de nuevos enfermos para la cohorte.
Celda PTSE + PTSNE ($T_1 + T_0$): total de personas de la cohorte (expuestos + no expuestos) para determinado tiempo de seguimiento.

En la tabla 3.10 se representa una tabla de contingencia 2x2 simple de un estudio de cohorte empleando la densidad de incidencia, para determinar la probable asociación a modo de ejemplo entre la variable independiente: presencia de exposición (expuesto o no) y la dependiente: presencia de nuevos enfermos donde se agrega la columna referente al total de tiempo por persona de seguimiento

Tabla 3.10. De contingencia para densidad de incidencia (por total tiempo por persona de seguimiento) en estudios de cohorte

Exposición a	Nuevos enfermos	Total de tiempo por persona de seguimiento
Exposición +	A	T_1
Exposición –	C	T_2 o T_0
Total	A + C	$T_1 + T_2$ o $T_1 + T_0$

Donde:
Celda A: nuevos enfermos previa exposición.
Celda B: nuevos enfermos sin previa exposición.
Celda T_1 (TSPE): tiempo total por persona de seguimiento en expuestos.
Celda T_2 o T0 (TSPNE): tiempo total por persona de seguimiento en no expuestos.
Celda A + C: total de nuevos enfermos (expuestos y no expuestos) de la cohorte.
Celda $T_1 + T_2$ o $T_1 + T_0$: total del tiempo por persona de seguimiento de la cohorte.

En la tabla 3.11 se representan los elementos a tomar en cuenta de un estudio de cohorte empleando la razón de mortalidad estandarizada, como parte del enfoque de riesgo al analizar la probable asociación entre un factor de riesgo como la exposición a algún agente o factor de riesgo y la presencia de un efecto como es la muerte.

Tabla 3.11. Para el estudio de la razón de mortalidad estandarizada para estudios de cohorte

Grupos de edad	Muertes observadas(A)	Años-persona de seguimiento población referencia (B)	Tasa mortalidad población referencia (valor sin 10^n) (C)	Muertes esperadas (C · B)/1 000
Totales				

Leyenda

(A): total de muertes observadas en la cohorte.

(B): total de años-persona de seguimiento en la población de referencia.

(C): tasa mortalidad en la población de referencia (valor sin 10^n).

(C · B)/1 000: muertes esperadas en la población de referencia.

Tablas 2x2 estratificadas

En el caso que se desee utilizar una tercera variable que permita construir estratos diferentes al determinar la relación entre dos variables con probable asociación como los ejemplos anteriores, se pueden emplear las tablas 2x2 en este caso estratificadas. Los datos que se agregan para este análisis a diferencia del anterior para las tablas de contingencia simple serán:

k: número de estratos (k > 1);

Tipo de tabla: kx4: que depende del tipo de estudio.

Estudios transversales

En la tabla 3.12 se representa una tabla de contingencia 2x2 estratificada de un estudio transversal, para determinar la probable asociación a modo de ejemplo entre la variable independiente: presencia de exposición (expuesto o no) y la dependiente: presencia de enfermedad (enfermo o sano) en los diferentes estratos (k > b_1) determinados por una tercera variable.

Tabla 3.12. De contingencia 2x2 estratificadas en estudios transversales

Estrato	Enfermos expuestos	Sanos expuestos	Enfermos no expuestos	Sanos no expuestos
1	a_1	b_1	c_1	d_1
2	a_2	b_2	c_2	d_2
...	...	...	...	...
K	a_k	b_k	c_k	d_k

Estudios caso-control

En la tabla 3.13 se representa una tabla de contingencia 2x2 estratificada de un estudio caso-control, para determinar la probable asociación a modo de ejemplo entre la variable independiente: presencia de exposición (expuesto o no) y la dependiente: presencia de enfermedad (casos o controles) en los diferentes estratos (k > 1) determinados por una tercera variable.

Tabla 3.13. De contingencia 2x2 estratificadas en estudios caso-control

Estrato	Casos expuestos	Controles expuestos	Casos no expuestos	Controles no expuestos
1	a_1	b_1	c_1	d_1
2	a_2	b_2	c_2	d_2
...	...	...	...	...
K	a_k	b_k	c_k	d_k

Estudios de cohorte con datos de tasas de incidencia

En la tabla 3.14 se representa una tabla de contingencia 2x2 estratificada de un estudio de cohorte empleando las tasas de incidencias, para determinar el comportamiento de los casos expuestos y no expuestos según el total de personas por tiempo de seguimiento para los diferentes estratos (k > 1) determinados por una tercera variable.

Tabla 3.14. De contingencia 2x2 estratificadas en estudios de cohorte por tasas de incidencia

Estrato	Casos expuestos	Personas-tiempo expuestas	Casos no expuestos	Personas-tiempo no expuestas
1	a_1	$T^1{}_1$	b_1	$T^0{}_1$
2	a_2	$T^1{}_2$	b_2	$T^0{}_2$
...	...	...	...	...
K	a_k	$T^1{}_k$	b_k	$T^0{}_k$

Estudios de cohorte con datos de incidencia acumulada

En la tabla 3.15 se representa una tabla de contingencia 2x2 estratificada de un estudio de cohorte empleando la incidencia acumulada, para determinar la probable asociación a modo de ejemplo entre la variable independiente: presencia de exposición (expuestos o no) y la dependiente: presencia de enfermedad (enfermos o sanos) en los diferentes estratos (k > 1) determinados por una tercera variable.

Tabla 3.15. De contingencia 2x2 estratificadas en estudios de cohorte por incidencia acumulada

Estrato	Enfermos expuestos	Sanos expuestos	Enfermos no expuestos	Sanos no expuestos
1	a_1	b_1	c_1	d_1
2	a_2	b_2	c_2	d_2
...	...	...	...	...
K	a_k	b_k	c_k	d_k

Tablas de contingencia 2xN simples

En el caso que se desee analizar mediante el enfoque de riesgo, la relación entre la variable independiente: presencia de exposición a un determinado factor de riesgo, a partir de sus diferentes niveles u órdenes de intensidad en su accionar y la variable dependiente: presencia de enfermedad u otro efecto se pueden emplear las tablas 2xN simples. Los datos que se agregan para este análisis son:

N: es el número de niveles de exposición al factor de riesgo (N > 2).

Estudios transversales

En la tabla 3.16 se representa una tabla de contingencia 2xN simple de un estudio transversal, para determinar la probable asociación a modo de ejemplo entre la variable independiente: nivel de exposición y la dependiente: presencia de enfermedad (enfermos o sanos).

Tabla 3.16. De contingencia 2xN simples en estudios transversales

Puntuación	Nivel de exposición			
	x_1	x_2	...	x_N
Enfermos	a_1	a_2	...	a_N
Sanos	c_1	c_2	...	c_N

Estudios caso-control

En la tabla 3.17 se representa una tabla de contingencia 2xN simple de un estudio caso-control, para determinar la probable asociación a modo de ejemplo entre la variable

independiente: nivel de exposición y la dependiente: presencia de enfermedad (casos o controles).

Tabla 3.17. De contingencia 2xN simples en estudios caso-control

Puntuación	Nivel de exposición			
	x_1	x_2	...	x_N
Casos	a_1	a_2	...	a_N
Controles	c_1	c_2	...	c_N

Estudios de cohorte con datos de tasas de incidencia

En la tabla 3.18 se representa una tabla de contingencia 2xN simple de un estudio de cohorte empleando las tasas de incidencias, para determinar la probable asociación a modo de ejemplo entre la variable independiente: nivel de exposición y la dependiente: presencia de enfermedad (casos) por el número de personas según el tiempo de su seguimiento.

Tabla 3.18. De contingencia 2xN simples, en estudios de cohorte por tasas de incidencia

Puntuación	Nivel de exposición			
	x_1	x_2	...	x_N
Casos	a_1	a_2	...	a_N
Personas-tiempo	T_1	T_2	...	T_N

Estudios de cohortes con datos de incidencia acumulada

En la tabla 3.19 se representa una tabla de contingencia 2xN simple de un estudio de cohorte empleando la incidencia acumulada, para determinar la probable asociación a modo de ejemplo entre la variable independiente: nivel de exposición y la dependiente: presencia de enfermedad (enfermos o sanos).

Tabla 3.19. De contingencia 2xN simples en estudios de cohorte por incidencia acumulada

Puntuación	Nivel de exposición			
	x_1	x_2	...	x_N
Enfermos	a_1	a_2	...	a_N
Sanos	c_1	c_2	...	c_N

Tablas de contingencia 2xN estratificadas

En el caso que se desee analizar mediante el enfoque de riesgo, la relación entre la variable independiente: presencia de exposición a un determinado factor de riesgo,

a partir de sus diferentes niveles u órdenes de intensidad en su accionar y la variable dependiente: presencia de enfermedad u otro efecto, en los diferentes estratos (k > 1), determinados por una tercera variable, se pueden emplear las tablas 2xN estratificadas. Los datos que se agregan para este análisis son:

K: número de estratos (K > 1);

N: número de niveles de exposición al factor de riesgo (N > 2).

Estudios transversales

En la tabla 3.20 se representa una tabla de contingencia 2xN estratificada de un estudio transversal, para determinar la probable asociación a modo de ejemplo entre la variable independiente: nivel de exposición y la dependiente: presencia de enfermedad (enfermos o sanos) en los diferentes estratos (k > 1) determinados por una tercera variable.

Tabla 3.20. De contingencia 2xN estratificadas en estudios transversales

Estrato	Puntuación	Nivel de exposición			
		x_1	x_2	...	x_N
1	Enfermos	$a^1{}_1$	$a^1{}_2$	...	$a^1{}_N$
	Sanos	$c^1{}_1$	$c^1{}_2$	...	$c^1{}_N$
...		...	...	...	...
K	Enfermos	$a^k{}_1$	$a^k{}_2$	...	$a^k{}_N$
	Sanos	$c^k{}_1$	$c^k{}_2$	...	$c^k{}_N$

Estudios caso-control

En la tabla 3.21 se representa una tabla de contingencia 2xN estratificada de un estudio de caso-control, para determinar la probable asociación a modo de ejemplo entre la variable independiente: nivel de exposición y la dependiente: presencia de enfermedad (casos o controles) en los diferentes estratos (k > 1) determinados por una tercera variable.

Tabla 3.21. De contingencia 2xN estratificadas en estudios caso-control

Estrato	Puntuación	Nivel de exposición			
		x_1	x_2	...	x_N
1	Casos	$a^1{}_1$	$a^1{}_2$	...	$a^1{}_N$
	Controles	$c^1{}_1$	$c^1{}_2$	...	$c^1{}_N$
...		...	...	...	...
K	Casos	$a^k{}_1$	$a^k{}_2$	...	$a^k{}_N$
	Controles	$c^k{}_1$	$c^k{}_2$	...	$c^k{}_N$

Estudios de cohorte con datos de tasas de incidencia

En la tabla 3.22 se representa una tabla de contingencia 2xN simple de un estudio de cohorte empleando las tasas de incidencias, para determinar la probable asociación a modo de ejemplo entre la variable independiente: nivel de exposición y la dependiente: presencia de enfermedad (casos) por el número de personas según el tiempo de su seguimiento en los diferentes estratos (k > 1) determinados por una tercera variable.

Tabla 3.22. De contingencia 2xN estratificadas en estudios de cohorte por tasas de incidencia

Estrato	Puntuación	Nivel de exposición			
		x_1	x_2	...	x_N
1	Casos	$a^1{}_1$	$a^1{}_2$	...	$a^1{}_N$
	Personas-tiempo	$c^1{}_1$	$c^1{}_2$	...	$c^1{}_N$
...		...	...	...	...
K	Casos	$a^k{}_1$	$a^k{}_2$	...	$a^k{}_N$
	Personas-tiempo	$c^k{}_1$	$c^k{}_2$	...	$c^k{}_N$

Estudios de cohorte con datos de incidencia acumulada

En la tabla 3.23 se representa una tabla de contingencia 2xN estratificada de un estudio de cohorte empleando la incidencia acumulada, para determinar la probable asociación a modo de ejemplo entre la variable independiente: nivel de exposición y la dependiente: presencia de enfermedad (enfermos o sanos) en los diferentes estratos (k > 1) determinados por una tercera variable.

Tabla 3.23. De contingencia 2xN estratificadas en estudios de cohortes por incidencia acumulada

Estrato	Puntuación	Nivel de exposición			
		x_1	x_2	...	x_N
1	Enfermos	$a^1{}_1$	$a^1{}_2$	...	$a^1{}_N$
	Sanos	$c^1{}_1$	$c^1{}_2$	...	$c^1{}_N$
...		...	...	...	...
K	Enfermos	$a^k{}_1$	$a^k{}_2$	...	$a^k{}_N$
	Sanos	$c^k{}_1$	$c^k{}_2$	...	$c^k{}_N$

Tablas MxN

Si se desea relacionar dos variables cualitativas nominales u ordinales que no estén contempladas en el enfoque de riesgo, con la particularidad de tener más de dos posibles

categorías cada una de ellas se puede elegir con ese propósito las tablas de contingencia denominadas MxN. Los datos que emplea para su análisis son:

M: número de filas de la tabla ($M \geq 2$);

N: número de columnas de la tabla ($N \geq 2$).

Este tipo de tabla de contingencia puede tener gran aplicación en la epidemiología clínica (Tabla 3.24). De ella puede derivar, a partir del análisis de sus datos, la descripción de medidas de frecuencia, de asociación y de impacto para dichas variables, el cálculo de estimadores muestrales para variables categóricas, así como puede brindar información necesaria en algunos casos para efectuar pruebas de hipótesis o de significación en la evaluación de los resultados que se obtengan.

Tabla 3.24. De contingencia 2x3 como ejemplo (2 columnas por 3 filas)

Variable independiente (más de dos categorías)	Variable dependiente (enfermedad)		Totales por filas
	Enfermos	Sanos	
Categoría 1 u orden 1	a	b	a + b
Categoría 2 u orden 2	c	d	c + d
Categoría 3 u orden 3	e	f	e + f
Totales por columnas	a + c + e	b + d + f	n

Eficacia de un procedimiento diagnóstico (prueba diagnóstica)

Se llamará prueba diagnóstica (Pd) a cualquier proceso, más o menos complejo, que pretenda determinar en un paciente la presencia de cierta condición, supuestamente patológica, no susceptible de ser observada directamente (con alguno de los cinco sentidos elementales).

Siempre que se vaya a aplicar una prueba diagnóstica es conveniente que sea sencilla de realizar, aceptada por los pacientes o la población general, que tenga los mínimos efectos adversos, que económicamente sea soportable y que tenga la eficacia requerida.

Las condiciones requeridas para una prueba diagnóstica son:

- Validez: es el grado en que una prueba diagnóstica mide lo que se supone que debe medir. Viene determinada por la sensibilidad y la especificidad. Responde a: ¿Con qué probabilidad el resultado de una prueba diagnóstica es confirmado por procedimientos diagnósticos más complejos y rigurosos?
- Seguridad: expresa la probabilidad de que un resultado positivo o negativo de una prueba indique la presencia o no de la enfermedad o un evento. Está determinada

por el valor predictivo de un resultado positivo o negativo. Responde a: ¿Con qué seguridad una prueba predecirá la presencia o ausencia de enfermedad?
- Reproductividad: es la capacidad de la prueba para ofrecer los mismos resultados cuando se repite su aplicación en circunstancias similares. La variabilidad biológica del hecho observado, la introducida por el propio observador y la derivada de la propia prueba, determinan su reproductividad.

Los aspectos a tomar en cuenta en la evaluación de una prueba diagnóstica resultan mejor comprendidos si se comienza con la ayuda de la representación tabulada de los factores que intervienen en ello al emplear dos variables dicotómicas: una que representa el resultado de la prueba diagnóstica a evaluar y la otra del resultado real en la prueba de referencia o prueba de oro (*gold test*) (Tabla 3.25).

Tabla 3.25. Tabla de contingencia 2x2 simple para el análisis de la eficacia de una prueba diagnóstica

Prueba diagnóstica específica a evaluar	**Prueba de oro o de referencia, con criterio de la verdad del resultado**		**Totales**
	Con resultado patológico (enfermos)	**Con resultado no patológico (sanos)**	
Resultado patológico (+)	a	b	a + b
Resultado no patológico (-)	c	d	c + d
Totales	a + c	b + d	a + b + c + d

Medidas o índices a estimar que permiten evaluar la eficacia de una prueba diagnóstica

Sensibilidad

La sensibilidad (S) mide la certeza de la prueba para detectar la enfermedad. Es la proporción de clasificados como positivos por la prueba diagnóstica del total de enfermos, es decir, el cociente de los verdaderos positivos (enfermos con prueba positiva) entre la suma de verdaderos positivos y falsos negativos para la prueba diagnóstica. Por tablas de contingencia 2x2 simple se expresa como:

$$S = \frac{\text{Verdaderos positivos}}{\text{Total de enfermos}} = \frac{VP}{VP + FN} = \frac{a}{a + c}$$

Especificidad

La especificidad (E) mide la capacidad de la prueba diagnóstica para detectar la ausencia de patología. Es la proporción de sanos clasificados como negativos para la

prueba diagnóstica del total de sanos, es decir, el cociente de los verdaderos negativos (sanos con prueba negativa) entre la suma de los verdaderos negativos y falsos positivos para la prueba diagnóstica). Por tablas de contingencia 2x2 simple se expresa como:

$$E = \frac{\text{Verdaderos negativos}}{\text{Total de no enfermos}} = \frac{VN}{VN + FP} = \frac{d}{b + d}$$

Valor predictivo positivo

El valor predictivo positivo (VPP) es la proporción de enfermos con la prueba diagnóstica positiva entre el total de pacientes con la prueba positiva. Por tablas de contingencia 2x2 simple se expresa como:

VPP = (a/a + b) o

VP(+) = P(Enf/T+)

Valor predictivo negativo

El valor predictivo negativo (VPN) es la proporción de sanos con prueba diagnóstica negativa entre el total de pacientes con prueba negativa. Por tablas de contingencia 2x2 simple se expresa como:

VPN = (d/c + b) o

VP(–) = P(No Enf/T–)

Razón de verosimilitudes, razón de probabilidad o cociente de probabilidades

La razón de verosimilitudes, razón de probabilidad o cociente de probabilidades mide cuánto más probable es un resultado concreto (positivo o negativo) según la presencia o ausencia de enfermedad.

Razón de verosimilitudes positiva o cociente de probabilidades positivo

La razón de verosimilitudes positiva o cociente de probabilidades positivo (RVP) se calcula dividiendo la probabilidad de un resultado positivo en los pacientes enfermos entre la probabilidad de un resultado positivo entre los sanos. Es el cociente de la fracción de verdaderos positivos (sensibilidad) y la fracción de falsos positivos (1-especificidad). Por tablas de contingencia 2x2 simple se expresa como:

$$RVP = \frac{a / a + c}{1 - (d / b + d)}$$ o

$RV+ = \frac{P(T+|Enf)}{P(T+|NoEnf)}$, expresa cuantas veces es más probable que la prueba sea positiva en los enfermos que en los no enfermos.

Razón de verosimilitudes negativa o cociente de probabilidades negativo

La razón de verosimilitudes negativa o cociente de probabilidades negativo (RVN) se calcula dividiendo la probabilidad de un resultado negativo en presencia de enfermedad entre la probabilidad de un resultado negativo en ausencia de la misma. Es el cociente de la fracción de falsos negativos (1 – sensibilidad) y la fracción de verdaderos negativos (especificidad). Por tablas de contingencia 2x2 simple se expresa como:

$$RVN = \frac{1-(a/a+c)}{(d/b+d)} \quad \text{o}$$

$RV- = \frac{P(T-|Enf)}{P(T-|NoEnf)}$, expresa cuantas veces es más probable que la prueba sea negativa en los enfermos que en los no enfermos.

Para un valor de RV– cercano a 0 y un RV+ alta se iría a favor de una mayor eficacia en una prueba diagnóstica.

En el contexto de las investigaciones clínicas al poder existir diferencias en la prevalencia de la enfermedad entre la población de la cual se extrajo la muestra y la población en general, se debe aplicar el teorema de Bayes, por medio del que, empleando diferentes expresiones matemáticas, se puede estimar la probabilidad *a posteriori* de enfermar teniendo una prueba positiva o la de no enfermar teniendo una prueba negativa, que se corresponden con los valores predictivos positivos y negativos, respectivamente. Estas formulaciones matemáticas emplean datos de sensibilidad, especificidad y la prevalencia de la enfermedad.

Probabilidad *a posteriori* en relación con el valor predictivo

Probabilidad *a posteriori* para el valor predictivo positivo

De tener cierta enfermedad o evento ante una prueba positiva, es para el valor predictivo positivo de la prueba diagnóstica y se denota por:

$$P(Enf|T+) = \frac{P(T+|Enf)\cdot P(Enf)}{P(T+|Enf)\cdot P(Enf) + P(T+|NoEnf)\cdot P(NoEnf)}$$

donde:

$$VP(+) = \frac{S\cdot P(Enf)}{S\cdot P(Enf) + (1-E)\cdot[1-P(Enf)]}$$

$$P(Enf) = \frac{a+c}{N}$$

$N = a + b + c + d$

Probabilidad *a posteriori* para valor predictivo negativo

De no tener cierta enfermedad o evento ante una prueba positiva, es para el valor predictivo negativo de la prueba diagnóstica y se denota por:

$$P(NoEnf|T-) = \frac{P(T-|NoEnf)\cdot P(NoEnf)}{P(T-|NoEnf)\cdot P(NoEnf) + P(T-|Enf)\cdot P(Enf)}$$

donde:

$$VP(-) = \frac{E\cdot[1-P(Enf)]}{E\cdot[1-P(Enf)] + (1-S)\cdot P(Enf)};$$

P(NoEnf) = 1 – P(Enf).

Nota: Con estas fórmulas pueden estimarse los valores predictivos de una prueba diagnóstica en cualquier contexto poblacional (del ámbito epidemiológico tradicional) o específico (de la clínica).

Índice de validez

El índice de validez (IV) es la proporción correcta de aciertos o de individuos clasificados correctamente. Su formulación matemática sería:

IV = (a + d)/N

donde:

a = $Sn_{1;}$

n_1 = a +c;

d= En_2;

n_2 = b +d;

P = n_1/N (prevalencia de la enfermedad);

P = Q = (1 – P) = n_2/N

IV = (Sn_1 + En_2)/N =S(n_1/N + E(n_2/N) = SP + E(1 – P) = SP + E – EP = P(S – E) + E

Índice de Youden

El índice de Youden (IJ) refleja la diferencia entre la tasa de verdaderos positivos y la de falsos positivos y no es afectada por los valores de prevalencia; a mayor valor, mejor es la prueba que se está evaluando:

IJ = S + E – 1 = S – (1 – E)

A modo de resumen, partiendo de la nomenclatura empleada en la tabla 3.23 los indicadores a tener en cuenta en la validación de una prueba diagnóstica serían:

– En el contexto epidemiológico tradicional: se asume coincidiendo la tasa de prevalencia en las poblaciones: general y la estimada para los datos muestrales:
  - Sensibilidad: S = a/a + c.
  - Especificidad: E = d/b + d.
  - Falsos positivos: FP = b/b + d.
  - Falsos negativos: FN = c/a + c
  - Valor predictivo positivo: VPP = a/a + b.
  - Valor predictivo negativo VPN = d/c + d.

- Prevalencia: P = a + c /a + b + c + d
- Certeza diagnóstica: CD = a + d/a + b + c + d
- Razón de verosimilitud positiva: RVP = (a/a + c)/1 − (d/b + d)
- Razón de verosimilitud negativa: RVN = 1 − (a/a + c)/(d/b + d).

– En el contexto epidemiológico de las investigaciones clínicas: se asume no coincidente la tasa de prevalencia en las poblaciones: general y la estimada para los datos muestrales:
  - Probabilidad posprueba o a posteriori para razón de verosimilitud positiva:

$$p(E+/P+) = \frac{p(P+/E+)p(E+)}{p(P+/E+)p(E+) + p(P+/E-)p(E-)}$$

  - Probabilidad posprueba o a posteriori para razón de verosimilitud negativa:

$$p(E-/P-) = \frac{p(P-/E-)p(E-)}{p(P-/E-)p(E-) + p(P-/E+)p(E+)}$$

Para recordar:

– Los valores de sensibilidad y especificidad, a pesar de definir completamente la validez de la prueba diagnóstica, presentan la desventaja de que no proporcionan información relevante a la hora de tomar una decisión clínica ante un determinado resultado de la prueba; sin embargo, tienen la ventaja adicional de que son propiedades intrínsecas a la prueba diagnóstica y definen su validez independientemente de cuál sea la prevalencia de la enfermedad en la población a la que se aplica.
– El concepto de valores predictivos, a pesar de ser de enorme utilidad para tomar decisiones clínicas y transmitir a los pacientes información sobre su diagnóstico, presenta la limitación de que dependen en gran medida de lo frecuente que sea la enfermedad a diagnosticar en la población objeto de estudio. Cuando la prevalencia de la enfermedad es baja, un resultado negativo permite descartar la enfermedad con mayor seguridad, siendo así el valor predictivo negativo mayor. Por el contrario, un resultado positivo no permite confirmar el diagnóstico, resultando en un bajo valor.

Concepto y el vínculo de estadígrafos muestrales y poblacionales

Los estadígrafos muestrales (estimadores) se expresan como algunos tipos de medidas muestrales específicas de una variable determinada, que permiten la estimación probabilística de las correspondientes medidas poblacionales para esa variable para un nivel de confianza determinado, por ejemplo, media aritmética, desviación estándar, varianza, coeficiente de variación, proporción, razón de varianzas, entre otras. Las características deseables de un buen estimador son:

- Carencia de sesgo: referida a cuando no existe diferencia entre la esperanza (valor esperado) del estimador y el verdadero valor del parámetro a estimar, es decir, el valor del sesgo es nulo. El valor medio que se obtiene de la estimación para diferentes muestras debe ser el valor del parámetro.
- Eficiencia: referida a la menor varianza del valor del estimador respecto al valor del parámetro, por lo que cuanto menor es la varianza del estimador mejor es su eficiencia o precisión. Aunque no puede exigírsele a un estimador que para una muestra cualquiera se obtenga el valor exacto del parámetro, sin embargo, su dispersión respecto al valor central (varianza) debe ser tan pequeña como sea posible.
- Consistencia: referida a que cuando el tamaño de la muestra crece arbitrariamente, el valor del estimador se aproxima al del parámetro desconocido.
- Robustez: referida a que la violación de los supuestos de partida en los que se basa la estimación, por ejemplo, atribuir a la población un determinado tipo de distribución que en realidad no es la correcta, no altera de manera significativa los resultados que el estimador proporciona.
- Suficiencia: referida a que el estimador es capaz de resumir toda la información relevante contenida en la muestra, por ejemplo, la media aritmética muestral es un estimador suficiente de la media poblacional, mientras que la moda no lo es.
- Invarianza: referida a que cada parámetro es estimado de forma invariante por el estimador específico muestral con similar denominación, por ejemplo, si para estimar la varianza poblacional se utiliza la varianza muestral, entonces para estimar la desviación típica poblacional se debe utilizar la desviación típica muestral.

Los estadígrafos poblacionales (parámetros) se expresan como los tipos de medidas poblacionales específicos para cada tipo de variable que pueden ser estimados para un nivel de confianza por intervalos a partir de estadígrafos muestrales correspondientes y con el empleo de las pruebas de hipótesis se evalúa la significación de su comportamiento mediante sus comparaciones o asociaciones de las variables que describen, por ejemplo, media aritmética poblacional, desviación estándar y varianza poblacional, proporción poblacional, razón de varianzas poblacionales, entre otras.

El vínculo entre un estimador y parámetro se puede ejemplificar a partir de:

- El estimador de la esperanza matemática para la media poblacional que plantea:

$$E\left[\overline{X}\right] = \mu ;$$

donde a partir de la media muestral:

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n) ;$$

se puede estimar el parámetro media poblacional (μ); ya que el valor del estimador la media de las medias muestrales ($\overline{X}$) tiende a ser o estimar el valor del parámetro media poblacional (por el teorema central del límite).

También la relación vendría dada porque una media muestral, con n ≥ 30, calculada para una variable X obtenida de una población, más y menos el error estándar estimado para la media muestral, conociendo la varianza poblacional y asumiendo una distribución normal de esa variable, lo que permite estimar la media poblacional (parámetro) por intervalos de confianza como:

$$IC(\mu) = \left[\overline{X}\right] \pm z_{\alpha/2}\left(\sigma / \sqrt{n}\right)$$

En caso de no conocer la desviación estándar poblacional (σ) o corresponder a una muestra pequeña (n < 30) se emplea la distribución teórica de probabilidades para t-student con el mismo propósito de estimar la media poblacional (parámetro) por intervalos de confianza como:

$$IC(\mu) = \left[\overline{X}\right] \pm t_{(\alpha/2;\, n-1)}\left(s / \sqrt{n}\right)$$

– El estimador para la varianza poblacional que plantea:

$$Var\left[\overline{X}\right] = \frac{\sigma^2}{n};$$

donde a partir del estimador máximo verosímil varianza muestral (S^2)para $Var\left[\overline{X}\right]$ o expresada como:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2;$$

se podría estimar la varianza poblacional según:
$\sigma^2 = S^2 \cdot n$, despejando σ^2 de la fórmula inicial de:

$$Var\left[\overline{X}\right] = \frac{\sigma^2}{n};$$

aunque se prefiere un estimador insesgado como la cuasivarianza muestral ($\hat{S}^2$) expresada como:

$$\hat{S}^2 = \frac{n}{n-1}S^2$$

para estimar la varianza poblacional según el siguiente razonamiento:

Si: $\sigma^2 = S^2 \cdot n$ y sabemos que: $S^2 = (n-1)\ \hat{S}^2 / n$; sustituyendo en la expresión para la σ^2 (varianza poblacional) tendríamos:

$$\sigma^2 = (n-1)\ \hat{S}^2 = (n-1)\left[\frac{n}{n-1}S^2\right]$$

Para recordar:

– La media muestral ($\overline{X}$) constituye un estimador insesgado de la media poblacional (μ) dado que el promedio de todas las posibles medias muestrales (media de las

medias muestrales) coincide exactamente con la media de la población, independientemente del tipo de población y del tamaño de la muestra.

- El promedio de las varianzas muestrales ($\bar{S}^2$) no constituye un estimador insesgado de la varianza poblacional (σ_N^2); por lo tanto la varianza muestral (S_n^2) no es un estimador insesgado de σ_N^2; más bien, S_{n-1}^2 definido de la forma:

$$S_{n-1}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1};$$

aunque tampoco es un estimador insesgado para σ_N^2, si lo es para σ_{N-1}^2, definido como:

$$\sigma_{N-1}^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N-1}$$

Cuando estimemos la varianza de una población grande (n ≥ 30), entonces σ_N^2, será prácticamente igual a σ_{N-1}^2 por lo que se puede emplear S_{n-1}^2 como el estimador de σ_N^2.

- El concepto de distribución muestral de un estadístico se ha definido como la distribución de frecuencias de los valores obtenidos al aplicar el estadístico a todas las muestras posibles de tamaño n extraídas aleatoriamente de una población con parámetros conocidos. Como se trata de muestras extraídas aleatoriamente, lo más frecuente es que los valores promedio observados (proporción p y media $\bar{x}$) sean valores cercanos a los verdaderos parámetros en la población (proporción π y media μ). No obstante el error aleatorio propio del muestreo, produce algunas muestras con valores de p y $\bar{x}$ alejadas de π y μ.
- La fuente principal de variabilidad da la distribución muestral no es la variabilidad de las observaciones en la población de sujetos, sino el error aleatorio del muestreo. Por esta razón la raíz cuadrada de la varianza del estimador no se llama *desviación estándar,* sino el *error estándar del estimador.*
- Se demuestra que el error estándar disminuye en la medida que aumenta el tamaño muestral.
- La construcción de intervalos de confianza para los parámetros poblacionales, parten del conocimiento de una muestra concreta de tamaño n, para la que se calculan los valores de interés p o $\bar{x}$ y resuelve un problema de estimación al inducir los límites del intervalo que tiene una probabilidad grande (1 – α) de contener al verdadero parámetro poblacional α o π.

  La fórmula general para construir un intervalo de confianza para un parámetro cualquiera de la población es:

  Intervalo de confianza = Valor puntual del estimador ± (coeficiente de confiabilidad o valor del percentil de la distribución correspondiente según alfa/2) (error estándar del estimador).

Estimación por intervalo de confianza de los parámetros (poblacionales) más frecuentemente empleados

A continuación se presenta un formulario resumen para la estimación por intervalo de confianza (IC) de los parámetros (poblacionales) que con más frecuencia se emplean, a partir de su vínculo con los estimadores (muestrales) respectivos.

La estimación puntual permite el acercamiento al conocimiento de un determinado parámetro de interés de la población, por ejemplo, la media de los valores que toma una variable en todos los individuos de la población, el percentil k-ésimo o la proporción de individuos que poseen determinada característica, a través de una medida apropiada (estimador) que se aplica a las observaciones de la muestra y proporciona un valor aproximado (estimación) del parámetro de interés.

Dar una estimación puntual sin indicar su precisión resulta de escasa utilidad e incluso puede resultar engañoso, por este motivo es recomendable proporcionar, junto con la estimación puntual del parámetro, los límites de un intervalo de valores entre los cuales puede hallarse el valor exacto del parámetro con una confianza elevada. Esta confianza se deriva de que el procedimiento usualmente empleado otorga una probabilidad igualmente alta de que los intervalos generados por su conducto contengan al parámetro en cuestión. El grado de confianza deseado debe ser prefijado por el investigador (habitualmente se utilizan valores como 0,90; 0,95 o 0,99); este se expresa como $1 - \alpha$ y se denomina *nivel de confianza*, donde α (nivel de significación) es un valor comprendido entre 0 y 1, usualmente muy pequeño (0,1; 0,05 o 0,01).

Como ejemplos de la construcción del Intervalo de confianza para algunos parámetros poblacionales partiendo de sus respectivos estimadores, empleados en la inferencia bioestadística se pueden citar:

1. Para la media poblacional estimada (μ) a partir de una muestra, se conoce la varianza poblacional (σ) y se cuenta con un tamaño muestral ($n \geq 30$), entonces:

$$IC(\mu) = \left[\bar{x} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right];$$

o también se expresa como:

$$\text{IC Inferior} = \bar{x} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}};$$

$$\text{CI Superior} = \bar{x} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}};$$

con valor de Z correspondiente con un nivel de confianza = $(1 - \alpha/2)\%$.

Un requerimiento para la estimación correcta de una media poblacional es el cálculo del tamaño muestral (n) necesario, para lo cual se requiere conocer la media muestral o de referencia a emplear, la desviación estándar poblacional, la anchura en cada

lado para el intervalo de la media muestral seleccionada por el investigador (FE) o el error de estimación a ambos lados determinado para la misma (e) seleccionado por el investigador, para un nivel de confianza = (1 – $\alpha/2$)% como:

$$n \geq \left(\frac{2 \cdot z_{1-\alpha/2} \cdot \sigma}{e} \right)^2 \quad o \quad n \geq \frac{4 \cdot z^2_{1-\alpha/2} \cdot \sigma^2}{e^2}$$

donde el error de estimación (e) es la amplitud del intervalo para la media poblacional estimada seleccionado por el investigador y se expresa como:

$$e \leq 2 \cdot z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Otra forma de estimar el tamaño muestral necesario (n) puede ser:

$$n = \frac{z^2_{\alpha/2} \cdot S^2}{(FE)^2} = \frac{z^2_{\alpha/2} \cdot S^2}{(e/2)^2} = 4\frac{z^2_{\alpha/2} \cdot S^2}{(e)^2} = \left(2\frac{z_{\alpha/2} \cdot S}{e} \right)^2$$

debido a que conceptualmente para el trabajo con muestras grandes (n ≥ 30) la desviación estándar poblacional (σ) puede ser estimada a partir de la muestral (S), donde:

a) FE = $Z_{\alpha/2}$ EE (factor de error o diferencia que selecciona el investigador respecto al valor puntual de la media muestral o de referencia a emplear), se puede expresar como: FE= e/2.

b) $EE = S/\sqrt{n}$ (error estándar para la media muestral o de referencia a emplear). Una forma abreviada de calcular el tamaño muestral necesario (n), basándonos en el redondeo de los valores del percentil respectivo de la distribución normal estándar Z, según el nivel de confianza seleccionado, puede ser como sigue:
Por ejemplo: cuando Z = 1,96 para un valor $\alpha/2$ = 0,025, si se redondea este valor de Z a 2; entonces (Z)2 = 4, el cual está en la fórmula inicial (antes enunciada), empleando este valor en la fórmula:
   - Si intervalo de confianza al 95 %; donde: α = 0,05; entonces:
     $n \approx 4S^2/FE^2 = 4/FEE^2$, si n < 75, n' = n + 0,05 · n; entonces se calcula el tamaño muestral necesario como: n' = n + 0,05 · n · FE.
   - Si intervalo de confianza al 90 %; donde α = 0,10; entonces: $n \approx 2,8/FEE^2$
   - Si intervalo de confianza al 99%; donde: α = 0,01; entonces: $n \approx 7,7/FEE^2$
     donde:
     FEE: factor de error estandarizado que interesa al investigador, es la anchura a cada lado del intervalo para la media poblacional estimada de la muestral, expresada en unidades de desviación estándar y se expresa como FE/s.

2. Para la media poblacional estimada (μ) a partir de una muestral, si se desconoce la varianza poblacional (σ) o se cuenta con un tamaño muestral (n < 30) entonces:

$$IC(\mu) = \left[\bar{x} - t_{(n-1,\ 1-\alpha/2)} \cdot \frac{S}{\sqrt{n}},\ \bar{x} + t_{(n-1,1-\alpha/2)} \cdot \frac{S}{\sqrt{n}} \right]$$

o también se expresa como:

$$\text{CI Inferior} = \bar{x} - t_{(n-1,\ 1-\alpha/2)} \cdot \frac{S}{\sqrt{n}};$$

$$\text{CI Superior} = \bar{x} + t_{(n-1,\ 1-\alpha/2)} \cdot \frac{S}{\sqrt{n}};$$

con valor de t-student con n – 1 grados de libertad y para un nivel de confianza de $(1 - \alpha/2)\%$, donde S (desviación estándar muestral) se asume que estima a la σ (desviación estándar poblacional).

Un requerimiento para la estimación correcta de una media poblacional es el cálculo del tamaño muestral (n) necesario a emplear en la misma; para lo cual se requiere conocer la media muestral, la desviación estándar muestral que estimaría la poblacional y la anchura o precisión del intervalo de cada lado para la media muestral (FE) seleccionado por el investigador y la t-student (n – 1 grados de libertad), para un nivel de confianza de $(1 - \alpha/2)\%$; siendo la expresión para el mismo como:

$n \approx t^2S^2/FE^2$

donde:

$FE = t_{n-1,\alpha/2}(S/\sqrt{n})$: factor de error o diferencia respecto al valor puntual de la media muestral o de referencia a emplear, precisión o anchura a cada lado, del intervalo para la media poblacional estimada a partir de la muestral o de referencia a emplear seleccionada por el investigador.

3. Para la varianza poblacional estimada (σ^2) a partir de una muestral, con media poblacional desconocida (μ^0) y se conoce la varianza muestral (S^2) para un número de muestra (n) preferentemente grande ($n \geq 30$), entonces:

$$IC(\sigma) = \left[\frac{(n-1)\hat{S}^2}{t^2_{n-1,\ 1-\alpha/2}},\ \frac{(n-1)\hat{S}^2}{t^2_{n-1,\ 1-\alpha/2}} \right]$$

o también se puede expresar como:

$$\text{IC Inferior}(\sigma^2) = \frac{(n-1)\hat{S}^2}{t^2_{n-1,\ 1-\alpha/2}}$$

$$\text{IC Superior}(\sigma^2) = \frac{(n-1)\hat{S}^2}{t^2_{n-1,\ 1-\alpha/2}}$$

4. Para la proporción poblacional estimada (P) a partir de una muestral de una característica de interés X para muestras grandes de preferencia ($n \cdot p \cdot q > 5$):

$$IC(P) = \left[\hat{P} - z_{1-\alpha/2}\sqrt{\frac{\hat{P}\cdot(100-\hat{P})}{n}},\ \hat{P} + z_{1-\alpha/2}\sqrt{\frac{\hat{P}\cdot(100-\hat{P})}{n}} \right]$$

o también se puede expresar como:

$$\text{IC Inferior} = \hat{P} - z_{1-\alpha/2}\sqrt{\frac{\hat{P}\cdot(100-\hat{P})}{n}}$$

$$\text{IC Superior} = \hat{P} + z_{1-\alpha/2}\sqrt{\frac{\hat{P}\cdot(100-\hat{P})}{n}}$$

Una forma resumida de su expresión pude ser:

IC (P) = p ± FE;

donde:

FE = $Z_{1\ -\ \alpha/2EE}$

$$EE = \sqrt{\frac{pq}{n}};$$

q = 1 − p;

$$\text{IC (P)} = p + z_{1-\alpha/2}\sqrt{\frac{pq}{n}};$$

donde:

$\hat{P}$ o p: valor puntual de la proporción muestral;

EE: error estándar de la proporción muestral que estimaría a una poblacional;

FE: es la amplitud de cada lado del intervalo de confianza seleccionada por el investigador y se puede expresar como FE = (e/2).

$Z_{1\ -\ \alpha/2}$: percentil de la distribución normal estándar N(0, 1) que deja una cola a la izquierda de 1 − $\alpha/2$.

Un requerimiento para la estimación correcta de una proporción poblacional es el cálculo del tamaño muestral (n) necesario con un error de estimación determinado (e):

$$n \geq \frac{4\cdot z^2_{1-\alpha/2}\cdot P\cdot(100-P)}{e^2};$$

donde:

e: error de estimación o diferencia que selecciona el investigador, es la amplitud del intervalo (de ambos lados, para la proporción muestral o de referencia a emplear que estimaría la poblacional) y se expresa como:

$$e \leq 2\cdot z_{1-\alpha/2}\cdot\sqrt{\frac{P\cdot(100-P)}{n}}$$

P: proporción conocida muestral que se utilizaría de referencia.

Otra forma de estimar el tamaño muestral necesario (n) es:

$$FE = z_{\alpha/2}\sqrt{\frac{pq}{n}};$$

quedando como:

$$n = \frac{4\cdot z^2_{\alpha/2}(pq)}{(2\cdot FE)^2};$$

FE = e/2.

5. Para la densidad de incidencia (ID) o de una tasa de incidencia poblacional estimada (TI), a partir de una muestra, empleando una tasa de incidencia de referencia (ID_0), para un número de nuevos eventos observados de cierta característica de interés, mayor o igual de 10 ($a \geq 10$) se expresa como:

$$IC\ (ID) = \frac{a \pm z_{1-\alpha/2}\sqrt{a}}{t'}$$

o también se puede expresar como:

$$IC\ Inferior = \frac{a - z_{1-\alpha/2}\sqrt{a}}{t'};$$

$$IC\ Superior = \frac{a - z_{1-\alpha/2}\sqrt{a}}{t'};$$

donde:

a: número de nuevos eventos observados de la característica o evento de interés;

t: número de personas-tiempo de seguimiento de la muestra;

Z: valor percentil de la distribución normal estándar N(0, 1) que deja una cola de probabilidad a la izquierda de $1 - \alpha/2$.

6. Para la diferencia de dos medias poblacionales estimadas ($\mu_1 - \mu_2$), a partir de las dos medias muestrales calculadas para la variable cuantitativa (dependiente) empleando dos nominaciones en la variable cualitativa (independiente) por medio de la expresión:

$$IC(\mu_1 - \mu_2) = \left[\left(\bar{x}_1 - \bar{x}_2\right) \pm z_{1-\alpha/2)}\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right], \text{ para muestras grandes } (n \geq 30).$$

$$IC\ Inferior = \left(\bar{x}_1 - \bar{x}_2\right) - z_{1-\alpha/2)}\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}};$$

$$IC\ Superior = \left(\bar{x}_1 - \bar{x}_2\right) + z_{1-\alpha/2)}\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}};$$

$$IC(\mu_1 - \mu_2) = \left[\left(\bar{x}_1 - \bar{x}_2\right) \pm t_{(n_1+n_2-2)(1-\alpha/2)} \cdot S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right] \text{ o también se puede expresar como:}$$

$$IC\ Inferior = \left(\bar{x}_1 - \bar{x}_2\right) - t_{(n_1+n_2-2)(1-\alpha/2)} \cdot S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}};$$

$$IC\ Superior = \left(\bar{x}_1 - \bar{x}_2\right) + t_{(n_1+n_2-2)(1-\alpha/2)} \cdot S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}};$$

donde:

$S^2 = \frac{(n_1-1)\cdot S_1^2 + (n_2-1)\cdot S_2^2}{(n_1+n_2-2)}$, para muestras pequeñas ($n < 30$), con: $\sigma_1^2 = \sigma_2^2$ (desconocidas).

$IC(\mu_1 - \mu_2) = \left(\overline{X}_1 - \overline{X}_2\right) \pm t_{f,1-\alpha/2}\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}$ o también se puede expresar como:

$IC\ Inferior = \left(\overline{X}_1 - \overline{X}_2\right) - t_{f,1-\alpha/2}\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}$

$IC\ Superior = \left(\overline{X}_1 - \overline{X}_2\right) + t_{f,1-\alpha/2}\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}$, para muestras pequeñas ($n < 30$), con: $\sigma_1^2 \neq \sigma_2^2$ (desconocidas) y donde:

$$f = \frac{\left(\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}\right)^2}{\frac{1}{n_1 + 1}\left(\frac{\hat{S}_1^2}{n_1}\right)^2 + \frac{1}{n_2 + 1}\left(\frac{\hat{S}_2^2}{n_2}\right)^2} - 2$$

7. Para la media poblacional estimada de las diferencias entre los pares de observaciones de dos muestras pareadas (μ_d) para una variable cuantitativa por medio de la expresión:

   $\overline{d} - E < \mu_d < \overline{d} + E$; o también se puede expresar como:

   $IC\ Inferior(\mu_d) = \overline{d} - E$;

   $IC\ Superior(\mu_d) = \overline{d} + E$;

   donde:

   $\overline{d}$: media muestral de todas las diferencias (d_i) entre las dos observaciones o las parejas de datos de las dos muestras pareadas y se expresa como:

$$d = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{\sum_{i=1}^{n} (x_i - x_2)}{n}$$

   $E = t_{\alpha/2}\frac{S_d}{\sqrt{n}}$: error estimado;

   n: es el número de parejas de datos u observaciones por parejas de datos;

   $S^2\overline{d} = \frac{\sum_{i=1}^{n}(d_1 - \overline{d})^2}{n-1}$; es la cuasivarianza para las diferencias de todas las observaciones pareadas correspondientes a las muestras pareadas que estimaría la varianza poblacional para los datos pareados.

   Un requerimiento para la estimación correcta de una diferencia de dos medias poblacionales estimadas de las dos muestrales respectivas, es el cálculo del tamaño muestral (n) necesario de cada muestra (se asumirían de igual tamaño para cada una), para una probabilidad de cometer error tipo 1 (α) y también del error tipo 2 (β), se requiere conocer:

a) Las dos desviaciones estándar poblacionales estimadas o de referencia esperadas respectivas para las dos medias, la diferencia de medias mínima que se desea detectar y determinar según tablas de referencia (véase el anexo 1) para la distribución de frecuencias normal Z ~ N(0, 1) el valor de riesgo para alfa ($Z_{\alpha/2}$) y el de beta (Zβ) para efectuar el contraste $\mu 1$ vs. $\mu 2$ como:

$n = \frac{2S^2(Z_{\alpha/2} + Z_{\beta})^2}{d^2}$ siendo necesario multiplicar el resultado de n por 2 para el total de individuos a incluir en el estudio.

b) En caso que se utilice un error bilateral alfa del 5 % ($\alpha/2$ = 0,025) y una potencia (1 – β) del 80 % (0,8) siendo β = 0,2, se empleará la fórmula:

$n = \frac{16}{(DE)^2}$

donde:

S: desviación estándar poblacional estimada común o de referencia a emplear a partir de ambas muestras (cuasivarianza muestral);

d: diferencia de medias mínima que se desea detectar y se expresa como:

$d = Z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} + Z_{\beta}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$;

$d = FE_1 + FE_2$;

$FE_1 = Z_{\alpha/2}$EEDM: diferencia entre dos medias que marcan el punto límite relacionado con la probabilidad de cometer el error tipo 1;

FE_2 = ZβEEDM: diferencia entre las dos medias que marcaría el punto límite relacionado con la probabilidad de cometer el error tipo 2;

EEDM: error estándar para la diferencia de las dos medias poblacionales estimadas;

(DE)2: diferencia estandarizada.

En la tabla 3.26 se presentan los valores frecuentemente empleados de la distribución normal estándar Z ~ N(0,1) según los valores de alfa y beta respectivos para la estimación del tamaño muestral a partir de la diferencia de dos medias poblacionales estimadas.

Tabla 3.26. Valores de Z según alfa (α) y beta (β) para la estimación del tamaño muestral a partir de la diferencia de dos medias poblacionales estimadas

α	$Z_{\alpha/2}$	β	Z_{β}
0,1	1,65	0,2	0,84
0,05	1,96	0,1	1,28
0,01	2,58	0,05	1,65

8. Para la razón entre dos varianzas poblacionales estimadas [$(\sigma_1^2)/(\sigma_2^2)$], a partir de las dos varianzas muestrales calculadas (S_1^2 y S_2^2) en la variable cuantitativa (dependiente) empleando las dos categorías o nominaciones en la variable cualitativa (independiente) por medio de la expresión:

$\frac{S_1^2 / S_2^2}{F_{\alpha/2}} > \frac{\sigma_1^2}{\sigma_2^2} > \frac{S_1^2 / S_2^2}{F_{1-(\alpha/2)}}$ o se puede expresar dicha razón por intervalo de confianza como:

$$\text{IC Inferior}(\sigma_1^2 / \sigma_2^2) = \frac{S_1^2 / S_2^2}{F_{1-(\alpha/2)}}$$

$$\text{IC Superior}(\sigma_1^2 / \sigma_2^2) = \frac{S_1^2 / S_2^2}{F_{\alpha/2}}$$

donde:

S_1^2 y S_2^2 : representan las varianzas muestrales 1 y 2, respectivamente;

$F_{1-(\alpha/2)}$ y $F_{\alpha/2}$: valor del percentil de la distribución F-Snedecor que deja una cola de probabilidad a la izquierda $1-\alpha/2$ y $\alpha/2$, respectivamente.

9. Para la diferencia de dos proporciones poblacionales estimadas ($P_1 - P_2$), a partir de dos proporciones muestrales calculadas empleando una nominación de referencia en la variable independiente y las dos nominaciones del evento en la dependiente (variables dicotómicas) por medio de la expresión:

$$IC(P_1 - P_2) = (\hat{P}_1 - \hat{P}_2) \pm Z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{P}_1 \cdot (100 - \hat{P}_1)}{n_1} + \frac{\hat{P}_2 \cdot (100 - \hat{P}_2)}{n_2}}$$ o

$$IC(P_1 - P_2) = (\hat{p}_1 - \hat{p}_2) \pm Z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}_1 \hat{q}_1)}{n_1} + \frac{\hat{p}_2 \hat{q}_2)}{n_2}}$$ o también se puede expresar como:

$$\text{IC Inferior} = (\hat{p}_1 - \hat{p}_2) - Z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}_1 \hat{q}_1)}{n_1} + \frac{\hat{p}_2 \hat{q}_2)}{n_2}};$$

$$\text{IC Superior} = (\hat{p}_1 - \hat{p}_2) + Z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}_1 \hat{q}_1)}{n_1} + \frac{\hat{p}_2 \hat{q}_2)}{n_2}}.$$

Un requerimiento para la estimación correcta de una diferencia de dos proporciones poblacionales estimadas o de referencia esperadas de las dos muestrales respectivas, es el cálculo del tamaño muestral (n) necesario de cada muestra (que se asumirían de igual tamaño para cada una), con una desviación estándar común y con una probabilidad de cometer error tipo 1 (α) y también del error tipo 2 (β), se debe conocer que las dos proporciones poblacionales estimadas de las dos muestrales o de referencia esperadas, las desviaciones estándar poblacionales estimadas o de referencia esperadas respectivas, la proporción promedio de las poblacionales estimadas, la diferencia de proporciones mínima que se desea detectar y el valor de riesgo para alfa ($Z_{\alpha/2}$) y para beta (Z_β) de acuerdo a la tabla de referencia de la distribución normal $Z \sim N(0,1)$, para efectuar el contraste P_1 *vs.* P_2 como:

$$n = \frac{2S^2(Z_{\alpha/2} + Z_\beta)^2}{d^2}$$ o también como:

$$n = \frac{2pq(Z_{\alpha/2} + Z_{\beta})^2}{(P_A - P_B)^2}$$

donde:

S^2= pq o $S = \sqrt{pq}$;

p: es una proporción media entre la proporción de la muestra o grupo A y la correspondiente a la muestra o grupo B y se expresa como:

p = PM = $(P_A + P_B)/2$; donde:

P_A y P_B: dos proporciones muestrales que estiman las dos poblacionales respectivas;

Si se asumen ambos grupos o muestras iguales se puede emplear la fórmula:

$n = \frac{2p_M q_M (Z_{\alpha/2} + Z_{\beta})^2}{(p_A - p_B)^2}$; es necesario multiplicar el resultado de n por 2 para el total de individuos a incluir en el estudio.

Los valores frecuentes de Z según la probabilidad para $\alpha/2$ y β, empleados en la estimación del tamaño de muestra para el contraste de la diferencia de dos proporciones poblacionales estimadas, se expresan en la tabla 3.24 para la diferencia de dos medias poblacionales estimadas.

10. Para la razón de dos tasas de incidencia poblacional estimada (IRR), a partir de dos grupos según exposición sin covariable de estratificación, empleando la expresión:

a) IC Inferior $(IRR) = e^{d_1}$

donde:

$$d_1 = \ln(\hat{IRR}) - Z_{1-\alpha/2} \cdot \sqrt{\frac{1}{a_1} + \frac{1}{a_2}} ;$$

$\hat{IRR} = \frac{a_1}{t_1} + \frac{a_2}{t_2}$: razón de tasas de incidencia para los dos grupos según exposición;

$\sqrt{\frac{1}{a_1} + \frac{1}{a_2}}$: error estándar para la razón de tasas de incidencia;

a_1: nuevos eventos observados en el grupo 1 (expuestos) o población 1;

a_2: nuevos eventos observados en el grupo 2 (no expuestos) o población 2;

t_1: total de personas-tiempo de grupo 1;

t_2: total de personas-tiempo de grupo 2.

b) IC Superior $(IRR) = e^{d_2}$

donde:

$$d_2 = \ln(\hat{IRR}) + Z_{1-\alpha/2} \cdot \sqrt{\frac{1}{a_1} + \frac{1}{a_2}} ;$$

$\hat{IRR} = \frac{a_1}{t_1} + \frac{a_2}{t_2}$: razón de tasas de incidencia para los dos grupos según exposición;

$\sqrt{\frac{1}{a_1}+\frac{1}{a_2}}$: error estándar para la razón de tasas de incidencia;

a_1: nuevos eventos observados en el grupo 1 (expuestos) o población 1;

a_2: nuevos eventos observados en el grupo 2 (no expuestos) o población 2;

t_1: total de personas-tiempo de grupo 1;

t_2: total de personas-tiempo de grupo 2;

e: valor de la base del logaritmo natural (ln) igual a 2,7182.

11. Para la razón de tasas de incidencia poblacional estimada en conjunto (IRR_C) para los diferentes estratos, a partir de dos grupos según exposición con estratificación por la presencia de una covariable:

a) IC Inferior $(IRRc) = e^{c_1}$

donde:

$$c_1 = \ln(\hat{IRR}) - Z_{1-\alpha/2} \cdot \sqrt{\frac{1}{\sum_{i=1}^{k} w_i}}$$

$$\ln(\hat{IRR}) = e^{c};$$

$$c = \frac{\sum_{i=1}^{k} w_i \ln(\hat{IRR})}{\sum_{i=1}^{k} w_i}$$: expresión del exponente para el antilogaritmo natural;

$\hat{IRR}_i = \frac{a_{1i} / t_{1i}}{a_{2i} / t_{2i}}$: razón de tasas de incidencia para cada estrato (i);

$w_i = \left(\frac{1}{a_{1i}} + \frac{1}{a_{2i}}\right)^{-1}$: peso promedio o el promedio ponderado estimado para el número de nuevos eventos observados en ambos grupos según exposición para cada estrato (i);

i: es el número de estrato específico.

b) IC Superior $(IRRc) = e^{c_2}$

donde:

$$c_2 = \ln(\hat{IRR}) + Z_{1-\alpha/2} \cdot \sqrt{\frac{1}{\sum_{i=1}^{k} w_i}};$$

$$\ln(\hat{IRR}) = e^{c};$$

$$c = \frac{\sum_{i=1}^{k} w_i \ln(\hat{IRR}i)}{\sum_{i=1}^{k} w_i}$$: expresión del exponente para el antilogaritmo natural;

$\hat{IRR}_i = \frac{a_{1i} / t_{1i}}{a_{2i} / t_{2i}}$: razón de tasas de Incidencia para cada estrato (i);

$w_i = \left(\frac{1}{a_{1i}} + \frac{1}{a_{2i}} \right)^{-1}$: peso promedio o el promedio ponderado estimado para el número de nuevos eventos observados en ambos grupos según exposición para cada estrato (i);

i: es el número de estrato específico;

e: es el valor de la base del logaritmo natural (ln) igual a 2,7182.

12. Para el coeficiente de correlación lineal simple poblacional estimado (ρ), empleando su estimador (r), a partir de la relación entre d variables cuantitativas en una muestra:

$\frac{e^{2w_I} - 1}{e^{2w_I} + 1} < p < \frac{e^{2w_D} - 1}{e^{2w_D} + 1}$ o también se puede expresar como:

$IC\ Inferior = \frac{e^{2w_I} - 1}{e^{2w_I} + 1}$

$IC\ Superior = \frac{e^{2w_D} - 1}{e^{2w_D} + 1}$

donde:

$W_I = \frac{1}{2} \ln\left(\frac{1+r}{1-r} \right) - z_{\alpha/2} \cdot \frac{1}{\sqrt{n-3}}$;

$W_D = \frac{1}{2} \ln\left(\frac{1+r}{1-r} \right) + z_{\alpha/2} \cdot \frac{1}{\sqrt{n-3}}$;

r: valor del coeficiente de regresión lineal simple muestral y se estima como:

$r = \frac{(xy)}{\sqrt{(xx)(yy)}}$ con:

$(xx) = \sum x_i^2 - \frac{\left(\sum x_i \right)^2}{n}$;

$(yy) = \sum y_i^2 - \frac{\left(\sum y_i \right)^2}{n}$;

$(xy) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$;

n: número de pares de datos presentes de las dos variables cuantitativas;

e: valor de la base del logaritmo natural (ln) igual a 2,7182.

13. Para el coeficiente de regresión lineal simple poblacional estimado (B), a partir de su estimador (b), a partir de la relación entre dos variables cuantitativas en una muestra:

$IC\ (B) = \left[b \pm \left(t_{1-2/2,n-2} \right) \sqrt{\frac{S^2}{(xx)}} \right]$

o también se puede expresar como:

$$\text{IC Inferior} = \left[b - \left(t_{1-2/2,n-2} \right) \sqrt{\frac{S^2}{(xx)}} \right]$$

$$\text{IC Superior} = \left[b + \left(t_{1-2/2,n-2} \right) \sqrt{\frac{S^2}{(xx)}} \right]$$

donde:

$b = \frac{(xy)}{(xx)}$ con

$(xy) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$;

$(xx) = \sum x_i^2 - \frac{\left(\sum x_i \right)^2}{n}$;

s: desviación estándar muestral;

n: número de pares de datos o valores de las dos variables x, y;

i: representa el orden de los pares de datos,

$t_{1-\alpha/2}$: es el percentil de la distribución de t-Student que deja una cola de probabilidad a la izquierda de $1^{-\alpha/2}$ con n – 2 grados de libertad.

Los elementos incluidos en cada expresión matemática se pueden consultar en el capítulo 4.

14. Para la ecuación de la recta de regresión lineal simple poblacional estimada (y); mediante su estimador ($\hat{y}$), a partir de la relación de dos variables cuantitativas en una muestra:

$\hat{y} - E < y < \hat{y} + E$ o también de puede expresar como:

IC Inferior = $\hat{y} - E$;

IC Superior = $\hat{y} + E$;

donde:

$E = t_{\alpha/2} S_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n\left(\sum x^2\right) - \left(\sum x^2\right)}}$;

$S_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$ o $S_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}}$;

$b_1 = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$;

$b_0 = \bar{y} - b_1 \bar{x}$;

E: margen de error para la recta de regresión;

Se: error estándar del estimado para la recta de regresión;

$\bar{x}$: media de los valores de x;

x_0: representa un valor x fijo a seleccionar para sustituir en la ecuación de la recta de regresión y calcular el valor predictivo de y específico;

x: valor de la variable independiente cuantitativa;

y: valor de la variable dependiente cuantitativa;

$\hat{y}$: valor predicho de y, obtenido por la ecuación de regresión;

b_0: intercepto de la recta de regresión con el eje y;

b_1: pendiente de la recta de regresión;

n: número de pares de datos de las variables x, y;

$t_{\alpha/2}$: percentil de la distribución de t-Student que deja una cola de probabilidad a la izquierda de $\alpha/2$ con n – 2 grados de libertad.

15. Para el intercepto β0 de una recta de regresión estimada ($y = \beta_0 + \beta_{1x}$):

$b_0 - E < \beta_0 < b_0 + E$

donde:

$$E = t_{\alpha/2} S_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}};$$

$$S_e = \sqrt{\frac{\sum\left(y - \hat{y}\right)^2}{n-2}}$$

16. Para la pendiente β_1 de una recta de regresión estimada ($y = \beta_0 + \beta_1 x$):

$b_1 - E < \beta_1 < b_1 + E$

donde:

$$E = t_{\alpha/2} \cdot \frac{S_e}{\sqrt{\sum x^2 + \frac{\left(\sum x\right)^2}{n}}};$$

$$S_e = \sqrt{\frac{\sum\left(y - \hat{y}\right)^2}{n-2}}$$

Los elementos incluidos en cada expresión matemática se pueden consultar en el en el capítulo 4.

17. Para el índice de posición (IP) de los individuos en una muestra, en relación al número de las categorías k de una escala en una variable cualitativa ordinal: su estimación puntual es:

$$IP = \frac{M-1}{k-1}$$ (Silva, 1997: 49-53);

mientras que por intervalo de confianza (IC) (Sánchez, 2004):

$$\text{IC Inferior} = IP - z_{1-\alpha/2}\sqrt{Var(IP)};$$

$$\text{IC Superior} = IP + z_{1-\alpha/2}\sqrt{Var(IP)};$$

donde:

k: número de categorías de la escala de la variable cualitativa ordinal;

$$M = \sum_{i=1}^{k} ip_i;$$

$p_i = \frac{x_i}{n}$: proporción de sujetos que eligen la categoría i con i = 1,..., k;

xi: es el número de sujetos que eligen la categoría i, con i = 1,..., k;

$n = \sum_{i=1}^{k} x_i$: tamaño de la muestra;

$$Var(IP) = \frac{1}{1(k-1)^2}\left\{\sum_{i=1}^{k} i^2 p_i(1-p_i) - 2\sum_{i=1}^{k}\sum_{j>1} ijp_ip_j\right\};$$

$Z_{1-\alpha/2}$: percentil de la distribución normal estándar N(0,1), que deja a la izquierda una cola de probabilidad $1 - \alpha/2$;

$1 - \alpha$: nivel de confianza.

18. Para un percentil a partir de los valores ordenados de una variable cuantitativa en una muestra su estimación puntual (Pk) es:

Pk = (1 – f)xr +fxr + 1 (Altman y Bland, 1994);

mientras que por intervalo de confianza (IC) (Royston, 1992: 12-15):

a) Método exacto:

IC Inferior = $x_l + g(x_{l+1} - x_l)$;

IC Superior = $x_S - h(x_S - x_{S-1})$;

Si $x_l + g(x_{l+1} - x_l) < \min_i\{x_i\}$ entonces el límite inferior del intervalo es $\min_i\{x_i\}$.

Si $x_l + g(x_{l+1} - x_l) > \max_i\{x_i\}$ entonces el límite superior del intervalo es $\max_i\{x_i\}$.

donde:

$x_1, x_2,..., x_3$: muestra ordenada de valores;

n: tamaño de la muestra;

k: orden del percentil;

r = [R]: parte entera de R, $0 \leq r \leq n$;

f = R – r: parte fraccionaria de R;

R = (n + 1)p;

$p = \frac{k}{100}$;

$X_0 = x_1$;

$X_n + 1 = x_n$;

l = máx{1, t};

máx: expresa el máximo valor;

S = mín {n, u + 1}

mín: expresa el mínimo valor;

i: expresa el número de orden de cada valor;

t = F–1(α/2): percentil de la distribución binomial con parámetros n y p, F~Bin(n, p) que deja a su izquierda una cola de probabilidad α/2;

u = F–1(1 – α/2): percentil de la distribución binomial con parámetros n y p, F~Bin(n, p) que deja a su izquierda una cola de probabilidad 1 – α/2;

1 – α: nivel de confianza;

xi: valor que ocupa la posición i en la muestra ordenada;

$$g = \frac{\alpha/2 - F_{I-1}}{F_I - F_{I-1}};$$

$$h = \frac{F_{S-1} - (1-\alpha/2)}{F_{S-1} - F_{S-2}};$$

Fy = Pr{X ≤ y} con X~Bin(n, p).

b) Método aproximado:

IC Inferior = Pk – z1 – α/2Sk

IC Superior = = Pk + z1 – α/2Sk

donde:

Pk: percentil de orden k;

$$S_k = \frac{\sqrt{k(100-k)}}{100nZ(P_k, \bar{x}, s)}$$

Z(y, μ, σ): función de densidad normal con media μ y desviación estándar σ, evaluada en el punto y;

$\bar{x}$: media de la muestra;

s: desviación estándar de la muestra;

n: tamaño de la muestra;

z1 – α/2: percentil de la distribución normal estándar N(0,1), que deja a la izquierda una cola de probabilidad 1 – α/2.

1 – α: nivel de confianza.

Para recordar (a modo de resumen)

- Siempre que en base a una muestra se quiera conocer un rasgo de la población con un margen de error especificado, se corre un riesgo de equivocación; pero este riesgo se puede controlar, llevándolo al valor que convenga, utilizando el tamaño de muestra adecuado; aunque otro elemento a tomar en cuenta es que se disponga del presupuesto y de los medios necesarios para tomar una muestra aleatoria del tamaño necesario.
- Se puede asumir al trabajar con una muestra aleatoria que los resultados obtenidos pueden ser capaces de predecir lo que ocurre en una población de

donde se extrajo; dado que la estimación del promedio de los valores de los estimadores muestrales coincide con el valor del parámetro poblacional respectivo y, además, la variabilidad del estimador disminuye a medida que aumenta el tamaño muestral.

La fórmula empleada para calcular el tamaño muestral n en general para poder estimar una media o proporción poblacional a partir de una muestral, es la que sigue:

$$n_0 = \frac{(C)^2 \sigma^2}{E_0^2};$$

que puede ser expresada como:

$$n = \left(\frac{Zc \cdot S}{E} \right)^2$$

o también como:

$$n = \left(\frac{tc \cdot S}{E} \right)^2$$

según la distribución de probabilidades (normal estándar o t-student) a emplear para la variable seleccionada en el caso de la media a estimar. En el caso de determinar el tamaño de la muestra para la estimación de la proporción es por aproximación a la distribución normal estándar de la binomial como:

$$n = \frac{Z_c^{\,2} \cdot p \cdot q}{E^2}$$

Al comparar n calculado con N (tamaño poblacional) si se cumple que $N \geq 10n$; entonces se calcula n, el tamaño muestral a seleccionar; en caso contrario, cuando $N < 10n$, entonces se calcula el tamaño muestral como:

$$n = \frac{n_0}{1 + \frac{n_0}{N}};$$

en la cual, n_0 correspondería con el valor de n anteriormente calculado por la fórmula anterior. Los valores de n calculados en caso de emplearse un muestreo simple aleatorio (MSA) se multiplican por 1 como valor del efecto de diseño (deff) para el mismo y se conserva el mismo valor de n; pero para el caso de diseños de estudio que empleen otros tipos de muestreo más complejos que el muestreo simple aleatorio, se requieren utilizar valores del efecto de diseño que pueden ir desde el valor de 1,5 a 3, generalmente.

N: tamaño poblacional;

$(C)^2$: se corresponde con el percentil para $(1 - \alpha/2)\%$ de la distribución de probabilidad empleada, ya sea Z_c o t_c;

σ^2: varianza poblacional, que es desconocida, pero se estima a partir de la muestral S^2;

E_0 o E: valor máximo de error definido por el investigador (expresado en por ciento), en relación al valor muestral del estimador seleccionado (media o proporción) y se emplea generalmente el valor del 1 % al 10 % del estimador a emplear.

– El tamaño poblacional no siempre se conoce; no obstante, el efecto de N sobre el tamaño muestral puede ser considerado a los efectos prácticos como despreciable; lo que realmente influye en la precisión que se obtenga no es la magnitud relativa de la muestra respecto a la población; sino el tamaño absoluto de n. Por lo tanto, en estos casos n_0 puede ser considerada como una aproximación aceptable del valor de n y a partir de n_0 es que se estimaría n para efectuar la estimación del parámetro en cuestión.

Los aspectos resumidos a tener en cuenta en las estimaciones de parámetros más frecuentemente empleados en los diferentes diseños de estudio son:

– Para el caso de la media muestral:
  - El error estándar de la media (EE) se define como:

$$EE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

  - El error de muestreo (e) o margen de error para la media se define como:

$$e = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

  - El intervalo de confianza para la media poblacional [IC(μ)] está dado por:

$$IC(\mu) = \bar{x} + e = \bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

  - El error relativo para la media muestral (ER) está dado por:
    ER = (Error de muestreo/Media muestral) · 100 %

– Para el caso de la proporción muestral:
  - El error estándar (EEp) de la proporción se define como:

$$EEp = \sqrt{\frac{p(1-p)}{n}}$$

  - El error de muestreo (e) o margen de error para la proporción se define como:

$$e = z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

  - El intervalo de confianza para la proporción poblacional [IC(π)] está dado por:

$$IC(\pi) \text{ o } IC(P) = p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

  - El error relativo para la proporción muestral (ER p) está dado por:
    ERp = (Error de muestreo/Proporción muestral) · 100 %

– Para el caso de la diferencia entre dos medias de dos poblaciones estimadas ($\mu_1 - \mu_2$):

Cuando se conoce la varianza de las dos poblaciones y las muestras son grandes (n_1 y $n_2 > 30$):

- El error estándar de la diferencia de dos medias [$EE(\mu_1 - \mu_2)$] se expresa como:

$$EE(\mu_1 - \mu_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- El error de muestreo [$e(\mu_1 - \mu_2)$] o margen de error de la diferencia de dos medias se expresa como:

$$e(\mu_1 - \mu_2) = z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- El Intervalo de confianza para la diferencia de dos medias poblacionales [IC ($\mu_1 - \mu_2$)] está dado por:

$$IC(\mu_1 - \mu_2) = (\bar{X}_1 - \bar{X}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Cuando no se conoce la varianza de las dos poblaciones estimadas y/o son pequeñas las muestras:

- El error estándar de la diferencia de dos medias [EE ($\mu_1 - \mu_2$)] se expresa como:

$$EE(\mu_1 - \mu_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$, para el caso que se suponen las varianzas diferentes.

$$EE(\mu_1 - \mu_2) = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$; para el caso que se suponen las varianzas iguales.

- El error de muestreo [$e(\mu_1 - \mu_2)$] o margen de error de la diferencia de dos medias se expresa como:

$$e(\mu_1 - \mu_2) = t_{1-\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$; para el caso que se suponen las varianzas diferentes.

$$e(\mu_1 - \mu_2) = t_{1-\alpha/2}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$; para el caso que se suponen las varianzas iguales.

- El Intervalo de confianza para la diferencia de dos medias poblacionales [$IC(\mu_1 - \mu_2)$] está dado por:

$$IC(\mu_1 - \mu_2) = (\bar{X}_1 - \bar{X}_2) \pm t_{1-\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$; para el caso que se suponen las varianzas diferentes.

$$IC(\mu_1 - \mu_2) = (\bar{X}_1 - \bar{X}_2) \pm t_{1-\alpha/2}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$; para el caso que se suponen las varianzas iguales.

En ambos casos se utiliza el percentil de la distribución t con ($n_1 + n_2 - 2$) grados de libertad. En el caso en que se suponen varianzas iguales se calcula la varianza

muestral de manera conjunta para ambas muestras y se denota por s^2_p, la cual se expresa como:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

– Para el caso de la diferencia entre dos proporciones de dos poblaciones estimadas $(\pi_1 - \pi_2)$:
  - El error estándar $[EE(\pi_1 - \pi_2)]$ de la diferencia de dos proporciones se expresa como:

$$EE(\pi_1 - \pi_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

  - El error de muestreo $[e(\pi_1 - \pi_2)]$ o margen de error de la diferencia de dos proporciones se expresa como:

$$e(\pi_1 - \pi_2) = z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

  - El intervalo de confianza para la diferencia de dos proporciones poblacionales $[IC(\pi_1 - \pi_2)]$ está dado por:

$$IC(\pi_1 - \pi_2) = (p_1 - p_2) \pm z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Otra forma de poder expresar aspectos sobre los intervalos de confianza en la inferencia estadística:

– Intervalos de confianza (una población, una variable):
  - $\bar{x} - E < \mu < \bar{x} + E$, una media (con σ conocida), donde:

$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, para poblaciones infinitas, grandes;

$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$, para poblaciones finitas, pequeñas.

  - $\bar{x} - E < \mu < \bar{x} + E$, una media (con σ desconocida), donde:

$$E = z_{\alpha/2}\frac{S}{\sqrt{n}}$$

  - $\hat{p} - E < \mu < \hat{p} + E$, una proporción, donde:

$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$, para población infinita o muy grande y se realiza muestreo con reemplazo;

$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\sqrt{\frac{N-n}{N-1}}$, para población finita o cuando se realiza muestreo sin reemplazo.

- $\frac{(n-1)s^2}{\chi^2_R} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L}$ una varianza, cuando n ≤ 101, o sea, para gl = n − 1 ≤ 100.

En caso contrario; n > 101, o sea para gl = n − 1 >100, se puede aproximar χ^2_I y χ^2_D utilizando la expresión:

$\chi^2 = \frac{1}{2}\left(\pm z_{\alpha/2} + \sqrt{2k-1}\right)^2$; donde k = n − 1 (grados de libertad).

– Intervalo de confianza (una población, dos variables cuantitativas):
  - Coeficiente de correlación lineal simple poblacional estimado (ρ).

    $\frac{e^{2w_I}-1}{e^{2w_I}+1} < \rho < \frac{e^{2w_D}-1}{e^{2w_D}+1}$

    donde:

    $w_I = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - z_{\alpha/2}\cdot\frac{1}{\sqrt{n-3}}$;

    $w_D = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) + z_{\alpha/2}\cdot\frac{1}{\sqrt{n-3}}$.

  - Coeficiente de regresión lineal simple poblacional estimado (β):

    $\beta = \left[b \pm \left(t_{1-\alpha/2,n-2}\right)\sqrt{\frac{s^2}{(xx)}}\right]$

  - Ecuación de la recta de regresión lineal simple poblacional estimada (y):

    $\hat{y} - E < y < \hat{y} + E$,

    donde:

    $E = t_{\alpha/2}s_e\sqrt{1+\frac{1}{n}+\frac{n(x_0-x)^2}{n\left(\sum x^2\right)-\left(\sum x\right)^2}}$;

    $s_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$ o $s_e = \sqrt{\frac{\sum y^2 - b_0\sum y - b_1\sum xy}{n-2}}$

  - Para el intercepto de una recta de regresión ($y = \beta_0 + \beta_1 x$):

    $b_0 - E < \beta_0 < b_0 + E$

    donde:

    $E = t_{\alpha/2}s_e\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}}$.

  - Para el intercepto de una recta de regresión ($y = \beta_0 + \beta_1 x$):

    $b_1 - E < \beta_1 < b_1 + E$

    donde:

$$E = t_{\alpha/2} \cdot \frac{s_e}{\sqrt{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}} \cdot$$

– Intervalos de confianza (dos poblaciones, una variable):

- Para la diferencia de dos medias no pareadas con σ^1 y σ^2 conocidas:
  $(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E$
  donde:

$$E = t_{\alpha/2}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} .$$

- Para la diferencia de dos medias no pareadas con σ^1 y σ^2 desconocidas, pero se suponen iguales:
  $(\bar{x}_{1-} \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_{1-} \bar{x}_2) + E$
  donde:

$$E = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}};$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} .$$

- Para la diferencia de dos medias no pareadas con σ_1 y σ_2 desconocidas, pero no se suponen iguales:
  $(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E$
  donde:

$$E = t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} .$$

- Para la diferencia de dos medias correspondientes a dos muestras pareadas:
  $\bar{d} - E < \mu d < \bar{d} + E;$
  donde:
  $\bar{d}$: media muestral de todas las diferencias (di) entre las dos observaciones o las parejas de datos de las dos muestras pareadas y se expresa como:

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{\sum_{i=1}^{n} (x_i - x_2)}{n}$$

$$E = t_{\alpha/2}\frac{S_d}{\sqrt{n}}$$

- Para la diferencia de dos proporciones:
  $(\hat{p}_1 - \hat{p}_2) - E < (p_1 - p_2) < (\hat{p}_1 - \hat{p}_2) + E;$
  donde:

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

– Determinación del tamaño de muestra para diseños de estudio empleando muestreo aleatorio simple (donde el efecto de diseño tiene valor de 1) que emplean estimadores como:

  • Una proporción con $\hat{p}$ y $\hat{q}$ desconocidas y estimadas para un valor de 0,5 para cada una, que aporta su multiplicación por 0,25, lo cual está en la fórmula:

  $$n = \frac{(z_{\alpha/2})^2 \cdot 0{,}25}{E^2};$$

  donde:

  E: margen de error para la proporción muestral que el investigador asume. Se emplea para una población infinita o muy grande o cuando el muestreo es con reeemplazo.

  • Una proporción con $\hat{p}$ y $\hat{q}$ conocidas y estimadas a partir de un estudio anterior, estudio piloto o criterio de experto:

  $$n = \frac{(z_{\alpha/2})^2 \hat{p}\hat{q}}{E^2}$$

  donde:

  E: margen de error para la proporción muestral que el investigador asume. Se emplea para una población infinita o muy grande o cuando el muestreo es con reemplazo.

  • Una proporción con $\hat{p}$ y $\hat{q}$ conocidas y estimadas a partir de un estudio anterior, estudio piloto o criterio de experto:

  $$n = \frac{N\hat{p}\hat{q}(z_{\alpha/2})^2}{\hat{p}\hat{q}(z_{\alpha/2})^2 + (N-1)E^2}$$

  donde:

  E: margen de error para la proporción muestral que el investigador asume. Se emplea para una población finita o pequeña o cuando el muestreo es sin reemplazo.

– Cálculo del estimado de proporción poblacional (P) y de margen de error (E) desde un intervalo de confianza:

  • Estimado puntual de proporción poblacional (P) a partir de la proporción muestral ($\hat{p}$):

  $$\hat{p} = \frac{(\text{Límite de confianza superior}) + (\text{Límite de confianza inferior})}{2}.$$

  • Estimado de margen de error (E) para la proporción poblacional estimada):

  $$E = \frac{(\text{Límite de confianza superior}) + (\text{Límite de confianza inferior})}{2}.$$

- Una media con σ conocida para una población infinita o muy grande o cuando el muestreo es con reemplazo:

$$n = \left(\frac{z_{\alpha} / 2\sigma}{E} \right)^2 ;$$

donde:

E: margen de error para la media poblacional estimada de la muestral que el investigador asume.

- Una media con σ conocida para una población finita o pequeña o cuando el muestreo es sin reemplazo:

$$n = \frac{N\sigma^2 (z_{\alpha/2})^2}{(N-1)E^2 + \sigma^2 (z_{\alpha/2})^2} ;$$

donde:

E: margen de error para la media poblacional estimada de la muestral que el investigador asume.

- Para la diferencia de dos medias poblacionales estimadas de dos muestrales, el valor de n corresponde a una de las muestras por lo que se debe multiplicar por 2 para calcular el n de ambas muestras de igual tamaño:

$$n = \frac{2s^2 (z_{\alpha/2} + z_{\beta})}{d^2}$$

- Para dos proporciones (para la diferencia de dos proporciones poblacionales estimadas de dos muestrales, el valor de n corresponde a una de las muestras por lo que se debe multiplicar por 2 para calcular el n de ambas muestras de igual tamaño:

$$n = \frac{2p_M q_M (z_{\alpha/2} + z_{\beta})^2}{(p_A - p_B)^2}$$

donde:

$$n = \frac{p_A - p_B}{2} .$$

Nota: Los procedimientos para calcular el tamaño muestral necesario para poder estimar σ^2 son mucho más complejos que los procedimientos resumidos anteriormente. En la tabla 3.27 se presentan los elementos a tomar en cuenta para dicho objetivo:

Por ejemplo, con un nivel de confianza del 95 % y un error del 10 % para la σ^2 se corresponde con la selección al azar de una muestra de 805 para permitir realizar inferencia a partir de la varianza poblacional estimada.

Tabla 3.27. Aspectos relacionados con el tamaño de muestra para poder estimar σ^2

Para tener una confianza del 95 % de que la varianza muestral (s^2) está dentro del:	Del valor de σ^2, el tamaño muestral n debe ser Al menos de:
1 %	77 207
5 %	3 148
10 %	805
20 %	210
30 %	97
40 %	56
50 %	37

La expresión: "un intervalo de confianza del 95 % es 29,5 % ± 3,2 %", significa que el valor 29,5 5 es el porcentaje que se obtiene de la muestra para cierta característica en una variable y lo se denomina *estimación puntual*; aunque si se tomar otra muestra de la misma población puede ser distinto este valor; pero parecido en torno al mismo. La cuantificación de ese entorno a este valor puntual, en este caso del ± 3,2 %, es lo que se llama *margen de error*. Si se añade y se resta el mismo al valor puntual, el porcentaje que se estima no es "exactamente" el 29,5; sino un valor probablemente comprendido entre 26,3 y 32,7 (29,5 % ± 3,2 %). Si el 95 % fuese el nivel de confianza la interpretación de este resultado es:

- Interpretación probabilística: si se repite el muestreo, aproximadamente el 95 % de los intervalos de confianza construidos mediante las fórmulas respectivas incluyen el valor del parámetro poblacional, en este caso la proporción poblacional.
- Interpretación práctica: se tiene un 95 % de confianza que el intervalo único calculado, contenga el valor del parámetro poblacional, en este caso la proporción poblacional.

Estadígrafos que permiten determinar relaciones o asociaciones entre variables

Como parte del procesamiento estadístico de la información recogida por el investigador se suele emplear estadígrafos que permiten determinar relaciones o establecer posibles asociaciones entre las variables. Entre los estadísticos (estadígrafos) o indicadores de asociación entre variables cuantitativas podemos mencionar:

- Coeficiente de correlación lineal simple muestral (momento producto de Pearson) (r).
- Coeficiente de determinación (r^2).
- Coeficiente de correlación por rangos de Spearman poblacional estimado (ρS) o (rS) poblacional muestral.
- Coeficiente de correlación lineal múltiple (R).

– Coeficiente de determinación múltiple o correlación parcial (R2).
– Coeficiente de regresión lineal simple muestral (b).
– Coeficiente de regresión lineal múltiple (β).

Coeficiente de correlación lineal simple muestral (momento producto de Pearson)

Se emplea para determinar el grado o intensidad de la asociación de la variable dependiente con la independiente, ambas cuantitativas, sus posibles valores son 1< r <1 y viene expresado matemáticamente por la expresión:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}};$$

$$r = \frac{\sum\left(x - \bar{x}\right)\left(y - \bar{y}\right)}{(n-1)s_x s_y} \quad \text{o} \quad r = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}}$$

donde:

n: representa el número de pares de datos presentes;

∑: denota la suma de los elementos indicados;

∑x: denota la suma de todos los valores de x;

$\sum x^2$: indica que cada valor de x debe elevarse al cuadrado y después deben sumarse esos cuadrados;

$(\sum x)^2$: indica que los valores de x deben sumarse y el total elevarse al cuadrado;

∑xy: indica que cada valor de x debe multiplicarse primero por su valor y correspondiente. Después de obtener todos estos productos se calcula su suma;

r: representa el coeficiente de correlación lineal de una muestra;

ρ: se usa para representar el coeficiente de correlación lineal poblacional;

sX: expresa la desviación estándar para los datos muestrales de x;

sy: expresa la desviación estándar para los datos muestrales de y.

La forma de interpretar su resultado sería:

– Valor 0: no correlación.
– Valores que se acercan a 1 o –1: hay intensidad en la correlación.
– Valor 1: correlación perfecta positiva, cuando una variable varía la otra lo hace en la misma intensidad en el mismo sentido.
– Valor –1: correlación perfecta negativa, cuando una variable varía la otra lo hace con la misma intensidad, pero en sentido inverso.
– Valor menor de 0,30: correlación o asociación débil.
– Valor mayor o igual a 0,30 y menor o igual a 0,70: correlación o asociación moderada.
– Valor mayor de 0,70: correlación o asociación fuerte.

El signo determina que si es positivo (+) las dos variables aumentan o disminuyen y si el signo es negativo (–) significa que a medida que una aumenta la otra disminuye.

Para evaluar la significación de su valor mediante el contraste de hipótesis se puede comparar su valor observado con el valor crítico (téorico-esperado) presente en la tabla de referencia específica (véase el anexo 1) en muestras pequeñas, así como se puede emplear la distribución de probabilidades t-Student, para comparar el valor observado-calculado del estadístico de prueba según la expresión:

$$\text{t-student r (observado-calculado)} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}},$$

con el respectivo valor téorico-esperado y tabulado de esta distribución para n – 2 grados de libertad según el valor de alfa preseleccionado.

Coeficiente de concordancia de rangos de Kendall

Si se desea examinar el grado de concordancia para n mayor o igual a dos conjuntos de datos cuantitativos expresados por sus rangos (correspondientes, por ejemplo, con la evaluación realizada por un número de calificadores) a partir del comportamiento de cierta característica mediante una variable cuantitativa (por ejemplo, tratamientos en un ensayo clínico) se puede emplear un índice de correlación denominado coeficiente de concordancia de rangos de Kendall, que se denota como r_{CK}. Este expresa un índice de la divergencia de la concordancia de los k conjuntos de datos rangueados analizados en relación con una concordancia perfecta, la máxima posible entre los distintos calificadores. Un valor cercano a 1 indica alta concordancia y poca divergencia. Se puede aplicar el coeficiente de concordancia de rangos de Kendall a los datos que se puedan disponer en una distribución bivariada de frecuencias, en que las filas rangueadas están representando a la variable: calificadores y las columnas están encabezadas por la variable: tratamientos (Tabla 3.28).

Tabla 3.28. Distribución bivariada de frecuencias de k columnas-tratamientos y filas rangueadas

Calificadores	**Tratamientos**			
	T_1	**T_2**	**…**	**T_k**
C_1	$rgf(X_{11})$	$rgf(X_{21})$	…	$rgf(X_{k1})$
C_2	$rgf(X_{12})$	$rgf(X_{22})$	…	$rgf(X_{k2})$
⋮	⋮	⋮	…	⋮
C_n	$rgf(X_{1n})$	$rgf(X_{2n})$	…	$rgf(X_{kn})$
Suma de rangos	R_1	R_2		R_k

donde:

rgf(Xij): indica rango por fila y representa el rango asignado por el calificador j al tratamiento i, para i = 1, 2,..., k y j = 1, 2,..., n;

$R_i = \sum_{i=1}^{n} rgf(x_{ij})$: la suma hasta la k-ésima columna después de rangueada la i-ésima fila, para i = 1, 2,..., n;

r_{CK}: coeficiente de concordancia de rangos de Kendall, que se basa en la comparación de las sumas Ri de los rangos de las columnas i = 1, 2,..., k; si varias de ellas son iguales o muy semejantes esto indica falta de concordancia entre los conjuntos rangueados referidos para la variable: calificadores.

La formulación matemática del coeficiente de concordancia de rangos de Kendall es:

$$r_{CK} = \frac{12S^2}{k^2(n^2-1)}$$

Donde:

k: número total de tratamientos, pruebas realizadas o de muestras formadas;

n: número total de los calificadores, evaluadores o analistas que se corresponde con el tamaño de cada muestra o grupo calificado;

$\bar{R}$: a media de las k columnas de rangos: $R_1, R_2,..., R_k$ después de ranguear todas las filas;

S^2: varianza de las sumas de las columnas Ri después de ranguear las filas.

El promedio de los r_{Sp} al emplear más de dos variables cuantitativas se puede utilizar también para evaluar la concordancia de rangos al considerarse un coeficiente de concordancia y se relaciona con el coeficiente de concordancia de rangos de Kendall mediante esta formulación matemática:

$$r_{CK} = \frac{(k-1)r_{sp}+1}{k}$$

Para el caso en que se presenten ligaduras o repeticiones en los rangos asignados de un mismo calificador, el valor del coeficiente de concordancia de rangos de Kendall disminuye, en caso de presentarse un número elevado de ligaduras se requiere emplear esta formulación corregida:

$$r_{CK} = \frac{12S^2}{k^2(n^2-1) - 12k\sum_{j=1}^{n} t_j(t_j^2-1)}$$

donde:

t_j: representa la cantidad de observaciones muestrales de la j-ésima ligadura o repetición del valor rangueado del calificador i-ésimo, para i = 1, 2,..., k; j = 1, 2,..., n;

Para evaluar la significación de la concordancia entre los calificadores o evaluadores a partir de un procedimiento realizado en la población de la que se extrajo la muestra por medio de este coeficiente, empleando el contraste de hipótesis se debe consultar la tabla

de niveles de significación para los valores críticos, específicamente de la varianza de las sumas de rangos (S^2) para n menor de ocho (véase el anexo 1); mientras que para n mayor de siete se apoya en la dócima o prueba de significación de rangos de Friedman con empleo de la distribución chi-cuadrado y el estadístico de prueba que involucra al coeficiente de concordancia de rangos de Kendall como:

$\chi^2 = k(n-1)\, r_{CK}$ con k – 1 grados de libertad para el alfa preseleccionado, donde la decisión de rechazar la hipótesis nula de no concordancia depende del análisis de la región critica determinada por la relación:

$Pr\{\chi^2(\text{observado-calculado}) > \chi^2_{k-1,\,1-\alpha}\ (\text{tabulado-esperado})\}$

En el primero de los casos el criterio para determinar la significación en la concordancia de los rangos asignados por los calificadores o evaluadores es:

Si S^2 (observada-calculada) > $S^2_{(k,\,n)\,1-\alpha}$ (tabulada-esperada), entonces se rechaza la hipótesis nula de no concordancia entre los grupos de datos analizados por calificadores o evaluadores a nivel poblacional.

Coeficiente de correlación punto biserial

El coeficiente de correlación punto biserial (rbis) constituye un índice de asociación que permite correlacionar a dos variables de las cuales una es cualitativa dicotómica y la otra es cuantitativa con expresión de sus datos en escala de intervalo. Su fórmula matemática es:

$$r_{bis} = \frac{\overline{Y}p - \overline{Y}q}{S_Y}\sqrt{pq}$$

donde:

Y: variable cuantitativa en escala de intervalo;

p: proporción de uno de los valores en la variable dicotómica X;

q: proporción del otro valor en la variable dicotómica X, q = 1 – p;

$\overline{Y}p$: media de los valores de Y correspondientes a los valores de X cuya proporción es p en la variable dicotómica X;

$\overline{Y}q$: media de los valores de Y correspondientes a los valores de X cuya proporción es q en la variable dicotómica X;

S_y: desviación estándar de los valores de la variable Y.

Para evaluar la significación de la correlación por medio de este coeficiente empleando el contraste de hipótesis se debe consultar la tabla de niveles de significación para los valores críticos para n – 2 grados de libertad y el nivel de confianza seleccionado (véase el anexo 1), donde la toma de decisión está determinada por este criterio: si r_{bis} (observado-calculado) > r_{bis} (n – 2) 1 – α (tabulada-esperada), entonces se rechaza la hipótesis nula de no correlación significativa a nivel poblacional entre ambas variables.

Coeficiente de determinación

Expresa el porcentaje de variación o cambios en la variable dependiente debido a la variación o influencia de la independiente (ambas variables cuantitativas), es decir, cuantifica el ajuste de la recta. Sus posibles valores son $0 < r^2 < 100$ y la forma de interpretarlo es si $(r^2) > 0{,}50$ o 50 %, el ajuste de la recta es bueno, es decir, más de la mitad de los cambios en la dependiente dependen de la independiente.

Coeficiente de correlación por rangos de Spearman poblacional estimado o poblacional muestral

Al igual que el momento producto de Pearson el coeficiente de correlación por rangos de Spearman poblacional estimado (ρ_S) o poblacional muestral (r_S) expresa el grado o intensidad de la asociación de la variable dependiente con la independiente, ambas cuantitativas (con distribución libre los datos), solo que los valores de cada variable son sustituidos por el rango o lugar que ocupan dentro de la serie ordenada de datos, es empleada la diferencia entre los rangos para las dos variables cuantitativas ordenadas en la expresión matemática:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

cuando no hay empates en los valores expresados en rangos de las variables, es decir, no hay valores repetidos en las variables;

$$r_s = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

cuando hay empates en los valores expresados en rangos de las variables, es decir, hay valores repetidos en las variables y los valores de las variables x, y están expresados en rangos en la fórmula.

donde:

n: número de pares de datos muestrales;

d: diferencia entre los rangos de los dos valores dentro de un par de datos de las variables.

La interpretación de sus resultados se toma a partir de la cercanía al valor 1, mientras mayor esta, mayor es la correlación entre ambas series de datos de las variables cuantitativas de distribución libre estudiadas.

Para evaluar la significación de su valor mediante el contraste de hipótesis se puede comparar su valor observado con el valor crítico (téorico-esperado) presente en la tabla

de referencia específica para muestras pequeñas (n ≤ 30) (véase el anexo 1); así como se puede emplear la distribución de probabilidades t-student para n > 30, al comparar el valor observado-calculado del estadístico de prueba según la expresión:

$$\text{t-student r (observado-calculado)} = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

con el respectivo valor téorico-esperado tabulado para esta distribución para n – 2 grados de libertad según el valor de alfa preseleccionado.

Coeficiente de correlación lineal múltiple

Expresa el grado o intensidad de la asociación (correlación) de la variable dependiente (cuantitativa) y todas las variables independientes (cuantitativas) tomadas en conjunto. Su valor es 0< R <1, cuanto más alto sea su valor las variables independientes están más correlacionadas con la variable dependiente y explican en mayor medida sus fluctuaciones (varianza). Sus posibles valores a interpretar son:

- Valor 0: no hay correlación entre las variaciones de las variables independientes y la dependiente.
- Valores que se acercan a 1 o –1: hay intensidad en la correlación de la dependiente con las independientes.
- Valor 1: correlación perfecta positiva, cuando varían las variables independientes en su conjunto lo hace en la misma intensidad y en el mismo sentido la dependiente.
- Valor –1: correlación perfecta negativa, cuando varían las variables independientes en su conjunto lo hace con la misma intensidad, pero en sentido inverso la dependiente.

El signo determina que si es positivo (+) las variables independientes y dependiente aumentan o disminuyen a la vez y si el signo es negativo (–) significa que a medida que unas aumentan la otra disminuye.

Coeficiente de determinación múltiple o correlación parcial

Expresa la varianza explicada de la variable dependiente por todas las independientes, es decir, el porcentaje de variación en la dependiente es debido a todas las independientes en conjunto consideradas en el análisis. Sus posibles valores son $0 < R^2 < 100$ y la forma de interpretarlo es si $(R^2) > 0{,}50$ o 50 % más de la mitad de los cambios en la variable dependiente se debe a las variaciones en las independientes de conjunto.

Coeficiente de regresión lineal simple muestral

Expresa cuanto cambia en promedio el valor de la variable dependiente por cada unidad que cambia la independiente. Es la pendiente o ángulo de inclinación de la recta

mínimo cuadrática con respecto al eje de las x que mejor ajusta o representa la relación gráfica entre ambas variables. Sus posibles valores son:

- Valores positivos: la recta asciende, las dos variables aumentan o disminuyen a la vez, es decir, el aumento o disminución de una unidad en la variable independiente, provoca el aumento o disminución respectiva de valores en la dependiente.
- Valores negativos: la recta desciende, el aumento en una unidad de la variable independiente provoca la disminución respectiva de valores de la variable dependiente.

Coeficiente de regresión lineal múltiple

Expresa cuanto puede variar la variable dependiente en término medio producto al cambio de valor en una unidad de una dependiente, si se controla o ajusta en un valor fijo el resto de las independientes.

En la tabla 3.29 se representa a modo de resumen los tipos de coeficientes más frecuentemente empleados en la determinación del tipo de relación bivariada.

Tabla 3.29. Relación entre los tipos de coeficientes y variables que emplean

Tipo de coeficiente	Tipo de variables a emplear	
De correlación lineal simple de Pearson	Cuantitativa con distribución normal en escala de intervalo	Cuantitativa con distribución normal en escala de intervalo
De determinación (lineal simple)	Cuantitativa con distribución normal en escala de intervalo	Cuantitativa con distribución normal en escala de intervalo
De correlación parcial lineal simple	Dos cuantitativas con distribución normal en escala de intervalo	Una o más cuantitativas con distribución normal en escala de intervalo
De correlación con rangos de Spearman	Cuantitativa en escala ordinal con rangos	Cuantitativa en escala ordinal con rangos
De correlación con rangos de Kendall	Cuantitativa en escala ordinal con rangos	Cuantitativa en escala ordinal con rangos
De correlación parcial con rangos de Kendall	Dos cuantitativas en escala ordinal con rangos	Una o más cuantitativas en escala ordinal con rangos
De concordancia de rangos de Kendall	Cuantitativa en escala ordinal con rangos	Cualitativa nominal
De correlación punto biserial	Cuantitativa en escala de intervalo	Cualitativa dicotómica
De correlación lineal múltiple	Cuantitativa en escala de intervalo	Dos o más cuantitativas en escala de intervalo
De determinación (lineal múltiple)	Cuantitativa en escala de intervalo	Dos o más cuantitativas en escala de intervalo

Tabla 3.29. Continuación...

Tipo de coeficiente	Tipo de variables a emplear	
De regresión lineal simple	Dependiente cuantitativa en escala de intervalo	Independiente cuantitativa en escala de intervalo
De regresión lineal múltiple	Dependiente Cuantitativa en escala de intervalo	Independientes cuantitativas en escala de intervalo
De correlación V de Cramer	Cualitativa nominal polinominal	Cualitativa polinominal o dicotómica
De correlación φ	Cualitativa dicotómica	Cualitativa dicotómica

Entre los posibles estadísticos o indicadores de asociación entre variables cualitativas están:

- OR (razón de momios, de disparidad, de desventaja, de productos cruzados, de chances, de oportunidades relativas de enfermedad, de oportunidades relativas de exposición u *odds ratio*): es una medida de asociación relativa entre variables cualitativas y expresa la relación por un cociente entre cierta oportunidad para un efecto y la oportunidad opuesta para otro efecto. Se interpreta como:
  - Si OR > 1: asociación positiva, la exposición o presencia del probable factor causal (variable independiente) está asociado, guarda relación o puede comportarse como una causa o factor de riesgo para el efecto relacionado (variable dependiente).
  - Si OR < 1: asociación negativa, la exposición o presencia del probable factor causal (variable independiente) está asociado, guarda relación o puede comportarse como una causa o factor negativo o protector para el efecto relacionado (variable dependiente).
  - Si OR = 1: no asociación, la exposición o presencia del probable factor causal (variable independiente) no está asociado, no guarda relación o no puede comportarse como una causa o factor para el efecto relacionado (variable dependiente).
- Riesgo relativo o razón de riesgo (RR) o razón de prevalencias en enfermos según exposición (RP): expresa una medida de la intensidad de la asociación entre el factor causal y el efecto correspondiente, representa la mayor cantidad de veces de tener un efecto ante la presencia del factor causal o también las veces más probables de tener el efecto para una condición o categoría del factor respecto a su otra condición.
- Diferencia de riesgo o exceso de riesgo o riesgo atribuible absoluto (RA): expresa una medida del impacto que provoca el factor causal para cierto efecto, significa el beneficio ganado en tasa, en cuantas unidades se reduciría la tasa en los expuestos si se elimina el factor de riesgo.
- Riesgo atribuible proporcional en expuestos (RAP exp): expresa una medida del impacto proporcional que provoca el factor causal para cierto efecto, significa el

beneficio proporcional ganado en tasa, en cuantas unidades se reduciría proporcionalmente la tasa en los expuestos si se elimina el factor de riesgo.

- Riesgo atribuible poblacional (RAP): expresa el exceso de mortalidad o morbilidad por la presencia del factor de riesgo, pero en toda la población, lo que reduce la tasa general de no existir el factor de riesgo.
- Riesgo atribuible proporcional poblacional (RAP pob): expresa una medida del impacto proporcional que provoca el factor causal en la población (para enfermos respecto a los no enfermos), significa el beneficio ganado en tasa, en cuantas unidades se reduciría la tasa de incidencia poblacional de enfermos proporcionalmente si se elimina el factor de riesgo.
- Riesgo total (RT): expresa una medida del impacto en toda la población que provoca el factor causal para cierto efecto.
- Fracción de prevención para expuestos (FPE) porcentual o porcentaje de protección puntual por la exposición) (FPE%) o porcentaje (puntual) de protección de expuestos en la población:
  - FPE% = PF_e = 1 – IR a partir de la densidad de incidencia en estudio de cohorte.
  - FPE% = PF_e = 1 – RR a partir de la incidencia acumulada en estudio de cohorte.
  - FPE% = PF_e = 1 – OR a partir del OR de exposición en estudio de caso-control.
- Fracción de prevención máxima en expuestos porcentual o porcentaje de protección máximo posible (%$FPE_{Máx}$) o porcentaje máximo posible de protección por la exposición al factor en los expuestos:
  - %$FPE_{Máx}$ = (1 – valor IC inferior FPE o PFe a partir del IR) · 100, para estudio de cohorte por densidad de incidencia con intervalo de confianza.
  - %$FPE_{Máx}$ = (1 – valor IC inferior FPE o PFe a partir del RR) · 100, para estudio de cohorte por incidencia acumulada.
  - %$FPE_{Máx}$ = (1 – valor IC inferior FPE o PFe a partir del OR) · 100, para estudio de caso-control.
- Fracción de prevención mínima en expuestos porcentual o porcentaje de protección mínimo posible (%$FPE_{Mín}$) o porcentaje mínimo posible de protección por la exposición al factor en los expuestos:
  - %$FPE_{Mín}$ = (1 – valor IC superior FPE o PFe a partir del IR) · 100, para estudio de cohorte por densidad de incidencia.
  - %$FPE_{Mín}$ = (1 – valor IC superior FPE o PFe a partir del RR) · 100 para estudio de cohorte por incidencia acumulada.
  - %$FPE_{Mín}$ = (1 – valor IC superior FPE o PFe a partir del OR) · 100 para estudio de caso-control.
- Fracción de prevención para la población porcentual (FPP%) o porcentaje de protección puntual para la población por la exposición al factor protector:
  - $FPP = PF_p = PF_e \frac{T_1}{T} \cdot 100$, a partir de la densidad de incidencia en estudios de cohorte.

- $FPP = PF_p = PF_e \frac{n_1}{n} \cdot 100$, a partir de la incidencia acumulada en estudios de cohorte
- $FPP = PF_p = \frac{p_1 PF_e}{p_1 PF_e + OR} \cdot 100$, a partir del OR de exposición en estudio de caso-control).

Nota: Las abreviaturas empleadas en las respectivas formulaciones pueden ser confrontadas al abordar los tipos de medidas para variables cualitativas al inicio de este capítulo.

Análisis de riesgo en estudios analíticos observacionales

Si tenemos una tabla de contingencia o cruzada (de filas por columnas) (Tabla 3.30), por ejemplo, al relacionar las variables presencia de exposición (independiente) y presencia de enfermedad (dependiente).

Tabla 3.30. De contingencia 2x2 simple con enfoque de riesgo de exposición

Variable independiente (exposición a:)	Variable dependiente (presencia de enfermedad)		Totales
	Presente (enfermo)	Ausente (sano)	
Expuestos	A	b	a + b
No expuestos	C	d	c + d
Total	a + c	b + d	n = a + b + c + d

Se pueden efectuar las siguientes determinaciones:

1. Cálculo de tasas de exposición según enfermedad o de tasas de enfermedad según exposición:
   a) Tasa de exposición en casos o enfermos (TExpEnf) para la población estimada:
   TExpEnf = a/(a + c)
   b) Tasa de exposición en controles o sanos (TExpSan) para la población estimada:
   TExpSan = b/(b + d)
   c) Tasa de prevalencia o incidencia acumulada (de enfermedad) en expuestos (TIncAcExp) para la población estimada:
   TIncAcExp = a/(a + b)
   d) Tasa de prevalencia o incidencia acumulada (de enfermedad) en no expuestos (TIncAcNoExp) para la población estimada:
   TIncAcNoExp = c/(c + d)

2. Prueba de hipótesis (bilateral) para la diferencia de dos proporciones ($p_1 - p_2$) empleando tablas de contingencia 2x2 simples (dos filas por dos columnas):
   a) De expuestos según la aparición de enfermedad o no (de estas proporciones) para muestras grandes (pqn > 5), con empleo de la aproximación de la distribución normal por el estadígrafo de prueba: Z ~ N(0,1) y con su aplicación a la diferencia de proporciones como:

      $p_1 - p_2 \sim N\,[0, p\,(1 - p)(1/n_1 + 1/n_2)]$.

      El posible planteamiento de hipótesis bilateral sería:

      - H_0: $P_1[a/(a + c)] = P_2[b/(b + d)]$ o también H_0: $P_1 - P_2 = 0$.
      - H_1: $P_1[a/(a + c)] \neq P_2[b/(b + d)]$ o también H_1: $P_1 - P_2 \neq 0$, para un nivel de significación alfa (de cometer error tipo 1), determinándose para dos colas en este caso (IC para Z~N(0, 1) entre $\alpha/2$ y $1 - \alpha/2$ bajo H_0), son los valores de percentiles de Z observados-calculados fuera de este intervalo, los que permiten rechazar H_0, es decir, hay diferencias estadísticamente significativas entre las dos proporciones comparadas.

        Se parte en los cálculos de:

$$Z = \frac{(p_1 - p_2)}{\sqrt{VARcomb}} = \frac{(p_1 - p_2)}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

        donde:

$$p_1 = \frac{a}{a + c}$$

$$p_2 = \frac{b}{b + d}$$

        VARcomb: varianza combinada para las proporciones p_1 y p_2;

$$VARcomb = pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = pq\left(\frac{1}{a + c} + \frac{1}{b + d}\right)$$

        p: proporción combinada para las proporciones p_1 y p_2;

$$p = \frac{a + b}{a + c + b + d} \quad \text{y } q = 1 - p.$$

        Para la toma de decisiones respecto al contraste de hipótesis se toma en cuenta:

- Si $p_1 - p_2 > 0$, se espera que Z observado-calculado > 0, entonces se compara el mismo con $Z_{1 - \alpha/2}$ teórico.
- Si $p_1 - p_2 < 0$, se espera que Z observado-calculado < 0, entonces se compara el mismo con $Z\alpha/2$ teórico.

Regla de decisión:

– Para Z observado-calculado ≥ 0 y el valor de alfa según el nivel de confianza seleccionado:
  - Si Z observado-calculado > Z esperado-tabulado según el valor de alfa (1 – $\alpha/2$ si es contraste bilateral o α si es contraste unilateral), entonces el valor de Z observado se sitúa en el área de rechazo para H_0, por lo que se rechaza H_0 y se acepta la alternativa, es decir, que existen evidencias para pensar en diferencias significativas entre las proporciones de los enfermos expuestos y los no enfermos expuestos, donde el valor de p bajo el criterio de H_0 correspondiente al valor de Z observado o la probabilidad del azar para este resultado es menor que el valor $\alpha/2$ o α según el tipo de contraste (bilateral o unilateral) seleccionado para el nivel de confianza elegido correspondiente, lo que disminuye la probabilidad de cometer el error tipo 1 (de equivocación al rechazar H_0 cuando realmente no es rechazable (p-valor < α).
  - Si Z observado-calculado ≤ Z esperado-tabulado según el valor de alfa (1 – $\alpha/2$, si es contraste bilateral o α si es contraste unilateral); entonces el valor de Z observado no se sitúa en el área de rechazo para H_0, por lo que no se rechaza H_0 y no puede aceptar la alternativa, es decir, que no existen evidencias para pensar en diferencias significativas entre las proporciones de los enfermos expuestos y los no enfermos expuestos, donde el valor de p bajo el criterio de H_0 correspondiente al valor de Z observado o la probabilidad del azar para este resultado es igual o mayor que el valor $\alpha/2$ o α según el tipo de contraste (bilateral o unilateral) seleccionado para el nivel de confianza elegido correspondiente, lo que aumenta la probabilidad de cometer el error tipo 1 (de equivocación al rechazar H_0 cuando realmente no es rechazable (p-valor ≥ α)).

– Para Z observado-calculado < 0 y el valor de alfa según el nivel de confianza seleccionado:
  - Si Z observado-calculado < Z esperado-tabulado según el valor de alfa (1 – $\alpha/2$ si es contraste bilateral o α si es contraste unilateral); entonces el valor de Z observado se sitúa en el área de rechazo para H_0, por lo que se rechaza H_0 y se acepta la alternativa, es decir, que existen evidencias para pensar en diferencias significativas entre las proporciones de los enfermos expuestos y los no enfermos expuestos, donde el valor de p bajo el criterio de H_0 correspondiente al valor de Z observado o la probabilidad del azar para este resultado es menor que el valor $\alpha 2$ o α según el tipo de contraste (bilateral o unilateral) seleccionado para el nivel de confianza elegido correspondiente, lo que disminuye la probabilidad de cometer el error tipo 1 (de equivocación al rechazar H_0 cuando realmente no es rechazable (p-valor < α)).

- Si Z observado-calculado ≥ Z esperado-tabulado según el valor de alfa (1 – $\alpha/2$ si es contraste bilateral o α si es contraste unilateral); entonces el valor de Z observado no se sitúa en el área de rechazo para H_0; por lo que no se rechaza H_0, y no puede aceptar la alternativa, es decir, que no existen evidencias para pensar en diferencias significativas entre las proporciones de los enfermos expuestos y los no enfermos expuestos, donde el valor de p bajo el criterio de H_0 correspondiente al valor de Z observado o la probabilidad del azar para este resultado es igual o mayor que el valor $\alpha/2$ o α según el tipo de contraste (bilateral o unilateral) seleccionado para el nivel de confianza elegido correspondiente, lo que aumenta la probabilidad de cometer el error tipo 1 (de equivocación al rechazar H_0 cuando realmente no es rechazable (p-valor ≥ α)).

Para recordar:

Al comparar el valor observado-calculado o experimental del estadístico de prueba o contraste *versus* el correspondiente con el valor teórico-esperado tabulado de la distribución de referencia se estima la probabilidad de que pueda ser cierta o no rechazable la hipótesis nula para el valor observado o mayor que el mismo del estadístico de contraste a partir de los datos muestrales empleados. Al valor de esta probabilidad se le llama "valor p" o p-valor y el mismo informa sobre el grado de compatibilidad de los datos con la hipótesis nula ($H_{0)}$.

b) De enfermos según exposición o no (de esas proporciones) para muestras grandes (pqn > 5), con empleo por aproximación a la normal del estadígrafo de prueba Z ~ N(0, 1) y con su aplicación a la diferencia de proporciones como:

$p_1 - p_2 \sim N\,[0, p\,(1 - p)(1/n_1 + 1/n_2)]$

El posible planteamiento de hipótesis bilateral sería:

- H_0: P_1 (a/(a + b)) = P_2 (c/(c + d) o también H_0: $P_1 - P_2 = 0$.
- H_1: P_1 (a/(a + b)) ≠ P_2 (c/(c + d), también H1: $P_1 - P_2 \neq 0$, para un nivel de significación alfa (de cometer error tipo 1), determinándose dos colas en este caso (IC para Z ~ N(0, 1) entre $\alpha/2$ y 1 – $\alpha/2$ bajo H_0), por aproximación a la distribución normal, los valores de percentiles de Z observados fuera de este intervalo son los que permiten rechazar H_0, es decir, hay diferencias estadísticamente significativas entre las proporciones comparadas.

Se parte en los cálculos de:

$$Z = \frac{(p_1 - p_2)}{\sqrt{\text{VARcomb}}} = \frac{(p_1 - p_2)}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

donde:

$$p_1 = \frac{a}{a + b}$$

$$p_2 = \frac{c}{c+d}$$

VARcomb: varianza combinada para las proporciones p_1 y p_2;

$$VARcomb = pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right) = pq\left(\frac{1}{a+b}+\frac{1}{c+d}\right);$$

p: proporción combinada para las proporciones p_1 y p_2;

$$p = \frac{a+c}{a+b+c+d} \quad \text{y } q = 1 - p.$$

Para la toma de decisiones respecto al contraste de hipótesis:

Si $p_1 - p_2 > 0$ se espera que Z observado-calculado > 0, entonces se compara el mismo con $Z_{1-\alpha/2}$ teórico.

Si $p_1 - p_2 < 0$, se espera que Z observado-calculado < 0, entonces se compara el mismo con $-Z_{1-\alpha/2}$ o $Z_{\alpha/2}$ teórico.

Regla de decisión:

– Para Z observado-calculado ≥ 0 y el valor de alfa según el nivel de confianza seleccionado:
  - Si Z observado-calculado > Z esperado-tabulado según el valor de alfa ($1 - \alpha/2$ si es contraste bilateral o α si es contraste unilateral), entonces el valor de Z observado se sitúa en el área de rechazo para H_0; por lo que se rechaza H_0 y se acepta la alternativa, es decir, que existen evidencias para pensar en diferencias significativas entre las proporciones de los enfermos expuestos y los enfermos no expuestos, donde el valor de p bajo el criterio de H_0 correspondiente al valor de Z observado o la probabilidad del azar para este resultado, es menor que el valor $\alpha/2$ o α según el tipo de contraste (bilateral o unilateral) seleccionado para el nivel de confianza elegido correspondiente, lo que disminuye la probabilidad de cometer el error tipo 1 (de equivocación al rechazar H_0 cuando realmente no es rechazable (p-valor < α)).
  - Si Z observado-calculado > Z esperado-tabulado según el valor de alfa ($1 - \alpha/2$ si es contraste bilateral o α si es contraste unilateral); entonces el valor de Z observado no se sitúa en el área de rechazo para H_0; por lo que no se rechaza H_0 y no puede aceptar la alternativa, es decir, que no existen evidencias para pensar en diferencias significativas entre las proporciones de los enfermos expuestos y los enfermos no expuestos, donde el valor de p bajo el criterio de H_0 correspondiente al valor de Z observado o la probabilidad del azar para este resultado, es igual o mayor que el valor $\alpha/2$ o α según el tipo de contraste (bilateral o unilateral) seleccionado para el nivel de confianza elegido correspondiente, lo que aumenta la probabilidad de cometer el error tipo 1 (de equivocación al rechazar H_0 cuando realmente no es rechazable (p-valor ≥ α)).

- Para Z observado-calculado < 0 y el valor de alfa según el nivel de confianza seleccionado:
  - Si Z observado-calculado $<$ Z esperado-tabulado según el valor de alfa ($1 - \alpha/2$ si es contraste bilateral o α si es contraste unilateral); entonces el valor de Z observado se sitúa en el área de rechazo para H_0, por lo que se rechaza H_0 y se acepta la alternativa, es decir, que existen evidencias para pensar en diferencias significativas entre las proporciones de los enfermos expuestos y los enfermos no expuestos, donde el valor de p bajo el criterio de H_0 correspondiente al valor de Z observado o la probabilidad del azar para este resultado es menor que el valor $\alpha/2$ o α según el tipo de contraste (bilateral o unilateral) seleccionado para el nivel de confianza elegido correspondiente, lo que disminuye la probabilidad de cometer el error tipo 1 (de equivocación al rechazar H_0 cuando realmente no es rechazable (p-valor < α)).
  - Si Z observado-calculado ≥ Z esperado-tabulado según el valor de α ($1 - \alpha/2$ si es contraste bilateral o α si es contraste unilateral), entonces el valor de Z observado no se sitúa en el área de rechazo para H_0, por lo que no se rechaza H_0 y no puede aceptar la alternativa, es decir, que no existen evidencias para pensar en diferencias significativas entre las proporciones de los enfermos expuestos y los enfermos no expuestos, donde el valor de p bajo el criterio de H_0 correspondiente al valor de Z observado o la probabilidad del azar para este resultado, es igual o mayor que el valor $\alpha/2$ o α según el tipo de contraste (bilateral o unilateral) seleccionado para el nivel de confianza elegido correspondiente, lo que aumenta la probabilidad de cometer el error tipo 1 (de equivocación al rechazar H_0 cuando realmente no es rechazable (p-valor ≥ α)).

3. Realizar prueba de hipótesis para independencia entre la variable independiente (factor de riesgo o exposición) y la dependiente (riesgo o enfermedad):
   a) Se lleva a cabo cuando se tiene una muestra de n individuos que se clasifican respecto a las dos variables cualitativas, por ejemplo: exposición o no y enfermedad o no, por lo que el planteamiento de hipótesis es:
      - H_0: existe (poblacionalmente) independencia entre las dos variables.
      - H_1: no existe (poblacionalmente) independencia entre las variables.

      Nota: puede servir de apoyo en su análisis el planteamiento de hipótesis para la diferencia entre dos proporciones cuando se emplean dos variables dicotómicas:
      - H_0: P_1exp-enf (a/a + b) = P_2noexp-enf (c/c + d) (son similares las dos proporciones, equivaldría a independencia entre las dos variables).
      - H_1: Pexp-enf (a/a + b) ≠ P_2noexp-enf (c/c + d) (no son similares las dos proporciones equivaldría a dependencia entre las dos variables).

b) Se construye la tabla de contingencia o cruzada (filas por columnas) con los datos correspondientes para las dos variables cualitativas y se realiza la comparación de los valores observados y esperados de cada celda tabulada; empleando el estadígrafo de prueba chi-cuadrado (χ^2) según la expresión:

- En tablas de contingencia (2x2) o cruzadas (filas por columnas para dos variables), la fórmula para el cálculo de chi-cuadrado sería:
  - Fórmula para chi-cuadrado de Pearson según la expresión:

    $\chi^2 Pearson = \frac{n(ad-bc)^2}{m_1 m_0 n_1 n_\partial}$, sigue una distribución teórica chi-cuadrado (χ^2_1) con un grado de libertad para (1 – α)% de nivel de confianza.
  - Fórmula para chi-cuadrado con corrección de Yates (χ^2 Yates) según la expresión:

    $\chi^2 Yates = \frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{m_1 m_0 n_1 n_\partial}$, sigue una distribución teórica chi-cuadrado (χ^2_1) con un grado de libertad para (1 – α)% de nivel de confianza.
- En tablas de contingencia en que el número de filas o columnas sea mayor de 2 es:
  - χ^2 observado-calculado $= \frac{\sum_{i=1}^{f}\sum_{j=1}^{c}(nij-eij)^2}{eij}$, sigue una distribución teórica chi-cuadrado ($\chi^2_{(m-1)(n-1),1-\alpha}$) con grados de libertad = (f – 1)(c – 1) para (1 – α)% de nivel de confianza.

    donde:

    nij: frecuencia absoluta observada en la celda i, j;

    eij: frecuencia absoluta esperada en la celda i, j = (n_{inj})/n;

    ni y nj: totales marginales correspondientes a la fila i-ésima y columna j-ésima;

    n: gran total o total correspondiente a toda la tabla.
  - χ^2 teórico-esperado-tabulado dependiendo de los grados de libertad (gl) = [columnas – 1] x [filas – 1]) y 1 – α prefijado se corresponde con el valor de probabilidad límite entre el área bajo la curva de chi-cuadrado, de rechazo o no para H_0.
  - Valor de p (p-valor bajo H_0) = Pr $\chi^2_{(f-1)(c-1),1-\alpha}$ teórico-tabulado ≥ χ^2 observado-calculado.

Recordar que:

- Respecto al valor p o p-valor: al comparar el valor observado-calculado o experimental del estadístico de prueba o contraste versus el correspondiente con el valor teórico-esperado tabulado de la distribución de referencia, se estima la probabilidad de que puede ser cierta o no rechazable la hipótesis nula

para el valor observado o mayor que el mismo del estadístico de contraste a partir de los datos muestrales empleados. Al valor de esta probabilidad se le llama "valor p" o p-valor e informa sobre el grado de compatibilidad de los datos con la hipótesis nula.

- Respecto a los grados de libertad: son el número de unidades independientes de información en una muestra, que son relevantes para la estimación de un parámetro o para el cálculo de un estadístico.
- La regla de decisión en el contraste de hipótesis es:
  - Si χ^2 observado-calculado > $\chi^2_{(f-1)(c-1),\,1-\alpha}$ teórico-esperado entonces se sitúa en el área de rechazo para H0 el chi-cuadrado observado(χ^2) por lo que se rechaza H_0, es decir, existe dependencia o asociación significativa entre las dos variables cualitativas al estimarse un p-valor bajo H_0 para el valor de chi-cuadrado (χ^2) observado o una probabilidad relacionada con el azar para el resultado obtenido, inferior al valor de αde referencia lo que disminuye la probabilidad de cometer el error tipo 1 (p-valor < α).
  - Si χ^2 observado-calculado ≤ $\chi^2_{(f-1)(c-1),\,1-\alpha}$ teórico-esperado, entonces no se sitúa en el área de rechazo para H_0 el chi-cuadrado observado (χ^2) por lo que no se puede rechazar H_0, es decir, existe dependencia o asociación significativa entre las dos variables cualitativas al estimarse un p-valor bajo H_0 para el valor de chi-cuadrado (χ^2) observado o una probabilidad relacionada con el azar para el resultado obtenido, igual o superior al valor de α de referencia lo que aumenta la probabilidad de cometer el error tipo 1 (p-valor ≥ α).

4. Realizar prueba de hipótesis para homogeneidad en el comportamiento de una variable cualitativa (dicotómica), por ejemplo, presencia o no de exposición en dos muestras independientes (grupos formados según la presencia o no de enfermedad).
   a) El posible planteamiento de hipótesis es como sigue:
      - H_0: no difiere la distribución de la variable en las "f" poblaciones subyacentes de las cuales se obtuvieron las muestras.
      - H_1: difiere la distribución de la variable en la "f" poblaciones subyacentes de las cuales se obtuvieron las muestras.

        Nota: puede servir de apoyo en su análisis el planteamiento de hipótesis para la diferencia entre dos proporciones cuando tenemos una variable dicotómica y dos muestras no pareadas (enfermos y no enfermos) o de casos y controles (muestras no pareados), como sigue:
      - H_0: P_1 enfermos-expuestos del total de enfermos (a/a + c) = P_2 no enfermos-expuestos del total de no enfermos (b/b + d) (son similares las dos proporciones, equivale a homogeneidad de la distribución de los valores de la variable según las muestras (enfermos y sanos).
      - H_1: P_1 enfermos-expuestos del total de enfermos (a/a + c) ≠ P_2 no enfermos-expuestos del total de no enfermos (b/b + d) (no son similares las dos

proporciones, equivale a heterogeneidad de la distribución de los valores de la variable según las muestras independientes (de enfermos y sanos).

b) Se construye la tabla de contingencia o cruzada (filas por columnas) con los datos correspondientes para la variable cualitativa según las muestras (enfermos o no enfermos) y se realiza la comparación de los valores observados y esperados de cada celda tabulada), empleando el estadígrafo de prueba chi-cuadrado (χ^2) según la expresión:
   - En tablas de contingencia (2x2) o cruzadas (filas por columnas para dos variables), la fórmula para el cálculo de chi-cuadrado es:
     - Fórmula para chi-cuadrado de Pearson según la expresión:

       $\chi^2 Pearson = \frac{n(ad-bc)^2}{m_1 m_0 n_1 n_\partial}$, sigue una distribución teórica chi-cuadrado (χ^2_1) con un grado de libertad para $(1-\alpha)$% de nivel de confianza.
     - Fórmula para chi-cuadrado con corrección de Yates (χ^2Yates) según la expresión:

       $\chi^2 Yates = \frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{m_1 m_0 n_1 n_\partial}$, sigue una distribución teórica chi-cuadrado (χ^2_1) con un grado de libertad para $(1-\alpha)$% de nivel de confianza.
   - En tablas de contingencia en que el número de filas o columnas sea mayor de 2 sería:
     - $\chi^2 Observado\text{-}calculado = \frac{\sum_{i=1}^{f}\sum_{j=1}^{c}(nij-eij)^2}{eij}$, cuya expresión matemática sigue una distribución teórica-esperada chi-cuadrado ($\chi^2_{(f-1)(c-1),\,1-\alpha}$ con grados de libertad = (f – 1)(c – 1), para $(1-\alpha)$% como nivel de confianza).
       donde:
       nij: frecuencia absoluta observada en la celda i, j;
       eij: frecuencia absoluta esperada en la celda i, j = (ninj)/n;
       ni y nj: totales marginales correspondientes a la fila i-ésima y columna j-ésima;
       n: gran total o total de sujetos en la tabla.
     - χ^2 teórico-esperado-tabulado, dependiendo de los grados de libertad (gl) = [columnas – 1] x [filas – 1]) y $1-\alpha$ prefijado como el valor de probabilidad límite entre el área bajo la curva de chi-cuadrado de rechazo o no para H_0.
     - Valor de p (p-valor bajo H_0) = Pr $\chi^2_{(f-1)(c-1),\,1-\alpha}$ teórico ≥ χ^2 observado.
   - La regla de decisión en el contraste de hipótesis es:
     - Si χ^2 observado-calculado > $\chi^2_{(f-1)(c-1),\,1-\alpha}$ teórico-esperado, entonces se sitúa en el área de rechazo para H_0 el chi-cuadrado observado (χ^2) por lo que se rechaza H_0, es decir, no existe homogeneidad significativa en la distribución de la variable cualitativa a partir de las proporciones estimadas para las muestras empleadas (enfermos o no enfermos), al estimarse un

p-valor bajo H_0 para el valor de chi-cuadrado (χ^2) observado o una probabilidad relacionada con el azar para el resultado obtenido, inferior al valor α de referencia, lo que disminuye la probabilidad de cometer el error tipo 1 (p-valor < α).

- Si χ^2 observado-calculado ≤ $\chi^2_{(f-1)(c-1),\,1-\alpha}$ teórico-esperado, entonces no se sitúa en el área de rechazo para H_0 el chi-cuadrado observado (χ^2) por lo que no se puede rechazar H_0; es decir, existe homogeneidad significativa en la distribución de la variable cualitativa a partir de las proporciones estimadas para las muestras empleadas (enfermos y no enfermos), al estimarse un p-valor bajo H_0 para el valor de chi-cuadrado (χ^2) observado o una probabilidad relacionada con el azar para el resultado obtenido, igual o superior al valor α de referencia, lo que aumenta la probabilidad de cometer el error tipo 1 (p-valor ≥ α).

En el caso de que la frecuencia esperada en un celda sea menor de 1 o que el 20 % de las frecuencias esperadas en las celdas sea menor de 5 de modo general, esto impide o invalida el empleo del estadígrafo de prueba chi-cuadrado y se requerirá la prueba de la probabilidad exacta de Fisher (PPEF) para evaluar la significación estadística en la asociación o independencia de las variables, donde se va directamente a estimar el valor de probabilidad exacta (p-valor) según el procedimiento de Fisher, para los datos de ambas variables cualitativas o para dos muestras independientes (con el evento o no) en relación con una variable cualitativa dicotómica, contenidos en la tabla de contingencia 2x2 simple.

Prueba de la probabilidad exacta o prueba exacta de Fisher

El criterio a seguir para su utilización en tablas de contingencia 2x2 simples es que al menos una de las frecuencias esperadas para las celdas sea menor de 5.

Puede ser empleado en la comparación de dos proporciones binomiales procedentes de dos muestras independientes en tablas de 2x2 con valores de frecuencias esperada menores de 5 en alguna celda.

Permite determinar la probabilidad (p-valor) exacta de obtener valores de la celda a fuera de las (a + b) observaciones para un número N (a + b + c + d = total de observaciones) de las cuales (a + c) se consideran con éxito dentro del evento estudiado, al comparar dos proporciones binomiales en tablas de contingencia 2x2 simples, esto se puede realizar empleando las probabilidades de la distribución hipergeométrica y estimar este valor p como (Rosner, 2000):

– Valor p unilateral:
  - p-valor para la cola derecha, se estima como probabilidades (+) a partir de la condición:

$$P(+) \text{ si } \frac{ad}{(b+1)(c+1)} > \frac{bc}{(a+1)(d+1)};$$

$$P(+) = \sum_{P(a+k)\leq P(a)} P(a+k)$$

cuando se quiere contrastar $p_1 > p_2$.

- p-valor para la cola izquierda, se estima como probabilidades (–) a partir de la condición como sigue:

$$P(-) \text{ si } \frac{ad}{(b+1)(c+1)} < \frac{bc}{(a+1)(d+1)};$$

$$P(-) = \sum_{P(a-k)\leq P(a)} P(a-k);$$

cuando se quiere contrastar $p_1 < p_2$.

– Valor p bilateral:

p-valor = P(+) + P(–) – P(a)

donde:

$P(+) = \sum_{P(a+k)\leq P(a)} P(a+k)$: quiere decir las probabilidades de las tablas estructuradas diferentes a la tabla observada inicial empleando valores mayores o iguales de los de la celda a (de la tabla inicial);

$P(-) = \sum_{P(a-k)\leq P(a)} P(a-k)$: quiere decir las probabilidades de las tablas estructuradas diferentes a la tabla observada inicial empleando valores menores o iguales de los de la celda a (de la tabla inicial);

P(a + k) y P(a – k) se calculan a partir de la función de probabilidad de una distribución hipergeométrica de parámetros N = a + b + c + d, R = a + b y n = a + c.

La descripción del procedimiento para obtener dicha probabilidad exacta de Fisher se explica a continuación, tomando elementos de la Ayuda de Tablas de Contingencia presente en el Software: EPIDAT versión 3.1 (Tabla 3.31).

Tabla 3.31. De contingencia 2x2 simple, que relaciona el hábito de fumar y la presencia de enfermedad cardiovascular

Hábito de fumar	Presencia de enfermedad cardiovascular	No enfermedad cardiovascular	Totales marginales
Fumador	A	b	a + b
No fumador	C	d	c + d
Totales	a + c	b + d	N = a + b + c + d

La probabilidad de obtener los valores para una tabla 2x2 simple observada se pueden calcular con la siguiente expresión matemática denominada *probabilidad exacta* (Pr) para los valores observados de las celdas a, b, c, d en una tabla 2x2 simple:

$$Pr(a,b,c,d) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$, donde el signo (!) significa factorial del valor o expresión que le antecede.

Si se asignan los siguientes valores a las celdas de la tabla 3.31 se representa como la tabla 3.32.

Tabla 3.32. De contingencia 2x2 simple, que relaciona el hábito de fumar y la presencia de enfermedad cardiovascular con valores en las celdas

Hábito de fumar	Presencia de enfermedad cardiovascular	No enfermedad cardiovascular	Totales marginales
Fumador	a = 2	b = 5	a + b = 7
No fumador	c = 3	d = 1	c + d = 4
Totales	a + c = 5	b + d = 6	N = a + b + c + d = 11

La probabilidad exacta para los valores de las celdas empleados para esta tabla 2x2 simple es:

$$Pr(2,5,3,1) = \frac{7!4!5!6!}{11!2!5!3!1!} = \frac{5\,040(24)(120(720)}{39\,916\,800(2)(120)(6)} = \frac{1{,}0450944 \cdot 10^{10}}{5{,}7480192 \cdot 10^{10}} = 0{,}182$$

Los pasos que se pueden seguir para el cálculo de la probabilidad exacta de Fisher, que incluye la sumatoria de las probabilidades para los distintos valores en las celdas para las tablas formadas, manteniendo sus valores totales marginales, a modo explicativo son:

- Teniendo en cuenta que la estrategia básica en el contraste de hipótesis (H_0: $p_1 = p_2$ vs. H_1: $p_1 \neq p_2$); por ejemplo si p_1 = (individuos con efecto negativo/total de individuos con factor de riesgo positivo) y p_2 = (individuos con efecto positivo/total de individuos con factor de riesgo positivo) se enumeran todas las posibles tablas a formar manteniendo constantes los totales marginales, tal y como se encuentran en la tabla observada con los valores iniciales y calcular la probabilidad exacta para cada una de las tablas formadas según la expresión matemática anterior (basada en las probabilidades de la distribución hipergeométrica), cuyo valor esperado E(X) y su varianza Var(X) para una variable aleatoria X (para la celda a o $C_{1,1}$), o sea, fila 1, columna 1, se expresa como:

  $$E(X) = \frac{M_1 N_1}{N}$$

  $$Var(X) = \frac{M_1 M_2 N_1 N_2}{N^2(N-1)}$$

- Se reestructuran las filas y columnas de la tabla observada con los valores iniciales de modo que el total marginal inferior por las filas quede en la primera fila y el total marginal por columna inferior en la primera columna.

– Se comienza asignando valor de 0 a la primara celda ($C_{1,1}$) o celda a tomando en lo siguiente cada celda restante los valores correspondientes que permitan sin variar los valores totales marginales por filas y columnas que presentó la tabla inicial observada y calcular la probabilidad para estos nuevos valores a partir de la distribución de probabilidades hipergeométrica.
– Se construyen nuevas tablas incrementando tabla a tabla los valores de la celda a, obteniéndose tablas hasta que alguno de los valores de las celdas restantes adquiera el valor 0 y entonces calcular igualmente la probabilidad para los valores de cada nueva tabla estructurada.
– Si se toma como ejemplo la tabla observada inicial con los valores de a, b, c, d con sus respectivos valores totales marginales: (a + b), (c + d), (a + c), (b + d) y el número total de observaciones (N = a + b + c + d) y denominamos tabla (a) a la observada (inicial) y la tabla (k) a la última posible obtenida. Se llega a esta tabla (k) cuando alguna otra celda toma un valor de 0 debido al aumento de la celda a, para el cálculo de probabilidad exacta positiva [Pr(+)] (Rosner, 2000) o por el contrario cuando la celda a disminuye hasta un valor de 0 o probabilidad exacta negativa [Pr(–)] (Rosner, 2000), cuando se mantienen fijos los totales marginales, se tiene que:
  - Para contrastar la hipótesis bilateral a partir de la estimación de la probabilidad exacta de Fisher se tiene:

    H_0: $p_1 = p_2$ *vs.* H_1: $p_1 \neq p_2$,

    p-valor Fisher (2 colas) = 2 · mín [Pr(0) + Pr(1)… + Pr(a) + Pr(a) + Pr(a + 1) + +Pr(k)] donde:

    mín: menor valor de las probabilidades;

    p-valor Fisher (2 colas): corresponde al duplicado de la menor área bajo la curva, bien de la cola derecha o de la izquierda);

    Otra forma que puede ser aplicada en la estimación de la Probabilidad exacta de Fisher bilateral es:

    $$\text{p-valor Fisher (2 colas)} = \sum_{i:\,Pr(i)\leq Pr(a)} Pr(i)\,;$$

    donde Pr (i) son las probabilidades obtenidas para los valores de cada tabla, y se sumarían todas las Pr(i) menores o iguales a la Pr(a) (de la tabla observada con los valores iniciales en cada celda).
  - Para contrastar la hipótesis unilateral:

    H_0: $p_1 = p_2$ vs. H_1: $p_1 < p_2$,

    p-valor Fisher = Pr(0) + Pr(1)… + Pr(a), corresponde al área bajo la curva para la cola izquierda).
  - Para contrastar la hipótesis unilateral:

    H_0: $p_1 = p_2$ vs. H_1: $p_1 \neq p_2$,

    p-valor Fisher = Pr(a) + Pr(a + 1) + … + Pr(k), corresponde al área bajo la curva para la cola derecha.

Nota: En este caso el valor de p (p-valor) expresa la probabilidad de obtener una tabla con valores tan extremos o más que los obtenidos con la tabla observada inicial.

Regla de interpretación:

- Si p-valor (observado-calculado) < valor del nivel de significación (α) de acuerdo con el tipo de contraste de hipótesis a realizar (bilateral o unilateral) y grados de libertad estimados; se rechaza H_0 y, por ende, se concluye que existen evidencias con los datos empleados de que existen diferencias significativas entre las dos proporciones comparadas para el nivel de confianza predeterminado por el investigador($p < \alpha$).
- Si p-valor (observado-calculado) ≥ valor del nivel de significación (α) de acuerdo al tipo de contraste de hipótesis a realizar (bilateral o unilateral) y grados de libertad estimados; no se puede rechazar H_0 y, por ende, se concluye que no existen evidencias con los datos empleados de que existen diferencias significativas entre las dos proporciones comparadas para un nivel de confianza predeterminado por el investigador($p \geq \alpha$).
- Si se tratara de un análisis estratificado (que ajustará la asociación entre las variables, según el factor de estratificación) puede utilizarse el correspondiente estadístico de prueba de Mantell-Haenzsel, es decir, el promedio del chi-cuadrado individual de cada tabla (cada parte del estrato) como:

  $\chi^2_{\text{Mantel Haenzsel (MH)}} = [\Sigma a_e - \Sigma E(a_e)]^2/\Sigma V(a_e)$

  $\chi^2_{\text{Mantel Haenzsel (MH)}} = \{[a_{e1} - E(a_{e1})]^2/V(a_{e1})\} + \{[a_{e2} - E(a_{e2})]^2/V(a_{e2})\}$

  donde:

  para cada parte del estrato (e_1: estrato 1 y e_2: estrato 2) o cada tabla del estrato se calculará:

  a_e: número de enfermos que están expuestos;

  $E(a_e)$: es el valor esperado de ae, se calcula como:

  $E(a_e)$ = total del renglón · total de columna/gran total = $(a_e + b_e) \cdot (a_e + c_e)/n_e$;

  $V(a_e)$: la varianza de a_e, se calcula como:

  $V(a_e) = [(a_e + b_e) \cdot (c_e + d_e) \cdot (a_e + c_e) \cdot (b_e + d_e)]/[ne^2(n_e - 1)]$

Regla de decisión para el análisis (sin estratificación) de la prueba de independencia entre variables:

- Si χ^2 observado-calculado > χ^2 esperado-tabulado), entonces el valor de chi-cuadrado observado se ubica a la derecha del valor crítico, lo que equivale a decir en el área de rechazo de H_0 y se acepta H_1, por lo que existe evidencia para aceptar una asociación significativa de dependencia entre las variables independiente (factor de riesgo o exposición) y dependiente (riesgo o enfermedad) para cierto nivel de confianza prefijado por el investigador; estimándose una probabilidad del azar o de cometer error tipo 1 menor que el valor alfa prefijado ($p < \alpha$).
- Si χ^2 observado-calculado ≤ χ^2 esperado-tabulado, entonces el valor de chi-cuadrado observado no lo ubica a la derecha del valor crítico, lo que no equivale a decir en el

área de rechazo de H_0, por lo que esta no es rechazable y existe evidencia para aceptar una independencia entre las variables independiente (factor de riesgo o exposición) y dependiente (riesgo o enfermedad) para cierto nivel de confianza prefijado por el investigador; estimándose una probabilidad del azar o de cometer el error tipo 1 igual o mayor que el valor alfa prefijado ($p \geq \alpha$).

Algunas situaciones a las cuales podemos enfrentarnos al evaluar la significación estadística de la asociación o dependencia entre las variables, después de la estratificación:

- Si existe una "marcada" diferencia entre los valores de chi-cuadrado de las pruebas crudas y ajustadas, se dice que hay evidencias para afirmar que la asociación de interés puede ser confundida por una tercera variable, en este caso el factor de estratificación.
- La asociación en estudio puede ser significante antes de ajustar para un confusor y no ser estadísticamente significante posteriormente a su estratificación.
- La asociación puede permanecer significante después de ajustar para un confusor, pero con una menor significación en el valor de p.
- Los estratos pueden mostrar resultados opuestos y en este caso es mejor presentar los resultados estratificados. Esto es interacción o modificación del efecto.
- El confusor puede esconder una relación que existe.

Estimación puntual y por intervalo de confianza de las medidas de frecuencia, asociación e impacto en los estudios epidemiológicos analíticos observacionales empleando tablas de contingencia

Los aspectos básicos conceptuales necesarios para poder realizar la estimación puntual y por intervalo de confianza de los tipos de medidas a emplear en los estudios epidemiológicos observacionales pueden ser abordados a partir del empleo de tablas de contingencia, las que facilitan el ordenamiento y posterior cálculo con los datos disponibles para esas medidas en los diseños de estudio que emplean dos variables cualitativas.

Los formularios a utilizar son presentados siguiendo la metodología empleada en el Software EPIDAT, versión 3,1 (Ayudas de Tabla de contingencia).

Estudio de corte o transversal

- Conformar una tabla de contingencia 2x2 simple para estudio de corte o transversal e introducir datos en las celdas (a, b, c, d).

		Enfermedad		
		Si	**No**	**Total**
Factor de riesgo	Expuestos	a	b	$n_1 = a + b$
	No expuestos	c	d	$n_0 = c + d$
	Total	$\mathbf{m_1 = a + c}$	$\mathbf{m_0 = b + d}$	**n**

- Calcular los totales marginales: por filas ($a + b = n_1$) y ($c + d = n_0$); así como los totales marginales por columnas ($a + c = m_1$) y ($b + d = m_0$) y el gran total: ($n = (a + b) + (c + d)$ o $n = n_1 + n_0$).
- Introducir los valores calculados en las celdas respectivas de la tabla anterior.
- Acceder a fórmulas para la estimación (cálculos) puntual y por intervalos de confianza de las medias de frecuencia y de asociación en estudio de corte o transversal para un valor de $\alpha = 0{,}05$ a partir de prevalencias de enfermedad y prevalencias de exposición.
- Calcular el valor puntual y por intervalos de confianza de las medidas de frecuencia, asociación e impacto.

Medidas de frecuencia

- Prevalencia de exposición:
  - Proporción grupo de enfermos:

$$pexp_1 = \frac{a}{m_1} = \frac{a}{a+c}$$

  - Proporción grupo de no enfermos (sanos):

$$pexp_0 = \frac{b}{m_0} = \frac{b}{b+d}$$

- Prevalencia de enfermedad:
  - Proporción grupo de expuestos:

$$p_1 = \frac{a}{n_1} = \frac{a}{a+b}$$

  - Proporción grupo de no expuestos:

$$p_0 = \frac{c}{n_0} = \frac{c}{c+d}$$

Medidas de asociación

- Razón de prevalencias de exposición (RPE):
  - Valor puntual (RPE):

$$RPE = \frac{pexp_1}{pexp_0} = \frac{a\ m_0}{b\ m_1} = \frac{a/(a+c)}{b/(b+d)}$$

  - IC inferior o límite inferior (RPE):

$$CInferior(RPE) = \frac{am_0}{bm_1} exp\left(-1{,}96\sqrt{\frac{c}{am_1} + \frac{d}{bm_0}}\right)$$

- IC superior o límite superior (RPE):

$$ICSuperior(RPE) = \frac{am_0}{bm_1} exp\left(1{,}96\sqrt{\frac{c}{am_1} + \frac{d}{bm_0}}\right)$$

– Razón de prevalencias de enfermedad (RP):

- Valor puntual (RP):

$$RP = \frac{p_1}{p_0} = \frac{an_0}{cn_1}$$

- IC inferior o límite inferior (RP):

$$ICInferior\ (RP) = \frac{an_0}{cn_1} exp\left(-1{,}96\sqrt{\frac{b}{an_1} + \frac{d}{cn_0}}\right)$$

Otra forma de efectuar su cálculo es:

- $$ICInferior\ (RP) = \frac{RP}{FE} = \frac{\frac{an_0}{cn}}{exp\left(1{,}96\sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}\right)}$$

donde:

$$FE = exp\left(1{,}96\sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}\right)$$

- IC superior o límite superior (RP):

$$ICSuperior(RP) = \frac{an_0}{cn_1} exp\left(1{,}96\sqrt{\frac{b}{an_1} + \frac{d}{cn_0}}\right)$$

Otra forma de efectuar su cálculo es:

- $$ICSuperior(RP) = RP \cdot FE = \frac{an_0}{cn} \cdot exp\left(1{,}96\sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}\right)$$

donde:

$$FE = exp\left(1{,}96\sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}\right)\ c$$

– Diferencia de riesgos (DR) o riesgo atribuible absoluto (RA) o exceso de riesgo (ER):

- Valor puntual (DR):

$$DR = p_1 - p_0 = \frac{a}{n_1} - \frac{c}{n_0}$$

- IC inferior o límite inferior (DR):

$$ICInferior(DR) = \left(\frac{a}{n_1} - \frac{c}{n_0}\right) - 1{,}96\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0}}$$

donde:

$p_1 = \frac{a}{n_1}$;

$p_0 = \frac{c}{n_0}$;

$n_1 = (a + b)$;

$n_0 = (c + d)$.

- IC superior o límite superior (DR):

$$ICSuperior(DR) = \left(\frac{a}{n_1} - \frac{c}{n_0}\right) + 1{,}96\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0}}$$

donde:

$p_1 = \frac{a}{n_1}$;

$p_0 = \frac{c}{n_0}$;

$n_1 = (a + b)$;

$n_0 = (c + d)$.

– *Odds ratio* (OR):

- Valor puntual (OR):

$$OR = \frac{ad}{bc}.$$

- IC inferior o límite Inferior (OR):

$$ICInferior(OR) = \frac{ad}{bc}\exp\left(-1{,}96\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right)$$

- IC superior o límite superior (OR):

$$ICSuperior(OR) = \frac{ad}{bc}\exp\left(1{,}96\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right)$$

Medidas de impacto

La mayoría de los autores no las calculan ya que este tipo de estudio no permite evaluar el impacto de la exposición retro ni prospectivamente, no obstante, en algunos diseños se han calculado como sigue:

– Fracción atribuible en los expuestos (FAE o PF_e): se calcula si RP ≥ 1:

- Valor puntual (FAE o AF_e):

$$AF_e = \frac{RP - 1}{RP};$$

donde:

$$RP = \frac{an_0}{cn_1} = \frac{p_1}{p_0}$$

- IC inferior o límite inferior (FAE) o PF_e:

$$ICInferior(AF_e) = \frac{\frac{ad}{bc}\exp\left(-1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)-1}{\frac{ad}{bc}\exp\left(-1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)}$$

- IC superior o límite superior (FAE o PFe):

$$ICSuperior(AF_e) = \frac{\frac{ad}{bc}\exp\left(1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)-1}{\frac{ad}{bc}\exp\left(1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)}$$

– Fracción de prevención en los expuestos (FAE o PFe) se calcula si RP < 1:

- Valor puntual (FP^E o PF_e):

  $PF_e = 1 - RP$;

  donde:

$$RP = \frac{an_0}{cn_1} = \frac{p_1}{p_0}.$$

- IC inferior o límite inferior (FP^E o PF_e):

$$ICInferior(PF_e) = 1 - \left[\frac{an_0}{cn_1}\exp\left(1{,}96\sqrt{\frac{b}{an_1}+\frac{d}{cn_0}}\right)\right]$$

  donde:

  $n_1 = (a + b)$;

  $n_0 = (c + d)$.

- IC superior o límite superior (FPE o PF_e):

$$ICSuperior(PF_e) = 1 - \left[\frac{an_0}{cn_1}\exp\left(-1{,}96\sqrt{\frac{b}{an_1}+\frac{d}{cn_0}}\right)\right]$$

  donde:

  $n_1 = (a + b)$;

  $n_0 = (c + d)$.

– Fracción atribuible poblacional (FAP o AF_p): se calcula si RP ≥ 1:

- Valor puntual (FAP o AF_p):

$$AF_p = AF_e \frac{a}{m_1};$$

donde:

$$AF_e = \frac{RP-1}{RP};$$

$$RP = \frac{an_0}{cn_1} = \frac{p_1}{p_0}.$$

- IC inferior o límite inferior (FAP o AF_P):

$$ICInferior(AF_p) = 1-\left(1-AF_p\right)\exp\left(1{,}96\sqrt{Var\left[Ln\left(1-AF_p\right)\right]}\right)$$

donde:

$$AF_p = AF_e \frac{a}{m_1};$$

$$Var\left[Ln\left(1-AF_p\right)\right] = \left(\frac{AF_p}{1-AF_p}\right)^2 \left\{\frac{Var(LnRP)}{(RP-1)^2} + \frac{2}{a(RP-1)} + \frac{c}{am_1}\right\};$$

$$Var\left[Ln(RP)\right] = \frac{b}{an_1} + \frac{d}{cn_0}.$$

$$ICInferior(AF_p) = 1-\left(1-AF_e\frac{a}{m_1}\right)\exp\left(1{,}96\sqrt{\left(\frac{AF_e\frac{a}{m_1}}{1-AF_e\frac{a}{m_1}}\right)^2 \left\{\frac{\frac{b}{an_1}+\frac{d}{cn_0}}{\left(\frac{an_0}{cn_1}-1\right)^2} + \frac{2}{a\left(\frac{an_0}{cn_1}-1\right)} + \frac{c}{am_1}\right\}}\right)$$

- IC superior o límite superior (FAP o AF_P):

$$ICSuperior(AF_p) = 1-\left(1-AF_p\right)\exp\left(-1{,}96\sqrt{Var\left[Ln\left(1-AF_p\right)\right]}\right)$$

donde:

$$AF_p = AF_e \frac{a}{m_1};$$

$$Var\left[Ln\left(1-AF_p\right)\right] = \left(\frac{AF_p}{1-AF_p}\right)^2 \left\{\frac{Var(LnRR)}{(RR-1)^2} + \frac{2}{a(RR-1)} + \frac{c}{am_1}\right\};$$

$$Var\left[Ln(RP)\right] = \frac{b}{an_1} + \frac{d}{cn_0}.$$

$$ICSuperior(AF_p) = 1-\left(1-AF_e\frac{a}{m_1}\right)\exp\left(-1{,}96\sqrt{\left(\frac{AF_e\frac{a}{m_1}}{1-AF_e\frac{a}{m_1}}\right)^2 \left\{\frac{\frac{b}{an_1}+\frac{d}{cn_0}}{\left(\frac{an_0}{cn_1}-1\right)^2} + \frac{2}{a\left(\frac{an_0}{cn_1}-1\right)} + \frac{c}{am_1}\right\}}\right)$$

– Fracción de prevención para la población (FPP o PF_p): se calcula si RP < 1:
  - Valor puntual (FPP o PF_p):

$$PF_p = PF_e \frac{n_1}{n};$$

donde:

$PF_e = 1 - RP$;

$$RP = \frac{an_0}{cn_1} = \frac{p_1}{p_0};$$

$n_1 = (a + b)$;

$n_0 = (c + d)$.

  - IC inferior o límite inferior (FPP o PF_p):

$$\left[L_I(PF_e)\frac{n_1}{n_1+n_0}\right] = \left[1-\left(\frac{an_0}{cn_1}\exp\left(1{,}96\sqrt{\frac{b}{an_1}+\frac{d}{cn_0}}\right)\ \frac{n_1}{n_1+n_0}\right)\right]$$

donde:

$$L_I(PF_e) = 1-\left[\frac{an_0}{cn_1}\exp\left(1{,}96\sqrt{\frac{b}{an_1}+\frac{d}{cn_0}}\right)\right];$$

$n_1 = (a + b)$;

$n_0 = (c + d)$.

  - IC superior o límite superior (FPP o PF_p):

$$\left[L_S(PF_e)\frac{n_1}{n_1+n_0}\right] = 1-\left[\frac{an_0}{cn_1}\exp\left(-1{,}96\sqrt{\frac{b}{an_1}+\frac{d}{cn_0}}\right)\frac{n_1}{n_1+n_0}\right]$$

donde:

$$L_S(PF_e) = 1-\left[\frac{an_0}{cn_1}\exp\left(-1{,}96\sqrt{\frac{b}{an_1}+\frac{d}{cn_0}}\right)\right];$$

$n_1 = (a + b)$;

$n_0 = (c + d)$.

Resultados en estudios transversales

Se realiza un análisis tomando como ejemplo un nivel de confianza 95 %.

Medidas de frecuencia

– Para Prevalencia de exposición:
  - Prevalencia en enfermos (riesgo en enfermos): valor puntual.

- Prevalencia en no enfermos (riesgo en no enfermos): valor puntual.

– Para Prevalencia de enfermedad:
  - Prevalencia en expuestos (riesgo en expuestos): valor puntual.
  - Prevalencia en no expuestos (riesgo en no expuestos): valor puntual.

Medidas de asociación

– Razón de prevalencias (de exposición): valor puntual y por intervalos de confianza.
– Razón de prevalencias (de enfermedad): valor puntual y por intervalos de confianza.
– Diferencia de riesgos: valor puntual y por intervalos de confianza.
– *Odds ratio*: valor puntual y por intervalos de confianza (Woolf).

Medidas de impacto

Aparecen solo los resultados de la fracción atribuible si RP ≥ 1:

– Fracción atribuible en expuestos: valor puntual y por intervalos de confianza.
– Fracción atribuible poblacional: valor puntual y por intervalos de confianza.

Aparecen solo los resultados de la fracción prevenible si RP< 1:

– Fracción de prevención en expuestos: valor puntual y por intervalos de confianza.
– Fracción de prevención poblacional: valor puntual y por intervalos de confianza.

Estudio de caso-control no apareado

– Conformar una tabla de contingencia 2x2 simple para estudio de caso-control no apareado (1 caso: 1 control) e introducir datos en las celdas (a, b, c, d).
– Calcular los totales marginales (por filas: $n_1 = (a + b)$ y $n_0 = (c + d)$, así como los totales marginales por columnas $m_1 = (a + c)$ y $m_0 = (b + d)$ y $n = (m_1 + m_0) = (a + c) + (b + d)$ o $n = n_1 + n_0 = (a + b) + (c + d)$.

		Enfermedad		
		Casos	**Controles**	**Total**
Factor de riesgo	Expuestos	a	b	$n_1 = a + b$
	No expuestos	c	d	$n_0 = c + d$
	Total	$\mathbf{m_1 = a + c}$	$\mathbf{m_0 = b + d}$	**n**

– Acceder a fórmulas para la estimación (cálculos) puntual y por intervalos de confianza de medidas de frecuencia, de asociación e impacto en estudio de caso y control no apareado para valor de $\alpha = 0,05$, a partir de prevalencias de exposición y oportunidad de exposición e introducir valores respectivos de las celdas en las mismas. Entonces seguir con:

		Enfermedad		
		Casos	**Controles**	**Total**
Factor de riesgo	Expuestos	a	b	$n_1 = a + b$
	No expuestos	c	d	$n_0 = c + d$
	Total	$m_1 = a + c$	$m_0 = b + d$	n

Medidas de frecuencia

– Proporción de expuestos en casos:

$$p_1 = \frac{a}{m_1}$$

– Proporción de expuestos en controles:

$$p_0 = \frac{b}{m_0}$$

Medidas de asociación

Odds ratio de exposición:

– Valor puntual (OR):

$$OR = \frac{ad}{bc}.$$

– Por intervalos de confianza para OR (Método de Woolf):

- IC inferior o límite inferior (OR):

$$ICInferior(OR) = \frac{ad}{bc}\exp\left(-1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)$$

- IC superior o límite superior (OR):

$$ICSuperior(OR) = \frac{ad}{bc}\exp\left(1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)$$

Medidas de impacto

– Fracción atribuible en los expuestos (FAE o AF_e): se calcula si OR ≥ 1:

- Valor puntual de (FAE o AF_e):

$$AF_e = \frac{OR-1}{OR};$$

donde:

$$OR = \frac{ad}{bc}$$

- IC inferior o límite inferior (FAE o AF_e):

$$ICInferior(AF_e) = \frac{\frac{ad}{bc}\exp\left(-1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)-1}{\frac{ad}{bc}\exp\left(-1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)}$$

- IC superior o límite superior (FAE o AF_e):

$$ICSuperior\ (AF_e) = \frac{\frac{ad}{bc}\exp\left(1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)-1}{\frac{ad}{bc}\exp\left(1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)}$$

– Fracción de prevención en los expuestos (FPE o PFe): se calcula si OR < 1:

- Valor puntual de (FPE o PF_e):

  PFe = 1 – OR

  donde:

$$OR = \frac{ad}{bc}$$

- IC inferior o límite inferior (FPE):

$$ICInferior\ (FPE) = \left[1-\frac{ad}{bc}\exp\left(1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)\right]$$

- IC superior o límite superior (FPE):

$$ICSuperior\ (FPE) = \left[1-\frac{ad}{bc}\exp\left(-1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)\right]$$

– Fracción atribuible poblacional (FAP o AF_p): se calcula si OR ≥ 1:

- Valor puntual (FAP o AF_p):

  $AFP = AFep_1$

  donde:

$$AF_e = \frac{OR-1}{OR};$$

$$OR = \frac{ad}{bc};$$

.

- IC inferior o límite inferior (FAP o AF_p):

$$L_l\left(AF_p\right) = 1-\left(1-AF_p\right)\exp\left(z_{1-\alpha/2}\sqrt{Var\left[Ln\left(1-AF_p\right)\right]}\right)$$

donde:

$AF_p = AFep_1$;

$$AF_e = \frac{OR-1}{OR};$$

$$Var\left[Ln\left(1-AF_p\right)\right] = Var\left[Ln\left(1-PF_p\right)\right] = \frac{a}{cm_1} + \frac{b}{dm_0};$$

$$L_I\left(AF_p\right) = 1-\left(1-[(\frac{(\frac{ad}{bc}-1)}{\frac{ad}{bc}})(p_1)]\right)\exp\left(1{,}96\sqrt{\frac{a}{cm_1}+\frac{b}{dm_0}}\right)$$

- IC superior o límite superior (FAP o AFp):

$$L_S\left(AF_p\right) = 1-\left(1-AF_p\right)\exp\left(-z_{1-\alpha/2}\sqrt{Var\left[Ln\left(1-AF_p\right)\right]}\right)$$

donde:

AFp = $AFep_1$;

$$AF_e = \frac{OR-1}{OR};$$

$$Var\left[Ln\left(1-AF_p\right)\right] = Var\left[Ln\left(1-PF_p\right)\right] = \frac{a}{cm_1} + \frac{b}{dm_0};$$

$$L_S\left(AF_p\right) = 1-\left(1-[\frac{(\frac{ad}{bc}-1)}{\frac{ad}{bc}}(p_1)]\right)\exp\left(-1{,}96\sqrt{\frac{a}{cm_1}+\frac{b}{dm_0}}\right);$$

– Fracción de prevención poblacional (FP^P o PF_P): se calcula si OR < 1:

- Valor puntual (FPP o PF_P):

$$PF_p = \frac{p_1 PF_e}{p_1 PF_e + OR}$$

donde:

$PF_e = 1 - OR$;

$$p_1 = \frac{a}{m_1};$$

$$OR = \frac{ad}{bc}.$$

- IC inferior o límite inferior (FPP o PF_P):

$$L_I\left(PF_p\right) = 1-\left(1-PF_p\right)\exp\left(z_{1-\alpha/2}\sqrt{Var\left[Ln\left(1-PF_p\right)\right]}\right)$$

donde:

$$Var\left[Ln\left(1-PF_p\right)\right] = \frac{a}{cm_1} + \frac{b}{dm_0};$$

$$PF_p = \frac{p_1 PF_e}{p_1 PF_e + OR};$$

PFe = 1 − OR;

$$p_1 = \frac{a}{m_1};$$

$$OR = \frac{ad}{bc};$$

$$L_I\left(PF_p\right) = 1 - \left\{\left(1 - \left[\frac{\left(\frac{a}{m_1}\right)\left(1 - \frac{ad}{bc}\right)}{\left[\left(\frac{a}{m_1}\right)\left(1 - \frac{ad}{bc}\right)\right] + \frac{ad}{bc}}\right]\right) \exp\left(1{,}96\sqrt{\frac{a}{cm_1} + \frac{b}{dm_0}}\right)\right\}$$

- IC superior o límite superior (FPP o PFP):

$$L_S\left(PF_p\right) = 1 - \left(1 - PF_p\right)\exp\left(-z_{1-\alpha/2}\sqrt{Var\left[Ln\left(1 - PF_p\right)\right]}\right)$$

donde:

$$Var\left[Ln\left(1 - PF_p\right)\right] = \frac{a}{cm_1} + \frac{b}{dm_0};$$

$$PF_p = \frac{p_1 PF_e}{p_1 PF_e + OR};$$

PF_e = 1 − OR;

$$p_1 = \frac{a}{m_1};$$

$$OR = \frac{ad}{bc}$$

$$L_S\left(PF_p\right) = 1 - \left\{\left(1 - \left[\frac{\left(\frac{a}{m_1}\right)\left(1 - \frac{ad}{bc}\right)}{\left[\left(\frac{a}{m_1}\right)\left(1 - \frac{ad}{bc}\right)\right] + \frac{ad}{bc}}\right]\right) \exp\left(-1{,}96\sqrt{\frac{a}{cm_1} + \frac{b}{dm_0}}\right)\right\}$$

Resultados en estudios de caso-control no apareado (sin *matching* o no apareamiento)

Se muestra ahora un ejemplo para un nivel de confianza de 95 %.

Medidas de frecuencia

- Proporción de expuestos en casos: valor puntual.
- Proporción de expuestos en controles: valor puntual.

Medidas de asociación

– *Odds ratio* de exposición: valor puntual y por intervalos de confianza (método Woolf).

Medidas de impacto

– Fracción atribuible en los expuestos: valor puntual y por intervalos de confianza.
– Fracción de prevención en los expuestos: valor puntual y por intervalos de confianza.
– Fracción atribuible poblacional: valor puntual y por intervalos de confianza.
– Fracción de prevención poblacional: valor puntual y por intervalos de confianza.

Estudio de cohorte con datos de incidencia acumulada

– Conformar una tabla de contingencia 2x2 simple para estudio de cohorte para incidencia acumulada e introducir datos en las celdas (a, b, c, d).

		Enfermedad		
		Si	**No**	**Total**
Factor de riesgo	Expuestos	a	b	$n_1 = a + b$
	No expuestos	c	d	$n_0 = c + d$
	Total	**$m_1 = a + c$**	**$m_0 = b + d$**	**n**

– Calcular los totales marginales: por filas: (a + b) y (c + d); así como, los totales marginales por columnas (a + c) y (b + d) y n = (a + b) + (c + d) o n = (a + c) + (b + d).
– Acceder a fórmulas para la estimación (cálculo) puntual y por intervalos de confianza de las medidas de frecuencia, asociación e impacto correspondientes, para valor $\alpha = 0{,}05$ e introducir los valores respectivos de las celdas (a, b, c, d) y totales marginales:

		Enfermedad		
		Si	**No**	**Total**
Factor de riesgo	Expuestos	a	b	$n_1 = a + b$
	No expuestos	c	d	$n_0 = c + d$
	Total	**$m_1 = a + c$**	**$m_0 = b + d$**	**n**

Medidas de frecuencia

– Proporción de expuestos enfermos (riesgo en expuestos):

$$p_1 = \frac{a}{n_1}$$

– Proporción de no expuestos enfermos (riesgo en no expuestos):

$$p_0 = \frac{c}{n_0}$$

Medidas de asociación

– Razón de riesgos o riesgo relativo (RR):

- Valor puntual riesgo relativo (RR):

$$RR = \frac{an_0}{cn_1} = \frac{p_1}{p_0}$$

- IC inferior o límite inferior (RR):

$$ICInferior(RR) = \frac{an_0}{cn_1}\exp\left(-1{,}96\sqrt{\frac{b}{an_1}+\frac{d}{cn_0}}\right)$$

Otra forma puede ser:

$$ICInferior(RR) = \frac{RR}{FE} = \frac{an_0}{cn}\exp\left(1{,}96\sqrt{\frac{1}{a}-\frac{1}{a+b}+\frac{1}{c}-\frac{1}{c+d}}\right)$$

donde:

$$FE = \exp\left(1{,}96\sqrt{\frac{1}{a}-\frac{1}{a+b}+\frac{1}{c}-\frac{1}{c+d}}\right)$$

- IC superior o límite superior (RR):

$$ICSuperior(RR) = \frac{an_0}{cn_1}\exp\left(1{,}96\sqrt{\frac{b}{an_1}+\frac{d}{cn_0}}\right)$$

Otra forma puede ser:

$$ICSuperior(RR) = RR \cdot FE = \frac{an_0}{cn}\exp\left(1{,}96\sqrt{\frac{1}{a}-\frac{1}{a+b}+\frac{1|}{c}-\frac{1}{c+d}}\right)$$

donde:

$$FE = \exp\left(1{,}96\sqrt{\frac{1}{a}-\frac{1}{a+b}+\frac{1}{c}+\frac{1}{c+d}}\right)$$

Regla de interpretación del riesgo relativo a partir de su valor puntual estimado y su intervalo de confianza estimado poblacional:

- Si RR ≠ 1 y el IC para RR ≠1 para 95 %, entonces el riesgo es estadísticamente significativo para $p < 0{,}05$.
- Si RR ≠ 1 y el IC para RR ≠1 para 99 %, entonces el riesgo es estadísticamente significativo para $p < 0{,}01$.
- Si RR > 1 y el IC >1 para 95 %, entonces se está en presencia de un factor de riesgo para $p < 0{,}05$.
- Si RR > 1 y el IC > 1 para 99 %, entonces se está en presencia de un factor de riesgo para $p < 0{,}01$.

- -Si RR < 1 y el IC < 1 para 95 %, entonces se está en presencia de un factor de protección para p < 0,05.
- Si RR < 1 y el IC < 1 para 99 %, entonces se está en presencia de un factor de protección para p < 0,01.

– Diferencia de riesgos (DR) o riesgo absoluto (RA) o exceso de riesgo (ER):
  - Valor puntual (DR):

$$DR = p_1 - p_0 = \frac{a}{n_1} - \frac{c}{n_0}$$

donde:

$n_1 = a + b$; $n_0 = c + d$; $n = n_1 + n_0$

  - IC inferior o límite inferior (DR):

$$ICInferior(DR) = \left(\frac{a}{n_1} - \frac{c}{n_0}\right) - 1{,}96\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0}}$$

  - IC superior o límite superior (DR):

$$ICSuperior(DR) = \left(\frac{a}{n_1} - \frac{c}{n_0}\right) + 1{,}96\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0}}$$

donde:

$$p_1 = \frac{a}{n_1};$$

$$p_0 = \frac{c}{n_0};$$

$n_1 = a + b$; $n_0 = c + d$.

– *Odds ratio* de enfermedad:
  - Valor puntual(OR):

$$OR = \frac{ad}{bc}.$$

Intervalo de confianza (OR) (Método de Woolf):

  - IC inferior o límite inferior (OR):

$$ICInferior(OR) = \frac{ad}{bc}\exp\left(-1{,}96\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right)$$

  - IC superior o límite superior (OR):

$$ICSuferior(OR) = \frac{ad}{bc}\exp\left(1{,}96\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right)$$

La regla de interpretación del *Odds ratio* en los estudios epidemiológicos se presenta en la tabla 3.33.

Tabla 3.33. Regla de interpretación del *Odds ratio* según sus valores puntual y por intervalo de confianza

Valor puntual del *odds ratio*	Intervalo de confianza		Tipo de asociación
	Inferior	Superior	
1			No evidencia de asociación
Mayor de 1	>1	>1	Significativa, riesgo
Mayor de 1	< 1	> 1	No significativa
Menor de 1	< de 1	< de 1	Significativa, protección
Menor de 1	< de 1	> de 1	No significativa

Medidas de impacto

– Fracción atribuible en los expuestos (FAE o PF_e): se calcula si RR ≥ 1:

- Valor puntual (FAE o AF_e):

$$AF_e = \frac{RR - 1}{RR};$$

donde:

$$RR = \frac{an_0}{cn_1} = \frac{p_1}{p_0}$$

- IC inferior o límite inferior (FAE o PF_e):

$$ICInferior(PF_e) = \frac{\left[\frac{an_0}{cn_1}\exp\left(-1{,}96\sqrt{\frac{b}{an_1}+\frac{d}{cn_0}}\right)-1\right]}{\frac{an_0}{cn_1}\exp\left(-1{,}96\sqrt{\frac{b}{an_1}+\frac{d}{cn_0}}\right)}$$

- IC superior o límite superior (FAE o PF_e):

$$ICSuperior(PF_e) = \frac{\left[\frac{an_0}{cn_1}\exp\left(1{,}96\sqrt{\frac{b}{an_1}+\frac{d}{cn_0}}\right)-1\right]}{\frac{an_0}{cn_1}\exp\left(1{,}96\sqrt{\frac{b}{an_1}+\frac{d}{cn_0}}\right)}$$

– Fracción de prevención en los expuestos (FPE o PF_e): se calcula si RR < 1:

- Valor puntual (FPE o PF_e):

  $FP_e = 1 - RR$

  donde:

$$RR = \frac{an_0}{cn_1} = \frac{p_1}{p_0}$$

- IC inferior o límite inferior (FPE o $PF_{e)}$:

$$ICInferior(FP_e) = \left[1 - \frac{an_0}{cn_1}\exp\left(1{,}96\sqrt{\frac{b}{an_1} + \frac{d}{cn_0}}\right)\right]$$

donde:

$n_1 = (a + b)$;

$n_0 = (c + d)$.

- IC superior o límite superior (FPE o PF_e):

$$ICSuperior(FP_e) = \left[1 - \frac{an_0}{cn_1}\exp\left(-1{,}96\sqrt{\frac{b}{an_1} + \frac{d}{cn_0}}\right)\right]$$

donde:

$n_1 = (a + b)$;

$n_0 = (c + d)$.

– Fracción atribuible poblacional (FAP o AF_p): se calcula si RR ≥ 1:

- Valor puntual (FAP o AF_p):

$$AF_p = AF_e\frac{a}{m_1};$$

donde:

$$AF_e = \frac{RR - 1}{RP};$$

$$RR = \frac{an_0}{cn_1} = \frac{p_1}{p_0}.$$

- IC inferior o límite inferior (FAP o AFP):

$$L_I\left(AF_p\right) = 1 - \left(1 - AF_p\right)\exp\left(z_{1-\alpha/2}\sqrt{Var\left[Ln\left(1 - AF_p\right)\right]}\right)$$

$$L_I\left(AF_p\right) = 1 - \left(1 - AF_e\frac{a}{m_1}\right)\exp\left(1{,}96\sqrt{\left(\frac{AF_e\frac{a}{m_1}}{1 - AF_e\frac{a}{m_1}}\right)^2\left\{\frac{\frac{b}{an_1} + \frac{d}{cn_0}}{\left(\frac{an_0}{cn_1} - 1\right)^2} + \frac{2}{a\left(\frac{an_0}{cn_1} - 1\right)} + \frac{c}{am_1}\right\}}\right)$$

donde:

$$AF_p = AF_e\frac{a}{m_1};$$

$$Var\left[Ln\left(1 - AF_p\right)\right] = \left(\frac{AF_p}{1 - AF_p}\right)^2\left\{\frac{Var(LnRR)}{(RR - 1)^2} + \frac{2}{a(RR - 1)} + \frac{c}{am_1}\right\};$$

Z = 1,96.

- IC superior o límite superior (FAP o AF_P)

$$L_S\left(AF_p\right) = 1 - \left(1 - AF_p\right)\exp\left(-z_{1-\alpha/2}\sqrt{Var\left[Ln\left(1 - AF_p\right)\right]}\right)$$

$$Ls\left(AF_p\right)=1-\left(1-AF_e\frac{a}{m_1}\right)\exp\left(-1{,}96\sqrt{\left(\frac{AF_e\frac{a}{m_1}}{1-AF_e\frac{a}{m_1}}\right)^2\left\{\frac{\frac{b}{an_1}+\frac{d}{cn_0}}{\left(\frac{an_0}{cn_1}-1\right)^2}+\frac{2}{a\left(\frac{an_0}{cn_1}-1\right)}+\frac{c}{am_1}\right\}}\right)$$

donde:

$$AF_p=AF_e\frac{a}{m_1};$$

$$Var\left[Ln\left(1-AF_p\right)\right]=\left(\frac{AF_p}{1-AF_p}\right)^2\left\{\frac{Var\left(LnRR\right)}{\left(RR-1\right)^2}+\frac{2}{a\left(RR-1\right)}+\frac{c}{am_1}\right\};$$

Z = 1,96.

– Fracción de prevención para la población (FPP o PF_p): se calcula si RR < 1:

  • Valor puntual (FPP o PF_p):

$$PF_p=PF_e\frac{n_1}{n},$$

donde:

$PF_e = 1 - RR$

$$RR=\frac{an_0}{cn_1}=\frac{p_1}{p_0};$$

$n_1 = (a + b)$; $n_0 = (c + d)$.

  • IC inferior o límite inferior (FPP o PFp):

$$\left[L_I\left(PF_e\right)\frac{n_1}{n_1+n_0}\right]=\left[1-\frac{an_0}{cn_1}\exp\left(1{,}96\sqrt{\frac{b}{an_1}+\frac{d}{cn_0}}\right)\frac{n_1}{n_1+n_0}\right]$$

donde:

$$L_I\left(PF_e\right)=\left[1-\frac{an_0}{cn_1}\exp\left(1{,}96\sqrt{\frac{b}{an_1}+\frac{d}{cn_0}}\right)\right];$$

$n_1 = (a + b)$; $n_0 = (c + d)$.

  • IC superior o límite superior (FPP o PF_p):

$$\left[L_I\left(PF_e\right)\frac{n_1}{n_1+n_0}\right]=\left[1-\frac{an_0}{cn_1}\exp\left(-1{,}96\sqrt{\frac{b}{an_1}+\frac{d}{cn_0}}\right)\frac{n_1}{n_1+n_0}\right]$$

donde:

$$L_I\left(PF_e\right)=\left[1-\frac{an_0}{cn_1}\exp\left(-1{,}96\sqrt{\frac{b}{an_1}+\frac{d}{cn_0}}\right)\right];$$

$n_1 = (a + b)$;

$n_0 = (c + d)$.

Resultados en estudios de cohorte, datos de la incidencia acumulada

Se muestra ahora un ejemplo para un nivel de confianza de 95 %.

Medidas de frecuencia

- Prevalencia en expuestos (riesgo en expuestos): valor puntual.
- Prevalencia en no expuestos (riesgo en no expuestos): valor puntual.

Medidas de asociación

- Riesgo relativo: valor puntual y por intervalos de confianza (Katz).
- Diferencia de riesgos: valor puntual y por intervalos de confianza.
- *Odds ratio* de enfermedad: valor puntual y por intervalos de confianza (Woolf).

Medidas de impacto

- Fracción atribuible en los expuestos: valor puntual e intervalo de confianza.
- Fracción prevenible en expuestos: valor puntual e intervalo de confianza.
- Fracción atribuible poblacional: valor puntual y por intervalos de confianza.
- Fracción de prevención poblacional: valor puntual y por intervalos de confianza.

Estudio de cohorte con datos de tasas de incidencia

Se tienen un total de T personas-tiempo, de las cuales T_1 están expuestas al factor y T_0 no están expuestas:

Factor de riesgo	Casos	Personas-tiempo
Expuestos	a	T_1
No expuestos	b	T_0
Total	n = a + b	$T = T_1 + T_0$

Medidas de frecuencia

Tasas de incidencia:

- Incidencia en grupo de expuestos (riesgo en expuestos):

$$I_1 = \frac{a}{T_1}$$

- Incidencia en grupo de no expuestos (riesgo en no expuestos):

$$I_0 = \frac{b}{T_0}$$

Medidas de asociación

– Razón de tasas de incidencia (RI o IR):

- Valor puntual (IR):

$$IR = \frac{aT_0}{bT_1}$$

- Por intervalos de confianza la razón de tasas de incidencia (RI o IR):
  Si V < 5 se calcula un intervalo de confianza exacto:
  Fórmula del IC inferior (RI o IR):

$$\text{IC inferior} = \left[\frac{p_I}{1-p_I}\frac{T_0}{T_1}\right]$$

  Fórmula del IC superior (RI o IR):

$$\text{IC superior} = \left[\frac{p_S}{1-p_S}\frac{T_0}{T_1}\right]$$

  Si V ≥ 5 se calcula un intervalo de confianza aproximado como sigue:
  Fórmula del IC inferior (RI o IR):

$$\text{IC inferior} = \left[IR\exp\left(-z_{1-\alpha/2}\sqrt{Var\left[Ln(IR)\right]}\right)\right]$$

  Fórmula del IC superior (RI o IR):

$$\text{IC superior} = \left[IR\exp\left(z_{1-\alpha/2}\sqrt{Var\left[Ln(IR)\right]}\right)\right]$$

  donde:

$$V = \frac{nT_1T_0}{T^2};$$

  $n = a + b$;
  $T = T_1 + T_0$.

$$p_I = \frac{a}{a+(n-a+1)F_{x,y,\alpha/2}};$$

$$p_S = \frac{a+1}{a+1+(n-a)\dfrac{1}{F_{z,t,\alpha/2}}};$$

$$Var\left[\ln(IR)\right] = \frac{1}{a}+\frac{1}{b}$$

  (p_I, p_S): intervalo de confianza exacto para la proporción a/n;
  $F_{r,s}$: variable con distribución F de Snedecor con r y s grados de libertad;
  $F_{r,s,\alpha}$: cuantil de la distribución F de Snedecor con r y s grados de libertad, que deja una cola α a la izquierda;
  $x = 2n - 2a + 2 = 2(b + 1)$;

y = 2a;

z = 2a + 2;

t = 2n − 2a = 2b;

1 − α: nivel de confianza;

zα: cuantil de la distribución normal que deja una cola α a la izquierda.

- Diferencia de tasas de incidencia (DI o ID):
  - Fórmula del valor puntual (DI o ID):

$$ID = \frac{a}{T_1} - \frac{b}{T_0}$$

  - Fórmula del IC inferior (DI o ID):

$$ICInferior(ID) = \left(ID - 1{,}96\sqrt{Var(ID)}\right)$$

donde:

$$Var(ID) = \frac{a}{(T_1)^2} + \frac{b}{(T_0)^2} = \left(\frac{a}{T_1} - \frac{b}{T_0}\right) - \left[1{,}96\sqrt{\frac{a}{(T_1)^2} + \frac{b}{(T_0)^2}}\right]$$

  - Fórmula del IC Superior (DI o ID):

$$ICSuperior(ID) = \left(ID + 1{,}96\sqrt{Var(ID)}\right)$$

donde:

$$Var(ID) = \frac{a}{(T_1)^2} + \frac{b}{(T_0)^2} = \left(\frac{a}{T_1} - \frac{b}{T_0}\right) + \left[1{,}96\sqrt{\frac{a}{(T_1)^2} + \frac{b}{(T_0)^2}}\right]$$

Medidas de impacto

- Fracción atribuible en los expuestos (FAE o AF_e): se calcula si IR ≥ 1:
  - Fórmula del valor puntual (FAE o AF_e):

$$AF_e = \frac{IR - 1}{IR};$$

donde:

$$IR = \frac{aT_0}{bT_1}.$$

  - IC inferior o límite inferior (FAE o AF_e):

$$ICInferior(AF_e) = \frac{L_I(IR) - 1}{L_I(IR)};$$

donde:

$$L_I(IR) = \left[\frac{aT_0}{bT_1}\exp\left(-1{,}96\sqrt{\frac{1}{a} + \frac{1}{b}}\right)\right]$$

$$ICInferior(AF_e) = \frac{\left[\frac{aT_0}{bT_1}\exp\left(-1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}}\right)\right]-1}{\left[\frac{aT_0}{bT_1}\exp\left(-1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}}\right)\right]}$$

- IC superior o límite superior (FAE o AF_e):

$$ICSuperior(AF_e) = \frac{L_S(IR)-1}{L_S(IR)}$$

donde:

$$L_s(IR) = \left[\frac{aT_0}{bT_1}\exp\left(1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}}\right)\right]$$

$$ICInferior(AF_e) = \frac{\left[\frac{aT_0}{bT_1}\exp\left(1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}}\right)\right]-1}{\left[\frac{aT_0}{bT_1}\exp\left(1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}}\right)\right]}$$

– Fracción de prevención en los expuestos (FPE o PF_e): se calcula si IR < 1:

- Valor puntual (FPE o PF_e):
  $PF_e = 1 - IR$
  donde:

$$IR = \frac{aT_0}{bT_1}.$$

- IC inferior o límite inferior (FPE o PF_e):
  $ICInferior(PF_e) = 1 - LS(IR)$
  donde:

$$L_S(IR) = \left[\frac{aT_0}{bT_1}\exp\left(1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}}\right)\right]$$

$$ICInferior(PF_e) = 1-\left[\frac{aT_0}{bT_1}\exp\left(1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}}\right)\right]$$

- IC superior o límite superior (FPE o PF_e):
  $ICSuperior(PF_e) = 1 - LS(IR)$
  donde:

$$L_I(IR) = \left[\frac{aT_0}{bT_1}\exp\left(-1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}}\right)\right]$$

$$L_I(IR) = 1-\left[\frac{aT_0}{bT_1}\exp\left(-1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}}\right)\right]$$

– Fracción atribuible poblacional (FAP o AF_p): se calcula si IR ≥ 1:
  - Valor puntual (FAP o AF_p):

$$AF_p = AF_e \frac{a}{n};$$

donde:

$$AF_e = \frac{IR-1}{IR};$$

$$IR = \frac{aT_0}{bT_1}.$$

  - IC inferior o límite inferior (FAP o AF_p):

$$L_I\left(AF_p\right) = 1-\left(1-AF_p\right)\exp\left(1{,}96\sqrt{Var\left[Ln\left(1-AF_p\right)\right]}\right).$$

donde:

$$AF_p = AF_e \frac{a}{n};$$

$$Var\left[Ln\left(1-AF_p\right)\right] = \left(\frac{AF_p}{1-AF_p}\right)^2 \left\{\frac{Var(LnIR)}{(IR-1)^2} + \frac{2}{a(IR-1)} + \frac{b}{an}\right\};$$

$$AF_e = \frac{IR-1}{IR};$$

$$IR = \frac{aT_0}{bT_1};$$

$$Var\left[ln(IR)\right] = \frac{1}{a} + \frac{1}{b};$$

n = a + b.

  - IC superior o límite superior (FAP o AF_p):

$$L_S\left(AF_p\right) = 1-\left(1-AF_p\right)\exp\left(-1{,}96\sqrt{Var\left[Ln\left(1-AF_p\right)\right]}\right).$$

Donde:

$$AF_p = AF_e \frac{a}{n};$$

$$Var\left[Ln\left(1-AF_p\right)\right] = \left(\frac{AF_p}{1-AF_p}\right)^2 \left\{\frac{Var(LnIR)}{(IR-1)^2} + \frac{2}{a(IR-1)} + \frac{b}{an}\right\};$$

$$AF_e = \frac{IR-1}{IR};$$

$$IR = \frac{aT_0}{bT_1};$$

$$Var\left[ln(IR)\right] = \frac{1}{a} + \frac{1}{b};$$

n = a + b.

– Fracción de prevención poblacional (FPP o PF_p): se calcula si IR < 1:
  - Valor puntual (FPP o PF_p):

$$PF_p = PF_e \frac{T_1}{T};$$

donde:

$PF_e = 1 - IR;$

$T = T_1 + T_0;$

$$IR = \frac{aT_0}{bT_1}.$$

  - IC inferior o límite inferior (FPP) o (PF_p):

$$ICInferior(PF_e) = (L_I(PF_e)\frac{T_1}{T}$$

donde:

$$L_I(PF_e) = \left[1 - \frac{aT_0}{bT_1}\exp\left(1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}}\right)\right];$$

$T = T_1 + T_0$

$$ICInferior(PF_e) = \left[1 - \frac{aT_0}{bT_1}\exp\left(1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}}\right)\right]\left(\frac{T_1}{T}\right)$$

  - IC superior o límite superior (FPP o PF_p):

$$ICSuperior(PF_e) = (L_S(PF_e)\frac{T_1}{T}$$

donde:

$$L_S(PF_e) = \left[1 - \frac{aT_0}{bT_1}\exp\left(-1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}}\right)\right];$$

$T = T_1 + T_0$

$$ICSuperior(PF_e) = \left[1 - \frac{aT_0}{bT_1}\exp\left(-1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}}\right)\right]\left(\frac{T_1}{T}\right)$$

Resultados en estudios de cohorte, datos por tasas de incidencia

Se muestra ahora un ejemplo para un nivel de confianza de 95 %.

Medidas de frecuencia

– Tasa de incidencia en expuestos (riesgo en expuestos): valor puntual.
– Tasa de incidencia en no expuestos (riesgo en no expuestos): valor puntual.

Medidas de asociación

- Razón de tasas de incidencia: valor puntual y por intervalos de confianza.
- Diferencia de tasas de incidencia: valor puntual y por intervalos de confianza.

Medidas de impacto

- Fracción atribuible en los expuestos: valor puntual y por intervalos de confianza.
- Fracción de prevención en los expuestos: valor puntual y por intervalos de confianza.
- Fracción atribuible poblacional: valor puntual y por intervalos de confianza.
- Fracción de prevención poblacional: valor puntual y por intervalos de confianza.

Para estudio de cohorte, datos de densidad de incidencia

- Tabla de contingencia 2x2 simple para estudio de cohorte para densidad de incidencia (tiempo total de seguimiento por persona) e introducir datos en las celdas (a, c, T_1 = TSPE, T_0 = TSPNE) de la misma.
- Calcular los totales marginales por columnas (A + C) y (T_1 + T_0) (Tabla 3.34).

Tabla 3.34. Ejemplo de tabla de contingencia para densidad de incidencia (tiempo por persona en seguimiento) en estudios de cohorte

Exposición a	Nuevos enfermos	Total de tiempo por persona de seguimiento
Exposición +	A	T_1
Exposición –	C	T_0
Total	**A + C**	**T_1 + T_0**

Donde:
Celda A: nuevos enfermos previa exposición.
Celda B: nuevos enfermos sin previa exposición.
Celda T_1 (TSPE): tiempo total por persona de seguimiento en expuestos.
Celda T_0 (TSPNE): tiempo total por persona de seguimiento en no expuestos.
Celda A + C: total de nuevos enfermos (expuestos y no expuestos) de la cohorte.
Celda T_1 + T_0: total del tiempo por persona de seguimiento de la cohorte.

- Acceder a fórmulas para la estimación (cálculo) puntual de las medidas de frecuencia y asociación correspondientes e introducir en las mismas los valores respectivos de las celdas (A, C, T_1, T_0) y totales marginales (A + C) y (T_1 + T_0) y n = (a + c) + (T_1 + T_0). Seguir con nuevo empleo de los datos en las celdas de la tabla 3.31.

Presentación de los resultados:

- Densidad de incidencia en expuestos (DIExp):
  $DIExp = A/T_1$
- Densidad de incidencia en no expuestos (DINoExp):
  $DINoExp = C/T_2$
- Densidad de incidencia total (*incidence rate*) (DIT):
  $DIT = (A + C)/(T_1 + T_2)$
- Razón de densidad de incidencia (RDI):
  $RDI = (A/T_1)/(C/T_2)$

Resultados en estudios de cohorte, datos de densidad de incidencia

- Densidad de incidencia en expuestos: valor puntual.
- Densidad de incidencia en no expuestos: valor puntual.
- Densidad de incidencia total: valor puntual.
- Razón de densidades de incidencia: valor puntual.

Para estudio de cohorte, datos de mortalidad estandarizada

- Conformar una tabla de contingencia para estudio de cohorte referido a la mortalidad estandarizada e introducir datos en las celdas (A, B, C) (Tabla 3.35).

Tabla 3.35. De contingencia para estudio de cohorte referido a la mortalidad estandarizada tomada como ejemplo

Grupos de edades	Muertes observadas (A)	Años-persona de seguimiento población referencia (B)	Tasa mortalidad población referencia por 10^n (C)	Muertes esperadas (C · B)/1 000
Totales	**Total de A**	**Total de B**	**Total de C**	**Total de (C · B)/1 000**

Donde:

(A): total de muertes observadas en la cohorte.

(B): total de años-persona de seguimiento en la población de referencia.

(C): tasa mortalidad en la población de referencia (10^n).

(C · B)/1 000: muertes esperadas en la población de referencia.

– Calcular el total de la celda A a partir de la sumatoria de sus valores.
– Calcular el total de la celda B a partir de la sumatoria de sus valores.
– Calcular el total de la celda C a partir de la sumatoria de sus valores.
– Calcular el total de la celda (C · B)/1 000 a partir de la multiplicación de (C · B)/1 000.
– Introducir los valores totales de A, B, C y (C · B)/ 1000 en la tabla.
– Acceder a fórmulas para la estimación (cálculo) puntual y por intervalos de confianza de la medida razón de mortalidad estandarizada.

Después se debe calcular:

– Valor puntual de:
  • Tasa de mortalidad estandarizada (TME):

$$\text{TME} = \frac{\text{Total de A}}{\text{Total de B} \cdot \text{C}} \cdot 100$$

  • Razón de mortalidad estandarizada (RME):

$$\text{RME} = \frac{\text{Total de A}}{\text{Total de B} \cdot \text{C}}$$

  • Exceso de mortalidad estandarizada (EME):

$$\text{EME} = \left(\frac{\text{Total de A}}{\text{Total de B} \cdot \text{C}} \cdot 100 \right) - 100$$

– IC inferior o límite inferior (RME):

$$\text{ICInferior(RME)} = \frac{\text{Total de A}}{\text{Total de B} \cdot \text{C}} - 1{,}96 \left(\frac{\frac{\text{Total de A}}{\text{Total de B} \cdot \text{C}}}{\sqrt{\text{A}}} \right)$$

– IC superior o límite superior (RME):

$$\text{ICSuperior(RME)} = \frac{\text{Total de A}}{\text{Total de B} \cdot \text{C}} + 1{,}96 \left(\frac{\frac{\text{Total de A}}{\text{Total de B} \cdot \text{C}}}{\sqrt{\text{A}}} \right)$$

– Si la tasa de mortalidad estandarizada (TME) > 100, entonces calcular:

$$\text{Exceso de riesgo de mortalidad} = \text{TME}(\text{calculado}) - 100 = \frac{\text{Total de A}}{\text{Total de B} \cdot \text{C}} \cdot 100$$

Resultados en estudios de cohortes, datos de mortalidad estandarizada

Medidas de frecuencia

– Total de muertes observadas en población de estudio (A): valor puntual.
– Total de muertes esperadas (población de referencia) (C B): valor puntual.
– Tasa de mortalidad estandarizada (A/B · C) · 100): valor puntual.

Medidas de asociación

– Razón de densidades de incidencia por mortalidad estandarizada: (valor puntual y por intervalos)

Para estudio de caso-control apareado (caso 1-control 1)

– Conformar una tabla de contingencia 2x2 simple para estudio de caso control apareado (caso 1-control 1) e introducir datos en las celdas (de pares caso-control) n_{11} (a), n_{10} (b), n_{01} (c), n_{00} (d) (Tabla 3.36).

Tabla 3.36. De contingencia 2x2 simples en estudio de casos y controles pareados (caso 1-control 1)

		Controles		
		Expuestos	**No expuestos**	
Casos	Expuestos	$(n_{11})(a)$	$(n_{10})(b)$	Número pares $n_{11} + n_{10}$ $(a + b)$
	No expuestos	$(n_{01})(c)$	$(n_{00})(d)$	Número pares $n_{01} + n_{00}$ $(c + d)$
		Número pares $n_{11} + n_{01}$ $(a + c)$	Número pares $n_{10} + n_{00}$ $(b + d)$	Total pares (n) $n_{11} + n_{10} + n_{01} + n_{00}$ o $(a + b) + (c + d)$

Donde:
Celda n_{11} (a): pares de caso y control expuestos.
Celda n_{10}(b): pares de casos expuestos y controles no expuestos.
Celda n_{01} (c): pares de caso no expuesto y control expuesto.
Celda n_{00} (d): pares de caso no expuesto y control no expuesto.
Celda $n_{11} + n_{01}$ (a + c): total de pares de controles expuestos.
Celda $n_{10} + n_{00}$ (b + d): total de pares de controles no expuestos.
Celda $n_{11} + n_{10}$ (a + b): total de pares de casos expuestos.
Celda $n_{01} + n_{00}$ (c + d): total de pares de casos no expuestos.
Celda n = (a + b) + (c + d) o $(n_{11} + n_{10}) + (n_{01} + n_{00})$: total de pares de la muestra.

– Calcular los totales marginales (por filas: $n_{11} + n_{10}$ (a + b) y $n_{01} + n_{00}$ (c + d), así como los totales marginales por columnas $n_{11} + n_{01}$ (a + c) y $n_{10} + n_{00}$ (b + d), así como: n = ($n_{11} + n_{10} + n_{01} + n_{00}$) o n = (a + b) + (c + d).
– Acceder a fórmulas para la estimación (cálculos) puntual de medidas de frecuencia, así como la estimación puntual y por intervalos de confianza de la medida de asociación en estudio de caso y control apareado para valor de α = 0,05 a partir de la introducir de los valores respectivos de las celdas.

Estudios caso-control pareados (caso 1-control 1)

En la tabla 3.38 se tienen n pares de casos y controles que se representan siguiendo la tabla 3.37.

Tabla 3.38. De contingencia 2x2 simples en estudio de casos y controles pareados (relación: caso 1, control 1) según exposición

		Controles	
		Expuestos	**No expuestos**
Casos	Expuestos	a(n_{11})	b(n_{10})
	No Expuestos	c(n_{01})	d(n_{00})

Medidas de frecuencia

– Proporción de casos expuestos:

$$p_1 = \frac{a+b}{n}$$

– Proporción de controles expuestos:

$$p_0 = \frac{a+c}{n}$$

donde:
Celda n_{11} (a): pares de caso y control expuestos.
Celda n_{10} (b): pares de casos expuestos y controles no expuestos.
Celda n_{01} (c): pares de caso no expuesto y control expuesto.
Celda n_{00} (d): pares de caso no expuesto y control no expuesto.
Celda $n_{11} + n_{01}$ (a + c): total de pares de controles expuestos.
Celda $n_{10} + n_{00}$ (b + d): Total de pares de controles no expuestos.
Celda $n_{11} + n_{10}$ (a + b): total de pares de casos expuestos.
Celda $n_{01} + n_{00}$ (c + d): total de pares de casos no expuestos.
Celda n = (a + b) + (c + d) o ($n_{11} + n_{10}$) + ($n_{01} + n_{00}$): total de pares de la muestra.

Medidas de asociación

– *Odds ratio* de exposición (OR):

- Valor puntual (OR):

$$OR = \frac{b}{c}$$

- IC inferior o límite inferior (OR):

$$L_I(OR) = \frac{L_I(p)}{1 - L_I(p)}$$

- IC superior o límite superior (OR):

$$L_S(OR) = \frac{L_S(p)}{1 - L_S(p)}$$

donde:

$L_I(p)$ y $L_{S(p)}$: límites inferior y superior respectivos de un intervalo de confianza para la proporción:

$$p = \frac{b}{m} = \frac{OR}{1 + OR}$$

m = b + c (que es el número de pares discordantes).

Si $mp(1 - p) \geq 5$, se calcula un intervalo de confianza aproximado para p como sigue:

$$L_I(p) = \frac{D - \sqrt{D^2 - 4AE}}{2A};$$

donde:

$A = m(m + 3{,}84)$;

$D = m[2(b - 1) + 3{,}84]$

$E = (b - 1)^2$

$m = b + c$

$$L_S(p) = \frac{B + \sqrt{B^2 - 4AC}}{2A};$$

donde:

$A = m(m + 3{,}84)$;

$B = m[2(b + 1) + 3{,}84]$

$C = (b + 1)^2$

$m = b + c$

$$p = \frac{b}{m} = \frac{OR}{1 + OR}$$

Si mp(1 – p) < 5, se calcula un intervalo de confianza exacto para p como sigue:

$$L_i(p) = \frac{b}{b + (m - b + 1)F_{z,t,\alpha/2}}$$

$$L_s(p) = \frac{b+1}{b + 1 + (m - b)\frac{1}{F_{r,s,\alpha/2}}}$$

donde:

$F_{u,v}$: variable con distribución F de Snedecor con u y v grados de libertad;

$F_{u,v,\alpha}$: cuantil de la distribución F de Snedecor con u y v grados de libertad, que deja una cola α a la izquierda;

z = 2m – 2b + 2;

t = 2b;

r = 2b + 2;

s = 2m – 2b.

Resultados en Estudios caso-control pareados (con apareamiento o *matching*), 1 caso para 1 control

Medidas de frecuencia

- Proporción de casos expuestos: valor puntual.
- Proporción de controles expuestos: valor puntual.

Medidas de asociación

- *Odds Ratio* de enfermedad: valor puntual e intervalos de confianza.

Estudios caso-control pareados, K controles por caso (1< K< 5)

Se tienen n pares de casos expuestos y no expuestos versus K controles que se representan siguiendo la tabla 3.39.

Tabla 3.39. De contingencia en estudio de casos y controles pareados, K controles por caso (1< K< 5)

		Número de controles expuestos			
		0	1	...	k
Casos	Expuestos	a_0	a_1	...	a_k
	No expuestos	c_0	c_1	...	c_k

Medidas de frecuencia

Proporción de expuestos:

– Casos:

$$p_1 = \frac{T_1}{n}$$

– Controles:

$$p_0 = \frac{T_0}{nk}$$

donde:

$T_1 = \sum_{i=0}^{k} a_i$: número de casos expuestos;

$T_0 = \sum_{i=1}^{k} i(a_i + c_i)$: número de controles expuestos.

Medidas de asociación

– *Odds ratio* (OR): estimador Mantel-Haenszel de la *odds ratio*:

- Valor puntual:

$$OR = \frac{\sum_{i=0}^{k}(k-i)a_i}{\sum_{i=0}^{k} i c_i}$$

- IC inferior o límite inferior (OR):

$$ICInferior(OR) = OR\exp\left(-z_{1-\alpha/2}\sqrt{Var\left[Ln(OR)\right]}\right)$$

- IC superior o límite superior (OR):

$$ICSuperior(OR) = OR\exp\left(z_{1-\alpha/2}\sqrt{Var\left[Ln(OR)\right]}\right)$$

donde:

$$Var\left[Ln(OR)\right] = \frac{1}{\chi^2_{M-H}};$$

$\chi^2_{M\text{-}H}$: estadístico chi-cuadrado del test de Mantel-Haenszel.

Resultados en Estudios caso-control pareados (con apareamiento o *matching*), relación caso: 1, control >1

Medidas de frecuencia

Proporción de expuestos, en casos y en controles.

Medidas de asociación

- *Odds ratio* (valor puntual e intervalo de confianza) (Mantel-Haenszel).
- Prueba estadística en el contraste de hipótesis.
- Test de asociación de Mantel-Haenszel.

Formulario resumen de las medidas de frecuencia, asociación e impacto empleadas en estudios epidemiológicos analíticos observacionales

Medidas de frecuencia

- Tasa de prevalencia (TPrev):

$$TPrev = \frac{a+c}{n} = \frac{a+c}{a+b+c+d} \cdot 100$$

- Tasa de enfermar los expuestos al factor (TEExp):

$$TEExp = \frac{a}{a+b} \cdot 100$$

- Tasa de enfermar los no expuestos al factor (TENoExp):

$$TENoExp = \frac{c}{c+d} \cdot 100$$

- Tasa de prevalencia relativa (TPrevR):

$$TPrevR = \frac{a/(a+c)}{c/(a+c+d)} \cdot 100$$

- Tasa de incidencia acumulada para la cohorte expuesta (TIncAC):

$$TIncAc = \frac{A}{A+B}$$

- Tasa de incidencia acumulada para la cohorte no expuesta (TIAcCohNoExp):

$$TIAcCohNoExp = \frac{C}{C+D} \cdot 100$$

- Tasa de incidencia acumulada para la cohorte total (TIncAcCohTot):

$$TIAcCohTot = \frac{A}{A+B} + \frac{C}{C+D} \cdot 100$$

- Densidad de incidencia en expuestos (por tasa de incidencia) (DIExpTI):

$$DIExpTI = \frac{A}{PTSE} \cdot 100$$

Donde:

PTSE: persona por tiempo de seguimiento.

– Densidad de incidencia en no expuestos (por tasa de incidencia) (DINoExpTI):

$$\text{DINoExpTI} = \frac{C}{\text{PTSE}} \cdot 100$$

– Densidad de incidencia en expuestos (por tiempo por persona seguimiento) (DIExp):

$$\text{DIExp} = \frac{A}{T_1} = \frac{A}{\text{TPSE}}$$

donde:

TPSE: tiempo por persona de seguimiento.

– Densidad de incidencia en no expuestos (por tiempo por persona seguimiento) (DINoExp):

$$\text{DINoExp} = \frac{C}{T_2} = \frac{C}{\text{TPSNE}}$$

Medidas de asociación

Riesgo relativo o razón de riesgo (RR), razón de prevalencia de enfermedad o prevalencia relativa de enfermedad (para estudio transversal y de caso-control, respectivamente) o razón de incidencia acumulada (para estudio de cohorte):

– En tablas de contingencia 2x2 simples se estima por:

- Estudio transversal:

$$RR = \frac{a/m_1}{c/m_2} = \frac{a(c+d)}{c(a+b)}$$

- Estudio caso-control no apareado:

$$RR = \frac{a/c}{b/d}$$

- Estudio cohorte por incidencia acumulada:

$$RR = \frac{A/(A+B)}{C/(C+D)}$$

– En tablas de contingencia 2x2 estratificadas se estima por:

- Razón de riesgo crudo puntual:

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{a/n_1}{cn_0} = \frac{an_0}{cn_1}$$

- Razón de riesgo crudo IC superior:

$$\text{IC RR superior} = \frac{an_0}{cn_1}\exp\left(1{,}96\sqrt{\frac{b}{an_1}+\frac{d}{cn_0}}\right) \quad \text{o}$$

$$\text{IC RR superior} = \text{RP} \cdot \text{FE} = \frac{an_0}{cn_1}\exp\left(1{,}96\sqrt{\frac{1}{a}-\frac{1}{a+b}+\frac{1}{c}-\frac{1}{c+d}}\right)$$

donde:

$$FE = \exp\left(1{,}96\sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}\right)$$

- Razón de riesgo crudo IC inferior:

$$\text{IC RR inferior} = \frac{an_0}{cn_1}\exp\left(-1{,}96\sqrt{\frac{b}{an_1} + \frac{d}{cn_0}}\right) \quad \text{o}$$

$$\text{IC RR inferior} = \frac{RP}{FE} = \frac{\frac{an_0}{cn_1}}{\exp\left(1{,}96\sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}\right)}$$

donde:

$$FE = \exp\left(1{,}96\sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}\right)$$

- Razón de riesgo estratificada puntual (Tabla 1):

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

- Razón de riesgo estratificada IC superior (Tabla 1):

$$\text{IC RR superior} = \frac{an_0}{cn_1}\exp\left(1{,}96\sqrt{\frac{b}{an_1} + \frac{d}{cn_0}}\right)$$

- Razón de riesgo estratificada IC inferior (Tabla 1):

$$\text{IC RR inferior} = \frac{an_0}{cn_1}\exp\left(-1{,}96\sqrt{\frac{b}{an_1} + \frac{d}{cn_0}}\right)$$

- Razón de riesgo estratificada puntual (Tabla 2):

$$RR = \frac{a_2/(a_2+b_2)}{c_2/(c_2+d_2)}$$

- Razón de riesgo estratificada IC superior (Tabla 2):

$$\text{IC RR superior} = \frac{an_0}{cn_1}\exp\left(1{,}96\sqrt{\frac{b}{an_1} + \frac{d}{cn_0}}\right)$$

- Razón de riesgo estratificada IC inferior (Tabla 2):

$$\text{IC RR inferior} = \frac{an_0}{cn_1}\exp\left(-1{,}96\sqrt{\frac{b}{an_1} + \frac{d}{cn_0}}\right)$$

- Razón de riesgo ponderado (RRp):

$$RRp = \exp\left[\frac{\sum_i W_i \text{Ln}(RR_i)}{\sum_i W_i}\right]$$

donde:

$$RRi = \frac{a(c+d)}{c(a+b)}$$

$$W_i = \frac{1}{Var\left[Ln(RR_i)\right]}$$

$$Var\left[Ln(RR)\right] = \frac{b}{a(a+b)} + \frac{d}{c(c+d)}$$

i: número de estratos (1, 2,..., k).

- Razón de riesgo combinada (Mantell-Haenszell):

$$RR = \frac{\left(\frac{a_1(c_1+d_1)}{n_1} + \frac{a_2(c_2+d_2)}{n_2}\right)}{\left(\frac{c_1(a_1+b_1)}{n_1} + \frac{c_2(a_2+b_2)}{n_2}\right)}$$

- Riesgo atribuible absoluto, diferencia de riesgo o exceso de riesgo (RA):

$$RA = \frac{a}{m_1} - \frac{c}{m_2} = \frac{A}{A+B} - \frac{C}{C+D}$$

- Riesgo atribuible absoluto porcentual (RA%) o porcentaje de exceso de riesgo (%ER):

$$RA\% = \left(\frac{A}{A+B} - \frac{C}{C+D}\right) \cdot 100$$

- Riesgo atribuible en expuestos, riesgo atribuible proporcional en expuestos (RAE) o fracción etiológica atribuible en expuestos (FEE):

$$RAE = \frac{\left(\frac{A}{A+B} - \frac{C}{C+D}\right)}{\left(\frac{A}{A+B}\right)}$$

RAE = RR – 1/ RR (para estudio cohorte por incidencia acumulada)
RAE = IR –1/IR (para estudio cohorte por tasa de incidencia)
RAE = OR – 1 / OR (para estudio de caso control no apareado)

- Riesgo atribuible en expuestos porcentual, riesgo atribuible proporcional en expuestos porcentual (%RAE) o fracción etiológica atribuible en expuestos porcentual (FEE%):

$$\%RAE = \frac{\left(\frac{A}{A+B} - \frac{C}{C+D}\right)}{\left(\frac{A}{A+B}\right)} \cdot 100$$

%RAE = (RR – 1/RR) · 100 (para estudio cohorte por incidencia acumulada)
%RAE = (IR –1/IR) · 100 (para estudio cohorte por tasa de incidencia)
%RAE = (OR – 1/OR) · 100 (para estudio de caso control no apareado)

- Riesgo atribuible poblacional, riesgo atribuible proporcional poblacional (RAP) o fracción etiológica atribuible poblacional (FAP):

$$RAP = \frac{\left(A + \frac{C}{A} + B + C + D\right) - \left(\frac{C}{C+D}\right)}{\left(A + \frac{C}{A} + B + C + D\right)}$$

- Riesgo Atribuible poblacional porcentual, riesgo atribuible proporcional poblacional porcentual (%RAP) o fracción etiológica atribuible poblacional porcentual (FAP%):

$$\%RAP = \frac{\left(A + \frac{C}{A} + B + C + D\right) - \left(\frac{D}{C+D}\right)}{\left(A + \frac{C}{A} + B + C + D\right)} \cdot 100$$

- Riesgo específico de enfermar los expuestos (REEExp):

$$REEExp = \frac{A}{A+B}$$

- Riesgo específico de enfermar los no expuestos (REENoExp):

$$REENoExp = \frac{C}{C+D}$$

- Riesgo total (RT):

$$RT = \frac{\left(\frac{A}{A+B} + \frac{C}{C+D}\right)}{(A+B+C+D)}$$

- Razón de disparidad o de desventaja o de productos cruzados (*odds ratio*) (OR):

$$OR = \frac{a \cdot d}{b \cdot c}$$

– En tablas de contingencia 2x2 simples la OR (*odds ratio*) se estima por:

- Razón de chances o razón de *Odds* (*odds ratio*) de exposición u oportunidad relativa de exposición estimada (OExp): se emplea en estudios de caso-control no apareado:

$$OExp = \frac{\left(\frac{a}{n_1}\right) / \left(\frac{c}{n_1}\right)}{\left(\frac{b}{n_2}\right) / \left(\frac{d}{n_2}\right)} = \frac{\left(\frac{a}{a+c}\right) / \left(\frac{c}{a+c}\right)}{\left(\frac{b}{b+d}\right) / \left(\frac{d}{b+d}\right)} = \frac{\left(\frac{a}{c}\right)}{\left(\frac{b}{d}\right)} = \frac{a \cdot d}{b \cdot c}$$

- Razón de *chances* o razón de *Odds* (*odds ratio*) de enfermedad u oportunidad relativa de enfermedad estimada (OREnf): se emplea para estudios de cohorte:

$$OREnf = \frac{\left(\frac{A}{A+B}\right) / \left(\frac{B}{A+B}\right)}{\left(\frac{C}{C+D}\right) / \left(\frac{D}{C+D}\right)} = \frac{\left(\frac{a}{b}\right)}{\left(\frac{c}{d}\right)} = \frac{a \cdot d}{b \cdot c}$$

- Razón de *chances* o de disparidad para estudios de caso-control apareado (ORpareada):

$$ORpareada = \frac{n_{10}}{n_{01}} = \frac{b}{b}$$

– En tablas de contingencia 2x2 estratificadas la OR (*Odds ratio*) se estima por:

- *Odds ratio* cruda puntual:

$$OR = \frac{a \cdot d}{c \cdot d} = \frac{a / c}{b / d}$$

- *Odds ratio* cruda IC superior:

$$RR = \left(\frac{a \cdot d}{c \cdot d}\right)\left[\exp\left(1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)\right]$$

- *Odds ratio* cruda IC inferior:

$$RR = \left(\frac{a \cdot d}{c \cdot d}\right)\left[\exp\left(1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)\right]$$

- *Odds ratio* estratificada 1 puntual:

$$OR = \frac{a \cdot d}{c \cdot d} = \frac{a / c}{b / c}$$

- *Odds ratio* estratificada 1 IC superior:

$$RR = \left(\frac{a \cdot d}{c \cdot d}\right)\left[\exp\left(1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)\right]$$

- *Odds ratio* estratficada 1 IC inferior:

$$RR = \left(\frac{a \cdot d}{c \cdot d}\right)\left[\exp\left(1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)\right]$$

- *Odds ratio* estratificada 2 puntual:

$$OR = \frac{a_2 \cdot d_2}{c_2 \cdot d_2} = \frac{a_2 / c_2}{b_2 / d_2}$$

- *Odds ratio* estratificada 2 IC superior:

$$OR = \left(\frac{a_2 \cdot d_2}{c_2 \cdot d_2}\right)\left[\exp\left(1{,}96\sqrt{\frac{1}{a_2}+\frac{1}{b_2}+\frac{1}{c_2}+\frac{1}{d_2}}\right)\right]$$

- *Odds ratio* estratificada 2 IC inferior:

$$OR = \left(\frac{a_2 \cdot d_2}{c_2 \cdot d_2}\right)\left[\exp\left(1{,}96\sqrt{\frac{1}{a_2}+\frac{1}{b_2}+\frac{1}{c_2}+\frac{1}{d_2}}\right)\right]$$

- *Odds ratio* ponderada (ORp):

$$ORp = \exp\left[\frac{\sum_i W_i Ln(OR_i)}{\sum_i W_i}\right]$$

donde:

$$ORi = \frac{a \cdot d}{b \cdot c}$$

$$W_i = \frac{1}{Var\left[Ln\left(OR_i\right)\right]}$$

$$Var\left[Ln\left(ORi\right)\right] = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

i: número de estratos (1, 2).

- Odds ratio combinada (Mantell-Haenszell) (ORCombM-H):

$$ORComM\text{-}H = \frac{\left(\frac{ad}{n_1} + \frac{a_2 d_2}{n_2}\right)}{\left(\frac{bc}{n_1} + \frac{b_2 c_2}{n_2}\right)}$$

- Razón de prevalencias o de prevalencias de enfermedad en expuestos y no expuestos (RPEnf) (similar al RR)

$$RPEnf = \frac{a/(a+b)}{c/(c+d)}$$

- Razón de prevalencia de exposición o de prevalencia de exposición en enfermos y no enfermos (RPNoEnf):

$$RPNoEnf = \frac{a/(a+c)}{b/(b+d)}$$

- Razón de incidencia acumulada o de incidencias acumuladas según exposición (RIAc):

$$RPIAc = \frac{A/(A+B)}{C/(C+D)}$$

- Razón de densidad de incidencia o de incidencias por tasas de incidencia (RDI):

$$RDI = \frac{A/PTSE}{C/PSE}$$

- Razón de densidades de incidencias por tiempo por persona de seguimiento [RDI(RR)]:

$$RDI(RR) = \frac{A/T_1}{C/T_2} = \frac{A/PTSE}{C/PTSNE}$$

- Razón de mortalidad estandarizada (RME):

$$RME = \frac{\sum A/\left(\sum B \cdot C\right)}{1000} \cdot 100 = \frac{\text{Total de } A/\left(\text{Total de } C \cdot B\right)}{1000} \cdot 100$$

Medidas de impacto

Medidas negativas por la exposición para la población

- Riesgo relativo o razón de riesgo (RR), razón de prevalencia de enfermedad o prevalencia relativa de enfermedad (para estudio transversal y de caso-control) o razón de incidencia acumulada (para estudio de cohorte):
  - Estudio transversal:

    $$RR = \frac{a / m_1}{c / m_2}$$

    $$RR = \frac{a \cdot (c + d)}{c \cdot (a + b)}$$

  - Estudio caso-control no apareado se estima como un OR:

    $$RR = \frac{a / c}{b / d}$$

  - Estudio cohorte:

    $$RR = \frac{A / (A + B)}{C / (C + D)}.$$

- Riesgo atribuible absoluto, diferencia de riesgo o exceso de riesgo (RA):

  $$RR = \frac{a}{m_1} - \frac{c}{m_2} = \frac{A}{A + B} - \frac{C}{C + D}$$

- Riesgo atribuible absoluto porcentual (RA%) o porcentaje de exceso de riesgo (%ER):

  $$RA\% = \left(\frac{A}{A + B} - \frac{C}{C + D} \right) \cdot 100$$

- Riesgo atribuible en expuestos, riesgo atribuible proporcional en expuestos (RAE) o fracción etiológica atribuible en expuestos (FEE):

  $$RAE = \left(\frac{A}{A + B} - \frac{C}{C + D} \right) / \left(\frac{A}{A + B} \right)$$

  - Para estudio cohorte por incidencia acumulada:
    RAE = RR – 1/RR
  - Para estudio cohorte por tasa de incidencia:
    RAE = IR –1/IR
  - Para estudio de caso control no apareado
    RAE = OR – 1/OR

- Riesgo atribuible en expuestos porcentual, riesgo atribuible proporcional en expuestos porcentual (%RAE) o fracción etiológica atribuible en expuestos porcentual (FEE%):

  $$\%RAE = \left[\left(\frac{A}{A + B} - \frac{C}{C + D} \right) / \left(\frac{A}{A + B} \right) \right] \cdot 100$$

- Para estudio cohorte por incidencia acumulada:
  %RAE = (RR – 1/RR) · 100
- Para estudio cohorte por tasa de incidencia:
  %RAE = (IR –1/IR) · 100
- Para estudio de caso control no apareado:
  %RAE = (OR – 1/OR) · 100

– Riesgo atribuible poblacional, riesgo atribuible proporcional poblacional (RAP) o fracción etiológica atribuible poblacional (FAP):

$$RAP = \left(\frac{A+C}{A+B+C+D} - \frac{C}{C+D}\right) / \left(\frac{A+C}{A+B+C+D}\right)$$

– Riesgo atribuible poblacional porcentual, riesgo atribuible proporcional poblacional porcentual (%RAP) o fracción etiológica atribuible poblacional porcentual (FAP%):

$$\%RAP = \left[\left(\frac{A+C}{A+B+C+D} - \frac{C}{C+D}\right) / \left(\frac{A+C}{A+B+C+D}\right)\right] \cdot 100$$

– Riesgo específico de enfermar los expuestos (REEExp):

$$REEExp = \frac{A}{A+B}$$

– Riesgo específico de enfermar los no expuestos (REENoExp):

$$REENoExp = \frac{C}{C+D}$$

– Riesgo total (RT):

$$RT = \frac{\left(\frac{A}{A+B} + \frac{C}{C+D}\right)}{A+B+C+D}$$

Medidas positivas, de prevención o protección para los expuestos

– Fracción de prevención proporcional para expuestos porcentual o porcentaje de protección puntual por la exposición (%FPE) o porcentaje (puntual) de protección de expuestos en la población:
  - En estudios de cohorte para incidencia acumulada:
    %FPE = (1 – RR) · 100
  - En estudios de cohorte por tasas de incidencia:
    %FPE = (1 – IR) · 100.
  - En estudios de caso-control no apareado:
    %FPE = (1 – OR) · 100

– Fracción de prevención máxima proporcional en expuestos porcentual o porcentaje de protección máximo posible ($\%FPE_{Máx}$) o porcentaje máximo posible de protección por la exposición al factor en la población:
  - Para estudio de cohorte por tasas de incidencia:

$\%FPE_{Máx}$ = (1 – valor IC Inferior de FPE o PFe, a partir del IR) · 100.

- Para estudio de cohorte por incidencia acumulada:
  $\%FPE_{Máx}$ = (1 – valor IC Inferior de FPE o PFe, a partir del RR) · 100
- Para estudio de caso-control:
  $\%FPE_{Máx}$ = (1 – valor IC Inferior de FPE o PFe, a partir del OR) · 100

– Fracción de prevención mínima proporcional en expuestos porcentual o porcentaje de protección mínimo posible ($\%PE_{Mín}$) o porcentaje mínimo posible de protección por la exposición al factor en la población:
  - Para estudio de cohorte por tasas de incidencia:
    %PEMín = (1 – valor IC Superior de FPE o PFe a partir del IR) · 100
  - Para estudio de cohorte por incidencia acumulada:
    %PEMín = (1 – valor IC Superior de FPE o PFe a partir del RR) · 100
  - Para estudio de caso –control:
    %PEMín = (1 – valor IC Superior de FPE o PFe a partir del OR) · 100

Medidas positivas, de prevención o de protección para la población

– Fracción de prevención proporcional para la población porcentual (FPP%) o % de protección puntual para la población por la exposición al factor protector:
  - A partir de la densidad de incidencia en estudios de cohorte:

$$FPP = PF_p = PF_e \frac{T_1}{T} \cdot 100$$

  - A partir de la incidencia acumulada en estudios de cohorte:

$$FPP = PF_p = PF_e \frac{n_1}{n} \cdot 100$$

  - A partir del OR de exposición en estudio de caso-control:

$$FPP = PF_p = \frac{p_1 PF_e}{p_1 PF_e + OR} \cdot 100$$

Teoría de probabilidades y las distribuciones de probabilidades de variables aleatorias

Una de las preocupaciones de los científicos ha sido construir modelos de distribuciones de probabilidad que puedan representar el comportamiento teórico de diferentes fenómenos aleatorios que aparecen en el mundo real. La pretensión de modelar lo observable ha constituido siempre una necesidad básica para el científico empírico, dado que, a través de esas construcciones teóricas, los modelos podían experimentar sobre lo que la realidad no le permitía. Por otra parte, un modelo resulta extremadamente útil, siempre que se corresponda con la realidad que pretende representar o predecir,

de manera que ponga de relieve las propiedades más importantes del mundo que nos rodea, aunque sea a costa de la simplificación que implica todo modelo.

Es necesario recordar algunos aspectos conceptuales generales y relacionables con la teoría matemática de probabilidades para poder comprender los diferentes aspectos que se describen más adelante:

- La probabilidad constituye por sí misma un concepto básico que refleja su relación con la faceta del mundo exterior que pretende estudiar: los fenómenos aleatorios, los que obedecen unas ciertas reglas de comportamiento. De alguna manera el concepto de probabilidad se relaciona o recuerda las propiedades de la frecuencia relativa.
- El cálculo de probabilidades suministra las reglas para el estudio de los experimentos aleatorios o de azar, lo que constituye la base para la estadística inductiva o inferencial.
- Las probabilidades constituyen la medida numérica de la verosimilitud o el grado de certidumbre de ocurrencia real de un fenómeno.
- Al analizar los resultados de un experimento o fenómeno aleatorio se establece la correspondencia entre los posibles resultados y la noción de probabilidad de ocurrencia de un determinado resultado o conjunto de resultados (sucesos aleatorios) de todos los posibles. Como consecuencia cada resultado o suceso aleatorio está asociado un número real, que es su probabilidad.
- Para aproximarse al conocimiento cuantitativo de una determinada parte de la realidad o fenómeno, es necesario construir un modelo de esta realidad particular objeto de estudio, partiendo de la premisa de que lo real es siempre más complejo y multiforme que cualquier modelo que se pueda construir.
- Los modelos teóricos para referirse a una realidad particular objeto de estudio se reducen en muchos casos a (o incluyen en su formulación) funciones de probabilidad, a partir de las cuales se obtienen las diferentes distribuciones de probabilidades para variables aleatorias, que se pueden dividir en: distribuciones de probabilidad para variables discretas y continuas.
- Una variable aleatoria puede definirse como cualquier característica medible que toma diferentes valores con probabilidades determinadas, es toda función que atribuye un único número real a cada suceso elemental del espacio muestral de un experimento aleatorio o es una cantidad variable que expresa el resultado de un experimento aleatorio.
- El modelo matemático que expresa las relaciones entre los valores de una variable aleatoria y las probabilidades de su ocurrencia o aparición se denomina *distribución de probabilidad*, por lo que toda variable aleatoria posee una distribución de probabilidad que describe su comportamiento. Si la variable es discreta, es decir, si toma valores aislados dentro de un intervalo, su distribución de probabilidad especifica todos los valores posibles de la variable junto con la probabilidad de que cada uno ocurra. En el caso continuo, es decir, cuando la variable puede tomar cualquier valor

de un intervalo, la distribución de probabilidad permite determinar las probabilidades correspondientes a subintervalos de valores.

- El concepto de variable aleatoria permite definir y denotar sucesos, así como calcular la probabilidad de estos, a través de la probabilidad de ocurrencia de los valores originales lo que coincide con el concepto de frecuencia relativa.
- Se puede pensar en las distribuciones de probabilidad como las formas teóricas o ideales de las distribuciones de frecuencia relativa, cuando el número de observaciones es grande (poblaciones).
- La distribución de frecuencias o distribución empírica de una variable viene dada por la frecuencia con que se observan realmente en la muestra estudiada cada uno de los posibles valores que puede tornar esa variable.
- Se denomina *distribución de probabilidad* a la que presenta el conjunto de todos los valores que teóricamente puede tomar una variable, junto con sus correspondientes probabilidades calculadas ordinariamente, siguiendo los principios de la ley de Laplace u otros análogos.
- La distribución de probabilidad de una variable aleatoria discreta es el conjunto de todos los valores de la variable en el eje de las abscisas (x), junto con sus correspondientes probabilidades en el eje de las ordenadas (y) si se representa en forma de gráfico, aunque puede ser representada tabularmente también. Esta distribución permite hacer afirmaciones (conclusiones) sobre la variable aleatoria en la población, hacer pronósticos y planificar medidas de prevención, entre otras acciones.
- La representación gráfica de la función de probabilidad se realiza mediante un diagrama de barras, donde la longitud de cada barra equivale al valor de probabilidad del correspondiente valor de x y todos los valores de P(X = x) son positivos, menores que 1 y la suma de los mismos es igual a 1. Estas características se cumplen para todas las distribuciones de probabilidad de variables discretas (Fig. 3.20).

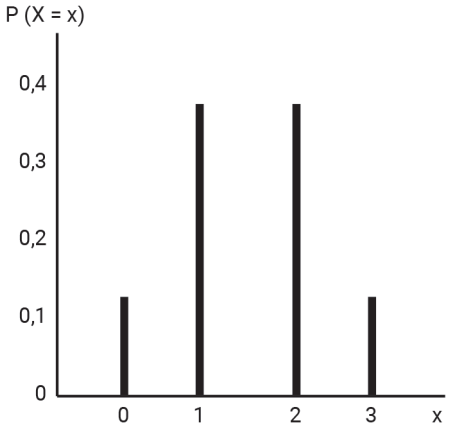


Fig. 3.20. Gráfico de barras simples para la representación de la función de probabilidad respecto al comportamiento de la variable aleatoria x.

– A una función no negativa f(x) que expresa los pares ordenados entre los valores distribuidos de x, y, la respectiva distribución de probabilidad para cada valor de x se le llama función de densidad de probabilidad para la variable aleatoria continua X [f(x)], si el área total delimitada por su curva y el eje de las x, es igual a 1 y si la subárea delimitada por la curva, el eje de las x y por las líneas perpendiculares levantadas sobre dos puntos cualesquiera a y b da la probabilidad de que X esté entre los puntos a y b y se denota por P(a < X < b) (Fig. 3.21).
– Por ser f una función integrable, la probabilidad de un punto es nula; es decir, la probabilidad de un valor específico de la variable aleatoria es cero, puesto que un valor específico se representa por un punto en el eje de las abscisas y el área encima de un punto es igual a cero. Por ello al calcular la probabilidad de un intervalo no afectará nada el que este sea abierto o cerrado por cualquiera de sus extremos, pues estos son puntos y por tanto de probabilidad nula: P(a < X < b) = P(a ≤ X ≤ b).

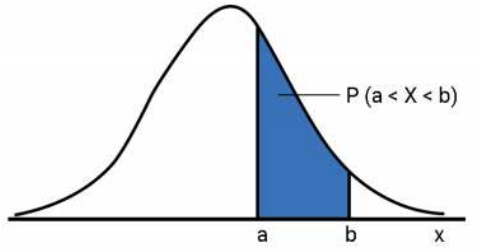


Fig. 3.21. Gráfico que representa la función de densidad de probabilidad para la variable aleatoria continua x [f(x)].

– Para calcular las áreas bajo la curva y obtener de ese modo las probabilidades de presentarse algún intervalo específico de valores de la variable se utiliza el proceso de integración, aunque al existir una serie de distribuciones teóricas, tanto para variables discretas como para variables continuas, que sirven de modelo para representar las distribuciones empíricas más frecuentemente encontradas no se requiere hallar ninguna integral, pues al alcance de todos existen los valores de probabilidad tabulados para cada distribución de probabilidad teórica.
– Una forma usual de describir la distribución de probabilidad de una variable aleatoria es mediante la denominada función de densidad en el caso de variables continuas y la función de masa de probabilidad en el caso de variables discretas, en tanto que lo que se conoce como *función de distribución* representa las probabilidades acumuladas (relacionada con el concepto de frecuencia relativa acumulada).
– La distribución de probabilidad de una variable aleatoria puede resumirse con la media, llamada en este caso *esperanza matemática* o *valor esperado* y la varianza, las que se denotan por E(X) y V(X), respectivamente. Su cálculo, análogo al de una distribución de datos, se efectúa con las probabilidades de cada valor, en lugar del número de veces (frecuencia) con que el valor aparece en la muestra. Estas medidas constituyen los parámetros de la distribución.

- Cuando se hace frente a una característica que varía de forma aleatoria y no se sabe a qué distribución se puede acudir para modelar su comportamiento entonces:
  - Si la variable es continua (puede tomar cualquier valor en un intervalo), presenta una distribución simétrica en tomo a su valor central y a medida que se aleja de la media disminuye la probabilidad de obtener valores, seguramente su comportamiento se puede modelar a través de una distribución normal.
  - La media muestral bajo el supuesto de muestreo sucesivo de tamaño n de una población con N elementos se convierte en una variable aleatoria que sigue una distribución normal.
  - Si todos los valores de la variable son igualmente probables, entonces el modelo de la distribución se llama *uniforme*.
  - Otras variables continuas tienen una distribución no simétrica. Esta circunstancia se da cuando los valores se alejan con más libertad hacia un lado que hacia el otro, en el que normalmente hay una frontera (puede ser lo que llamamos "cero natural") que impide el paso. En alguna de estas situaciones el logaritmo de los datos sigue una distribución que puede considerarse normal, por lo que esta distribución se llama *lognormal*.
  - Si fuese el tiempo transcurrido para dos sucesos consecutivos el fenómeno estudiado entonces los datos siguen una distribución exponencial y si el caso fuese aplicado para modelar el tiempo de vida se puede seleccionar una distribución de Weibull.
  - Si la variable es discreta (solo puede tomar valores a saltos sin decimales), cuando todos los valores tienen la misma probabilidad de aparecer, la distribución se llama *uniforme discreta*.
  - Si el caso fuese para n pruebas o experimentos independientes en que el resultado posible fuese: éxito o fracaso para un evento, siendo las probabilidades de éxito (p) y las de fracaso (1 – p) las que se mantendrían constantes a lo largo de todas las pruebas; en este caso el número de éxitos X al realizar n experimentos es una variable aleatoria que sigue una distribución binomial.
- En la práctica existe una serie de distribuciones teóricas de probabilidad para diferentes estadígrafos de prueba empleados en el contraste de hipótesis estadísticas como parte de la inferencia bioestadística, los que cumplen con el atributo de variable aleatoria, donde para cada valor teórico de dicho estadístico se corresponde un valor respectivo esperado-tabulado de probabilidad. Al realizar el contraste de hipótesis estadísticas son comparados el valor experimental u observado-calculado de dicho estadístico a partir de las muestras de estudio *versus* el valor teórico-esperado respectivo, para un cierto nivel de confianza preseleccionado por el investigador, asumiendo el caso que la hipótesis nula fuese cierta, lo cual permite la toma de decisiones respecto a las hipótesis contrastadas, esto constituye una aplicación de los aspectos antes relacionados de la teoría de probabilidades y variables aleatorias, por lo que es imprescindible su comprensión y aprendizaje para el adecuado procedimiento en la inferencia bioestadística.

Ejemplos de distribuciones de probabilidades discretas:

- Uniforme discreta.
- Binomial
- Multinomial.
- Hipergeométrica.
- Geométrica.
- Binomial negativa.
- Pascal.
- Poisson.

Ejemplos de distribuciones de probabilidades continuas:

- Uniforme o rectangular.
- Normal.
- Normal bivariante.
- Lognormal.
- Logística.
- Beta.
- Gamma.
- Exponencial.
- Chi-cuadrado.
- T de Student.
- F de Snedecor.
- Cauchy.
- Weibull.
- Laplace.
- Pareto.
- Triangular.

Los aspectos conceptuales esenciales de los diferentes tipos de distribuciones de probabilidad más frecuentemente empleadas para variables aleatorias se presentan a continuación.

Distribuciones discretas

Emplean la función de masa de probabilidad para cada distribución:

Distribución uniforme discreta

La distribución uniforme discreta (a, b) (Fernández Abascal, 1994: 338-391) describe el comportamiento de una variable discreta que puede tomar n valores distintos con la misma probabilidad cada uno de ellos.

Esta distribución asigna igual probabilidad a todos los valores enteros entre el límite inferior y el límite superior que definen el recorrido de la variable. Si la variable puede

tomar valores entre a y b, debe ocurrir que b sea mayor que a, y la variable toma los valores enteros empezando por a, a + 1, a + 2, entre otros hasta el valor máximo b, por ejemplo, cuando se observa el número obtenido tras el lanzamiento de un dado perfecto, los valores posibles siguen una distribución uniforme discreta en {1, 2, 3, 4, 5, 6}, y la probabilidad de cada cara es 1/6:

- Valores:
  k: a, a + 1, a + 2,..., b, números enteros
- Parámetros:
  a: mínimo, a entero
  b: máximo, b entero con a < b.
- Función de masa de probabilidad:

$$f(k) = \frac{1}{N}, \quad a \le k \le b, \text{ con k entero.}$$

- Valores característicos:
  - Media:

$$\bar{x} = \frac{a+b}{2}$$

  - Varianza:

$$Var(x) = \frac{N^2 - 1}{12}$$

  - Asimetría: 0.
  - Curtosis:

$$Kurt(x) = \frac{-6(N^2+1)}{5(N^2-1)}$$

    donde:
    N = b − a + 1: número de enteros entre a y b.

Distribución binomial

La distribución binomial (n, p) (Canavos, 1998: 89-99) es una distribución discreta muy importante que surge en muchas aplicaciones bioestadísticas. Fue obtenida por Jakob Bernoulli (1654-1705) y publicada en su obra póstuma *Ars Conjectandi* en 1713.

Esta distribución aparece de forma natural al realizar repeticiones independientes de un experimento que tenga respuesta binaria, generalmente clasificada como "éxito" o "fracaso"; este experimento recibe el nombre de experimento de Bernoulli. Ejemplos de respuesta binaria pueden ser el hábito de fumar (sí/no), si un paciente hospitalizado desarrolla o no una infección, o si un artículo de un lote es o no defectuoso. La variable discreta que cuenta el número de éxitos en n pruebas independientes de ese experimento, cada una de ellas con la misma probabilidad de "éxito" igual a p, sigue una distribución binomial de parámetros n y p, que se denota por Bi(n, p). Este modelo se aplica a

poblaciones finitas de las que se toman elementos al azar con reemplazo y también a poblaciones conceptualmente infinitas.

Cuando se tiene un número de pruebas "grande", la distribución binomial se aproxima a una distribución normal de media: np y varianza: np(1 – p).

Una variable aleatoria binomial es el número de éxitos obtenidos en n repeticiones independientes de un experimento Bernoulli, la cual sigue una ley de distribución binomial con parámetros n y p que se denota como:

$X \sim Bin\ (n, p)$

Esta variable tomará valores entre 0 y n, es decir el número de éxitos puede ser ninguno (cero), uno, dos, y así sucesivamente hasta n (todos). Su valor esperado es E(X) = np y la varianza V(x) = npq están dados por:

- Valores:
  k: 0, 1, 2,..., n
- Parámetros:
  n: número de pruebas o repeticiones, n ≥ 1 entero.
  p: probabilidad de ocurrencia de un éxito, 0 < p < 1, la cual se mantiene constante a través de las n repeticiones.
- Función de masa de probabilidad o ecuación básica para la distribución binomial:

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad 0 \le k \le n, \text{ con k entero siendo:}$$

,

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- Valores característicos:
  - Media:
    $\bar{x} = np$
  - Varianza:
    Var(x) = np(1 – p)
  - Asimetría:

$$A(x) = \frac{1-2p}{\sqrt{np(1-p)}}$$

  - Curtosis:

$$Kurt(x) = \frac{1-6p(1-p)}{np(1-p)}$$

La distribución binomial es una distribución de probabilidad discreta que cuenta el número de éxitos en una secuencia de n ensayos de Bernoulli independientes entre sí, con una probabilidad fija p de ocurrencia del éxito entre los ensayos. Un experimento de Bernoulli se caracteriza por ser dicotómico, esto es, solo son posibles dos resultados. A uno de estos se denomina éxito y tiene una probabilidad de ocurrencia p y al otro, fra-

caso, con una probabilidad q = 1 – p. En la distribución binomial el anterior experimento se repite n veces, de forma independiente, y se trata de calcular la probabilidad de un determinado número de éxitos. Para n = 1, la binomial se convierte, de hecho, en una distribución de Bernoulli.

Para representar que una variable aleatoria sigue una distribución binomial de parámetros n y p, se escribe:

$X \sim Bin(n, p)$

Se demuestra que una variable aleatoria discreta con distribución binomial, $X \sim Bin(n, p)$ se puede aproximar mediante una distribución normal si n es suficientemente grande y p no está ni muy próximo a 0 ni a 1. Como el valor esperado y la varianza de X son respectivamente: (np) y (npq), la aproximación consiste en decir que: $X \sim N(np, npq)$

El convenio que se suele utilizar para poder realizar esta aproximación es:

$X \sim Bin(n, p)$

donde:

para: $n > 30$, $np > 4$, $nq > 4$ se cumple: $X \sim N(n\,p, n\,p\,q)$.

Cuando se trabaja con una variable aleatoria que sigue la distribución de Poisson en una muestra se puede tener por objetivos:

– Determinar la probabilidad (Pr) para un número de éxitos o sucesos observados (k) en una muestra con observaciones o repeticiones de un experimento, contando con el valor estimado-esperado de la proporción poblacional (π o P), empleando la expresión para la función de masa de probabilidad de la distribución binomial:

$$Pr(k) = \pi^k (1-\pi)^{n-k} \frac{n!}{(n-k)!k!}$$

Otra forma de estimar esta probabilidad es empleando la aproximación a la distribución normal de la binomial, mediante la expresión:

$$z = \frac{x-\mu}{\sigma}$$

donde:

$\mu = n\pi$;

$\sigma = n\pi(1-\pi)$

Por lo que quedaría como:

$$z = \frac{x-n\pi}{\sqrt{n}\,\pi(1-\pi)}$$

Cuyo valor de Z se corresponde tabularmente con el valor de probabilidad a estimar.

– Determinar la probabilidad acumulada de la distribución binomial en relación con el valor de k seleccionado, lo cual se puede estimar por la sumatoria de las probabilidades estimadas para los valores de k menores que el valor de k seleccionado empleando la expresión anterior de la distribución binomial, lo cual comprende el área de la cola a la izquierda para ese valor de dicha distribución o para los valores

de k mayores que el valor de k seleccionado, lo que comprende el área de la cola a la derecha.

– Determinar la significación estadística en relación a que la proporción muestral (p) que estimaría una poblacional π o P guarda diferencias con la proporción poblacional de referencia π_o o P_o, para un nivel de confianza preseleccionado, mediante el empleo del método exacto para muestras pequeñas o la prueba Z por la aproximación a la distribución normal estándar de la binomial.

Distribución hipergeométrica

La distribución hipergeométrica (N, R, n) (Canavos, 1988: 108-115) suele aparecer en procesos muestrales sin reemplazo, en los que se investiga la presencia o ausencia de cierta característica. Esta distribución es la equivalente a la binomial, pero cuando el muestreo se hace sin reemplazo, de forma que la probabilidad de éxito no permanece constante a lo largo de las n pruebas, a diferencia de la distribución binomial. Esta distribución se puede ilustrar del modo siguiente: se tiene una población finita con N elementos, de los cuales R tienen una determinada característica que se llama "éxito" (diabetes, obesidad, hábito de fumar, entre otras). El número de "éxitos" en una muestra aleatoria de tamaño n, extraída sin reemplazo de la población, es una variable aleatoria con distribución hipergeométrica de parámetros N, R y n. Cuando el tamaño de la población es grande, los muestreos con y sin reemplazo son equivalentes, por lo que la distribución hipergeométrica se aproxima en tal caso a la binomial:

– Valores:
  k: m*á*x{0, n – (N – R)},..., min{R, n}, donde máx{0, n – (N – R)} indica el valor máximo entre 0 y n – (N – R) y m*í*n{R, n} indica el valor mínimo entre R y n.
– Parámetros:
  - N: tamaño de la población, N ≥ 1 entero.
  - R: número de éxitos en la población; 1 ≤ R ≤ N, N entero.
  - n: número de pruebas; 1 ≤ n ≤ N, n entero.
– Función de masa de probabilidad:

$$f(k)=\frac{\binom{R}{k}\binom{N-R}{n-k}}{\binom{N}{n}}$$ máx(0, n – (N – R) ≤ mín(R, n), con k entero.

– Valores característicos:
  - Media:

$$\bar{x}=np$$

  - Varianza:

$$Var(x)=np(1-p)\frac{N-n}{N-1}$$

- Asimetría:

$$A(x) = \frac{\frac{N-2n}{N-2}(1-2p)}{\sqrt{\frac{N-2n}{N-2}np(1-p)}}$$

- Curtosis:

$$Kurt(x) = A + \frac{BC}{np(1-p)} - 3$$

donde:

$$p = \frac{R}{N}$$

$$A = \frac{3(N-1)(N+6)}{(N-2)(N-3)}$$

$$B = \frac{(N-1)n(N+1)}{(N-n)(N-2)(N-3)}$$

$$C = 1 - \frac{6N\left(p(1-p) + n\frac{N-n}{N^2}\right)}{N+1}$$

Distribución geométrica

La distribución geométrica (p) (Fernández-Abascal, 1994: 403-407) permite calcular la probabilidad de que tenga que realizarse un número k de repeticiones antes de obtener un éxito por primera vez; esta probabilidad decrece a medida que aumenta k con lo que la función de masa de probabilidad es siempre decreciente. Así pues, se diferencia de la distribución binomial en que el número de repeticiones no está predeterminado, sino que es la variable aleatoria que se mide y, por otra parte, el conjunto de valores posibles de la variable es ilimitado.

Se utiliza en la distribución de tiempos de espera, de manera que si los ensayos se realizan a intervalos regulares de tiempo, esta variable aleatoria proporciona el tiempo transcurrido hasta el primer éxito. Por ejemplo: la variable aleatoria "intentos fallidos en la aplicación del medicamento antes del primer éxito" sigue una distribución geométrica de parámetro p. Otro ejemplo de variable geométrica es el número de hijos hasta el nacimiento de la primera niña:

- Valores:
  k: 0, 1, 2,...
- Parámetros:
  p: probabilidad de éxito, $0 < p < 1$.
- Función de masa de probabilidad:
  $f(k) = p(1 - p)^k$, $k \geq o$, con k entero.

– Valores característicos:
  - Media:

$$\bar{x} = \frac{1-p}{p}$$

  - Varianza:

$$Var(x) = \frac{1-p}{p^2}$$

  - Asimetría:

$$A(x) = \frac{2-p}{\sqrt{1-p}}$$

  - Curtosis:

$$Kurt(x) = \frac{p^2 - 6p + 6}{1-p}$$

Distribución binomial negativa

La distribución binomial negativa (r, p) (Canavos, 1988: 115-121) aparece en un estudio de Pierre Rémond de Montmort (1678-1719) sobre los juegos de azar en 1714, pero años antes ya había sido descrita por Blaise Pascal (1623-1662). Más adelante esta distribución fue propuesta como una alternativa a la distribución de Poisson para modelar el número de ocurrencias de un suceso cuando los datos presentan lo que se conoce como variación extra Poisson o sobredispersión. La primera aplicación en bioestadística la realizó Student (William Sealy Gosset (1876-1937)) a principios de siglo cuando propuso esta distribución para modelar el número de glóbulos rojos en una gota de sangre.

Una generalización obvia de la distribución geométrica aparece si se supone que un experimento se continúa hasta que un determinado suceso, de probabilidad p, ocurre por r-ésima vez. La variable aleatoria que proporciona la probabilidad de que se produzcan k fracasos antes de obtener el r-ésimo éxito sigue una distribución binomial negativa de parámetros r y p, BN(r, p).

Esta distribución fue propuesta como una alternativa a la distribución de Poisson para modelar el número de ocurrencias de un suceso cuando los datos presentan lo que se conoce como variación extra Poisson o sobredispersión.

La distribución binomial negativa es más adecuada que la de Poisson para modelar, por ejemplo, el número de accidentes laborales ocurridos en un determinado lapso. La distribución de Poisson asume que todos los individuos tienen la misma probabilidad de sufrir un accidente y que esta permanece constante durante el periodo de estudio; sin embargo, es más plausible la hipótesis de que los individuos tienen probabilidades constantes en el tiempo, pero que varían de unos sujetos a otros:

– Valores:
  k: 0, 1, 2,...
– Parámetros:

r: número de éxitos, r ≥ 1 entero.

p: probabilidad de éxito, 0 <p < 1.

– Función de masa de probabilidad:

$$f(k) = \binom{k+r-1}{r-1} p^r (1-p)^k \quad k \geq 0 \text{ con k entero.}$$

,

– Valores característicos:

• Media:

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

• Varianza:

$$Var(x) = \frac{r(1-p)}{p^2}$$

• Asimetría

$$A(x) = \frac{2-p}{\sqrt{r(1-p)}}$$

• Curtosis:

$$Kurt(x) = \frac{p^2 - 6p + 6}{r(1-p)}$$

Distribución Pascal

La distribución de Pascal (r, p) (Meyer, 1994: 178-179; Canavos, 1988: 115-121) debe su nombre al matemático francés Blaise Pascal (1623-1662), uno de los matemáticos que creó las bases de la teoría de la probabilidad.

El número de pruebas necesarias para obtener r éxitos, siendo p la probabilidad de éxito, una variable aleatoria que sigue una distribución Pascal de parámetros r y p, por tanto, esta distribución está relacionada con la binomial negativa de idénticos parámetros de este modo:

Pascal(r, p) = BN(r, p) + r

– Valores:

k: r, r + 1, r + 2,...

– Parámetros:

r: número de éxitos, r ≥ 1 entero.

p: probabilidad de éxito, 0 <p < 1.

– Función de masa de probabilidad:

$$f(k) = \binom{k-1}{r-1} p^r (1-p)^{k-r} \quad k \geq r, \text{ con k entero.}$$

,

– Valores característicos:

• Media:

$$\bar{x} = \frac{r}{p}$$

- Varianza:

$$Var(x) = \frac{r(1-p)}{p^2}$$

- Asimetría:

$$A(x) = \frac{2-p}{\sqrt{r(1-p)}}$$

- Curtosis:

$$Kurt(x) = \frac{p^2 - 6p + 6}{r(1-p)}$$

Distribución Poisson

La distribución de Poisson (λ) (Canavos, 1988: 100-108) debe su nombre al matemático francés Simeón Denis Poisson (1781-1840), aunque ya había sido introducida en 1718 por Abraham De Moivre (1667-1754). Poisson la da a conocer en 1838 en su trabajo: "Recherches sur la probabilité des jugements en matières criminelles et matière civile" (Investigación sobre la probabilidad de los juicios en materias criminales y civiles).

Surge como una forma límite de la distribución binomial o una aproximación de la misma que se aplica a sucesos con una probabilidad muy baja de ocurrir (evento raro) después de un número grande de repeticiones. En general, la distribución de Poisson (Poi) de parámetro λ se puede utilizar como una aproximación de la binomial, Bin(n, p), si el número de pruebas n es grande ($n \geq 50$); pero la probabilidad de éxito p es pequeña ($p \leq 0{,}1$), siendo la media o promedio de esta distribución $\lambda = np \leq 5$, en ese caso. Se puede considerar que la aproximación Poisson binomial es "buena" si $n \geq 20$ y p 0,05 y "muy buena" si $n \geq 100$ y $p \leq 0{,}01$. La variable asociada es el número de ocurrencias del evento en un intervalo o espacio continuo, por tanto, es una variable aleatoria discreta que toma valores enteros de 0 en adelante (0, 1, 2,...).

Una variable que se distribuye Poisson se define como el número de veces que ocurre un evento en un intervalo de tiempo, un área, un volumen o en un espacio dado y se denota como: $X \sim Poi(\lambda)$.

La distribución de Poisson es el resultado de un conjunto de suposiciones acerca de un proceso implícito para formar un conjunto de observaciones numéricas. Este proceso surge al contar el número de ocurrencias, eventos o sucesos, distribuidos al azar, que ocurren en un espacio o intervalo de tiempo dado. Para su consideración como tal debe cumplir con estos requisitos:

– Las ocurrencias de los eventos son independientes, es decir, la ocurrencia de un evento en un intervalo de tiempo o espacio, no tiene efecto en la probabilidad de una segunda ocurrencia del evento en el mismo o en algún otro intervalo.

– Es posible la ocurrencia del evento, teóricamente, un número infinito de veces dentro del intervalo.
– La probabilidad de una sola ocurrencia del evento en un intervalo dado es proporcional a la dimensión del intervalo.
– En cualquier fracción infinitesimal (muy, pero que muy pequeña) del intervalo, la probabilidad de más de una ocurrencia del evento es insignificante.

Las características generales de la distribución de Poisson serían:

– Valores:
  k: 0, 1, 2,...
– Parámetros:
  λ: tasa de ocurrencia, λ > 0 (es el número promedio de ocurrencias del evento por unidad de tiempo. A la vez que constituye el valor esperado E(x) = λ y la varianza Var(X) = λ de la distribución).
– Función de masa de probabilidad:
  $f(k) = \frac{\lambda^k e^{-\lambda}}{k!}$, k ≥ r, con k entero.
– Valores característicos:
  • Media:
    $\bar{x} = \lambda$
  • Varianza:
    Var(X) = = λ
  • Asimetría:
    $A(x) = \frac{1}{\sqrt{\lambda}}$
  • Curtosis:
    $Kurt(x) = \frac{1}{\lambda}$
    donde:
    k: es el número de ocurrencias del evento o fenómeno.
    λ: es un parámetro positivo que representa el número de veces que se espera que ocurra el fenómeno durante un intervalo dado. Por ejemplo, si el suceso estudiado tiene lugar en promedio 3 veces por minuto y estamos interesados en la probabilidad de que ocurra k veces dentro de un intervalo de 10 minutos, usaremos un modelo de distribución de Poisson con λ = 10 · 3 = 30.
    e: es la base de los logaritmos naturales (e = 2,71828...)

El concepto de evento "raro" o poco frecuente debe ser entendido en el sentido de que la probabilidad de observar k eventos decrece rápidamente a medida que k aumenta.

La distribución de Poisson tiene iguales la media y la varianza. Si la variación de los casos observados en una población excede a la variación esperada por la Poisson, se

está ante la presencia de un problema conocido como sobredispersión y en ese caso la distribución binomial negativa es más adecuada.

El parámetro de la distribución λ, representa el número promedio de eventos esperados por unidad de tiempo o de espacio, por lo que también se suele hablar de λ como "la tasa de ocurrencia" del fenómeno que se observa.

Para valores de λ mayores de 20 la distribución de Poisson se aproxima a una distribución normal de media y varianza iguales a λ.

En teoría de probabilidad y estadística, la distribución de Poisson es una distribución de probabilidad discreta que expresa, a partir de una frecuencia de ocurrencia media, la probabilidad de que ocurra un determinado número de eventos durante cierto periodo de tiempo. Concretamente se especializa en la probabilidad de ocurrencia de sucesos con probabilidades muy pequeñas o sucesos "raros", que ocurren a lo largo del tiempo. El tiempo entre cada par de eventos consecutivos tiene una distribución exponencial con el parámetro λ.

Este tipo de leyes se aplican a sucesos con probabilidad muy baja de ocurrir, obteniéndose como la distribución límite de una sucesión de variables binomiales: Bin(n, p), donde: n p = λ y n → ∞ (por tanto p → +) y donde su valor esperado y la varianza coinciden: E[X] = Var[X] = λ.

La distribución de Poisson difiere de una distribución binomial en que:

- La distribución binomial es afectada por el tamaño de la muestra n y la probabilidad p, mientras que la distribución de Poisson solo se ve afectada por la media m, n este caso λ.
- En una distribución binomial, los valores posibles de la variable aleatoria x son 0, 1,..., n; pero los valores posibles x de una distribución de Poisson son 0, 1, 2,..., sin límite superior.

En general se utiliza la distribución de Poisson como aproximación de experimentos binomiales cuando el número de pruebas es muy alto, pero la probabilidad de éxito muy baja. A veces se suele utilizar como criterio de aproximación: n ≥ 100, np ≤ 10 → Bin(n, p) ~ Poi (n, p).

Cuando se trabaja con una variable aleatoria que sigue la distribución de Poisson en una muestra se puede tener por objetivos:

- Determinar la probabilidad (Pr) para un número de ocurrencias de un evento o suceso raro (k) en una población estimada que tiene un valor promedio esperado para dicho número de ocurrencias del evento raro en un periodo de tiempo o espacio dado, también llamado: tasa de ocurrencia (λ), empleando la expresión para la función de masa de probabilidad de la distribución de Poisson:

  $$\Pr(k) = \frac{\lambda^k}{k!} e^{-k} \quad k > 0, \text{ k entero.}$$

  Otra forma de estimar esta probabilidad es por medio de la aproximación a la distribución normal de la de Poisson, siguiendo la expresión:

$$z = \frac{x - \mu}{\sigma}$$

que al sustituir los parámetros de la distribución de Poisson quedaría como:

$z = \frac{x - \lambda}{\sqrt{\lambda}}$ (al ser la media y desviación estándar igual a λ en la distribución de Poisson), cuyo valor de Z, se corresponde tabularmente con el valor de probabilidad a estimar.

- Determinar la probabilidad acumulada de la distribución de Poisson en relación con el valor de k y λ seleccionados; lo que se puede estimar por la sumatoria de las probabilidades estimadas para los valores de k menores e iguales que el valor de k seleccionado en la población estimada, empleando la expresión anterior de la función de masa de distribución de Poisson, lo cual comprende el área de la cola a la izquierda para este valor de dicha distribución o para los valores de k mayores que el valor de k seleccionado lo cual comprende el área de la cola a la derecha.
- Determinar la significación estadística en relación a que el valor promedio esperado de la ocurrencia del evento raro k muestral que estima el poblacional para un tiempo o espacio dado λ, no guarda diferencia con el de la ocurrencia del evento raro k en una población de referencia o hipótetica λ_o, para un nivel de confianza preseleccionado, empleando el método exacto por el p-valor para una muestra pequeña o por prueba de chi-cuadrado para muestra grande.

Distribuciones continuas

Son las que emplean la función de densidad.

Distribución uniforme o rectangular

La distribución uniforme (a, b) (Canavos, 1988: 143-147) es útil para describir una variable aleatoria con probabilidad constante sobre el intervalo (a, b) en el que está definida y se denota por U(a, b). También es conocida con el nombre de distribución rectangular por el aspecto de su función de densidad.

Una peculiaridad importante de esta distribución es que la probabilidad de un suceso depende exclusivamente de la amplitud del intervalo considerado y no de su posición en el campo de variación de la variable.

Cualquiera que sea la distribución F de cierta variable X, la variable transformada Y = F(X) sigue una distribución uniforme en el intervalo (0, 1). Esta propiedad es fundamental por ser la base para la generación de números aleatorios de cualquier distribución en las técnicas de simulación, y recibe el nombre de método de inversión:

- Campo de variación:
  $a < x < b$
- Parámetros:
  a: mínimo, $-\infty < a < \infty$
  b: máximo, $-\infty < b < \infty$ con $a < b$

– Función de densidad:

$$f(x) = \frac{1}{n-a}, \quad a < x < b$$

– Valores característicos:

- Media = Mediana:

$$\bar{x} = Me = \frac{a+b}{2}$$

- Moda: intervalo (a, b).
- Varianza:

$$Var(x) = \frac{(a-b)^2}{12}$$

- Asimetría: 0.
- Curtosis: −6/5

Distribución normal

Aunque la distribución normal (μ, σ) (Canavos, 1988: 130-143) no es una ley inherente a todas las características mesurables de la naturaleza, muchas de estas características tienen una distribución aparentemente "normal", por lo que es, sin duda, la distribución de probabilidad más importante de la estadística.

Constituye un modelo matemático que se expresa mediante la ecuación descubierta por Abrahan de Moivre en 1773:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

y su representación gráfica produce la muy conocida curva en forma de campana o campana de Gauss.

La importancia de la distribución normal queda totalmente consolidada por ser la distribución límite de numerosas variables aleatorias, discretas y continuas, como se demuestra a través del teorema central del límite, en el cual se describe que para estas variables en la distribución de frecuencias de sus valores es más frecuente encontrar los valores intermedios o próximos al valor central que los de los extremos, los cuales son más raros. Las consecuencias de estos teoremas implican la casi universal presencia de la distribución normal en todos los campos de las ciencias empíricas: biología, medicina, psicología, física, economía, entre otras. En particular, muchas medidas de datos continuos en medicina y en biología (talla, presión arterial, entre otras) se aproximan a la distribución normal. De ella derivan, entre otras, tres distribuciones (chi-cuadrado, t de Student y F de Snedecor), de importancia clave en el campo de la contrastación de hipótesis estadísticas. De modo que cualquier distribución, aunque no siga una distribución normal, tiende a ella en su distribución muestral, es decir, aunque la población de la que proceden las muestras no siga una distribución normal para la variable en

estudio, los estimadores calculados en las sucesivas muestras sí la siguen, siempre que las muestras tengan suficiente tamaño (n > 30).

La distribución normal queda totalmente definida mediante dos parámetros: la media (μ) y la desviación estándar o desviación típica (σ) o varianza poblacional (σ^2) y se denota como $N(\mu, \sigma^2)$. Su función de densidad es simétrica respecto a la media y la desviación estándar, indica el mayor o menor grado de apertura de la curva.

La distribución normal es una "familia de distribuciones" en la que cada miembro se distingue de los otros por la combinación de valores de μ y σ que lo caracterice. Esto puede constituir una dificultad porque sería necesario construir la curva particular de los valores de la variable específica en cada población estudiada, para calcular las probabilidades de ocurrencia de estos valores. En la práctica esta dificultad queda resuelta trabajando con el "miembro más importante de esta familia": la distribución normal estándar o unitaria, que tiene $\mu = 0$ y $\sigma = 1$. Esta distribución se puede obtener a partir de la ecuación general de la distribución normal, creando la variable normalizada, tipificada o estandarizada:

$$z = \frac{x - \mu}{\sigma y}$$

entonces cambiaría la ecuación para la distribución normal, en este caso para la distribución normal estándar o tipificada sería:

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Las características generales de la distribución normal serían:

- Campo de variación de la distribución normal:
  $-\infty < x < \infty$
- Parámetros:
  μ: media, $-\infty < \mu < \infty$.
  σ: desviación estándar, $\sigma > 0$.
- Función de densidad:

$$f(k) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right], \quad -\infty < x < \infty$$

- Valores característicos:
  - Media, mediana, moda:
    $\bar{x} = Me = M = \mu$
  - Varianza:
    $Var(x) = \sigma^2$
  - Asimetría: 0.
  - Curtosis: 0.
- La gráfica de su función de densidad tiene una forma acampanada y es simétrica respecto de un determinado parámetro estadístico, la media poblacional. Esta curva se conoce como campana de Gauss y es el gráfico de una función gaussiana.

Algunas propiedades de la distribución normal

- La distribución normal está definida como la función de distribución de probabilidad de una variable aleatoria X (Fig. 3.20) y se denota por:
  $X \sim N(\mu, \sigma^{2)}$
  Donde:
  μ: media o valor esperado poblacional;
  σ^2: varianza poblacional.

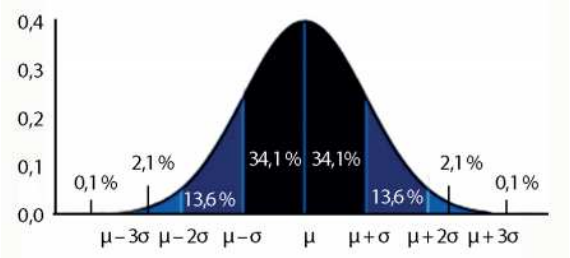


Fig. 3.22. Gráfico para representar la distribución de probabilidad alrededor de la media en una distribución normal [$N(\mu, \sigma^2)$].

- Es simétrica respecto de su media (μ).
- Tiene su punto máximo en la media (μ).
- La moda y la mediana son ambas iguales a la media (μ).
- El área total bajo la curva es igual a 1 y como es simétrica, el 50 % del área está a la derecha e izquierda de la media.
- Siempre es positiva esta distribución.
- Las colas de la curva están cada vez más próximas al eje x a medida que se alejan de la media, pero nunca lo alcanzan.
- Los puntos de inflexión de la curva se dan para $x = \mu - \sigma$ y $x = \mu + \sigma$.
- En el intervalo $[\mu - \sigma, \mu + \sigma]$ se encuentra comprendida, aproximadamente, el 68,26 % de la distribución; en el intervalo $[\mu - 2\sigma, \mu + 2\sigma]$ se encuentra, aproximadamente, el 95,44 % de la distribución y en el intervalo $[\mu - 3\sigma, \mu + 3\sigma]$ se encuentra comprendida, aproximadamente, el 99,74 % de la distribución. Estas propiedades son de gran utilidad para el establecimiento de intervalos de confianza.
- Esta distribución normal queda definida por dos parámetros que son la media poblacional (μ) y la desviación estándar poblacional (σ) o varianza poblacional (σ^2).
- Para expresar que los valores de una variable aleatoria cuantitativa X sigue una distribución normal con (μ) y la varianza poblacional (σ^2) se expresa como:
  $X \sim N(\mu, \sigma^2)$.
- Los valores de una variable aleatoria X con esta distribución tiene su máxima frecuencia alrededor de μ y en este valor coinciden el valor de su media, mediana y moda.
- Esta distribución es completamente simétrica respecto al eje vertical que corta al eje de las abscisas (eje x), en el valor μ.

– Es asintótica respecto al eje de las abscisas (nunca llega a cruzar este eje) y a medida que los valores del eje x se acercan hacia $-\infty$ o $+\infty$, el gráfico de la distribución normal se aproxima más y más a este eje x.
– Los parámetros μ (mu) y σ (sigma), determinan completamente la forma de la distribución. Para cada valor de μ y σ hay una distribución normal diferente, es decir, que diferentes valores de μ trasladan la curva horizontalmente sobre el eje X, mientras que σ determina el grado de apuntamiento de la curva a menor desviación estándar, menos dispersión alrededor de la media y por lo tanto más alta la curva (Figs. 3.23. y 3.24).

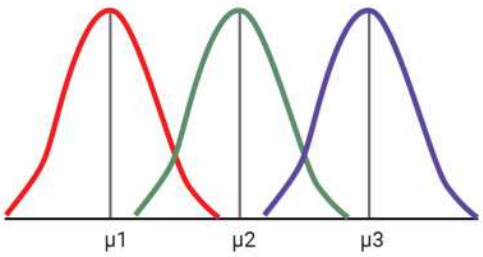


Fig. 3.23. Gráfico para representar tres distribuciones normales con diferentes medias poblacionales ($\mu_1 \neq \mu_2 \neq \mu_3$); pero igual variabilidad ($\sigma_1 = \sigma_2 = \sigma_3$).

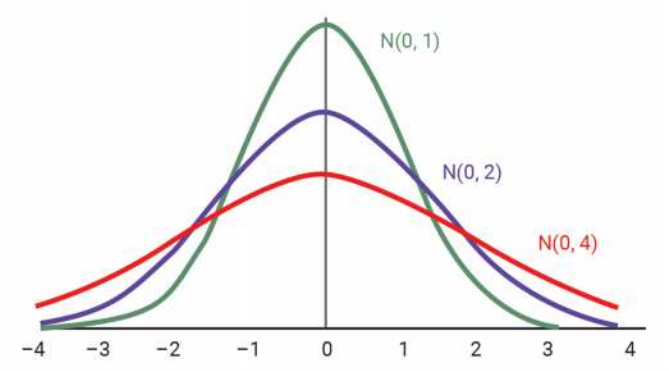


Fig. 3.24. Gráfico para representar tres distribuciones normales con iguales medias poblacionales ($\mu_1 = \mu_2 = \mu_3$); pero diferente variabilidad ($\sigma_1 \neq \sigma_2 \neq \sigma_3$).

– El hecho de que prácticamente la totalidad de la distribución se encuentre a tres desviaciones típicas de la media justifica los límites de las tablas empleadas habitualmente en la normal estándar.
– Cuando la variable aleatoria X en estudio no sigue una distribución N(0, 1), es decir, para cualquier variable que se distribuya: N(μ, σ^2), se debe a la misma convertirla o transformarla en una distribución normal estándar o tipificada para poder calcular las probabilidades correspondientes a su distribución de frecuencias relativas mediante el siguiente procedimiento matemático:

$X \sim N(\mu, \sigma^2) \Rightarrow Z = X - \mu/\sigma$, es decir, ahora X equivale matemáticamente a Z, entonces se emplean las tablas de probabilidades de Z (distribución teórica de probabilidades normal tipificada) para evaluar la distribución de los valores de X.

– Como no existen tablas para cada una de las posibles distribuciones de tipo Normal para las variables aleatorias X que se pueden presentar en la práctica, la solución a esta problemática es la existencia de una distribución Normal especial, llamada Normal estándar o típica, que se caracteriza por tener media (μ) = 0 y varianza (σ^2) = 1 y se denota por $X \sim N(\mu, 1)$ y se puede representar gráficamente (Fig. 3.25).

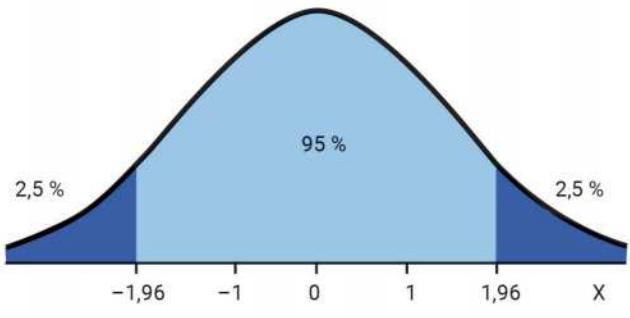


Fig. 3.25. Gráfico para representar las áreas bajo la curva correspondiente a la distribución normal estándar.

– A la distribución $Z \sim N(0, 1)$ se le conoce como distribución normal estándar o tipificada y se representa por la letra Z. Existen tablas que permiten el cálculo de forma sencilla de los valores de las probabilidades correspondientes a la distribución N ($\mu = 0$; $\sigma^2 = 1$) para cada Z, es decir, $Z \sim N(0, 1)$ (véase el anexo 1).

– Para el cálculo de probabilidades (área bajo la curva correspondiente a la distribución de probabilidades) en la distribución normal, como en toda distribución continua, la probabilidad para un valor específico de X es cero, lo que se halla es la probabilidad para intervalos entre dos valores de X luego de ser tipificados o estandarizados en Z, por ejemplo, las probabilidades (P) de que la variable X tome valores en un intervalo (a, b), $P(a \leq X \leq b)$ o que $P(X \leq a)$, $P(X \geq a)$, $P(X \leq b)$, $P(X \geq b)$ y se puede representar gráficamente (Fig. 3.26).

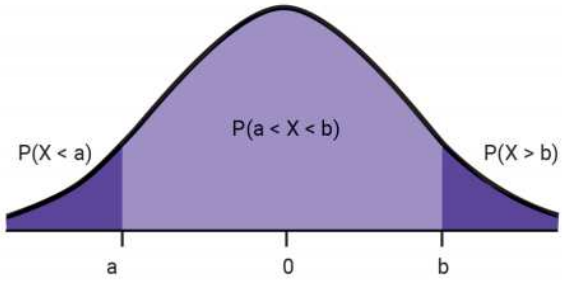


Fig. 3.26. Gráfico de una distribución normal para representar los intervalos (según a, b, correspondientes con los diferentes colores) que delimitan diferentes áreas bajo la curva para una variable X.

– Cuando se trabaja con las tablas de referencia para la distribución normal estándar que relaciona los valores de probabilidad en la distribución para cada valor de Z (percentiles correspondientes a la distribución normal estándar) se debe tener en cuenta algunas reglas:
  • Siendo z el valor tipificado-estandarizado o el observado-calculado-empírico (obtenido a partir del valor X de la variable transformado, correspondiente de la muestra) y Z el valor teórico-esperado de la distribución normal-estándar teórico-tabulado, para estimar el valor de probabilidad de z tipificado que cumpla la condición $P(Z \geq z)$ se consulta la tabla de referencia y se debe seleccionar el valor de Z en la tabla y hacerlo corresponder con la respectiva casilla de probabilidad que le corresponde según las filas y columnas.
  • S $P(Z \leq z)$, equivale a decir: $1 - P(Z \geq z)$.
  • Si $P(Z \geq -z)$ equivale a decir: $P(Z \leq z)$.
  • Si $P(Z \leq -z)$ equivale a decir: $P(Z \geq z)$.
– Una distribución binomial con tamaño de muesta (n) grande y si p y q no están muy cercanas a cero, esta puede aproximarse a la distribución normal y en la medida que n aumente y p se mantenga constante, la aproximación es cada vez mejor. De forma general hay buena aproximación cuando n es mayor o igual que 30; en la medida que p se aleja de 0,5, se necesitan mayores tamaños muestrales.
– La distribución de Poisson es una aproximación de la binomial cuando n es grande y P o π es pequeña, a su vez la de Poisson puede ser aproximada a la distribución normal bajo la condición de λ y n grandes.

Los aspectos resumidos en los dos puntos anteriores se ejemplifican gráficamente (Fig. 3.27).

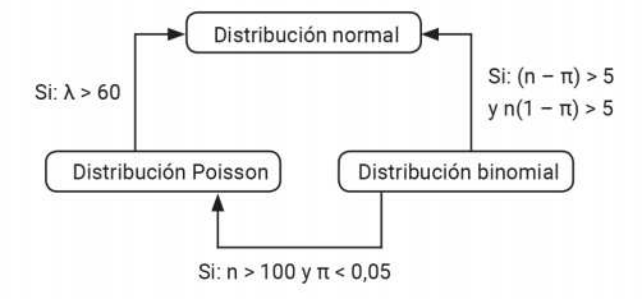


Fig. 3.27. Aproximaciones de una a otra distribución, según el sentido de las flechas y las condiciones dadas por los valores de los estadígrafos.

– Los dos parámetros μ y σ^2 coinciden con la media (esperanza) y la varianza respectivamente de la distribución de los estimadores muestrales respectivos:
  $E(X) = \mu$
  $Var(X) = \sigma^2$

– En probabilidad, la distribución normal aparece como el límite de varias distribuciones de probabilidad continuas y discretas.
– La distribución normal también es importante por su relación con la estimación por mínimos cuadrados, uno de los métodos de estimación más simples y antiguos.
– La importancia fundamental de esta distribución radica en que permite "modelar" numerosos fenómenos naturales, sociales y sicológicos. Mientras que los mecanismos que subyacen a gran parte de este tipo de fenómenos son desconocidos, por la enorme cantidad de variables incontrolables que en ellos intervienen, el uso del modelo normal puede justificarse asumiendo que cada observación se obtiene como la suma de unas pocas causas independientes.

Algunos ejemplos de variables asociadas a fenómenos naturales que siguen el modelo de la normal son:
– Caracteres morfológicos de individuos como la estatura.
– Caracteres fisiológicos como el efecto de un fármaco.
– Caracteres sociológicos como el consumo de cierto producto por un mismo grupo de individuos.
– Caracteres psicológicos como el cociente intelectual.
– Nivel de ruido en telecomunicaciones.
– Errores cometidos al medir ciertas magnitudes.

La distribución normal también aparece en muchas áreas de la propia estadística, por ejemplo, la distribución muestral de las medias muestrales es aproximadamente normal cuando la distribución de la población de la cual se extrae la muestra no es normal. Además, la distribución normal maximiza la entropía (correspondencia) entre todas las distribuciones con media y varianza conocidas, lo que la convierte en la elección natural de la distribución subyacente a una lista de datos resumidos en términos de media muestral y varianza. La distribución normal es la más extendida en estadística y muchas pruebas estadísticas están basados en una supuesta "normalidad" en la distribución de sus datos.

Distribución teórica de probabilidades normal estándar o tipificada Z~N(0,1)

– Calcular las probabilidades o proporción en porcentaje (área bajo la curva) que representa un valor x de una variable aleatoria de estudio según su distribución de frecuencias y estandarización normal (por su tipificación en Z).
– Determinar diferencias de probabilidades, entre dos valores de una variable x de estudio, según su distribución de frecuencias y estandarización normal (por su tipificación en Z).
– Determinar probabilidades (área bajo la curva) mayores o menores, que se corresponden con un valor de la variable aleatoria X de estudio a partir de su distribución de frecuencias y la respectiva tipificación en Z de sus valores.

– Realizar la estimación de parámetros de una población por intervalos de confianza a partir de los estimadores muestrales para variables aleatorias cuantitativas continuas y discretas (estas últimas a partir de la transformación de variables cualitativas dicotómicas).
– Realizar pruebas de significación o del contraste de hipótesis para la comparación de parámetros poblacionales estimados a partir de los muestrales para variables cuantitativas y cualitativas dicotómicas con salida como discretas, estas últimas por aproximación de la distribución binomial a la normal estándar: $Z \sim N(0, 1)$ para muestras grandes.
– Al realizar pruebas de significación no paramétricas con tamaños grandes de dos muestras relacionadas se puede aproximar a la distribución normal estándar los datos poblacionales estimados.

Distribución lognormal

La distribución lognormal (μ, σ) (Fernández-Abascal, 1994) es la variable resultante de aplicar la función exponencial a una variable que se distribuye normal con media μ y desviación estándar σ, sigue una distribución lognormal con parámetros μ (escala) y σ (forma). Dicho de otro modo, si una variable X sigue una distribución lognormal entonces la variable lnX se distribuye normalmente, es decir, como una distribución normal.

Esta variable aleatoria fue propuesta por Francis Galton (1822-1911) en 1879, como consecuencia del estudio de la media geométrica de n variables aleatorias independientes.

La distribución lognormal es útil para modelar datos de numerosos estudios médicos tales como el periodo de incubación de una enfermedad, los títulos de anticuerpo a un virus, el tiempo de supervivencia en pacientes con cáncer o SIDA entre otros.

Las características generales de la distribución lognormal serían:

– Campo de variación:
  $0 < x < \infty$
– Parámetros:
  μ: escala, $-\infty < \mu < \infty$.
  σ: forma, $\sigma > 0$.
– Función de densidad:

$$f(k) = \frac{1}{\sigma x\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2\right], \quad x > 0$$

– Valores característicos:
  • Media:

$$\bar{x} = e^{\mu+\frac{\sigma^2}{2}}$$

  • Mediana:

$$Me = e^{\mu}$$

- Moda:

  $M = e^{\mu - \sigma^2}$

- Varianza:

  $Var(x) = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)$

- Asimetría:

  $A(x) = \sqrt{\left(e^{\sigma^2} - 1\right)}\left(e^{\sigma^2} + 2\right)$

- Curtosis:

  $Kurt(x) = e^{4\sigma^2} + 2e^{3\sigma^2} + 3e^{2\sigma^2} - 6$

Distribución logística

La distribución logística (a, b) (Fernández-Abascal, 1994: 464-466) fue descrita por primera vez por Pierre François Verhulst (1804-1849) en un trabajo publicado en 1845, que versaba sobre las investigaciones matemáticas en las leyes que gobiernan el crecimiento de la población.

La distribución logística se utiliza en el estudio del crecimiento temporal de variables, en particular, demográficas. En biología se ha aplicado, por ejemplo, para modelar el crecimiento de células de levadura y para representar curvas de dosis-respuesta en bioensayos.

La más conocida y generalizada aplicación de la distribución logística en ciencias de la salud se fundamenta en esta propiedad:

Si U es una variable uniformemente distribuida en el intervalo (0, 1), entonces la variable:

$$x = \ln\left(\frac{U}{1-U}\right)$$

sigue una distribución logística. Esta transformación, denominada logit, se utiliza para modelar datos de respuesta binaria, especialmente en el contexto de la regresión logística.

Los parámetros asociados a esta distribución son situación (a) y escala (b). Su función de densidad es simétrica respecto al parámetro a y presenta un perfil más apuntado que el de la distribución normal con la misma media y desviación estándar (distribución leptocúrtica).

– Campo de variación:

  $-\infty < x < \infty$

– Parámetros:

  a: situación, $-\infty < a < \infty$.

  b: escala, $b > 0$.

– Función de densidad:

$$f(k) = \frac{1}{b} \frac{e^{-\frac{x-a}{b}}}{\left[1+e^{-\frac{x-a}{b}}\right]^2}, \; -\infty < x < \infty$$

– Valores característicos:

- Media, mediana, moda:

  $\bar{x} = Me = M = a$

- Varianza:

  $Var(x) = \frac{\pi^2}{3} b^2$

- Asímetría:

  A(x) = 0

- Curtosis:

  Kurt(x)= 6/5

Distribución beta

La distribución beta (p, q) (Canavos, 1988: 147-151) es adecuada para variables aleatorias continuas que toman valores en el intervalo (0, 1), lo que la hace muy apropiada para modelar proporciones. En la inferencia bayesiana, por ejemplo, es muy utilizada como distribución *a priori* cuando las observaciones tienen una distribución binomial.

Uno de los principales recursos de esta distribución es el ajuste a una gran variedad de distribuciones empíricas, pues adopta formas muy diversas dependiendo de cuáles sean los valores de los parámetros de forma p y q, mediante los que viene definida la distribución, denotada por beta (p, q):

– Campo de variación:

  $0 < x < 1$

– Parámetros:

  p: forma, p > 0.

  q: forma, q > 0.

– Función de densidad:

$$f(k) = \frac{x^{p-1}(1-x)^{q-1}}{B(p,\ q)}, \quad 0 < x < 1$$

  donde:

  B: función beta:

$$B(p,q) \int_0^1 t^{p-1}(1-t)^{q-1} dt$$

– Valores característicos:

- Media:

$$\bar{x} = \frac{p}{p+q}$$

- Mediana: no tiene expresión explícita.
- Moda:

$$M = \frac{p-1}{p+q-2}, \text{ para } p > 1 \text{ y } q > 1$$

- Varianza:

$$Var(x) = \frac{pq}{(p+q)^2(1+p+q)}$$

- Asimetría:

$$A(x) = \frac{2(q-p)\sqrt{p+q+1}}{(p+q+2)\sqrt{pq}}$$

- Curtosis:

$$Kurt(x) = 6\frac{p(p+1)(p-2q)+q(q+1)(q-2p)}{pq(p+q+2)(p+q+3)}$$

Distribución gamma

La distribución gamma $\Gamma(a, p)$ (Fernández-Abascal, 1994: 448-452) se puede caracterizar de este modo: si se está interesado en la ocurrencia de un evento generado por un proceso de Poisson de media λ, la variable que mide el tiempo transcurrido hasta obtener n ocurrencias del evento sigue una distribución gamma con parámetros $a = n\lambda$ (escala) y $p = n$ (forma) y se denota por $\Gamma(a, p)$.

Por ejemplo, la distribución gamma aparece cuando se realiza el estudio de la duración de elementos físicos (tiempo de vida).

Cuando p es un número entero positivo se tiene un caso particular de la distribución gamma que se denomina *distribución de Erlang*. Otros casos particulares de la distribución gamma son la distribución exponencial $[\Gamma(\lambda, 1)]$ y la distribución chi-cuadrado $[\Gamma(1/2, n/2)]$.

Según los valores que tome el parámetro de forma (p), la función de densidad presenta perfiles muy diversos. Con valores de p menores o iguales que 1, la función de densidad muestra un perfil decreciente, en cambio, si p es mayor que la unidad, la función de densidad crece hasta el valor $x = (p - 1)/a$ y decrece a partir de este valor:

– Campo de variación:
  $0 < x < \infty$
– Parámetros:
  a: escala, $a > 0$.
  p: forma, $p > 0$.
– Función de densidad:

$$f(k) = \frac{a^p}{\Gamma(p)} e^{-ax} x^{p-1}, \; x > 0$$

donde:

Γ: función gamma:

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt$$

y si n es un entero:

$\Gamma(n) = (n - 1)!$

– Valores característicos:
  - Media:

    $$\bar{x} = \frac{p}{a}$$

  - Mediana: no tiene expresión explícita.
  - Moda:

    $$M = \frac{p-1}{a} \quad \text{para } p > 1$$

  - Varianza:

    $$Var(x) = \frac{p}{a^2}$$

  - Asimetría:

    $$A(x) = \frac{2}{\sqrt{p}}$$

  - Curtosis:

    $$Kurt(x) = \frac{6}{p}$$

Distribución exponencial

La distribución exponencial (λ) (Fernández-Abascal, 1994: 452-455) es un caso particular de la distribución gamma y el equivalente continuo de la distribución geométrica discreta. Esta ley de distribución describe procesos en los que interesa saber el tiempo hasta que ocurre determinado evento, en particular se utiliza para modelar tiempos de supervivencia, un ejemplo es el tiempo que tarda una partícula radiactiva en desintegrarse. Una característica importante de esta distribución es la propiedad conocida como "falta de memoria". Esto significa, por ejemplo, que la probabilidad de que un individuo de edad t sobreviva x años más, hasta la edad x + t, es la misma que tiene un recién nacido de sobrevivir hasta la edad x. Dicho de manera más general, el tiempo transcurrido desde cualquier instante dado t_0 hasta que ocurre el evento, no depende de lo que haya ocurrido antes del instante t_0.

Se cumple que variable aleatoria que tome valores positivos y que verifique la propiedad de "falta de memoria" sigue una distribución exponencial.

Esta distribución se puede caracterizar como la distribución del tiempo entre sucesos consecutivos generados por un proceso de Poisson, por ejemplo, el tiempo que transcurre entre dos heridas graves sufridas por una persona. La media de la distribución de Poisson (λ) que representa la tasa de ocurrencia del evento por unidad de tiempo es el parámetro de la distribución exponencial y su inversa es el valor medio de la distribución:

- Campo de variación:

  $0 < x < \infty$
- Parámetros:

  λ: tasa.

  $\lambda > 0$.
- Función de densidad:

  $F(x) = \lambda e^{-\lambda x}, x > 0$
- Valores característicos:
  - Media:

    $$\bar{x} = \frac{1}{\lambda}$$
  - Mediana:

    $$Me = \frac{\ln 2}{\lambda}$$
  - Moda: no definida.
  - Varianza:

    $$Var(x) = \frac{1}{\lambda^2}$$
  - Asimetría: 2.
  - Curtosis: 6.

Distribución chi-cuadrado

La distribución chi-cuadrado (χ^2) (Fernández-Abascal, 1994: 473, 478) es un caso especial y muy importante de la distribución gamma y se obtiene cuando a = 1/2 y p = n/2 y es conocida por el nombre de *distribución chi-cuadrado de Pearson* con *n grados de libertad* (Fig. 3.28).

Esta distribución debe su nombre al matemático inglés Karl Pearson (1857-1936), es fundamental en inferencia estadística y en las pruebas estadísticas de bondad de ajuste. Se emplea, entre otras muchas aplicaciones, para realizar la prueba de hipótesis de homogeneidad, de independencia o la prueba de bondad de ajuste (todas denominadas pruebas ch-cuadrado) y para determinar los límites de confianza de la varianza muestral de una población normal.

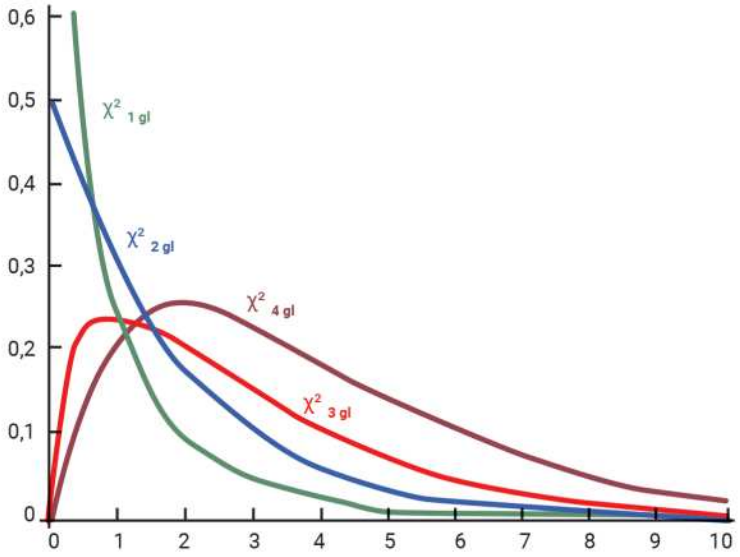


Fig. 3.28. Gráfico para la representación de varias curvas (con diferentes colores según grados de libertad) para la función de distribución de probabilidad correspondiente a la distribución chi-cuadrado.

La distribución chi-cuadrado deriva de la distribución normal. Surge cuando tenemos n variables independientes (X_1, X_2,.., X_n), que siguen aproximadamente una distribución normal estándar Z ~ N (0, 1) o que son variables estandarizadas, donde la variable resultante de sumar los cuadrados de dichas n variables sigue una distribución chi-cuadrado con n grados de libertad. Se denota como:

$X \sim \chi^2$ (n gl)

Si tenemos n variables aleatorias independientes $Z \rightarrow N(0, 1)$, la suma de sus cuadrados respectivos es una distribución que se denomina ley de distribución chi-cuadrado (χ^2) con n grados de libertad (χ^2_n):

$$\{Zi\}_{i=1}^{n} \rightarrow N(0,\ 1) \Rightarrow \sum_{i=1}^{n} Z_i^2 \rightarrow \chi_n^2$$

En consecuencia, si se tiene: (X_1,..., X_n) variables aleatorias independientes, donde cada:

$$X_i \rightarrow N(\mu_i,\ \sigma_i^2)$$

Se tiene:

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 \rightarrow \chi_n^2$$

- Se trata de una función positiva, no puede tomar valores negativos, al derivarse de la suma de valores elevados al cuadrado.
- Su forma depende de los grados de libertad, con tendencia a aproximarse a la curva normal en la medida que aumentan los grados de libertad.

- Al igual que la distribución normal y cualquier distribución continua, el área bajo la curva es igual a 1, pero no es simétrica.
- Su valor esperado es $E(\chi^2) = n$ y la varianza es $V(\chi^2) = 2n$.
- Los grados de libertad se refieren al número de términos independientes que es necesario, para obtener el valor de la variable χ^2.
- Esta distribución es muy empleada para probar hipótesis, cuando los datos están en forma de frecuencias. Estos procedimientos se estudian bajo el título de *pruebas de bondad de ajuste* y preferentemente con variables categóricas.
- La cantidad χ^2, es una medida del grado de congruencia entre las frecuencias observadas y las esperadas, bajo el cumplimiento de una hipótesis dada.
- Existen tablas teóricas de probabilidades para los percentiles χ^2 de esta distribución según el tamaño o magnitud de n y los grados de libertad.
- Campo de variación:
  $0 < x < \infty$
- Parámetros:
  n: grados de libertad.
  $n \geq 1$ entero.
  chi-cuadrado (n) = gamma (1/2, n/2)].
- Función de densidad:

$$f(x) = \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}$$

  donde:
  Γ: función gamma:

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt$$

  y si n es un entero:
  $\Gamma(n) = (n-1)!$
- Valores característicos:
  - Media:
    $\bar{x} = n$
  - Mediana: no tiene expresión explícita.
  - Moda:
    $M = n - 2$ para $n > 2$
  - Varianza:
    $Var(x) = 2n$
  - Asimetría:

$$A(x) = \sqrt{\frac{8}{n}}$$

- Curtosis:

$$Kurt(x) = \frac{12}{n}$$

En estadística, la distribución de Pearson, llamada también j-cuadrada o chi-cuadrado (χ^2) es una distribución de probabilidad continua con un parámetro que representa los grados de libertad de la variable aleatoria:

$$X = Z_1^2 + ... + Z_k^2$$

Donde:

Z_i: variables aleatorias normales independientes de media cero y varianza uno.

El que la variable aleatoria X tenga esta distribución se representa habitualmente así:

$$X \sim \chi_k^2$$

Si se considera una variable aleatoria Z → N (0, 1), la variable aleatoria $X = Z^2$ se distribuye según una ley de probabilidad: distribución (χ^2) con un grado de libertad, χ_n^2, lo cual se representa como: $X \sim \chi_1^2$.

La distribución chi-cuadrado (χ^2) tiene muchas aplicaciones en inferencia estadística. La más conocida es la de la denominada prueba χ^2 utilizada como prueba de independencia, como prueba de bondad de ajuste y en la estimación de varianzas. Pero también está involucrada en el problema de estimar la media de una población normalmente distribuida y en el problema de estimar la pendiente de una recta de regresión lineal, a través de su papel en la distribución t de student.

Aparece también en todos los problemas de análisis de varianza por su relación con la distribución F de Snedecor, que es la distribución del cociente de dos variables aleatorias independientes con distribución χ^2.

Distribución t de Student

La distribución t de Student (n) (Fernández-Abascal, 1994: 478-481) fue propuesta y tabulada por William Sealy Gosset (1876-1937), más conocido por el seudónimo de Student, como resultado de un estudio sobre la estimación de la media cuando el tamaño de muestra es pequeño, estos resultados fueron publicados en 1908 en el artículo: "The Probable Error of a Mean" (El error probable de la media).

La distribución t de Student queda completamente definida por medio de sus grados de libertad (n) y se denota por t_n.

Surge cuando se plantea estudiar el cociente entre una variable aleatoria con distribución normal estándar y la raíz cuadrada del cociente entre una variable aleatoria con distribución chi-cuadrado y sus grados de libertad (n), siendo las dos variables independientes.

Como características de la distribución t-Student se tiene:

– Es muy similar a la curva normal estándar, pero más dispersa, su media también es cero y es simétrica respecto a la misma.

- La varianza decrece hasta 1 cuando el número de grados de libertad aumenta.
- Es simétrica respecto a la media, su forma depende de los grados de libertad, así en la medida que aumentan los grados de libertad mayor es la aproximación a la normal.
- Para un número alto de grados de libertad se puede aproximar la distribución t de Student por la normal, es decir, $t_n \xrightarrow{n \to \infty} N(0, 1)$.
- Los grados de libertad se refieren al número de términos independientes que es necesario para obtener el valor de la variable tn.
- Puede tomar valores tanto positivos como negativos y al igual que toda distribución continua, el área bajo la curva es igual a uno.
- Es empleada como la distribución de probabilidad del cociente:

$$\text{t-student}(T) = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

cuando se cuenta con una muestra pequeña y se desconoce la desviación estándar poblacional, la cual es estimada de la cuasivarianza muestral (S^2).
- Este modelo teórico, al igual que la chi-cuadrado se deriva del modelo normal y se denota por la letra t minúscula (Fig. 3.29).

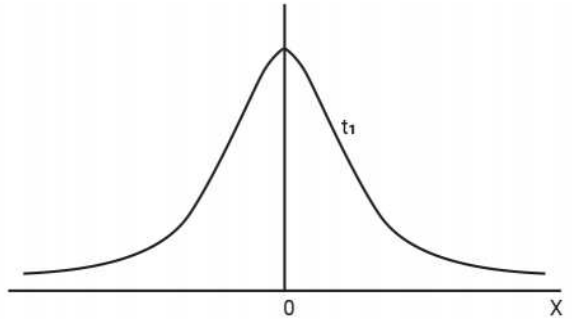


Fig. 3.29. Gráfico para representar la función de densidad de probabilidad para la distribución de la t de Student con un grado de libertad.

- Su valor esperado es E(t) = 0 y la varianza V(t) = n/(n – 2).
- Al igual que en la distribución chi-cuadrado, existen tablas teóricas de probabilidades para los percentiles de esta distribución (tStudent) según el tamaño muestral y los grados de libertad, los que son empleados en las pruebas de hipótesis y en la construcción de intervalos de confianza.
- Esta distribución desempeña un papel muy importante en la inferencia estadística asociada a la teoría de muestras pequeñas y es usada habitualmente en el contraste de hipótesis para la media de una población o para comparar medias de dos poblaciones.
- Campo de variación:

  $-\infty < x < \infty$
- Parámetros:

n: grados de libertad con n ≥ 1 entero.

– Función de densidad:

$$f(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad -\infty < x < \infty$$

,

– Valores característicos:

- Media = Mediana = Moda: 0.
- Varianza:

$$Var(x)=\frac{n}{n-1} \text{ para } n > 2.$$

- Asimetría: 0.
- Curtosis:

$$Kurt(x)=\frac{6}{n-4} \text{ para } n > 4.$$

La distribución t–Student se construye como un cociente entre una normal y la raíz de una χ^2 independientes. Se le denomina distribución t-Student con n grados de libertad (t_n) a la de una variable aleatoria (T):

$$T=\frac{Z}{\sqrt{\frac{1}{n}\chi^2_n}} \rightarrow t_n$$

donde:

$$Z \rightarrow N(0, 1), \quad \chi^2_n \rightarrow \chi^2_n$$

Este tipo de distribuciones aparece cuando se tiene n + 1 variables aleatorias independientes $X \rightarrow N(\mu, \sigma^2)$, $X_i \rightarrow N(\mu_i, \sigma^2_i)$, i = 1,..., n e interesa la distribución de:

$$T=\frac{\frac{x-\mu}{\sigma}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2}} \rightarrow t_n$$

En cuanto a la forma que presenta su función de densidad cabe destacar las similitudes que mantiene con la función de densidad de la distribución normal estándar: forma acampanada, simétrica y centrada en el origen; la única diferencia existente entre ambas distribuciones es que la función de densidad de la t-Student presenta unas colas más pesadas (mayor dispersión) que la normal.

Distribución F de Snedecor-Fisher

La distribución F de Snedecor-Fisher (n, m) (Fernández-Abascal, 1994:482-486) es otra de las distribuciones importantes asociadas a la normal y se define como el

cociente de dos variables aleatorias independientes con distribución chi-cuadrado divididas entre sus respectivos grados de libertad, n y m; la variable aleatoria resultante sigue una distribución F de Snedecor de parámetros n y m (denotada por $F_{n,m}$). Hay muchas aplicaciones de la distribución F en estadística y, en particular, tiene un papel importante en las técnicas del análisis de la varianza (ANOVA) y del diseño de experimentos. Debe su nombre al matemático y estadístico americano George Waddel Snedecor (1881-1974):

- Campo de variación:
  $0 < x < \infty$
- Parámetros:
  n: grados de libertad del numerador, n ≥ 1 entero.
  m: grados de libertad del denominador, m ≥ 1 entero.
- Función de densidad:

$$f(x) = \frac{\Gamma\left(\frac{n+m}{2}\right)\left(\frac{n}{m}\right)^{\frac{n}{2}} x^{\frac{n-2}{2}}}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)\left(1+\frac{nx}{m}\right)^{\frac{n+m}{2}}} \quad x > 0,$$

- Valores característicos:
  - Media:

$$\bar{x} = \frac{m}{m-2} \quad \text{para } m > 2.$$

  - -Mediana: no tiene expresión explícita.
  - Moda:

$$M = \frac{m(n-2)}{n(m+2)} \quad \text{para } n > 2.$$

  - Varianza:

$$\text{Var}(x) = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)} \quad \text{para } m > 4.$$

  - Asimetría:

$$A(x) = \frac{(2n+m-2)\sqrt{8(m-4)}}{(m-6)\sqrt{n(n+m-2)}} \quad \text{para } m > 6.$$

  - Curtosis:

$$\text{Kurt}(x) = \frac{12\left[(m-2)^2(m-4) + n(n+m-2)(5m-22)\right]}{n(m-6)(m-8)(n+m-2)} \quad \text{para } m > 8.$$

La distribución F es una distribución de probabilidad continua. También se le conoce como distribución F de Snedecor (por George Snedecor) o como distribución F de Fisher-Snedecor (Fig. 3.30).

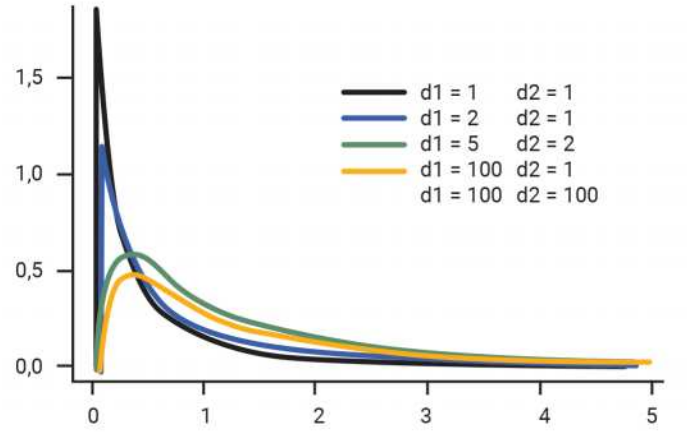


Fig. 3.30. Gráfico para la representación de varias curvas (con diferentes colores según grados de libertad d_1 y d_2 en el numerador y denominador, respectivamente) para la función de densidad de probabilidad correspondiente a la distribución F de Snedecor-Fisher.

Esta distribución de probabilidad está también asociada a la distribución Normal y se define como el cociente de distribuciones χ^2 independientes. Si $X \to \chi^2_n$, $Y \to \chi^2_m$ son variables aleatorias independientes, se dice entonces que la variable:

$$F = \frac{\frac{1}{n}X}{\frac{1}{m}Y} = \frac{m}{m}\frac{Y}{x} \to F_{m,n}$$

sigue una distribución de probabilidad de Snedecor, con n, m grados de libertad.

La forma más habitual en que nos encontraremos esta distribución sería en el caso en que tengamos n + m variables aleatorias independientes:

$X_i \to N(\mu_i, \sigma^2_i)$, $i = 1,\ldots, n$

$Y_j \to N(m_j, s^2_j)$, $j = 1,\ldots, n$

por lo que:

$$F = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2}{\frac{1}{m}\sum_{j=1}^{m}\left(\frac{Y_j - m_j}{s_j}\right)^2} \to F_{n,m}$$

La distribución F aparece frecuentemente como la distribución nula de una prueba estadística, especialmente en el análisis de varianza.

Distribución Cauchy

La distribución Cauchy (μ, θ) (Fernández-Abascal,1994: 461-463) fue introducida por Simeón Denis Poisson (1781-1840) en 1824, aunque debe su nombre al matemático

francés Augustin Louis Cauchy (1789-1857) quien la reintrodujo en 1853. En el ámbito de la física también es conocida con el nombre de *distribución de Lorentz* o *distribución de Breit-Wigner*.

La distribución de Cauchy depende de dos parámetros: escala (μ) y situación (θ), en el caso particular de que $\mu = 1$ y $\theta = 0$, recibe el nombre de *distribución de Cauchy estándar*.

Una característica destacable de esta distribución es que carece de momentos, por lo que no existen la media, varianza, asimetría y curtosis de esta distribución. Su función de densidad es simétrica respecto al parámetro de situación θ:

- Campo de variación:
  $-\infty < x < \infty$
- Parámetros:
  μ: escala, $\mu > 0$.
  θ: situación, $-\infty < \theta < \infty$.
- Función de densidad:

$$f(x) = \frac{\mu}{\pi}\frac{1}{\left[\mu 2 + (x-\theta)^2\right]} \quad -\infty < x < \infty.$$

- Valores característicos:
  - Media: no definida.
  - Mediana: 0.
  - Moda: 0.
  - Varianza: no definida.
  - Asimetría: no definida.
  - Curtosis: no definida.

  Nota: con $\mu = 1$ y $\theta = 0$ se tiene la distribución de Cauchy estándar.

Distribución Weibull

La distribución Weibull (a, b) (Canavos, 1988: 159-163) debe su nombre al físico sueco Waloddi Weibull (1887-1979) quien la usó en un artículo publicado en 1939 sobre resistencia de los materiales: "A Statistical Theory of the Strength of Materials", aunque ya era conocida de años antes. Se utiliza para modelar situaciones del tipo tiempo-fallo, modelar tiempos de vida o en el análisis de supervivencia, a parte de otros usos como, por ejemplo, caracterizar el comportamiento climático de la lluvia en un año determinado.

La distribución Weibull queda totalmente definida mediante dos parámetros: forma (a) y escala (b). En el caso particular de que a = 1, se tiene la distribución exponencial y si a = 2 y $b = \sqrt{2\sigma}$ recibe el nombre de distribución de Rayleigh.

El perfil de la función de densidad presenta formas muy variadas dependiendo del valor que tome su parámetro de forma a. Si a es menor o igual que 1, la función de densidad es siempre decreciente, en caso de tomar valores mayores que la unidad su función de densidad muestra una forma más acampanada, pero no simétrica, de forma que crece hasta alcanzar el máximo y luego decrece:

– Campo de variación:
  $0 < x < \infty$
– Parámetros:
  a: forma, a > 0.
  b: escala, b > 0.
– Función de densidad:

$$f(x) = \frac{a}{b}\left(\frac{x}{b}\right)^{a-1} \exp\left[-\left(\frac{x}{b}\right)^{a}\right], \; x > 0$$

– Valores característicos:
  • Media:

$$\bar{x} = b\Gamma\left(\frac{1}{a}+1\right)$$

  • Mediana:

$$Me = b(\ln 2)^{\frac{1}{a}}$$

  • Moda:

$$M = b\left(\frac{a-1}{a}\right)^{\frac{1}{a}}, \text{ para } a > 1.$$

  • Varianza:

$$Var(x) = b\left[\Gamma\left(\frac{2}{a}+1\right) - \left[\Gamma\left(\frac{1}{a}+1\right)\right]^2\right]$$

  • Asimetría:

$$A(x) = \frac{2\Gamma^3\left(\frac{1}{a}+1\right) - 3\Gamma\left(\frac{1}{a}+1\right)\Gamma\left(\frac{2}{a}+1\right) + \Gamma\left(\frac{3}{a}+1\right)}{\left[\Gamma\left(\frac{2}{a}+1\right) - \Gamma^2\left(\frac{1}{a}+1\right)\right]^{\frac{3}{2}}}$$

  • Curtosis:

$$Kurt(x) = \frac{-6\Gamma^4\left(\frac{1}{a}+1\right) + 12\Gamma^2\left(\frac{1}{a}+1\right)\Gamma\left(\frac{2}{a}+1\right) - 3\Gamma^2\left(\frac{2}{a}+1\right) - 4\Gamma\left(\frac{1}{a}+1\right)\Gamma\left(\frac{4}{a}+1\right)}{\left[\Gamma\left(\frac{2}{a}+1\right) - \Gamma^2\left(\frac{1}{a}+1\right)\right]^{\frac{3}{2}}}$$

Distribución de Laplace

La distribución de Laplace fue descubierta en 1774 por Pierre-Simon Laplace (1749-1827), a quien debe su nombre, aunque también es conocida por el nombre de *distribución doble exponencial*.

Esta distribución viene determinada por dos parámetros, uno de situación (a) y otro de escala (b).

Su función de densidad es simétrica y el parámetro de situación determina su eje de simetría, además de ser el punto donde la función alcanza su valor máximo en forma de pico afilado. Independientemente de los valores que tomen sus parámetros, es una distribución leptocúrtica, lo que quiere decir que su función de densidad es más apuntada que la función de densidad de la normal con la misma media y desviación estándar:

- Campo de variación:
  $-\infty < x < \infty$
- Parámetros:
  a: situación, $-\infty < a < \infty$.
  b: escala, $b > 0$.
- Función de densidad:

$$f(x) = \frac{1}{2b}\exp\left(-\frac{x|z-a|}{b}\right),\ -\infty < x < \infty$$

- Valores característicos:
  - Media = mediana = moda:
    $\bar{x} = Me = M = a$
  - Varianza:
    $Var(x) = 2b^2$
    - Asimetría: 0.
    - Curtosis: 3.

Distribución de Pareto

La distribución de Pareto fue introducida por el economista italiano Vilfredo Pareto (1848-1923) como un modelo para explicar la distribución de las rentas de los individuos de una población, siempre y cuando se partiera de dos supuestos: la existencia de un umbral inferior (x_0) de forma que no haya rentas inferiores a dicho umbral y el decrecimiento de manera potencial del porcentaje de individuos con una renta superior o igual a un cierto valor de renta a medida que dicho valor de renta crece. El uso de esta distribución se ha ido ampliando a diferentes ámbitos de estudio.

Se trata de una distribución biparamétrica, con parámetros de forma (α) y de situación (x_0). El parámetro x_0 es un indicador de posición (valor mínimo) que, en términos económicos, puede interpretarse como el ingreso mínimo de la población. El parámetro α está asociado con la dispersión, donde a mayor valor se obtienen densidades de Pareto más concentradas en las proximidades de x_0, es decir, menos dispersas:

- Campo de variación:
  $x_0 \leq x < \infty$
- Parámetros:
  α: forma, $\alpha > 0$.
  x_0: situación, $x_0 > 0$.

– Función de densidad:

$$f(x) = \frac{\alpha x_0^{\alpha}}{x^{\alpha+1}},\ x \geq x_0$$

– Valores característicos:

- Media:

$$\bar{x} = \frac{\alpha x_0}{\alpha - 1},\ \text{para } \alpha > 1.$$

- Mediana:

$$Me = x_0 2^{\frac{1}{\alpha}}$$

- Moda:

$$M = x_0$$

- Varianza:

$$Var(x) = \frac{\alpha x_0^2}{(\alpha-2)(\alpha-1)^2},\ \text{para } \alpha > 2$$

- Asimetría:

$$A(x) = \frac{2(1+\alpha)}{\alpha-3}\sqrt{\frac{\alpha-2}{\alpha}},\ \text{para } \alpha > 3$$

- Curtosis:

$$Kurt(x) = \frac{6(\alpha^3+\alpha^2-6\alpha-2}{\alpha(\alpha-3)(\alpha-4)},\ \text{para } \alpha > 4$$

Distribución triangular

El nombre de la distribución triangular (a, c, b) (Herrerías y Palacios, 2007) viene dado por la forma de su función de densidad. Este modelo proporciona una primera aproximación cuando hay poca información disponible, de forma que solo se necesita conocer el mínimo (valor pesimista), el máximo (valor optimista) y la moda (valor más probable).

Estos tres valores son los parámetros que caracterizan a la distribución triangular y se denotan por a, b y c, respectivamente.

Un ejemplo del uso de esta distribución se encuentra en el análisis del riesgo, donde la distribución más apropiada es la beta, pero dada su complejidad, tanto en su comprensión como en la estimación de sus parámetros, se utiliza la distribución triangular como proxy para la beta:

– Campo de variación:

$a \leq x \leq b$

– Parámetros:

a: mínimo, $-\infty < a < \infty$.

c: moda, $-\infty < c < \infty$ con $a \leq c \leq b$.

b: máximo, $-\infty < b < \infty$ con $a < b$.

– Función de densidad:

$$f(x) = \frac{2(x-a)}{(b-a)(c-a)}, \text{ para } a \le x \le c$$

$$f(x) = \frac{2(b-x)}{(b-a)(c-a)}, \text{ para } c \le x \le b$$

– Valores característicos:

- Media:

$$\bar{x} = \frac{a+b+c}{3}$$

- Mediana:

$$Me = b - \sqrt{\frac{(b-a)(b-c)}{2}} \text{ en caso de que: } c \le \frac{a+b}{2}$$

$$Me = a + \sqrt{\frac{(b-a)(c-a)}{2}} \text{ en caso de que: } c \ge \frac{a+b}{2} \; c \ge \frac{a+b}{2}$$

- Moda:

  M = c

- Varianza:

$$Var(x) = \frac{(b-c)^2 + (c-a)^2 + (b-c)(c-a)}{18}$$

- Asimetría:

$$A(x) = \frac{\sqrt{2}(a+b-2c)(b+c-2a)(2b-c-a)}{5\left[(b-a)^2 - (c-a(b-c)\right]^{\frac{3}{2}}}$$

- Curtosis:

$$Kurt(x) = -\frac{3}{5}$$

Para recordar

A modo de resumen se presenta en las tablas 3.40 y 3.41 las características esenciales y de más amplio uso en las distribuciones de probabilidad: discretas y continuas.

Tabla 3.40. Resumen de las distribuciones de probabilidad discretas

Distribución	Valores	Parámetros	Definición de la variable	Observaciones
Uniforme discreta	a, a + 1, a + 2,...,b	a: mínimo b: máximo	Variable que puede tomar n valores distintos con la misma probabilidad cada uno de ellos	

Distribución	Valores	Parámetros	Definición de la variable	Observaciones
Binomial	0, 1, 2,..., n	n: número de pruebas p: probabilidad de éxito	Número de éxitos en n pruebas independientes de un experimento con probabilidad de éxito constante	Esta distribución se aplica a poblaciones finitas cuando los elementos se toman al azar y con reemplazo y a poblaciones conceptualmente infinitas cuando el proceso es estable y sin memoria
Multinomial	X_i: 0, 1, 2,... (i = 1, 2,..., m)	n: número de pruebas m: número de resultados posibles p_i: probabilidad del suceso i	Número de veces que ocurren m sucesos disjuntos en n pruebas independientes	Se aplica cuando se tiene un proceso estable y sin memoria
Hipergeométrica	De máximo: {0, n-(N-R)} A mínimo: {R, n}	N: tamaño de la población R: número de éxitos n: número de pruebas	Número de éxitos en una muestra de tamaño n, extraída sin reemplazo de una población de tamaño N que contiene R éxitos	Es equivalente a la distribución binomial cuando el muestreo se hace sin reemplazo. Si el tamaño de la población es grande ambas distribuciones se pueden considerar prácticamente iguales
Geométrica	0, 1, 2,...	p: probabilidad de éxito	Número de fracasos antes de obtener un éxito por primera vez	Se utiliza en la distribución de tiempos de espera y tiene la propiedad de "falta de memoria"
Binomial negativa	0, 1, 2,...	r: número de éxitos p: probabilidad de éxito	Número de fracasos antes de obtener el r-ésimo éxito	Cuando r = 1 se obtiene la distribución geométrica.

Tabla 3.40. Continuación...

Distribución	Valores	Parámetros	Definición de la variable	Observaciones
Pascal	r, r + 1, r + 2,...	r: número de éxitos p: probabilidad de éxito	Número de pruebas necesarias para obtener r éxitos	Se relaciona con la binomial negativa de la siguiente manera: Pascal (r, p) = BN(r, p) + r
Poisson	0, 1, 2,...	λ: tasa de ocurrencia	Número de ocurrencias de un evento "raro" o poco frecuente en un intervalo o espacio continuo de tiempo.	El proceso que genera una distribución de Poisson es estable y no tiene memoria La distribución binomial se aproxima por la Poisson si n es grande y p pequeña, siendo λ = np

Tabla 3.41. Resumen de las distribuciones de probabilidad continuas

Distribución	Campo de variación	Parámetros	Observaciones
Uniforme	(a, b)	a: mínimo b: máximo	Distribución clave en la generación de distribuciones
Normal	$(-\infty, \infty)$	μ: media σ: desviación estándar	Si $\mu = 0$ y $\sigma = 1$ se denomina *distribución normal estándar*. De ella derivan las distribuciones chi-cuadrado, t de Student y F de Snedecor
Normal bivariante	$X_1 \epsilon (-\infty, \infty)$ $X_2 \epsilon (-\infty, \infty)$	$\mu = (\mu_1, \mu_2)$: media $\sigma = (\sigma_1, \sigma_2)$: desviación estándar ρ: coeficiente de correlación	
Lognormal	$(0, \infty)$	μ: escala σ: forma	Si X sigue una distribución lognormal entonces su logaritmo neperiano sigue una distribución normal

Distribución	Campo de variación	Parámetros	Observaciones
Logística	$(-\infty, \infty)$	a: situación b: escala	Si U sigue una distribución uniforme en el intervalo (0, 1) entonces $X = \ln(U/(1-U))$ sigue una distribución logística
Beta	(0, 1)	p: forma q: forma	Es adecuada para modelar proporciones Si p = q = 1 se obtiene la distribución uniforme en (0, 1)
Gamma	$(0, \infty)$	a: escala p: forma	Es adecuada para modelar tiempos de vida Si p es un número entero se denomina *distribución de Erlang*
Exponencial	$(0, \infty)$	λ: tasa	Equivalente continuo de la distribución geométrica, también posee la propiedad de "falta de memoria"
Chi-cuadrado	$(0, \infty)$	n: grados de libertad	Distribuciones importantes en la contrastación de hipótesis estadísticas
T de Student	$(-\infty, \infty)$	n: grados de libertad	
F de Snedecor	$(0, \infty)$	n: grados de libertad m: grados de libertad	
Cauchy	$(-\infty, \infty)$	μ: escala θ: situación	Si $\mu = 1$ y $\theta = 0$ se denomina *distribución de Cauchy estándar*
Weibull	$(0, \infty)$	a: forma b: escala	Si a = 1 se tiene la distribución exponencial Otro caso particular es la distribución de Rayleigh
Laplace	$(-\infty, \infty)$	a: situación b: escala	
Pareto	$[x_0, \infty)$	α: forma x_0: situación	
Triangular	[a, b]	a: mínimo c: moda b: máximo	Se emplea cuando hay poca información disponible de la variable

Capítulo 4

Algunos aspectos conceptuales de la inferencia bioestadística

Es en una práctica cotidiana para los seres humanos enfrentarse a la toma de decisiones, que pueden ser simples si se cuenta, por ejemplo, con la experiencia anterior suficiente para esa situación o pueden ser complejas cuando de la decisión, por ejemplo, depende una acción que implica la vida de una persona o de un grupo de ellas. La toma de decisiones está siempre asociada a un determinado riesgo, es decir, tomar una decisión que sea desacertada o errónea. A veces la persona que tiene que tomar la decisión no tiene una clara percepción del riesgo a que se expone al tomar la decisión, en otras lo conoce, pero no es capaz de medirlo.

En este capítulo se desarrollan procedimientos estadísticos relacionados con la toma de decisiones ante la contrastación de hipótesis estadísticas. Uno de los métodos estadísticos utilizado con más frecuencia que emplean la toma de decisiones acerca de los parámetros de la distribución de una población sobre la base de la información de una o varias de sus muestras reciben el nombre de contraste de hipótesis estadísticas, prueba de hipótesis, dócimas de hipótesis o prueba de significación y constituyen el núcleo de la inferencia bioestadística. Por lo general los métodos estadísticos proporcionan un algoritmo en el procedimiento para realizar la, utilizando los medios digitales informáticos (computacionales) que en la actualidad se ha puesto al alcance de muchos profesionales, por contar con la informática y el *software* estadístico que lo facilita. En este capítulo se hace también una introducción que abarca la terminología general y el desarrollo del modelo de prueba de hipótesis, particularmente se profundiza en el tratamiento de hipótesis en el que se decide según el valor de una probabilidad muy relacionada con el nivel de significación, que se denomina *probabilidad crítica*, el resultado de este procedimiento facilita una acertada toma de decisión respecto a las hipótesis estadísticas contrastadas, como paso previo de preparación teórica a la utilización práctica de las tecnologías informáticas digitales actuales para este tema en particular.

Contraste de hipótesis estadísticas como parte del procesamiento estadístico de la información y bases para su interpretación

El contraste de hipótesis, también llamado *prueba de significación*, es uno de los recursos más extendidos dentro de la inferencia estadística. Fue propuesto por Ronald Fisher en 1925 y desarrollado por Jerzy Neyman y Egon Pearson en un artículo publicado en 1928, titulado "On the use and interpretation of certain test criteria for purposes of statistical inference".

Las pruebas estadísticas encaminadas a verificar la veracidad de una hipótesis estadística se denominan *pruebas de hipótesis*. La hipótesis a contrastar recibe el nombre de hipótesis nula y se simboliza por H_0. Junto a esta se plantea siempre su hipótesis complementaria, que recibe el nombre de *hipótesis alternativa* y se simboliza por H_1. Esta última no es sometida directamente a prueba, pero resulta conveniente plantearla, pues es la más verosímil cuando la prueba conduce a rechazar la H_0.

Las pruebas de hipótesis son construidas en función de la hipótesis nula, por consiguiente, se debe expresar sus conclusiones en términos de rechazo o de no rechazo de H_0, o en los términos equivalentes de valor significativo o no significativo estadísticamente. No se debe emplear el término *acepto* porque implica demostración y un resultado significativo no es una demostración de la veracidad de la H_0.

Un resultado no significativo estadísticamente solo indica que es compatible con la hipótesis nula porque la discrepancia observada puede ser explicada por el error aleatorio del muestreo. Pero la significación estadística depende tanto de la magnitud de la diferencia observada como del tamaño de la muestra estudiada.

Estudios con muestras pequeñas difícilmente obtendrán significación estadística aunque las diferencias sean importantes. De la misma manera, estudios con muestras grandes obtendrán significación estadística aunque las diferencias sean poco importantes a nivel clínico o epidemiológico. Por ello, afirmar simplemente que existe o no significación estadística aporta poca información en sí misma. El intervalo de confianza para un determinado parámetro constituye un conjunto aceptable de hipótesis y proporciona información sobre la magnitud de dicho parámetro que ayuda al clínico o al epidemiólogo experto a valorar la importancia clínica de ese resultado.

Comparando el valor obtenido del estadístico de prueba y la probabilidad que representa con la distribución teórica de probabilidades (de referencia) se calcula la probabilidad de que se dé un valor como el obtenido o mayor en el caso de que la hipótesis nula sea cierta. Al valor de esta probabilidad se le llama "valor p" o "p-valor".

Esta prueba de significación permite, tras plantearse cierta hipótesis estadística (hipótesis nula: H_0) sobre la población en estudio, tomar la decisión de si se rechaza o

no, a partir de la información recogida en la muestra. Esta decisión se toma en virtud del llamado valor p, que es la probabilidad de obtener un resultado muestral como el que se está observando (medido a través del estadístico del contraste) u otro más extremo que este, bajo el supuesto de que fuera cierta la hipótesis nula, es decir, para los posibles valores que quedarían de probabilidad más hacia los extremos que el valor del estadístico de contraste experimental. El estadístico de contraste o de prueba, que se puede comportar como una variable aleatoria, representa el resultado numérico de una determinada formulación matemática que contiene y utiliza los valores de los respectivos parámetros poblacionales estimados de los estimadores según el tipo de contraste específico a realizar. Este estadístico permite medir la discrepancia entre la hipótesis nula y las observaciones de la muestra y cuya distribución de probabilidad es conocida según sus posibles valores bajo H_0 (si se asume que la hipótesis nula fuese cierta).

El contraste de hipótesis, también denominado *prueba de hipótesis* o *prueba de significación*, es un procedimiento para juzgar si una propiedad que se supone en una población estadística es compatible con lo observado en una muestra de esa población.

Mediante esta teoría se aborda el problema estadístico considerando una hipótesis determinada H_0 (nula o de nulidad) y una hipótesis alternativa H_1, y se intenta dirimir cuál de las dos es la hipótesis verdadera, tras aplicar el problema estadístico a un cierto número de experimentos.

Para la contrastación de las hipótesis se emplean técnicas paramétricas que utilizan parámetros poblacionales estimados de los muestrales en el cálculo del estadístico de prueba o contraste y no paramétricas, sin parámetros poblacionales estimados, en su lugar se utilizan estadígrafos o estadísticos específicos en cada caso que siguen una distribución de frecuencias libre.

Se denomina hipótesis nula siempre a la hipótesis que se desea contrastar. El nombre de "nula" significa "sin valor, efecto o consecuencia", lo que sugiere que debe identificarse con la hipótesis de no cambio o en el sentido de "neutra"; mientras que la hipótesis alternativa expresa lo contrario a la nulidad.

El contraste de hipótesis está fuertemente asociada a los considerados errores de tipo I y II en estadística, que definen, respectivamente, la posibilidad de tomar un suceso como falso siendo verdadero o uno como verdadero siendo falso, respectivamente.

Los errores de contraste al tomar la decisión respecto al problema estudiado se pueden dar en los cuatro casos (Tabla 4.1).

Tabla 4.1. Tipos de errores relacionados con el contraste de hipótesis estadísticas

	H_0 es cierta	**H_1 es cierta**
Se escogió H_0	No hay error	Error de tipo II
Se escogió H_1	Error de tipo I	No hay error

Si la probabilidad (P) de cometer un error de tipo I está unívocamente determinada, su valor se suele denotar por la letra griega α y en las mismas condiciones se denota por β la probabilidad de cometer el error de tipo II, esto se expresa como:

P (escoger H_1 |H_0 es cierta) = α

P (escoger H_0 |H_1 es cierta) = β

Para interpretar las inferencias estadísticas, el investigador debe especificar los niveles aceptables de error estadístico. El modo de aproximación más común es determinar el nivel de error tipo I, también conocido como alfa. El error tipo I es la probabilidad de rechazar la hipótesis nula cuando es cierta o, expresado en términos más sencillos, la posibilidad de que la prueba muestre significación estadística cuando en realidad no está presente (el caso de un "positivo falso"). Estableciendo un nivel de significación alfa el investigador fija los márgenes o límites de referencia admisibles de error aclarando la probabilidad de concluir que la significación existe cuando en realidad no existe.

Al especificar el nivel o valor del error de tipo I o alfa, el investigador también determina un error asociado, denominado el error de tipo II o beta. El error de tipo II es la probabilidad de fallar al no rechazar la hipótesis nula cuando es realmente falsa y por tanto rechazable.

Una probabilidad más interesante es 1 – β, denominado la potencia de la prueba de inferencia estadística. Potencia es la probabilidad de rechazar correctamente la hipótesis nula cuando debe ser rechazada. Por tanto, la potencia es la probabilidad de que la inferencia estadística se indique cuando esté presente (Tabla 4.2).

Tabla 4.2. Posibles resultados en el contraste de hipótesis estadísticas a partir de la relación entre la decisión estadística del investigador y los resultados estadísticos reales del estudio, según las opciones para la hipótesis nula (H_0)

Decisión estadística	**En la realidad**	
	H_0: cierta	**H_0: falsa**
H_0: aceptar	Probabilidad = 1 – α (hay acierto)	Riesgo β (error de tipo II)
H_0: no aceptar	Riesgo α (error de tipo I)	Potencia = 1 – β (hay acierto)

Para recordar:

Basados en la información que brinda la tabla 4.2 se debe tomar en cuenta que cuando se realiza una prueba de hipótesis la decisión acerca de la veracidad de la hipótesis nula puede conducir a cometer dos tipos de errores:

- Error de tipo I: consiste en rechazar la hipótesis nula cuando es cierta. La probabilidad de incurrir en este error se denomina *nivel de significación* y se denota por α. El valor 1 – α expresa, por tanto, la probabilidad de no rechazar una hipótesis que es cierta y se conoce como el *nivel de confianza de la prueba*.
- Error de tipo II: consiste en aceptar la hipótesis nula cuando esta es en realidad falsa. La probabilidad de cometer este error se denota por β y el valor 1 – β que expresa la probabilidad de rechazar la hipótesis nula cuando es falsa se conoce como *la potencia de la prueba*.

Lo ideal es que estas dos posibilidades se puedan hacer muy pequeñas, pero no se puede disminuir una de ellas sin que aumente inmediatamente la otra. La solución consiste en fijar el valor de α lo suficientemente pequeño y tratar de hacer mínimo el valor de β, una vez fijado el valor de α. Como la probabilidad de equivocación β no se controla, entonces se fija el contenido de las hipótesis, de modo que el error tipo 1 sea el de consecuencias más graves, así la probabilidad de cometerlo queda convenientemente fijada. Un criterio equivalente al anterior es el de poner en la hipótesis alternativa a lo que se quiere probar. Por sobre todo esto predomina el hecho de que en la hipótesis nula siempre debe ir la igualdad, menor o igual o mayor o igual.

Antes de proseguir con la presentación del procedimiento estandarizado y aceptado en general por la literatura especializada para efectuar el contraste de hipótesis, se deben retomar algunos conceptos:

- Estadígrafo: es la fuente de información muestral, una función de la misma, representa el valor, por ejemplo, de una medida de resumen para variables o de cierta combinación aritmética de sus datos: media, moda, percentil, desviación estándar, proporción, coeficiente de correlación, entre otras, todos muestrales.
- Estimador: es un estadígrafo bueno para estimar al cumplir las propiedades de ser insesgado, o sea, que su valor esperado coincide con el parámetro que estima y su varianza es mínima en el conjunto de todos los estimadores insesgados del parámetro en cuestión, se entiende como el valor esperado a la media de la variable aleatoria en cuestión.
- Parámetro: constituye la fuente de información poblacional, es un indicador numérico de la población y puede ser estimado a partir de estadígrafos que cumplen con las propiedades de buenos estimadores: media, moda, percentil, desviación estándar, proporción, coeficiente de correlación, entre otros, todos poblacionales.
- Estadígrafo o estadístico calculado, observado, experimental, de prueba, de contraste: expresa el valor calculado a partir de la combinación en una fórmula matemática de los respectivos estimadores o estadígrafos. Se emplea para ser comparado con el valor esperado, teórico, crítico o de referencia que se corresponde con un nivel de confianza de la distribución de una variable aleatoria correspondiente al contraste de hipótesis empleado.
- Estadígrafo o estadístico esperado, teórico, de referencia, tabular, crítico: expresa el valor de uno de los percentiles pertenecientes a una distribución teórica para una variable aleatoria que se corresponde con cierto nivel de confianza escogido por el investigador para poder efectuar el contraste de hipótesis correspondiente.
- Decisión: consiste en rechazar o no la hipótesis H_0 sobre la base de la información muestral, es decir, a partir de los resultados obtenidos utilizando los datos muestrales y mediante la comparación de los estadísticos de prueba observado-experimental *versus* el estadístico teórico-esperado según el nivel de confianza seleccionado.
- Región crítica: es la región de rechazo para la hipótisis H_0, expresa un intervalo limitado de valores de percentiles dentro del conjunto total de ellos que pertenecen a la

distribución teórica de una variable aleatoria correspondiente, cuyo intervalo de valores está determinado por el nivel de confianza seleccionado por el investigador o su contraparte que es el nivel de significación α y que queda representa en un gráfico de coordenadas (x, y) para los que si se obtuviera un valor del estadístico de prueba calculado-experimental, en este intervalo o región, lo que permite rechazar la hipótesis H_0.

– Probabilidad crítica (PrC): es la probabilidad de que el estadígrafo o estadístico de prueba experimental o calculado caiga o se ubique en el intervalo limitado de valores correspondientes con la región crítica, es decir, expresa el valor de probabilidad para la variable aleatoria o estadígrafo seleccionada puntual o más extremo (modularmente mayor) que el correspondiente al valor observado con los datos muestrales del estadígrafo o estadístico de prueba correspondiente. También es reconocido como p-valor o valor p y se puede expresar matemáticamente como:
p-valor = Probabilidad {|valores de la variable aleatoria seleccionada o valores del estadígrafo teórico| > |valor observado del estadígrafo de prueba experimental seleccionado|}, su resultado se emplea de forma comparativa con el nivel de significación α para la toma de decisión en los contrastes de hipótesis respecto al rechace o no de la hipótesis H_0.

Pasos generales a seguir en el contraste de hipótesis

Primero se deben considerar los distintos aspectos generales implicados en este contraste, algunos de los cuales se presentan relacionados siguiendo un algoritmo (Fig. 4.1).

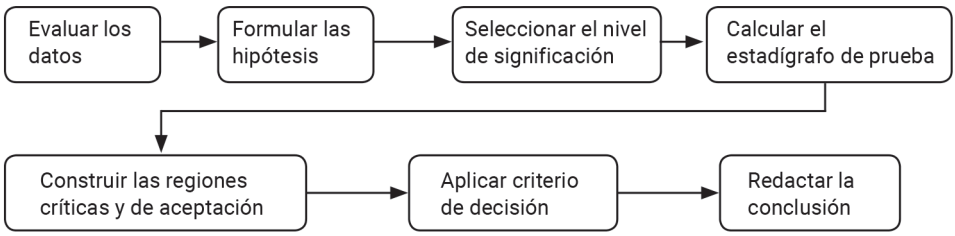


Fig. 4.1. Algoritmo propuesto para la realización de las pruebas de hipótesis.

Es válido recordar que una prueba de hipótesis se denomina *bilateral o de dos colas* cuando la hipótesis alternativa está planteada sin especificar el sentido de la diferencia, mientras que en el caso *unilateral o de una cola* se especifica el sentido o la dirección de la diferencia esperada. En el primer caso la zona de rechazo está dividida en dos partes, a ambos extremos de la distribución teórica de probabilidades seleccionada, siendo las probabilidades a cada lado igual a $\alpha/2$, para que entre sí sumen un total de α, mientras que en el caso unilateral la zona crítica se encuentra hacia un extremo o el otro de esta distribución, cuya área es igual a α. La decisión de cuál contraste emplear depende del

objetivo de la investigación. Los pasos propiamente a seguir son los que se presentan a continuación.

Primer momento

Se realiza el planteamiento de la relación a probar entre los parámetros (partiendo de sus estadígrafos estimadores) de asociación-comparación por medio de hipótesis estadísticas:

1. Hipótesis nula o de nulidad (H_0): se plantea en ella lo que el investigador espera que no deba suceder o lo que está deseando rechazar, es decir, por ejemplo: la existencia de no diferencias entre el comportamiento de parámetros de una variable; entre parámetros de dos o más muestras o que los estadígrafos poblacionales de asociación (coeficientes) tienen valor igual a cero. Este planteamiento debe ir acompañado de un valor probabilístico alfa (nivel de significación) que representa el que el investigador está dispuesto a aceptar como equivocación (error tipo I), al rechazar esta hipótesis nula, siendo realmente cierta la misma. Son varios los ejemplos de formas a plantear este contraste de hipótesis:
   a) Primer ejemplo (contraste bilateral) con dos medias:
      - H_0: $\mu_1 = \mu_2$ o $\mu_1 - \mu_2 = 0$, para dos colas, siendo el intervalo de percentiles de $Z \sim N(0, 1)$ bajo H_0 (área de no rechazo desde $Z_{\alpha/2}$ a $Z_{1-\alpha/2}$), para un nivel de significación alfa prefijado por el investigador.
   b) Segundo ejemplo (contraste unilateral) con dos medias:
      - H_0: $\mu_1 > \mu_2$, para una cola, siendo el intervalo de percentiles de $Z \sim N(0, 1)$ bajo H_0 (área de no rechazo $\leq Z_{1-\alpha}$), para un nivel de significación alfa prefijado por el investigador.
      - H_0: $\mu_1 < \mu_2$, para una cola, siendo el intervalo de percentiles de $Z \sim N(0, 1)$ bajo H_0 (área de no rechazo) $\geq -Z_{\alpha}$, para un nivel de significación alfa prefijado por el investigador.
2. Alternativa (H_1) (o contrario a H_0), es decir, lo que el investigador espera o sospecha que debería suceder a partir de su observación del fenómeno. Se define el área o región bajo la curva de la distribución teórica de probabilidades del estadístico de prueba (región critica o de rechazo) para la cual se rechazaría la H_0, y se cumple o es válida la H_1, es decir, un área donde se cumpliría H0 con un valor menor que el valor alfa prefijado o de incertidumbre.
   1. Primer ejemplo:
      - H_1: $\mu_1 \neq \mu_2$ (contraste bilateral), para dos colas, siendo el intervalo de percentiles de Z bajo H_1 o (área de rechazo de H_0) los mayor de $Z_{1-\alpha/2}$ y los menores de $-Z_{\alpha/2}$, con nivel de significación $\alpha/2$.
        Debe plantearse el área bajo la curva de rechazo de H_0 y por ende de aceptación para H1, es decir, el área fuera del intervalo de confianza para los valores teóricos del estadístico de prueba (Z teórico-tabulado) según el alfa seleccionado. Se emplea para comparar la ubicación del percentil correspondiente al estadístico

de prueba observado-calculado (Z observado-calculado) y su relación con el valor teórico-esperado y tabulado del estadístico de prueba (Z teórico-tabulado) según el alfa prefijado por el investigador (tomados en forma modular en caso necesario), lo que a su vez permite determinar el valor de probabilidad (p-valor) donde es igual o mayor modularmente el valor teórico respecto al observado del estadístico de prueba; o sea, donde se cumpliría H_0 o relacionado con el azar para los datos muestrales o experimentales empleados según el valor de alfa prefijado, esto es empleado para estimar la significación estadística del resultado obtenido.

2. Segundo ejemplo:
   - Si H_1: $\mu_1 > \mu_2$, es para una cola (contraste unilateral), por ejemplo, con $\alpha = 0,05$ (si nos interesa el valor de μ_1 comparativamente), entonces se debe determinar el valor positivo de Z, es decir, $Z_{1-\alpha}$.
   - Si H_1: $\mu_1 < \mu_2$, es para una cola (contraste unilateral) por ejemplo con $\alpha = 0,05$ (si interesa el valor de $\mu 1$ comparativamente), entonces se debe determinar el valor negativo de Z, es decir, $-Z_{\alpha}$.

     La presencia de un valor de probabilidad (p-valor) donde se cumple H_0; es decir, donde el valor de Z (teórico-esperado) es mayor o igual al Z (observado-calculado), en relación con el valor obtenido correspondiente de Z (observado-calculado) específico con los datos muestrales, menor que el valor de alfa prefijado por el investigador, representa un resultado donde es mínima la probabilidad de cumplirse H_0, siendo menor que el valor de incertidumbre que tenía el investigador en un inicio lo cual da razones suficientes para rechazar H_0 y aceptar la alternativa (H_1), concluyendo que existen diferencias significativas entre las dos medias aritméticas comparadas en el nivel de confianza (correspondiendo con el intervalo de valores teóricos de la distribución de probabilidades) seleccionado previamente por el investigador (Figs. 4.2. y 4.3).

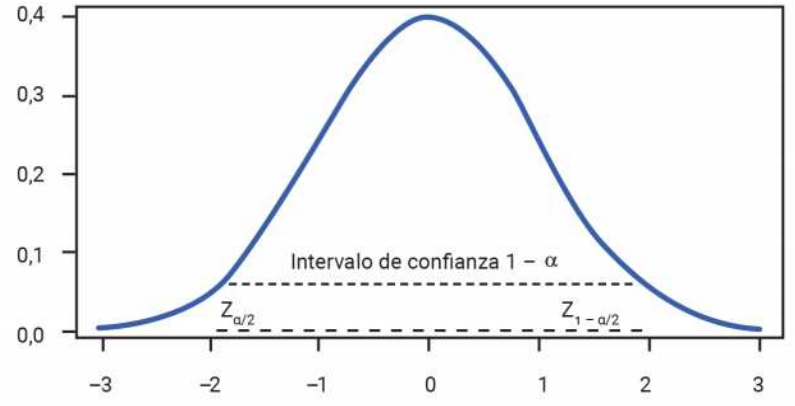


Fig. 4.2. Gráfico de la distribución normal estándar para representar el intervalo de dos valores de Z teóricos empleados en el contraste de hipótesis, donde no se rechazaría H_0 si se incluyese en el mismo el valor de Z observado.

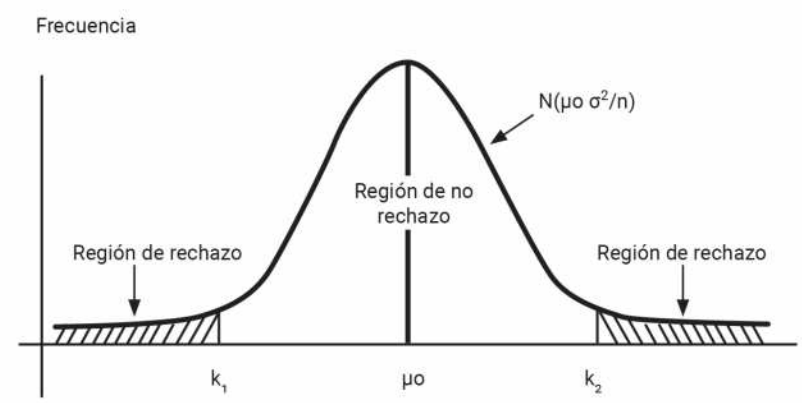


Fig. 4.3. Gráfico de la distribución normal para representar las áreas de rechazo y no rechazo (aceptación) en el contraste de hipótesis estadísticas, empleando como guía para los intervalos a formar los valores k_1 y k_2, que representan valores de percentiles de la distribución teórica normal.

Segundo momento

1. Cálculo relacionado con el estadístico de prueba, su comparación con el teórico-tabulado e interpretación de la significación de su resultado para el contraste de hipótesis empleando:
   a) Primer ejemplo: con H_0: $\mu_1 = \mu_2$ o $\mu_1 - \mu_2 = 0$, para dos colas, o sea, bilateral, se emplean los percentiles (Z teórico $1 - \alpha/2$ si Z observado > 0):
      - Intervalo del estadístico de prueba teórico en su distribución de probabilidades donde no se rechaza H_0: de $Z_{\alpha/2}$ a $Z_{1-\alpha/2}$), correspondiendo en el figura 4.3 a: $Z_{\alpha/2}$ o k_1 y $Z_{1-\alpha/2}$ o k_2, respectivamente.
      - Intervalo del estadístico de prueba teórico en su distribución de probabilidades donde se rechaza H_0 (región critica): de $-\infty$ a $Z_{\alpha/2}$ y de $Z_{1-\alpha/2}$ a $+\infty$, en el caso de la figura 4.3 para Z ~ N (0,1), así mismo, de $-\infty$ a t-student $_{n-1}$ a $\alpha/2$ y de t-student $_{n-1,\,1-\alpha/2}$ a $+\infty$ (Fig. 4.4).

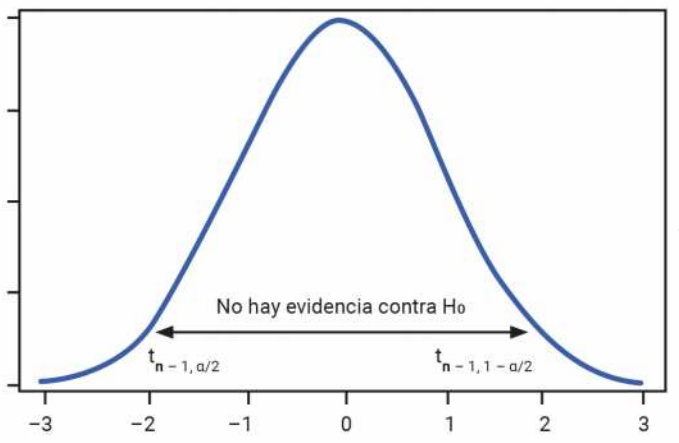


Fig. 4.4. Gráfico de la distribución teórica de probabilidades de la t-student para representar el intervalo de dos valores de T teóricos empleados en el contraste de hipótesis, donde no se rechazaría H_0 si se incluyese en el mismo, el valor de t-student observado.

b) Segundo ejemplo:
   - Cuando H_0: $\mu_1 > \mu_2$, una cola, unilateral, es decir, contraste unilateral derecho:
     - Intervalo del estadístico de prueba teórico en su distribución de probabilidades donde no se rechaza H_0: $(-\infty, Z_{1-\alpha})$ con $Z \sim N(0, 1)$, correspondiendo $(Z_{1-\alpha}$ a $k_2)$ en la figura 4.3 o para $(-\infty$, t-student $_{n-1,\,1-\alpha})$ (Fig. 4.4).
     - Intervalo del estadístico de prueba teórico en su distribución de probabilidades donde se rechaza H_0 (región critica): $(Z_{1-\alpha,}\ +\infty)$ (Fig. 4.3) o para (t-student $_{n-1,\,1-\alpha}$, $+\infty$) (Fig. 4.5).
   - Cuando H_0: $\mu_1 < \mu_2$, para una cola, es decir, contraste unilateral izquierdo:
     - Intervalo del estadístico de prueba teórico en su distribución de probabilidades donde no se rechaza H_0: $(+\infty, -Z_\alpha)$, correspondiendo $(-Z_\alpha$ a $k_1)$ (Fig. 4.3) o para $(+\infty, -t_{n-1,\,\alpha})$ (Fig. 4.6).
     - Intervalo del estadístico de prueba teórico en su distribución de probabilidades donde se rechaza H0 (región critica): $(-Z_\alpha, -\infty)$. correspondiendo $(-Z_\alpha$ a $k_1)$ (Fig. 4.3) o para $(-t_{n-1;\,\alpha,\,-\infty})$ (Fig. 4.6).

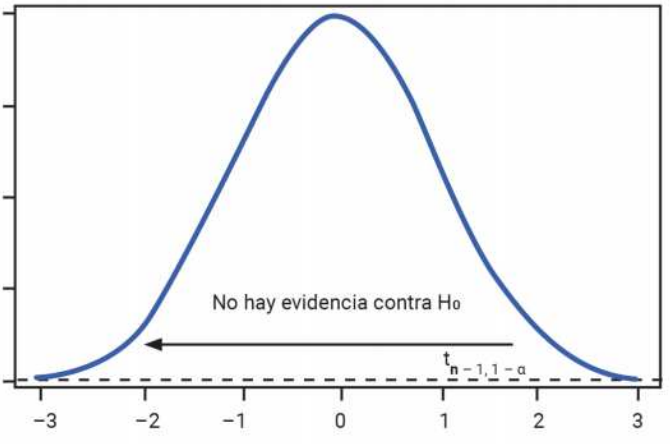


Fig. 4.5. Gráfico de la distribución teórica de probabilidades de la t-student para representar el intervalo de valores de t-student teórico en el contraste de hipótesis; donde se rechazaría H_0 para t-student observado > t-student $_{n-1,\,1-\alpha}$ y donde no, para t-student observado ≤ t-student $_{n-1,\,1-\alpha}$.

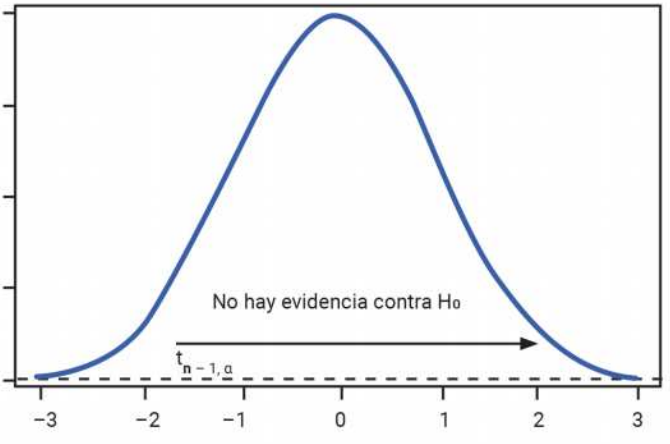


Fig. 4.6. Gráfico de la distribución teórica de probabilidades de la t-student para representar el intervalo de valores de t-student teórico en el contraste de hipótesis, donde se rechazaría H_0 para t-student observado < t-student $_{n-1,\,\alpha}$ y donde no, para t-student observado ≥ t-student $_{n-1,\,\alpha}$.

De forma general para recordar

Empleo de las respectivas áreas bajo la curva de rechazo para H_0, según los valores observados-calculados del estadístico de contraste y el tipo de contraste de hipótesis realizada:

- Para contraste bilateral:
  - Si Z o t-student (observado) ≥ 0, entonces emplear Z o t-student teório-tabulado-esperado para 1 – $\alpha/2$.
  - Si Z o t-student (observado) < 0, entonces emplear Z o t-student teório-tabulado-esperado para $\alpha/2$.
- Para contraste unilateral:
  - Si Z o t-student (observado) ≥ 0, entonces emplear Z o t-student teório-tabulado-esperado para 1 – α.
  - Si Z o t-student (observado) < 0, entonces emplear Z o t-student teório-tabulado-esperado para α.

Regla de decisión (primer ejemplo)

Si el valor modular de Z calculado-observado > Z esperado-tabulado, se ubica en el área de rechazo para H_0 del gráfico, por lo cual se rechaza H_0 y se acepta H_1, es decir, que hay diferencias estadísticamente significativas entre las medias poblacionales 1 y 2, con una probabilidad igual o mayor de que se exceda el valor observado de Z (p-valor) bajo H_0 o una probabilidad de que el resultado esté relacionado con el azar (aleatoriedad), inferior al nivel de significación alfa, o sea, de cometer el error tipo I ($p < 0,05$, para $\alpha = 0,05$).

En caso contrario, que no se puede rechazar H_0, se concluye que las medias de ambas poblaciones no son significativamente diferentes y su diferencia estaría fundamentalmente explicada por el efecto del error de muestreo ($p \geq 0,05$, para $\alpha = 0,05$).

Regla de decisión (segundo ejemplo)

Si el valor positivo de Z (+Zobservado-calculado) > (+Z esperado-tabulado), entonces se rechaza H0 y se acepta H1, es decir, la media de la población 1 es mayor que la de la población 2 de forma estadísticamente significativa, con una probabilidad de cumplirse H_0 (p-valor) para los datos empleados en la muestra, inferior al nivel de significación alfa o de cometer del error tipo I ($p < 0,05$, para $\alpha = 0,05$) en un nivel de confianza del 95 %. En caso contrario no se puede rechazar H_0 y, por ende, se concluye que la media de la población 1 no es significativamente mayor que la de la población 2 y sus diferencias están relacionadas con el efecto del error de muestreo ($p \geq 0,05$).

Si el valor negativo de Z (–Z (observado-calculado)) < (–Z esperado-tabulado), entonces se rechaza H0 y se acepta H_1, es decir, que la media de la población 1 es menor que la de la población 2 de forma estadísticamente significativa, con una probabilidad de cumplirse H_0 (p-valor) para los datos empleados en la muestra, inferior al nivel de significación alfa o de cometer del error tipo I ($p < 0,05$, para $\alpha = 0,05$) en un nivel de confianza del 95 %. En caso contrario no se puede rechazar H_0 y por ende, se concluye que la media de la población 1 no es significativamente mayor que la de la población 2 y sus diferencias están relacionadas con el efecto del error de muestreo ($p \geq 0,05$).

A modo de conclusión en la interpretación del contraste de hipótesis se debe recordar:

– La regla de decisión a partir del planteamiento de la hipótesis alternativa (Tabla 4.3).

Tabla 4.3. Relación entre el tipo de contraste (según hipótesis alternativa) y la decisión a tomar respecto a H_0, según el resultado de la comparación del estadístico de contraste observado versus el teórico.

Hipótesis alternativa	Se rechaza H_0 si
Bilateral (a ≠ b)	Estadístico de prueba observado > Percentil teórico para $1 - \alpha/2$ o |Estadístico de prueba observado| > Percentil teórico para $\alpha/2$
Unilateral (a > b)	Estadístico de prueba observado > Percentil teórico para $1 - \alpha$
Unilateral (a < b)	|Estadístico de prueba observado| > Percentil teórico para α

– Valor de p (valor de probabilidad) asociado al valor del estadístico de contraste observado-calculado: es la probabilidad de obtener en una muestra el resultado obtenido o más extremo que el mismo, bajo el cumplimiento de H_0, es decir, de que el valor del estadístico de prueba teórico-esperado fuese mayor o igual (en término modular) que el valor experimental observado-calculado a partir de los datos muestrales (que estimarían los poblacionales), esto se corresponde también en términos prácticos con la probabilidad de que no se cumpliera la hipótesis alternativa. También puede expresar el p-valor, la probabilidad que tendría una región crítica de comenzar exactamente en el valor del estadístico experimental obtenido de la muestra o la probabilidad de tener una muestra que discrepe aún más que la empleada para H_0, lo que equivale a la probabilidad de que por puro azar se obtenga una muestra más "extraña" que la empleada para este resultado. En este caso se compara el valor de p (p-valor) con el valor de α según el nivel de confianza preseleccionado por el investigador y el tipo de contraste (bilateral o unilateral que se trate), en caso de que:

- El p-valor (bajo H_0) < α, entonces hay evidencias para el rechazo de H_0, dado que la probabilidad estimada en una muestra de estudio del resultado obtenido o más extremo que el mismo si se cumpliera H_0, sería muy baja (menor que alfa), es decir, la probabilidad estimada de cumplirse H0 es menor que la probabilidad de equivocación al rechazar H_0 por ser cierta la misma (del error de tipo I) o que las probabilidades estimadas del azar o por el error de muestreo para el resultado obtenido serían muy bajas y como tal, ello es estadísticamente significativo (Fig. 4.7).

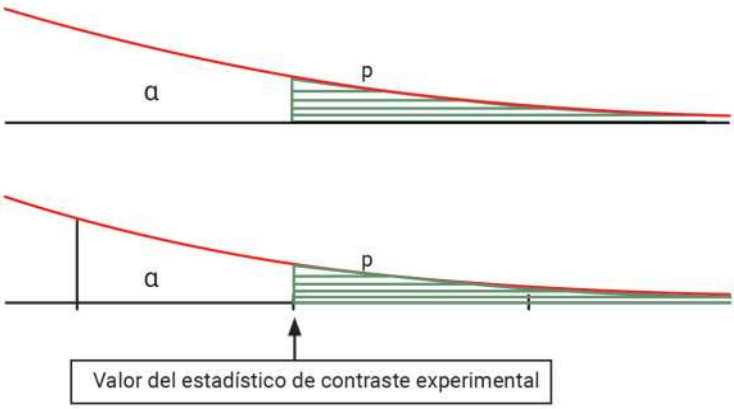


Fig. 4.7. Representación del área correspondiente al p-valor (p) y el nivel de significación alfa (α), en el caso del rechazo de H_0.

- El p-valor informa sobre el grado de compatibilidad de los datos con la hipótesis nula. Así, cuando el p-valor es grande, por ejemplo, 25 %, se entiende que la muestra no proporciona argumentos para dudar de la hipótesis nula. Cuando el p-valor es muy pequeño puede interpretarse como un indicador de incompatibilidad entre la hipótesis nula y los datos observados, pues está diciendo que si la hipótesis nula fuera cierta, es muy raro obtener unos datos como los que obtuvimos en la población de la cual se extrae la muestra.

Aunque a todos les gustan las reglas claras y sencillas, no es sensato tomar un p-valor como frontera universal y aplicarlo siempre con independencia del contexto en que esté situado. Fijar un valor frontera es lo mismo que decidir la probabilidad de equivocación si se rechaza la hipótesis nula y la probabilidad de error que es razonable asumir depende, sin duda, de la situación en la que se esté y de las consecuencias de cometer el error.

Una diferencia altamente significativa puede ser irrelevante a efectos prácticos y también puede ocurrir que una diferencia importante no sea significativa, por tanto, hay que andar con cuidado con las diferencias significativas, porque el hecho de que lo sean no implica que también sean importantes. Y no hay que confundir diferencia

no significativa con no existencia de diferencias. Es posible que el pequeño tamaño de muestra o la variabilidad que existe en los datos no deje ver una diferencia que sí existe y que además puede ser importante.

En la figura 4.8.se presenta una forma algorítmica de cómo se estimaría el valor de p (p-valor) en los diferentes tipos de contraste de hipótesis según el número y posición de las colas, así como el valor observado del estadístico de contraste o prueba a emplear:

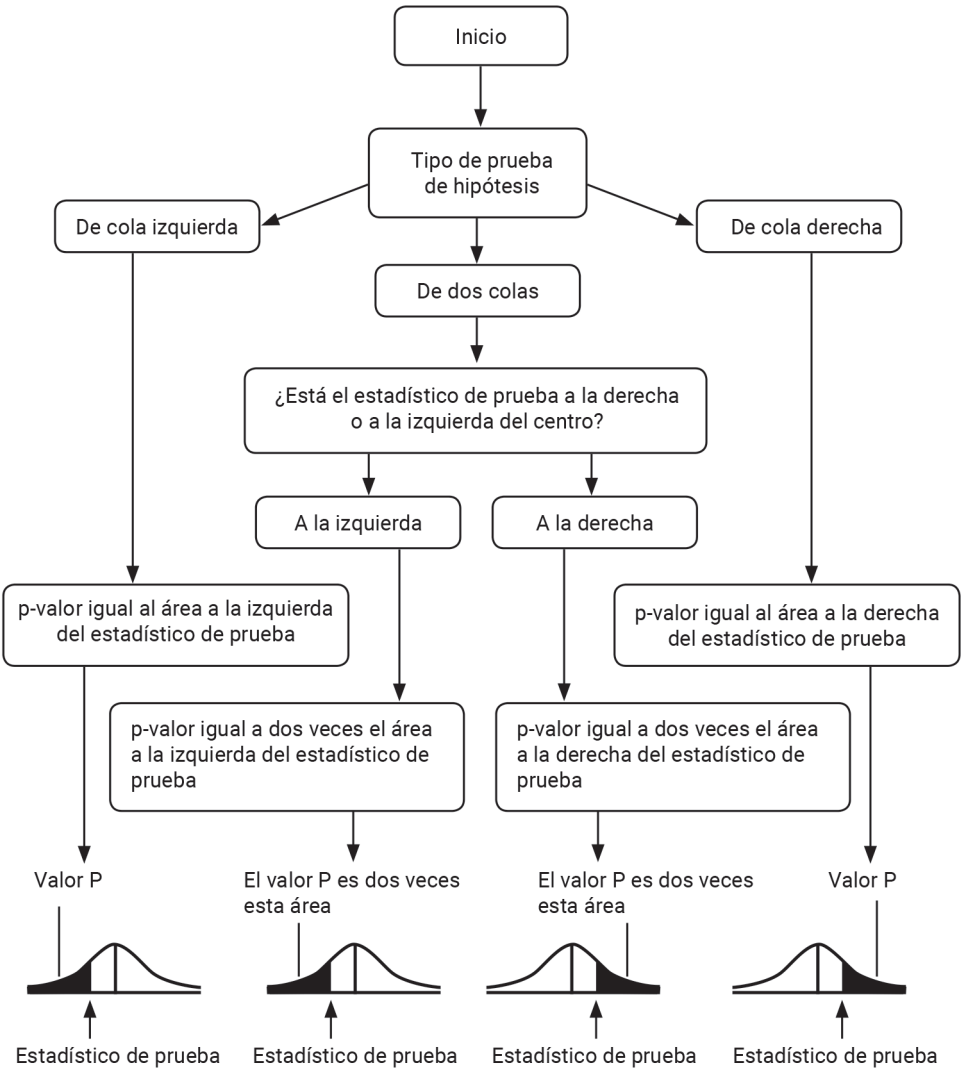


Fig. 4.8. Aspectos algorítmicos relacionados con la estimación del valor de p (p-valor) en los contrastes de hipótesis según el número y posición de colas y el resultado del estadístico de prueba a emplear en el método frecuentista.

Contraste de hipótesis estadísticas como parte del procesamiento estadístico de la información

A continuación se relacionan los pasos a seguir en el desarrollo de un proyecto investigativo aplicando e contraste de hipótesis estadísticas como parte del procesamiento estadístico de la información:

- Se parte de la observación de un fenómeno y la identificación del problema de investigación, el cual es caracterizado y se deben establecer hipótesis en relación a este problema y con vistas a comprobarlas se deben trazar los objetivos correspondientes.
- Se elabora un diseño de estudio específico para dar respuesta al problema de investigación que incluya el material y método a seguir para poder cumplir los objetivos correspondientes.
- Se conforma una base de datos, como parte del material en el que se debe apoyar el investigador, que incluya toda la información requerida en cuanto a las variables con su operacionalización involucradas en este fenómeno observado y las muestras necesarias, como eslabones indispensables para el procesamiento estadístico de los datos.
- Apoyados en el material a emplear y como parte del método seleccionado a seguir se pasa a realizar el procesamiento estadístico de la información que conlleva la vertiente descriptiva e inferencial, en lo cual se plantean las hipótesis estadísticas de trabajo.
- Se realiza la descripción del as variables en las muestras empleadas mediante las medidas de resumen adecuadas a los objetivos.
- Se seleccionan los estadígrafos de interés según los objetivos del estudio, que son algunas de las medidas de resumen de las variables o estadísticos elaborados a partir de la transformación de estos datos, que se comportan como estimadores muestrales de los respectivos parámetros poblacionales para métodos paramétricos o no para métodos no paramétricos.
- Se determina el modelo de distribución de frecuencias específica de los datos observados en las variables empleando los respectivos estimadores muestrales seleccionados para el número y tipo de variables; así como el número, tamaño y el tipo de muestras necesarios, como una forma de aproximación para evaluar las posibilidades a tomar en cuenta de sus valores, haciendo uso de la teoría de probabilidades.
- Una vez determinado el tipo de distribución de frecuencias de los datos observados de las variables empleando los respectivos estimadores muestrales o estadísticos elaborados, se seleccionan según los objetivos en el procesamiento estadístico las combinaciones en expresiones o formulaciones matemáticas de estos estimadores muestrales o estadísticos que expresan el valor observado-calculado del estadístico de contraste específico respectivo a emplear en las pruebas de hipótesis como parte de las opciones estadísticas, en este caso por el método paramétrico, mediante el contraste de hipótesis a efectuar y siguiendo el manejo de muestras y variables en

la inferencia bioestadística, con el uso del paradigma frecuentista en este caso, aunque también se puede emplear el enfoque bayesiano (véase el capítulo 5).

- Paralelamente se determina el valor teórico-esperado o crítico para este estadístico de prueba o contraste, según el nivel de confianza, que es lo que espera que ocurra el investigador o H_1 y los grados de libertad si fuese necesario por el tipo de estadístico empleado para los datos observados o experimentales.
- Se comparan los valores observado-calculado del estadístico de prueba seleccionado *versus* su valor teórico-esperado o crítico que queda contenido en las tablas de referencia para la distribución teórica de probabilidades según los posibles valores numéricos de este estadístico, las que pueden ser en el anexo 1 para la mayoría de las pruebas de hipótesis paramétricas, así como se pueden encontrar las tablas de valores críticos para estadígrafos esperados, empleadas en la mayoría de las pruebas de hipótesis no paramétricas.
- Determinar el valor de probabilidad (p-valor) igual o más extremo de que se cumpla la hipótesis nula para los datos observados de las variables en las muestras de estudio y compararlos con el respectivo nivel de significación (alfa) según la prueba de hipótesis seleccionada para una cola (unilateral) o dos colas (bilateral) de probabilidades.
- Realizar la toma de decisiones respecto al rechace o no de la hipótesis nula como parte del análisis de los resultados estadísticos lo que facilita responder a los objetivos del estudio a partir de su diseño metodológico.
- Las pruebas de hipótesis paramétricas por medio del intervalo de confianza, consisten, básicamente, en calcular un intervalo de confianza para el parámetro que se desea contrastar, puede ser un valor único o la diferencia entre dos valores, empleando un determinado nivel de confianza seleccionado por el investigador. La regla de decisión se basa en si el intervalo observado-calculado con los datos de las muestrales contiene o no al valor puntual del parámetro hipotetizado bajo la hipótesis nula H_0, por ejemplo, para un contraste de una diferencia de dos medias igual a 0 bajo H_0, si el valor 0 no se pueda ubicar entre los valores numéricos que comprende este intervalo estimado, entonces como decisión se tiene que se puede rechazar la H_0 y se acepta la hipótesis alternativa H_1 para la existencia estadísticamente significativa de la diferencia de dos medias poblacionales estimadas de las dos muestrales, para el nivel de confianza seleccionado por el investigador.

Relación entre las distribuciones teóricas de probabilidades de variables aleatorias conocidas y los elementos presentes en los contrastes de hipótesis en la Inferencia estadística

El empleo de los aspectos conceptuales fundamentales de la teoría de probablidades (véase el capítulo 3) resulta de gran importancia y encuentra una aplicación muy precisa en la inferencia bioestadística.

La pretensión de modelar lo observable, es decir, de poder hallar y establecer una regla en la ocurrencia de un fenómeno observable ha constituido siempre una necesidad básica para el científico empírico, dado que a través de esas construcciones teóricas (los modelos), puede experimentar sobre lo que la realidad no le permite. Los especialistas en cibernética-matemática han intentado construir modelos, por ejemplo, de las distribuciones de probabilidad que puedan representar el comportamiento teórico de diferentes fenómenos aleatorios que aparecen en el mundo real.

La experimentación con diferentes tipos de funciones matemáticas, que expresan la relación entre un conjunto de posibles valores numéricos teóricos ordenados y la frecuencia o probabilidad correspondiente acumulada que cada uno representan dentro del total de ellos, ha conllevado, mediante la sustitución de estos últimos, en esta funciones matemáticas correspondientes a la formación de pares ordenados que se pueden representar mediante un eje de coordenadas cartesianas x, y, donde al unir esos pares se logra obtener un gráfico o curva de distribución teórica-esperada de probabilidades o frecuencias, lo que puede ser presentado también en forma tabular.

Ante la necesidad de comparar, correlacionar, asociar, predecir, entre otros, el comportamiento de las variables de las muestras de estudio a partir de sus datos recogidos y con el desarrollo alcanzado en la bioestadística se ha logrado combinar algunos parámetros estimados de los muestrales para construir nuevos estadísticos expresados mediante formulaciones matemáticas, lo que se emplea en las pruebas de hipótesis para caracterizar o ser expresión directa de aspectos como la diferencia entre parámetros, la asociación entre variables, la homogeneidad en las muestras, la correlación o regresión entre variables, entre otras.

El conjunto de los posibles valores numéricos a obtener variando los valores de los paramétros estimados en esas formulaciones matemáticas y ser sustituidos en funciones matemáticas específicas desarrolladas por los especialistas en cibernética-matemática han permitido estimar la probabilidad que representa cada valor y con esto formar pares ordenados que se pueden ubicar mediante un sistema de ejes de coordenadas cartesianos x, y para representar la curva gráfica teórica de distribución de probabilidad o frecuencia para este estadístico. Basados en esto se ha podido asumir ciertos tipos específicos de distribución teórica de probabilidad para las formulaciones matemáticas que expresen los estadísticos de prueba o contraste específicos a emplear en los diferentes contrastes de hipótesis o pruebas de significación en la inferencia estadística, teniendo como referencia los valores téoricos-tabulados esperados de las funciones matemáticas téoricas. De ahí derivan los conceptos de estadísticos de prueba o contraste experimental-observado-calculado que expresan los resultados obtenidos a partir de las sustituciones en estos estadísticos de los valores de los estimadores u otros estadísticos muestrales elaborados y paralelamente el concepto de estadístico de prueba o contraste esperado-teórico-tabulado o de referencia, que no es más que la expresión de la combinación teórica de los valores numéricos ordenados y el resultado de su evaluación o sustitución en una función matemática cuyo gráfico de distribución

de probabilidades resultante es asumido como de referencia, para el estadístico de contraste experimental-observado según los datos muestrales, durante el proceso inferencial de las pruebas de significación.

En el procesamiento estadístico de la información, una vez ejecutada la perspectiva descriptiva de los datos recogidos de las variables en las muestras de estudio puede llevar a la necesidad de realizar un proceder analítico inmediato como parte de la perspectiva Inferencial de la bioestadística para esta información. En la gran mayoría de los diseños de estudio analíticos desde la fase de planificación del método a seguir para el procesamiento estadístico de la información, el investigador prevé que de ocurrir lo antes mencionado, se selecciona un método que le permita evaluar la significación estadística en los resultados obtenidos para los diferentes parámetros poblacionales estimados en un nivel de confianza seleccionado por el mismo. En este momento es cuando se debe prestar toda la atención a las posibles pruebas de significación o del contraste de hipótesis que se pudieran seleccionar para cada caso específico. La ruta a seguir para poder efectuar las pruebas de significación o de contraste de hipótesis en la inferencia estadística si empleamos un enfoque frecuentista, como parte medular del procesamiento estadístico inferencial de la información recogida en relación a las muestras y variables empleadas, comienza por tener un acercamiento a el enunciado de los objetivos que faciliten dar una respuesta a las hipótesis emitidas en relación con el problema de investigación identificado.

Cada objetivo que lleve implícito la necesidad de realizar un proceder analítico de tipo comparativo, asociativo, predictivo, de tendencia entre variables o de homogeneidad en el comportamiento de las muestras, entre otros, se puede hacer corresponder con algunos tipos específicos de contraste de hipótesis estadísticas: nula (H_0) *versus* alternativa (H_1). A su vez estos contrastes de hipótesis requieren de la representación en formulaciones matemáticas que expresan una combinación de los parámetros estimados e involucrados en el problema de investigación y que se denominan *estadísticos de prueba o contraste*.

Apuntes imprescindibles acerca de los pasos a seguir al realizar las pruebas de significación

- El investigador selecciona los tipos de pruebas de significación a emplear, las cuales se deben corresponder con o facilitar el cumplimiento del objetivo específico de tipo analítico enunciado.
- Seleccionar el tipo de combinación de muestras y variables requerido para poder efectuar la prueba de significación seleccionada.
- Seleccionar los tipos de medidas de resumen a emplear según el tipo de variable cuantitativas o cualitativas necesarios para efectuar las pruebas de significación seleccionada.
- Determinar el tipo de distribución de los datos obtenidos para los tipos de medidas de las variables a emplear en esas pruebas de significación, para el caso de las varia-

bles cuantitativas que siguen una distribución normal o se pueden aproximar a esta, elegir pruebas de significación paramétricas, en caso contrario seleccionar pruebas de significación no paramétricas.
- Seleccionar de ser posible, los parámetros estimados de los muestrales relacionados con dichas pruebas de significación para el caso de las paramétricas o el estadístico específico elaborado por la transformación de los datos de las variables en el caso de las pruebas de significación no paramétricas.
- Llevar a cabo el planteamiento del contraste de hipótesis estadísticas específico: nula *versus* alternativa, para cada prueba de significación paramétrica o no paramétrica seleccionada, con el respectivo nivel de confianza seleccionado por el investigador para el resultado a obtener.

Para el caso de pruebas paramétricas (emplean parámetros estimados)
- Se seleccionan los tipos de estadísticos de prueba o contraste específicos para las pruebas de significación, los cuales expresan una combinación de los parámetros estimados de los muestrales mediante una relación de igualdad o fórmula matemática. Se sustituyen los valores de los parámetros estimados a utilizar en esta fórmula matemática y se calculan los valores experimentales observados-calculados de dichos estadísticos, así como conjuntamente se calcula el valor de probabilidad (p-valor) para el que se cumple la hipótesis nula para este resultado del estadístico de contraste. Como ejemplos de estadísticos de contraste a calcular (experimental-observado) están: Z~N (0, 1), t-student, F-Snedecor, entre los más empleados.
  La distribución ordenada de los posibles valores numéricos teóricos-esperados de cada uno de los estadísticos de contraste o de prueba antes mencionados, con la respectiva probabilidad acumulada que cada uno de estos representan en esa distribución se encuentran tabulados (véase el anexo 1).
- Se compara el valor del estadístico de contraste o de prueba experimental observado-calculado versus el valor esperado-teórico de referencia según el nivel de confianza seleccionado y se estima el valor de probabilidad bajo H_0 (p-valor) correspondiente con el valor experimental observado-calculado para compararlo con el nivel de significación alfa, según el tipo de contraste de hipótesis bilateral o unilateral seleccionado, para poder tomar la decisión de rechazar o no a H_0.

Para el caso de pruebas no paramétricas (no emplean parámetros estimados)
- Se seleccionan los tipos de estadísticos de prueba o contraste específicos para dichas pruebas de significación, los que se calculan a partir de la transformación de los valores de las variables en rangos ordenables o del conteo de signos para diferencias de los valores en las variables seleccionadas según sea el caso de la prueba de significación específica escogida, pero sin poder emplear parámetros estimados. Algunos estadísticos de contraste requieren para obtener su valor experimental-observado-calculado del empleo de fórmulas matemáticas.

– Los posibles valores numéricos críticos (teóricos-esperados) distribuidos según el nivel de significación alfa y el tamaño muestral para cada uno de los estadísticos de contraste o prueba seleccionados se encuentran tabulados (véase el anexo 1).
– Se compara el valor experimental observado-calculado de estos estadísticos de contraste o prueba versus el valor referencial teórico-esperado del mismo, según el nivel de confianza seleccionado, así como conjuntamente se calcula el valor de probabilidad (p-valor) para el cual se cumple la hipótesis nula para este resultado del estadístico de contraste, preferiblemente cuando el tamaño de muestra empleado es pequeño. Para el caso del tamaño de muestra grande se puede emplear la aproximación de la distribución normal y emplear su estadístico Z en fórmulas matemáticas específicas que involucran estadísticos o estadígrafos muestrales en las distintas pruebas no paramétricas y para lo cual en el análisis del contraste de hipótesis estadísticas se aplica el mismo procedimiento explicado para las pruebas paramétricas que utilizaban precisamente este estadístico de contraste. Como ejemplos de estadísticos de contraste a calcular (experimental-observado) están: U de Mann Whitney, T+ de Wilcoxon, K de Kruskal Wallis, Q de Friedman entre los más empleados.

El método frecuentista emplea la prueba de significación en la evaluación del resultado estimado, englobando un índice para medir la fuerza de la evidencia llamado valor p o p-valor y un procedimiento de elección de hipótesis, denominado prueba de hipótesis para este resultado, se pronuncia sobre los datos a partir de supuestos (distribuciones de probabilidades), o sea, expresa como la probabilidad de obtener ciertos datos supuesta la veracidad de H_0 [$p = Pr (D \geq d_0 \mid H_0$]. El método frecuentista obliga a una decisión dicotómica (rechazo o no) y sus resultados dependen crucialmente de un elemento exógeno a los datos, que es el tamaño de muestra.

En la figura 4.9. se representan de forma resumida y esquematizada los aspectos que vinculan el objetivo investigativo y la inferencia estadística expresada mediante las pruebas de significación por el método o enfoque frecuentista.

Cuando no se puede asumir la normalidad en la distribución de frecuencias para los datos de una variable y se desea utilizar una prueba de significación como parte de la inferencia estadística que exija la normalidad como supuesto, se pueden realizar dos procedimientos para enfrentar esta situación:

– El primero de ellos es intentar una transformación logarítmica de la variable que suele conseguir hacer normal la distribución de frecuencias de esa variable.
– El segundo es emplear una prueba de significación por método no paramétrico y comparar los resultados con los de una prueba de significación por el método paramétrico para el mismo objetivo y empleando la misma variable. En caso de no hallarse diferencias entre ambos resultados respecto a la significación estadística

para ambos métodos, se puede asumir la prueba de significación por el método paramétrico, pese a no cumplir estrictamente con la normalidad en la distribución de frecuencias los datos de la variable estudiada.

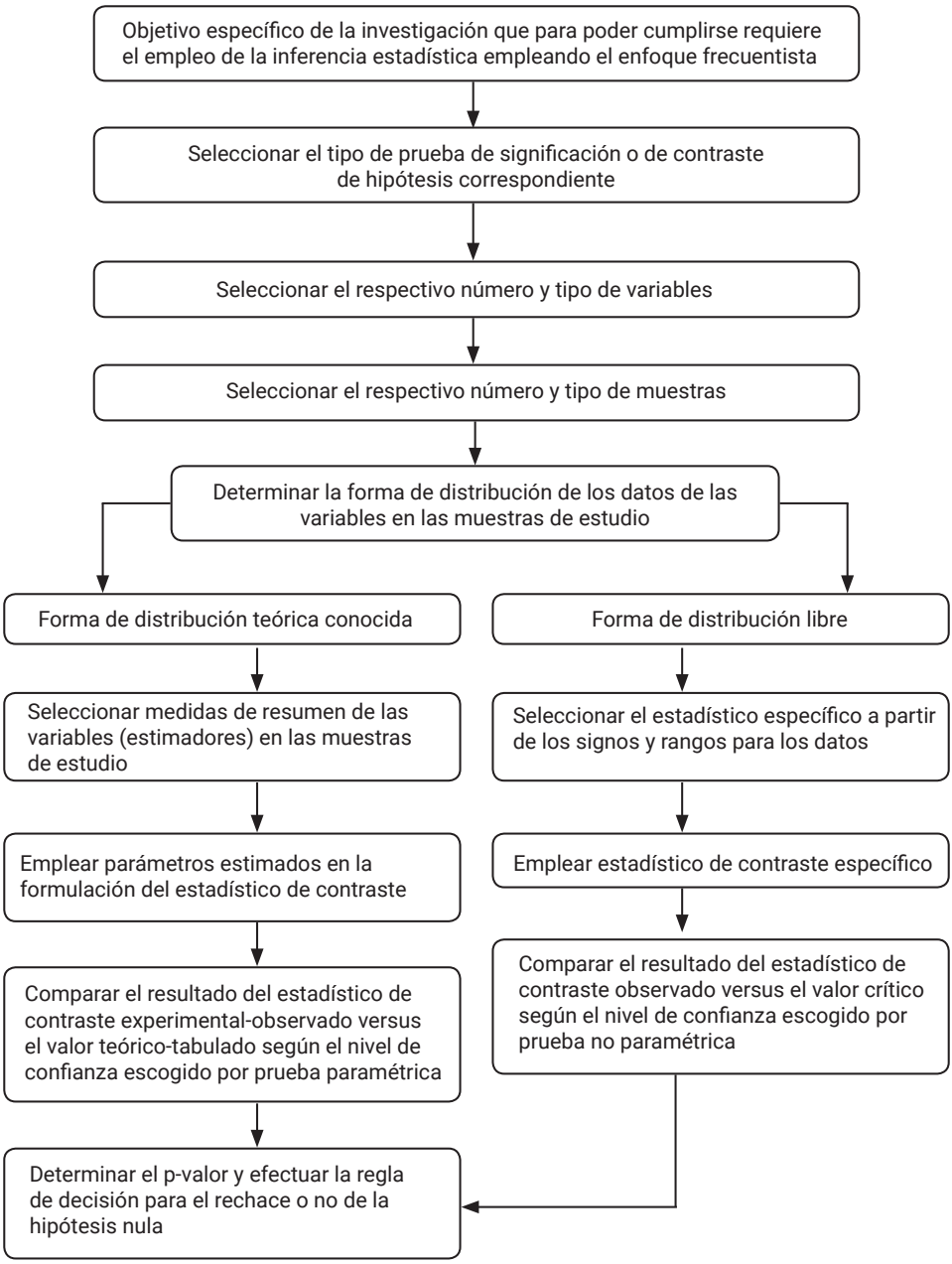


Fig. 4.9. Relación entre el objetivo investigativo y la inferencia estadística con el empleo de pruebas de significación y mediante el enfoque frecuentista.

Algunos ejemplos de planteamientos de contraste de hipótesis con el uso de los estadísticos de prueba o de contraste más frecuentemente empleados en las respectivas opciones de las pruebas de hipótesis en la Inferencia bioestadística mediante el método frecuentista y su relación con la adecuada combinación de muestras y variables, como parte del procesamiento estadístico de la información que faciliten emitir una respuesta ante diferentes objetivos investigativos:

1. Para pruebas de hipótesis paramétricas:
   a) Una muestra con una variable cuantitativa que sigue o se aproxima a una distribución de frecuencias normal estándar, si se desea comparar una media poblacional estimada a partir de una muestral *versus* una media poblacional conocida, un valor preestablecido, de referencia, hipotético o definido por un procedimiento teórico, un posible planteamiento de hipótesis estadísticas para el contraste sería:
      - Contraste bilateral:
        H_0: $\mu = \mu_0$ *versus* H_1: $\mu \neq \mu_0$ o H_0: $\mu - \mu_0 = 0$ *versus* H_1: $\mu - \mu_0 \neq 0$.
      - Contraste unilateral:
        - H_0: $\mu = \mu_0$ *versus* H_1: $\mu > \mu_0$ o H_0: $\mu = \mu_0$ *versus* H_1: $\mu < \mu_0$.
        - H_0: $\mu \geq \mu_0$ *versus* H_1: $\mu < \mu_0$ o H_0: $\mu \leq \mu_0$ *versus* H_1: $\mu > \mu_0$.

          Si se conoce la varianza poblacional (σ^2) y n ≥ 30, el estadístico de prueba es Z~ N (0, 1) para un nivel de significación alfa, adoptando la expresión:

          $$Z(\text{observado-calculado}) = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

          donde:

          μ: es la media poblacional conocida, un valor preestablecido, de referencia, hipotético o definido por un procedimiento teórico para contrastar con la media poblacional estimada de la muestral;

          σ: es la desviación poblacional conocida de referencia.

        Para recordar:

        A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste: siendo: T_i el valor inferior del estadístico de prueba o contraste y T_s el valor superior es el estadístico empleado para el contraste de hipótesis, entonces:
      - Para contraste bilateral:
        - Intervalo donde no se rechaza $H_0 = (T_i, T_s)$

          Región crítica = $(-\infty, T_i] \cup [T_s, +\infty)$ donde se rechaza H_0
      - Para contraste unilateral:

- Intervalo donde no se rechaza $H_0 = (T_i, +\infty)$
  Región crítica = $(-\infty, T_i)$ donde se rechaza H_0
- Intervalo donde no se rechaza $H_0 = (-\infty, TS)$
  Región crítica = $(T_{S, +\infty})$ donde se rechaza H_0

– Valor de p (p-valor), bajo H_0:
  p-valor = [Pr (Z (observado-calculado)) ≤ Z (teórico-esperado) según el valor de alfa)].

Regla de decisión

Se compara Z (observado-calculado) *versus* Z (teórico-esperado) según valor de alfa para el nivel de confianza seleccionado y si el contraste es unilateral o bilateral:

– Para Z (observado-calculado) < 0:
  - Si Z (observado-calculado) < Z (teórico-esperado) según valor α, entonces Z (observado-calculado) se ubica en la región de rechazo para H_0, disminuyendo la probabilidad de cometer el error tipo I, por lo que existe evidencia suficiente para rechazar hipótesis nula y se acepta la alternativa, por lo que las diferencias entre la media poblacional estimada (de la muestral) y la de referencia o hipotética son estadísticamente significativas, al estimarse un p-valor bajo H_0 para Z (observado-calculado) o una probabilidad relativa al azar para este resultado, inferior a la probabilidad alfa para el nivel de confianza seleccionado ($p < \alpha$).
  - Si Z (observado-calculado) ≥ Z (teórico-esperado) según valor de alfa, entonces no se puede rechazar H_0.

– Para Z (observado-calculado) ≥ 0:
  - Si Z (observado-calculado) > Z (teórico-esperado) según valor α, entonces Z (observado-calculado) se ubica en la región de rechazo para H_0, disminuyendo la probabilidad de cometer el error tipo I, por lo que existe evidencia suficiente para rechazar hipótesis nula y se acepta la alternativa, por lo que las diferencias entre la media poblacional estimada (de la muestral) y la de referencia o hipotética son estadísticamente significativas, al estimarse un p-valor bajo H_0 para Z (observado-calculado) o una probabilidad relativa al azar para este resultado, inferior a la probabilidad alfa para el nivel de confianza seleccionado ($p < \alpha$).
  - Si Z (observado-calculado) ≤ Z (teórico-esperado) según valor de alfa, entonces no se puede rechazar H_0.
    De forma tabulada se pueden resumir los elementos esenciales teóricos de este Contraste de Hipótesis (Tabla 4.4.):

Tabla 4.4. Resumen de la Dócima o Prueba de Hipótesis para una media poblacional.

Hipótesis	Estadígrafo	Región Crítica
H_0: $\mu = \mu_0$ contra H_1: $\mu \neq \mu_0$	$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$	$\lvert Z \rvert > Z_{1-\alpha/2}$
H_0: $\mu \leq \mu_0$ contra H_1: $\mu > \mu_0$		$Z > Z_{1-\alpha}$
H_0: $\mu \geq \mu_0$ contra H_1: $\mu < \mu_0$		$Z < -Z_{1-\alpha}$

b) Una muestra con una variable cuantitativa que sigue o se aproxima a una distribución de frecuencias: t-student, si se desconoce la varianza poblacional (σ^2) de referencia o n < 30, el estadístico de prueba es: t-student con (n – 1) grados de libertad y un nivel de significación alfa (α) prefijado, adoptando la siguiente expresión:

$$\text{t-student(observado-calculado)} = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$; con n – 1 grados de libertad y según el valor de alfa prefijado por el investigador.

donde:

μ: representa a la media poblacional conocida, un valor preestablecido, de referencia, hipotético o definido por un procedimiento teórico para contrastar con la media poblacional estimada de la muestral;

$\hat{S}$: denotado como S en el denominador de la fórmula es la cuasi desviación muestral típica que estimaría a la desviación estándar poblacional de referencia (σ) y se expresa por la fórmula:

$$\hat{S} = \sqrt{\frac{n}{n-1}} S$$

S: desviación estándar muestral o típica (no presente en el denominador de la fórmula para el estadístico de prueba t-student);

n: es el tamaño muestral.

Para recordar:

Los grados de libertad son el número de unidades independientes de información en una muestra, que son relevantes para la estimación de un parámetro o para el cálculo de un estadístico de contraste.

Para el caso de la varianza (S^2) presente en la fórmula del estadístico de prueba existe una restricción en el número de unidades independientes de información en la muestra, por lo que los grados de libertad se reducen a n – 1, y el número de unidades de información que se considera a efectos de la estimación de la varianza poblacional son n – 1. Si X~ N (μ, σ) y S es un estimador de σ calculado a partir de una muestra con n – 1 grados de libertad, entonces la variable (X – μ) /S se

distribuye según una t-student con n – 1 grados de libertad (los de S) según alfa prefijado.

Para recordar

A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste, siendo Ti el valor inferior del estadístico de prueba o contraste y Ts el valor superior de dicho estadístico empleado para el contraste de hipótesis, entonces:

- Para contraste bilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, T_s)$
    Región crítica = $(-\infty, T_i \cup [T_s, -\infty)$, donde se rechaza H_0.
- Para contraste unilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, +\infty)$
    Región crítica = $(-\infty, T_{i)}$, donde se rechaza H_0.
  - Intervalo donde no se rechaza $H_0 = (-\infty, T_S)$
    Región crítica = $(T_S, +\infty)$, donde se rechaza H_0.
- Valor de p (p-valor), bajo H_0:
  p-valor = Pr [t-student (observado-calculado) ≤ t-student (teórico-esperado) según el valor de alfa y n – 1 grados de libertad].

Regla de decisión

– Para t-student observado-calculado < 0:

- Si t-student observado-calculado < t-student (teórico-esperado) según los grados de libertad y valor alfa correspondientes, entonces t-student observado se ubica en la región de rechazo para $H_{0,}$ disminuyendo la probabilidad de cometer el error tipo 1, por lo que existe evidencia suficiente para rechazar hipótesis nula y se acepta la alternativa, por lo que las diferencias entre las media poblacionales estimada (de la muestral) y la de referencia o hipotética son estadísticamente significativas, al estimarse un p-valor bajo H_0 para t-student observado o una probabilidad relativa al azar para este resultado inferior a la probabilidad alfa para el nivel de confianza seleccionado ($p < \alpha$).
- Si t-student (observado-calculado) ≥ t-student (teórico-esperado), según los grados de libertad y valor de alfa correspondientes no se puede rechazar H_0.

– Para t-student (observado-calculado) ≥ 0:

- Si t-student (observado-calculado) > t-student (teórico-esperado), según los grados de libertad y valor alfa correspondientes; entonces t-student observado se ubica en la región de rechazo para H_0, disminuyendo la probabilidad de cometer el error tipo 1, por lo que existe evidencia suficiente para rechazar hipótesis nula y se acepta la alternativa, por lo que las diferencias entre las media poblacionales estimada (de la muestral) y la de referencia o

hipotética son estadísticamente significativas, al estimarse un p-valor bajo H_0 para t-student observado o una probabilidad relativa al azar para este resultado inferior a la probabilidad alfa para el nivel de confianza seleccionado ($p < \alpha$).

- Si t-student (observado-calculado) ≤ t-student (teórico-esperado), según los grados de libertad y valor alfa correspondientes, no se puede rechazar H_0.

De forma tabulada se pueden resumir los elementos esenciales teóricos de este contraste de hipótesis (Tabla 4.5).

Tabla 4.5. Resumen de la dócima o prueba de hipótesis para una media poblacional para muestras pequeñas y sin conocer la varianza poblacional

Hipótesis	Estadígrafo	Región Crítica
H_0: $\mu = \mu_0$ contra H_1: $\mu \neq \mu_0$	$t\text{-student} = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$	$\lvert t \rvert > t_{1-\alpha/2,\, n-1}$
H_0: $\mu \leq \mu_0$ contra H_1: $\mu > \mu_0$		$t > t_{1-\alpha/2,\, n-1}$
H_0: $\mu \geq \mu_0$ contra H_1: $\mu < \mu_0$		$t < -t_{1-\alpha/2,\, n-1}$

c) Una muestra con una variable cuantitativa que sigue o se aproxima a una distribución de frecuencias normal, si se desea comparar la varianza muestral (s^2) que estima una poblacional (σ^2) versus una varianza poblacional conocida (σ_0^2), un valor preestablecido, de referencia, hipotético o un valor definido por un procedimiento teórico. El posible planteamiento de hipótesis estadística para el contraste sería:

- Contraste bilateral:
  H_0: $(\sigma^2) = (\sigma_0^2)$ *vs.* H_1: $(\sigma^2) \neq (\sigma_0^2)$ o H_0: $(\sigma^2) / (\sigma_0^2) = 0$ *vs.* H_1: $(\sigma^2) / (\sigma_0^2) \neq 0$.
- Contraste unilateral:
  - H_0: $(\sigma^2) = (\sigma_0^2)$ *vs.* H_1: $(\sigma^2) > (\sigma_0^2)$ o H0: $(\sigma^2) = (\sigma_0^2)$ *vs.* H_1: $(\sigma^2) < (\sigma_0^2)$.
  - H_0: $(\sigma^2) > o = (\sigma_0^2)$ *vs.* H_1: $(\sigma^2) < (\sigma_0^2)$ o H_0: $(\sigma^2) < o = (\sigma_0^2)$ *vs.* H_1: $(\sigma^2) > (\sigma_0^2)$.
    Se emplea como estadístico de prueba o contraste chi-cuadrado para la varianza a partir del teorema de Cochran, que adopta la forma:
    χ^2 (observado-calculado) = $[(n - 1)\, \hat{S}^2] / (\sigma_0^2)$ que sigue una distribución teórica de probabilidades con n – 1 grados de libertad y para un nivel de significación alfa según el nivel de confianza preseleccionado.
    También sustituyendo la cuasivarianza muestral típica ($\hat{S}^2$) por su expresión que contiene a la desviación estándar muestral o típica (S) se obtiene:
    χ^2 (observado-calculado) = $(n - 1)\, [n / (n - 1)\, (S)\,] / (\sigma_0^2)$

donde:

$\hat{S}^2$: presente en el numerador de la fórmula del estadístico de contraste chi-cuadrado, es la cuasivarianza muestral típica y se calcula elevando al cuadrado la cuasidesviación muestral típica como sigue:

$$\hat{S} = \sqrt{n/(n-1)}S$$

S: desviación estándar muestral o típica.

Para recordar

Los grados de libertad son el número de unidades independientes de información en una muestra, que son relevantes para la estimación de un parámetro o para el cálculo de un estadístico de contraste. Para el caso del estadístico varianza, existe una restricción en el número de unidades independientes de observación en una muestra, por lo que los grados de libertad se reducen a n – 1, y el número de unidades de información que se considera a efectos de la estimación de la varianza poblacional son n – 1. El número de grados de libertad también forma parte de la descripción de algunas distribuciones de probabilidad, que toman esta característica de su relación con ciertos estadísticos. Para describir la variabilidad que presenta la varianza muestral S^2 se utilizala distribución chi-cuadrado, con n – 1 grados de libertad (los de S^2).

A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo T el valor del estadístico de prueba o contraste empleado (chi-cuadrado), entonces:

- Intervalo donde no se rechaza H_0 para cualquier valor de T donde:
  T observado ≤ T esperado
- Región crítica para cualquier valor de T donde:
  T observado > T esperado (donde se rechaza H_0)
- Valor de p (p-valor), bajo H_0:
  p-valor = Pr [chi-cuadrado (calculado-observado) ≤ t-student $_{n-1,\alpha}$ (teórico-esperado)].

Regla de decisión

Se compara chi-cuadrado observado-calculado *versus* chi-cuadrado teórico-esperado según los grados de libertad correspondientes y valor de alfa preseleccionado por el investigador:

- Si chi-cuadrado observado-calculado > chi-cuadrado teórico-esperado según sus grados de libertad y valor de alfa prefijado, esto se corresponde con la ubicación de chi-cuadrado calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que disminuye la probabilidad de cometer el error tipo 1, entonces hay evidencias suficientes para rechazar H_0 y por ende se acepta H_1 (la hipótesis alternativa), siendo, por

tanto, estadísticamente significativo el resultado obtenido en la comparación de varianzas al estimarse un p-valor bajo Ho para chi-cuadrado observado o una probabilidad relativa al azar inferior a la correspondiente con el nivel de significación alfa prefijado ($p < \alpha$).

- Si chi-cuadrado observado-calculado ≤ chi-cuadrado teórico-esperado según los grados de libertad y valor de alfa prefijado, esto no se corresponde con la ubicación de chi-cuadrado calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que aumenta la probabilidad de cometer el error tipo 1, entonces no hay evidencias suficientes para rechazar H_0 y, por ende, no se acepta H_1 (la hipótesis alternativa), siendo, por tanto, estadísticamente no significativo el resultado obtenido en la comparación de varianzas al estimarse un p-valor bajo H_0 para chi-cuadrado observado o una probabilidad relativa al azar igual o superior a la correspondiente con el nivel de significación alfa prefijado ($p \geq \alpha$).

d) Una muestra con una variable cualitativa que sigue una distribución de frecuencias binomial y que se aproxima a la normal estándar, si se desea comparar una proporción poblacional estimada (P), a partir de una muestral *versus* con una proporción poblacional conocida (P_0) de referencia, hipotética o definida por un procedimiento teórico para muestras preferiblemente grandes ($npq \geq 5$), donde se puede asumir que la variable X sigue una distribución de frecuencias binomial con aproximación de la normal estándar para un cierto nivel de significación alfa, el posible planteamiento de hipótesis estadísticas para el contraste sería:

- Contraste bilateral:
  $H_0: P = P_0$ *vs.* $H_1: P \neq P_0$
- Contraste unilateral:
  - $H_0: P = P_0$ *vs.* $H_1: P > P_0$ o $H_0: P = P_0$ *vs.* $H_1: P < P_0$.
  - $H_0: P > o = P_0$ *vs.* $H_1: P < P_0$ o $H_0: P < o = P_0$ *vs.* $H_1: P > P_0$.
- El estadígrafo de la prueba o contraste es Z para un nivel de confianza de $(1 - \alpha/2)$ % que adopta la forma:

  $Z = (p - pH_0) / \sigma p \sim N(0,1)$

  donde:

  $\sigma_p = \sqrt{\frac{pH_0qH_0}{n}}$ es la desviación poblacional para la proporción poblacional de la característica X conocida de referencia;

  pH_0qH_0: proporción conocida o de referencia poblacional (pH_0) multiplicada por su complemento (qH_0) = 1 – (pH_0) empleada para el contraste.

  P = X/n: proporción muestral de la característica x de la variable en estudio que estimaría una proporción poblacional.

Otra forma de plantear el cálculo del valor de Z utilizando otra nomenclatura empleada en diferentes textos, pero representando los mismos aspectos incluidos en la ecuación es:

$$Z(\text{observado-calculado}) = \frac{\hat{P} - P}{\sqrt{\frac{P \cdot (100 - P)}{n}}} \sim N(0, 1)$$

donde:

P: representa la proporción poblacional conocida de referencia para contrastar (en algunos textos se expresa como P_0);

$\hat{P}$: proporción poblacional estimada de la muestral (se utiliza el valor empírico para la proporción muestral).

Para recordar:

A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste (Z) siendo $T_i = (Z_{\alpha/2})$ el valor inferior del estadístico de prueba o contraste y $Ts = (Z_{1-\alpha/2})$ el valor superior de este estadístico empleado para el contraste de hipótesis, entonces:

- Para contraste bilateral:
  Intervalo donde no se rechaza $H_0 = (T_i, T_s)$
  Región crítica = $(-\infty, T_i] \cup [T_s, +\infty)$, donde se rechaza H_0.
- Para contraste unilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, +\infty)$:
    Región crítica = $(-\infty, T_i)$ donde se rechaza H_0.
  - Intervalo donde no se rechaza $H_0 = (-\infty, T_S)$:
    Región crítica = $(T_S, +\infty)$ donde se rechaza H_0.
- Valor de p (p-valor), bajo H_0 donde se cumpliría H_0:
  - Si contrastamos: $P \leq P_0$, entonces:
    p-valor = 2 (probabilidad para $Z_{\alpha/2}$) dos veces el área a la derecha de –Z, bajo la curva normal estándar N (0, 1).
  - Si contrastamos: $P > P_0$, entonces:
    p-valor = 2 (1 – probabilidad para $Z_{\alpha/2}$) dos veces el área a la izquierda de Z, bajo la curva normal estándar N (0, 1).

Regla de decisión

Se compara el valor de Z calculado (observado) *versus* el valor de Z teórico-tabulado para el valor de significación de α prefijado:

- Para Z (observado-calculado) < 0:
  - Si Z (observado-calculado) < Z (teórico-esperado) según el valor de alfa prefijado, esto se corresponde con la ubicación de Z calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo

teórico de referencia, lo que aumenta la probabilidad de cometer el error tipo 1, entonces hay evidencias suficientes para rechazar H_0 y, por ende, se acepta H_1 (la hipótesis alternativa), siendo estadísticamente significativo el resultado obtenido al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa a la azar inferior a la correspondiente con el valor de significación alfa prefijado ($p < \alpha$).

- Si Z (observado-calculado) ≥ Z (teórico-esperado) según el valor de alfa prefijado no se puede rechazar H_0.

– Para Z (observado-calculado) ≥ 0:

- Si Z (observado-calculado) > Z (teórico-esperado) según el valor de alfa prefijado, esto no se corresponde con la ubicación de Z calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que aumenta la probabilidad de cometer el error tipo 1, entonces no hay evidencias suficientes para rechazar H_0 y, por ende, no se acepta H_1 (la hipótesis alternativa), siendo estadísticamente significativo el resultado obtenido al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa a la azar inferior a la correspondiente con el valor de significación alfa prefijado ($p < \alpha$).
- Si Z (observado-calculado) ≤ Z (teórico-esperado) según el valor de alfa prefijado, no se puede rechazar H_0.

De forma tabulada se pueden resumir los elementos esenciales teóricos de este contraste de hipótesis (Tabla 4.6).

Tabla 4.6. Resumen de la dócima o prueba de hipótesis para una proporción poblacional por aproximación a la distribución normal estándar

Hipótesis	Estadígrafo	Región crítica
$H_0: \hat{P} = P$ *vs.* $H_1: \hat{P} \neq P$	$Z = \dfrac{\hat{P} - P}{\sqrt{\dfrac{P \cdot (100 - P)}{n}}}$	$\lvert Z \rvert > Z_{1-\alpha/2}$
$H_0: \hat{P} \leq P$ *vs.* $H_1: \hat{P} > P$		$Z > Z_{1-\alpha}$
$H_0: \hat{P} \geq P$ *vs.* $H_1: \hat{P} < P$		$Z < -Z_{1-\alpha}$

Para el empleo del método exacto si se desea comparar una proporción poblacional estimada (P), a partir de una muestral (p) con una proporción poblacional conocida (P_0) de referencia, hipotética o definida por un procedimiento teórico para muestras pequeñas con $npq < 5$, donde no se puede asumir la aproximación

normal a la binomial de la variable X, para cierto nivel de significación alfa, el posible planteamiento de hipótesis estadísticas es:

- Contraste bilateral:
  H_0: P = P_0 *vs.* H_1: P ≠ P_0
- Contraste unilateral:
  - H_0: P = P_0 *vs.* H_1: P > P_0 o H_0: P = P_0 *vs.* H_1: P < P_0.
  - H_0: P ≥ P_0 *vs.* H_1: P < P_0 o H_0: P ≤ P_0 *vs.* H_1: P > P_0.
- El cálculo del p-valor se realiza a partir de la relación de los valores de P y P_0 como sigue:
  - Si P ≤ P_0, entonces:

$$\text{p-valor} = 2\Pr(X \le x) = \text{mín}\left[2\sum_{k=0}^{x}\binom{n}{k}p_0^k(1-p_0)^{n-k}, 1\right]$$

  - Si P > P_0, entonces:

$$\text{p-valor} = 2\Pr(X > x) = \text{mín}\left[2\sum_{k=x}^{n}\binom{n}{k}p_0^k(1-p_0)^{n-k}, 1\right]$$

  donde:
  mín: el mínimo valor para la expresión que le sigue.

Regla de decisión

Se compara el p-valor estimado con el valor de alfa predeterminado:

- Si el p-valor estimado bajo H_0 < valor de alfa predeterminado, entonces se rechaza H_0 y se concluye que existen diferencias significativas entre las proporciones poblacionales comparadas.
- Si el p-valor estimado bajo ≥ valor de alfa predeterminado, entonces no se puede rechazar H_0 y se concluye que no existen diferencias significativas entre las proporciones poblacionales comparadas.

e) Empleo de la distribución F de Fisher en el método exacto del contraste de hipótesis para comparar una proporción muestral que estima la poblacional (P) *versus* una proporción poblacional de referencia (P_0) para el trabajo con muestras pequeñas. Se fundamenta el método en que si X es una variable aleatoria de distribución binomial con parámetros n y p se puede probar que P{X < k} = P{F < F_0}, donde F representa una variable aleatoria con distribución F de Fisher con m y r como grados de libertad, k es el número de veces que ocurre el fenómeno A y n es el tamaño muestral. Teniendo en cuenta que se aproxima una distribución discreta, como la binomial por una continua como la F de Fisher-Snedecor, se utiliza la corrección por continuidad y por ello se agrega el valor 0,5 en los cálculos a realizar en este contraste, por lo que m = 2 (n − k + 0,5) y r = 2 (k + 0,5) para los grados de libertad esta corrección.

Se emplea el estadístico de prueba $F = \frac{rQ_0}{mP_0}$, cuyo valor calculado se compara con el valor teórico-esperado-tabulado correspondiente con la expresión F(r, m)

α, donde: r expresa los grados de libertad del denominador y m los del numerador respectivo a utilizar (calculados con la formulación referida previamente), según el valor de alfa seleccionado por el investigador.

Para el caso que se quiera probar que $P\{X > k\} = P\{F > F_0\}$, se emplea la para ese estadígrafo o estadístico de prueba o contraste la fórmula $F = \frac{rQ_0}{mQ_0}$, cuyo valor calculado se compara con el valor teórico-esperado-tabulado correspondiente con la expresión $F_{(r, m)1 - \alpha}$, donde m expresa el valor del numerador y r el del denominador respectivo a utilizar, según el valor de alfa seleccionado. En este caso $m = 2(k + 0{,}5)$ y $r = 2(n - k + 0{,}5)$.

A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste (F de Fisher):

Siendo $T_i = (F_{r, m, \alpha})$ el valor inferior del estadístico de prueba o contraste y $T_s = (F_{m, r, 1 - \alpha})$ el valor superior de dicho estadístico empleado para el contraste de hipótesis, entonces:

- Para contraste bilateral:
  Intervalo donde no se rechaza $H_0 = (T_i, T_s)$
  Región crítica = $(-\infty, T_i] \cup (Ts, +\infty)$ donde se rechaza H_0
- Para contraste unilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, +\infty)$
    Región crítica = $(-\infty, T_i)$ donde se rechaza H_0
  - Intervalo donde no se rechaza $H_0 = (-\infty, T_S)$
    Región crítica = $(T_S, +\infty)$ donde se rechaza H_0
- Valor de p (p-valor), bajo H_0 donde se cumpliría H_0:
  - Si contrastamos: $P \leq P_0$, entonces:
    p-valor = Probabilidad $\{F(\text{observada}) \geq F_{r, m, \alpha}$ (teórica-tabulada)$\}$
  - Si contrastamos: $P > P_0$, entonces:
    p-valor = Probabilidad $\{F(\text{observada-calculada}) \leq F_{m, r, 1 - \alpha}$ (teórica-tabulada)$\}$

Regla de decisión

Se compara el valor de F (calculado-observado) *versus* el valor de F (teórico-tabulado) para el valor de significación de alfa prefijado:

- Si |F(observado-calculado)| > F (teórico-esperado) según los grados de libertad m, r ara el valor de alfa prefijado, esto se corresponde con la ubicación de F (calculado) en el intervalo de los valores del área critica o de rechazo de H_0 según el modelo teórico de referencia, lo que aumenta la probabilidad de cometer error tipo 1; entonces hay evidencias suficientes para rechazar H_0, y por ende se acepta H_1 (la hipótesis alternativa), por lo que es estadísticamente significativo el resultado obtenido al estimarse un p-valor bajo H_0 para F observado o una probabilidad relativa a la azar inferior a la correspondiente con el valor de significación alfa prefijado ($p < \alpha$).

– Si |F(observado-calculado)| ≤ F(teórico-esperado) según los grados de libertad m, r para el valor de α prefijado; no se puede rechazar H_0.

De forma tabulada se pueden resumir los elementos esenciales teóricos e este contraste de hipótesis (Tabla 4.7).

Tabla 4.7. Resumen de la dócima o prueba de hipótesis para una proporción poblacional empleando la aproximación por el estadígrafo F de Fisher-Snedecor

Hipótesis	Estadígrafo	Región Crítica
H_0: $P = P_0$ *vs.* H_1: $P \neq P_0$	Si $P < P_0$ el caso se trata como el caso 1 en la siguiente columna para rechazar H_0: $F = \frac{rP_0}{mQ_0}$ donde: $m = 2(n - k + 0{,}5)$ y $r = 2(k + 0{,}5)$	$F_{exp} < F_{\alpha/2;\, m,\, r}$
	Si $P > P_0$ el caso se trata como el caso 2 en la siguiente columna para rechazar H_0: $F = \frac{rP_0}{mQ_0}$ donde: $m = 2(k + 0{,}5)$ y $r = 2(n - k + 0{,}5)$	$F_{exp} > F_{1-\alpha/2;\, m,\, r}$
	Si $P = P_0$, no se rechaza H_0	
H_0: $P \leq P_0$ contra H_1: $P > P_0$	$F = \frac{rP_0}{mQ_0}$ donde: $m = 2(k + 0{,}5)$ y $r = 2(n - k + 0{,}5)$	$F_{exp} > F_{1-\alpha;\, m,\, r}$
H_0: $P \geq P_0$ contra H_1: $P < P_0$	$F = \frac{rP_0}{mQ_0}$ donde: $m = 2(n - k + 0{,}5)$ y $r = 2(k + 0{,}5)$	$F_{exp} < F_{\alpha;\, m,\, r}$

f) Una muestra con una variable cualitativa aleatoria que sigue una distribución de frecuencias de tipo Poisson; si se desea comparar el número de eventos esperados (de baja frecuencia en general dicho evento o evento raro) para una población (estimado de una muestra = μ), respecto al número de eventos esperados de esa misma variable para una población de referencia, hipótetica o definida por un procedimiento teórico (μ_0) para un determinado nivel de significación alfa, el posible planteamiento de hipótesis estadísticas para el contraste sería:

– Contraste bilateral:

H_0: $\mu = \mu_0$ *vs.* H_1: $\mu \neq \mu_0$

donde:

$$\mu_0 = \sum_{i=1}^{n} p_i$$

μ_0: parámetro poblacional (valor esperado del número de eventos en la población de referencia);
X_i: número total de eventos observados en la muestra;
pi: probabilidad del evento observado en la muestra; es el número total de eventos en individuos entre el número de muestra;
n: tamaño de la muestra;
μ: parámetro poblacional estimado del muestral (valor esperado del número de eventos en la muestra).

– Método por la prueba del valor crítico: pasos a seguir:
  - Determinar el número de eventos observados en la muestra estudiada.
  - Obtención del intervalo de confianza (IC) de los valores de μ, donde C_i es el valor inferior y C_s es el valor superior, y para estimarlos hay que auxiliarse del valor observado del evento en la muestra y se hace corresponder con los respectivos valores en la tabla de referencia empleada para la distribución de Poisson de variables aleatorias según el valor de alfa prefijado.
  - Comparación de μ_0 *vs.* intervalo de confianza para μ.

Regla de decisión

- Si $\mu_0 < C_i$ o $\mu_0 > C_s$, entonces se rechaza la hipótesis nula de no diferencias significativas entre los valores esperados del evento para las poblaciones comparadas (para dos colas).
  - Si $\mu_0 \geq C_i$ o $\mu_0 \leq C_s$, entonces no se puede rechazar la hipótesis nula de no diferencias significativas entre los valores esperados del evento para las poblaciones comparadas (para dos colas).

– Método por la prueba del p-valor para muestra pequeña: el cálculo del p-valor exacto (valor de probabilidad inferior extremo relacionado con el azar para el resultado obtenido) para ambos lados (colas) se realiza como sigue:
  - Si $x < \mu_0$:

$$\text{p-valor} = \text{mín}\left(2\sum_{k=0}^{x}\frac{e^{-\mu_0}\mu_0^k}{k!}, 1\right)$$

  - Si: $x \geq \mu_0$:

$$\text{p-valor} = \text{mín}\left(2\sum_{k=0}^{x-1}\frac{e^{-\mu_0}\mu_0^k}{k!}, 1\right)$$

donde:
mín: significa el mínimo valor para la expresión que le sigue.

Regla de decisión

• Si p-valor estimado< valor de alfa, se rechaza H_0 y se acepta la alternativa, o sea, existen diferencias significativas entre los valores de los parámetros comparados: el estimado a partir de la muestra (μ) y el de la población de referencia (μ_0).

- Si p-valor estimado ≥ valor de alfa, no se puede rechazar H_0 y no se acepta la alternativa, o sea, no existen diferencias significativas entre los valores de los parametros comparados: el estimado a partir de la muestra (μ) y el de la población de referencia (μ_0) para el nivel de confianza prefijado por el investigador.

– Método por la prueba del p-valor para muestra grande: pasos a seguir:

- Se empleará solo si μ_0> 10.
- Determinar el número de eventos observados (x) en la muestra de estudio (estimaría una poblacional).
- Sustituir los valores de x y μ_0 en la fórmula del estadístico de prueba o contraste: chi-cuadrado según:

  $\chi^2 = \frac{(x-\mu_0)^2}{\mu_0}$, que sigue una distribución teórica de probabilidades chi-cuadrado con 1 grado de libertad, para (1 – α) % bajo H_0.
- Plantear el valor de p (p-valor) bajo la H_0:
  p-valor = Pr ($\chi^2_{1,\,1-\alpha}$ (téorico-esperado) ≥ χ^2 (observado-calculado).
- Comparar valores de chi-cuadrado: χ^2 (observado-calculado) *versus* χ^2 (esperado-teórico) para 1 grado de libertad, para (1 – α) % como nivel de confianza.

Regla de decisión

- Si chi-cuadrado (observado-calculado) > chi-cuadrado (esperado-teórico) con 1 grado de libertad, para 1 – α, esto se corresponde con la ubicación de chi-cuadrado calculado en el intervalo de los valores del área crítica o de rechazo de H0 según el modelo teóricode referencia, lo que disminuye la probabilidad de cometer el error tipo 1, entonces hay evidencias suficientes para rechazar H_0 y, por ende, se acepta H_1 (la hipótesis alternativa), o sea, existen diferencias significativas entre los valores de los eventos comparados: el estimado a partir de la muestra (μ) y el de la población de referencia (μ_0), observandose un p-valor bajo H_0 para chi-cuadrado observado o una probabilidad relativa al azar inferior a la correspondiente con el valor de significación alfa prefijado (p < α).
- Si chi-cuadrado (observado-calculado) ≤ chi-cuadrado (esperado-teórico) con 1 grado de libertad, para 1 – α, esto no se corresponde con la ubicación de chi-cuadrado calculado en el intervalo de los valores del área crítica o de

rechazo de H_0 según el modelo teóricode referencia, lo que aumenta la probabilidad de cometer el error tipo 1, entonces no hay evidencias suficientes para rechazar H_0 y, por ende, no se acepta H_1 (la hipótesis alternativa), o sea, no existen diferencias significativas entre los valores de los eventos comparados: el estimado a partir de la muestra (μ) y el de la población de referencia (μ_0), observandose un p-valor bajo H_0 para chi-cuadrado observado o una probabilidad relativa al azar igual o mayor a la correspondiente con el valor de significación alfa prefijado ($p \geq \alpha$).

f) Una muestra con una variable cualitativa expresada en forma de tasas de incidencias (nuevos eventos según personas-tiempo) que sigue una distribución de frecuencias que se aproxima a la de Poisson, si se desea comparar la densidad de incidencia (DI) o tasa de incidencia (TI) de una población estimada a partir de una muestra respecto a la densidad de incidencia (DI_0) o tasa de incidencia (TI_0) de una población de referencia o hipotética, a partir de los valores observados de incidencia para un número de personas-tiempo de seguimiento en una muestra, para un nivel de significación alfa, el posible planteamiento de hipótesis estadísticas para el contraste es:

– Método de inferencia para datos de tasas de incidencia (prueba para muestra grande): se emplea solo si $\mu_0 > 10$.

Contraste bilateral:

H_0: DI = DI_0 *vs.* H_1: DI ≠ DI_0

Pasos a seguir:

- Determinar el número de eventos observados (a) en la muestra de estudio (estimaría una poblacional).
- Determinar el número de eventos esperados (parámetro) en la población de referencia:

  $\mu_0 = T\ DI_0$

  donde:

  T: tiempo de seguimiento de la muestra en años;

  DI_0: densidad de incidencia o tasa de incidencia para el evento en la población de referencia (es estimada a partir del valor de los datos empleados en la muestra de estudio, debido a que bajo H_0, es asumido que no tienen diferencias ID e ID_0 en el contraste).
- Sustituir los valores de a y μ_0 en la fórmula del estadístico de prueba o contraste chi-cuadrado:

  chi-cuadrado (observado-calculado) = $(a - \mu_0)^2/\mu_0$, sigue una distribución teórica de probabilidades chi-cuadrado ($\chi^2_{1,\ 1\ -\ \alpha}$) con 1 grado de libertad para el valor de significación alfa prefijado bajo la H_0.
- Plantear valor de p (p-valor) bajo la H_0 como:

p-valor = Pr[$\chi^2_{1,\,1-\alpha}$ (teórico-esperado) ≥ χ^2 (observado-calculado)]

- Comparar valores de chi-cuadrado χ^2 (observado-calculado) *versus* χ^2 (esperado-teórico) para 1 grado de libertad y nivel de confianza (1 − α) %.

Regla de decisión

- Si chi-cuadrado (observado-calculado) > χ^2 (esperado-teórico) para 1 grado de libertad (1 − α), esto se corresponde con la ubicación de chi-cuadrado calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teóricode referencia, lo que disminuye la probabilidad de cometer el error tipo 1, entonces hay evidencias suficientes para rechazar H_0 y, por ende, se acepta H_1 (la hipótesis alternativa), o sea, existen diferencias significativas entre los valores de los eventos esperados según tasas de incidencia: el estimado a partir de la muestra (μ) y el de la población de referencia (μ_0), observandose un p-valor bajo H_0 para chi-cuadrado observado o una probabilidad relativa al azar inferior a la correspondiente con el valor de significación alfa prefijado (p < α).
- Si chi-cuadrado (observado-calculado) > χ^2 (esperado-teórico) para 1 grado de libertad (1 − α), esto no se corresponde con la ubicación de chi-cuadrado calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teóricode referencia, lo que aumenta la probabilidad de cometer el error tipo 1, entonces no hay evidencias suficientes para rechazar H_0 y, por ende, no se acepta H_1 (la hipótesis alternativa), o sea, existen diferencias no significativas entre los valores de los eventos esperados según tasas de incidencia: el estimado a partir de la muestra (μ) y el de la población de referencia (μ_0), observandose un p-valor bajo H_0 para chi-cuadrado observado o una probabilidad relativa al azar igual o superior a la correspondiente con el valor de significación alfa prefijado (p ≥ α).

– Método exacto de inferencia para datos de tasas de incidencia (prueba para muestra pequeña):
Contraste bilateral:
H_0: $\mu = \mu_0$ *vs.* H_1: $\mu \neq \mu_0$.
El cálculo del p-valor exacto (valor de probabilidad inferior extremo relacionado con el azar para el resultado obtenido) para ambos lados (colas) se realiza como:

- Si a < μ_0, $\text{p-valor} = \text{mín}\left(2\sum_{k=0}^{a}\frac{e^{-\mu_0}\mu_0^k}{k!}, 1\right)$
- Si a ≥ μ_0, $\text{p-valor} = \text{mín}\left[2\left(1-\sum_{k=0}^{a-1}\frac{e^{-\mu_0}\mu_0^k}{k!}\right), 1\right]$

donde:
mín: significa el mínimo valor para la expresión que le sigue.

Regla de decisión

- Si p-valor estimado < α, se rechaza H_0 y se acepta la alternativa, o sea, existen diferencias significativas entre los valores de los eventos comparados según las tasas de incidencia: el estimado a partir de la muestra (μ) y el de la población de referencia (μ_0) para cierto nivel de significación alfa prefijado.
- Si p-valor estimado $\geq \alpha$ no se puede rechazar H_0 y no se acepta la alternativa, o sea, no existen diferencias significativas entre los valores de los eventos comparados según las tasas de incidencia: el estimado a partir de la muestra (μ) y el de la población de referencia (μ_0) para cierto nivel de significación alfa prefijado.

g) Una muestra con dos variables cuantitativas, si se desea evaluar la significación de la Correlación lineal simple entre las mismas por medio del coeficiente de correlación lineal simple poblacional (estimado del muestral), empleando el estadístico de prueba o contraste: t-student, según los grados de libertad correspondientes para un nivel de significación alfa prefijado entonces el posible planteamiento de hipótesis estadísticas para el contraste sería:

- Contraste bilateral:

  H_0: (ρ) = 0 *vs.* H_1: (ρ) ≠ 0

  donde:

  (ρ): es el coeficiente de correlación lineal simple poblacional estimado.

  El estadístico de prueba o contraste a emplear es t-student (método del pivote) adoptando la expresión:

  $$\text{t-student (observado-calculado)} = \sqrt{\frac{(n-2)r^2}{1-r^2}} \sim t_{n-2}$$, sigue una distribución t-student para n – 2 grados de libertad y el valor alfa según el nivel de confianza seleccionado

  donde:

  r: coeficiente de correlación lineal simple muestral que estima al poblacional (ρ) y se expresa como:

  $$r = \frac{(xy)}{\sqrt{(xx)\cdot(yy)}}$$

  donde:

  sumas de cuadrados:

  $$(xx) = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

  $$(yy) = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

  $$(xy) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

n: número de pares de observaciones para ambas variables cuantitativas continuas X y Y.

Para recordar:

A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo T_i el valor inferior del estadístico de prueba o contraste y T_s el valor superior de dicho estadístico empleado para el contraste de hipótesis, entonces para contraste bilateral:

- Intervalo donde no se rechaza $H_0 = (T_i, T_s)$
  Región crítica = $(-\infty, T_i] \cup [T_s, +\infty)$ donde se rechaza H_0.
- Valor de p (p-valor) bajo H_0:
  p-valor = Pr[t-student (teórico-esperado) ≥ t-student (observado-calculado)]

Regla de decisión

- En el caso de muestras pequeñas (n < 30) se prefiere el empleo de la tabla de referencia con los valores esperados-teóricos o críticos para el coeficiente de correlación lineal de Pearson (r) (véase el anexo 1) y se procede a comparar el valor observado-calculado o experimental del mismo versus el valor teórico-esperado-tabulado según el nivel de confianza $(1 - \alpha)$ % preseleccionado, y es rechazada H_0 cuando r observado-calculado-experimental > r tabulado-esperado o crítico.
- En el caso de muestras grandes (n ≥ 30), se puede emplear la aproximación de la distribución de probabilidad de t-student, según los grados de libertad y el valor de alfa correspondientes como:

  $$\text{t-student (observado-calculado)} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$ sigue una distribución t-student con

  n – 2 grados de libertad para el nivel de confianza preseleccionado por el investigador.
- Se compara t-student (observado-calculado) *versus* t-student (teórico-esperado) según valor de alfa prefijado y n – 2 grados de libertad.
  Para t-student (observado-calculado) < 0:
- Si t-student (observado-calculado) < t-student (teórico-esperado) según valor de alfa prefijado y n – 2 grados de libertad, esto se corresponde con la ubicación de t-student (calculado) en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que disminuye la probabilidad de cometer el error tipo 1: entonces hay evidencias suficientes para rechazar H_0 y, por ende, se acepta H_1 (la hipótesis alternativa), la que es estadísticamente significativa la correlación lineal simple entre ambas variables cuantitativas estimándose un p-valor bajo H_0 para el t-student (observado-calculado) o la probabilidad relacionada con el azar de los mismos, inferior a la esperada del modelo teórico de referencia para alfa ($p < \alpha$).

- Si t-student (observado-calculado) ≥ t-student (teórico-esperado) según valor de alfa prefijado y n – 2 grados de libertad no se puede rechazar H_0.

  Para t-student (observado-calculado) ≥ 0:
- Si t-student (observado-calculado) > t-student (teórico-esperado) según el valor de alfa prefijado y n - 2 grados de libertad, esto se corresponde con la ubicación de t-student calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teóricode referencia;lo cual disminuiría la probabilidad de cometer el error tipo 1: entonces hay evidencias suficientes para rechazar H_0 y, por ende, se acepta H_1 (la hipótesis alternativa), siendo estadísticamente significativa la correlación lineal simple entre ambas variables cuantitativas estimándose un p-valor bajo H_0 para el t-student (observado-calculado) o la probabilidad relacionada con el azar de los mismos, inferior a la esperada del modelo teórico de referencia para alfa ($p < \alpha$).
- Si t-student (observado-calculado) ≤ t-student (teórico-esperado) según valor de alfa prefijado y n – 2 grados de libertad, no se puede rechazar H_0.

Para recordar

El coeficiente de correlación lineal simple muestral o momento producto de Pearson (r) expresa el grado, fuerza o intensidad de la asociación de la variable dependiente con la independiente, ambas cuantitativas. Sus posibles valores son ($-1 < r < 1$) y la forma de interpretarlos es:

- Valor 0: no correlación entre variables cuantitativas.
- Valores que se acercan a 1 o –1: hay intensidad en la correlación.
- Valor 1: correlación perfecta positiva, cuando una variable varía la otra lo hace en la misma intensidad en el mismo sentido.
- Valor –1: correlación perfecta negativa, cuando una variable varía la otra lo hace con la misma intensidad, pero en sentido inverso.

  El signo determina que si es positivo (+) las dos variables aumentan o disminuyen y si el signo es negativo (–) significa que a medida que una aumenta la otra disminuye. Su expresión matemática es:

$$r = \frac{(xy)}{\sqrt{(xx)\cdot(yy)}}$$

  El coeficiente de determinación (r^2) es un indicador de la bondad del ajuste que proporciona la recta de regresión, cuantifica el ajuste de la recta, expresa la proporción o porcentaje de la variabilidad total de Y o cambios en la variable dependiente que es capaz de explicar la variable X, es decir, debido a la variación o influencia de la independiente. Sus posibles valores son entre 0 y 1 o si se expresa en por ciento: $0 < r^2 < 100$. La forma de interpretarlos es:

- Si (r^2) > 0,50 o 50 %, el ajuste de la recta es bueno, es decir, más de la mitad de los cambios en la variable dependiente (Y) dependen de la independiente (X).
- Si (r^2) es cercano a 0, la variabilidad observada en la dependiente (Y) no se explica por su relación con X.
- Si (r^2) es cercano a 1, la variabilidad observada en la dependiente (Y) se explica por su relación con X.

h) Una muestra, dos variables cuantitativas que siguen o se aproximan en su distribución de frecuencias a la normal-estándar, si se desea evaluar la significación de la regresión lineal simple por medio del coeficiente de regresión lineal simple poblacional (B), estimado del muestral (b) mediante un contraste de independencia empleando el estadístico de prueba o contraste t-student (método del pivote) para un nivel de significación alfa y los grados de libertad correspondientes al caso.

– El posible planteamiento de hipótesis estadísticas para contrastar sería contraste bilateral:
  - H_0: (B) = 0 *vs.* H_1: (B) ≠ 0 o H_0: B = B_0 *vs.* H_1: B ≠ B_0.

– Si fuese un contraste de desigualdad podría emplearse el contraste unilateral:
  - H_0: B = 0 *vs.* H_1: B > 0 o H_0: B = 0 *vs.* H_1: B < B_0.
  - H_0: B > o = B_0 *vs.* H_1: B < B_0 o H_0: B < o = B_0 *vs.* H_1: B > B_0.

    El estadístico de prueba o contraste a emplear es t-student que adopta la expresión:

$$\text{t-student (observado-calculado)} = \frac{b - B}{\sqrt{\frac{S^2}{(xx)}}} \sim t_{n-2}$$, sigue una distribución t-student,

con n – 2 grados de libertad y para un nivel de significación alfa prefijado.

donde:

B: coeficiente de regresión lineal simple poblacional conocido o de referencia para contrastar, que bajo H_0 toma valor B = 0;

b: coeficiente de regresión lineal simple muestral o pendiente de la recta que estima al poblacional:

$$b = \frac{(xy)}{(xx)}$$

donde:

$$(xy) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

$$(xx) = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

S^2: varianza típica residual muestral para la recta, que estimaría la varianza típica residual poblacional o el error (ε) para el ajuste de la recta y se expresa por la fórmula:

$$s^2 = \frac{(yy) - \left(\frac{xy}{xx}\right)^2}{n-2}$$

donde:

$$(yy) = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

$$(xy) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

$$(xx) = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

n: número de pares x – y, la sumatoria de las parejas de observaciones de cada valor de las dos variables cuantitativas continuas (una dependiente y la otra independiente).

Para recordar

A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo Ti el valor inferior del estadístico de prueba o contraste y Ts el valor superior de este estadístico empleado para el contraste de hipótesis, entonces:

- Para contraste bilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, T_s)$:
    Región crítica = $(-\infty, T_i] \cup [T_s, +\infty)$ donde se rechaza H_0.
- Para contraste unilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, +\infty)$:
    Región crítica = $(-\infty, T_i)$ donde se rechaza H_0.
  - Intervalo donde no se rechaza $H_0 = (-\infty, T_S)$:
    Región crítica = $(T_S, +\infty)$ donde se rechaza H_0.
    El coeficiente de regresión lineal simple (b) expresa cuanto cambia en promedio el valor de la variable dependiente por cada unidad que cambia la independiente, es la pendiente o ángulo de inclinación de la recta mínimo cuadrática con respecto al eje de las x, que mejor ajusta o representa la relación gráfica entre ambas variables. Sus posibles valores son:
  - Valores positivos: la recta asciende, las dos variables aumentan o disminuyen a la vez; es decir el aumento o disminución de una unidad en la variable independiente, provoca el aumento o disminución respectiva de (b) valores en la dependiente.

- Valores negativos: la recta desciende, el aumento en una unidad de la variable independiente provoca la disminución en la variable dependiente (b) valores.
- Valor de p o p-valor bajo $H_{0:}$
  p-valor = Pr [t-student (teórico-esperado) ≥ t-student (observado-calculado)]

Regla de decisión

Para t-student (observado-calculado) < 0:

- Si t-student (observado-calculado) < t-student (teórico-esperado) según los grados de libertad y valor alfa correspondientes, entonces t-student (observado) se ubica en la región de rechazo para H_0, por lo que existe evidencia suficiente para rechazar la hipótesis nula y se acepta la alternativa, por lo que la regresión lineal simple de ambas variables cuantitativas estimada por medio del coeficiente de regresión lineal simple poblacional estimado es estadísticamente significativa al estimarse una probabilidad relativa al azar y una probabilidad de cometer el error tipo 1 para t-student observado menor que la probabilidad alfa para el nivel de confianza seleccionado ($p < \alpha$).
- Si t-student (observado-calculado) ≥ t-student (teórico-esperado) según grados de libertad y valor alfa correspondientes, no se puede rechazar H_0.

Para t-student (observado-calculado) ≥ 0:

- Si t-student (observado-calculado) > t-student (teórico-esperado) según grados de libertad y valor alfa correspondientes, entonces t-student (observado) se ubica en la región de rechazo para H_0, por lo que existe evidencia suficiente para rechazar la hipótesis nula y se acepta la alternativa, por lo que la regresión lineal simple de ambas variables cuantitativas estimada por medio del coeficiente de regresión lineal simple poblacional estimado es estadísticamente significativa al estimarse una probabilidad relativa al azar y una probabilidad de cometer el error tipo 1 para t-student (observado) menor que la probabilidad alfa para el nivel de confianza seleccionado ($p < \alpha$).
- Si t-student (observado-calculado) < t-student (teórico-esperado) según los grados de libertad y el valor alfa correspondientes; no se puede rechazar H_0.

i) Una muestra, dos variables cuantitativas que siguen o se aproximan en su distribución de frecuencias a la normal-estándar, si se desea evaluar la significación en la regresión lineal simple para el ajuste del modelo de regresión (ecuación de regresión) a los datos muestrales o experimentales mediante la aplicación de la prueba del ANOVA simple de Fisher este es un posible planteamiento de hipótesis estadísticas:

- H_0: no existe un buen ajuste del modelo de regresión lineal simple mediante la recta de regresión a los datos muestrales o experimentales.
- H_1: existe un buen ajuste del modelo de regresión lineal simple mediante la recta de regresión a los datos muestrales o experimentales.

  Esta prueba se basa en el cálculo del cuadrado medio (CM) para los errores relacionados con la regresión y no relacionados. El cuadrado medio de cada tipo de error se obtiene como el cociente de la suma de cuadrados (SC) del error específico y los grados de libertad correspondientes para cada uno:

- $S^2_{X,Y} = \frac{SC_{NE}}{n-2} = \frac{\sum\left[y_i - \hat{y}_i\right]^2}{n-2} CM_{NE}$, cuadrado medio no explicado para la recta de la regresión lineal simple con grados de libertad:
  $(\theta) = n - 2$
- $S^2_{\hat{Y}} = \frac{SC_E}{\theta_e} = \sum\left[\hat{y}_i - \bar{y}\right]^2 = CM_E$, cuadrado medio explicado para la recta de la regresión lineal simple con grados de libertad $(\theta_e) = 1$.

  Dado que S_{YX} tiene propiedades de la desviación estándar, los cuadrados medios tendrán propiedades de varianzas, de esto se deriva que se puede aplicar la prueba de razón de varianzas (Fisher), como método de comprobación de la diferencia entre los cuadrados medios explicado y no explicado por la regresión. A modo de resumen se presentan de forma tabulada a continuación los elementos estadísticos involucrados en el ANOVA simple para el análisis de regresión lineal simple (Tabla 4.8).

  El estadístico de contraste o prueba es:

  $$F \text{ observado-calculado} = \frac{CM_e}{CM_{ne}}$$, sigue una distribución $F_{1,\, n-2,\, 1-\alpha}$.

Tabla 4.8. Elementos del ANOVA simple para el análisis de regresión lineal simple

Fuente de variación	Suma de los cuadrados (SC)	Grados de libertad (θ)	Cuadrado medio (CM = SC/θ)	F
Error explicado (θ_E)	$\sum\left[\hat{y}_i - \bar{y}_i\right]^2$	1	$\sum\left[\hat{y}_i - \bar{y}_i\right]^2$	$\frac{CM_e}{CM_{ne}}$
Error no explicado (θ_{NE})	$\sum_{i=1}^{n}\left[y_i - \hat{y}_i\right]^2$	$n - 2$	$\frac{\sum_{i=1}^{n}\left[y_i - \hat{y}_i\right]^2}{n-2}$	

– Valor de p, bajo H_0:

  p-valor = Pr[$F_{1, n-2, 1-\alpha}$ (teórico-esperado) ≥ F (observado-calculado)].

  La regla de decisión respecto al contraste de hipótesis tendría como criterio que si F(observado-calculado) > $F_{1, n-2, 1-\alpha}$ (teórico-esperado) para la tabla de la distribución de probabilidades de F de Snedecor, entonces se ubica el valor observado de F en la región de rechazo de H_0; por lo que existirían evidencias para rechazar H_0 y plantear que el cuadrado medio explicado por la regresión (CM_e) es significativamente mayor que el cuadrado medio no explicado por la regresión (CM_{ne}), por lo que el ajuste del modelo de regresión lineal simple a los datos muestrales (experimentales) es suficientemente bueno.

Para recordar

Cuando se realiza este tipo de contraste de hipótesis en que se desea determinar lo significativo o no en el ajuste del modelo teórico o recta de regresión estimada con los datos muestrales aportados por las dos variables x, y que permita describir la relación de regresión entre las dos variables, se deben seguir los pasos que quedaron implícitos en la explicación anterior, pero se hace necesario tomar en cuenta:

– Si se observa que por medio de una tabla con los valores de las dos variables o mediante un gráfico de dispersión existe un patrón sugerente de una relación lineal entre las dos variables que puede ser descrita por una ecuación de una recta del tipo y = bx + a, donde b representa la pendiente de la recta, que equivale a decir el coeficiente de regresión lineal muestral de esa ecuación y a el valor del intercepto del gráfico de esa recta con eje y u ordenada, es decir, para el origen de la abscisa x = 0. Los valores de b y a deben ser calculados partiendo de los valores xi, yi que se corresponden con cada variable respectivamente empleando el método de mínimos cuadrados, que permite encontrar la recta de mínimos cuadrados o que mejor pueda describir la relación entre ambas variables. La fórmula matemática para obtener el valor de a y b es:

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}$$

$$a = \frac{\sum_{i=1}^{n} x_i \cdot y_i - \frac{b\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

– Con los valores de a y b se puede construir la ecuación para la recta de mínimos cuadrados o modelo teórico estimado que mejor describe la relación entre las dos variables y a partir de la evaluación de los valores x_i muestrales o experimentales en esa ecuación, se obtienen sucesivamente los valores de y, que son denominados como $\hat{y}_i$, los que mediante su unión sucesiva permiten conformar el gráfico correspondiente. Conjuntamente debe ser ubicada una recta paralela al eje de las abscisas, correspondiente con el valor del intercepto calculado previamente cuyo valor fijo se denota como $\bar{y}$. Con todos estos valores ubicados en el eje y (y_i; $\hat{y}_i$; $\bar{y}$) se calculan las diferencias entre los mismos, que se denominan errores (total, explicado y no explicado por la regresión) que son empleadas en el cálculo de la suma de cuadrados, necesarios para efectuar la aplicación del procedimiento del ANOVA en la regresión lineal simple. La forma de calcular cada uno de estos errores o diferencias entre los valores de y para este procedimiento se expresan a continuación (Fig. 4.10):

  - Error no explicado por la regresión o de falta de adaptación del modelo (e_{NE}): Distancia vertical entre cada valor experimental (muestral) yi de la variable Y y el correspondiente ŷi, obtenido a partir de la sustitución o a evaluación del xi específico en la ecuación de la recta mínimo cuadrado o modelo teórico estimado. Se expresa como $e_{NE} = y_i - \hat{y}_i$ y representa la porción de la desviación total que no se ha tomado en cuenta por la introducción de la recta de regresión.
  - Error explicado por la regresión o de adaptación del modelo (e_E): Distancia vertical entre los valores de $\hat{y}_i$, del modelo teórico o recta mínimo cuadrado y el valor correspondiente al intercepto o con el valor de la recta $\bar{y}$. Se expresa como $e_E = \hat{y}_i - \bar{y}$ y muestra en cuanto disminuye la desviación total a medida que la recta de regresión se ajusta a los puntos experimentales.
  - Error total (e_i): distancia vertical entre cada valor experimental y_i y el valor correspondiente al intercepto o con el valor de la recta $\bar{y}$ se expresa como $e_i = y_i - \bar{y}$.

    La relación de estos tipos de errores con la fórmula matemática de los cuadrados medios se presenta como:

$$\sum_{i=1}^{n} e_{NE}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = SC_{NE}$$

$$\sum_{i=1}^{n} e_E^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = SC_E$$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SC_T$$

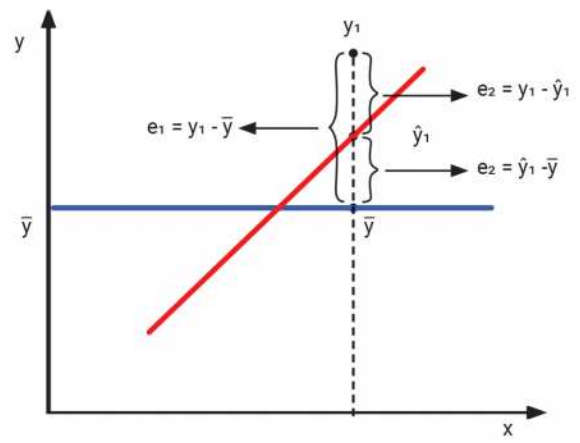


Fig. 4.10. Representación gráfica de las desviaciones o errores del valor y_i, correspondiente a la variable Y con respecto a los valores del modelo teórico o recta de regresión estimada con $\hat{y}_i$ y a la recta con valor $\bar{y}$.

– Si una ecuación de regresión lineal representada gráficamente por una recta como modelo teórico describe correctamente la relación entre dos variables, la suma de cuadrados explicada o de adaptación al modelo, debe constituir una gran proporción de la suma total de cuadrados. Esta proporción puede ser calculada y representa su resultado, el valor del coeficiente de determinación (r^2) para las dos variables en la muestra y se expresa matemáticamente por la fórmula:

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SC_E}{SC_T}$$

Sus posibles valores $0 \leq r^2 \leq 1$ son consecuencia de que las sumas de cuadrados son magnitudes positivas y $SC_E \leq SC_T$. Su interpretación a partir de sus resultados sería:

- Para $r^2 = 0$, entonces $SC_E = 0$, la regresión lineal simple o modelo teórico estimado no explica la dependencia entre las variables X y Y.
- Para $r^2 = 1$, entonces $SC_E = SC_T$, el modelo teórico estimado de regresión lineal simple se ajusta totalmente a los datos. La ecuación de regresión refleja la perfecta dependencia entre las variables X y Y.
- Para $0 < r^2 < 1$, entonces $SC_E < SC_T$, el modelo teórico estimado de regresión lineal simple explica parte del error, quedan errores no explicados o residuos, constituye lo más frecuente observado en la práctica.
  Se puede afirmar que este coeficiente permite medir la "bondad del ajuste" del modelo de regresión, en tanto más cercano sea su valor a 1, esto significa

que hay un mejor ajuste entre el modelo teórico estimado o recta de regresión estimada y los datos recogidos en la muestra para ambas variables. Esto implica que la recta de regresión estimada describe más eficientemente la relación entre las dos variables en el estudio.

j) Una muestra, una variable cuantitativa aleatoria dependiente que sigue o se aproxima a una distribución normal estándar y otra variable cualitativa dicotómica, con interés en el parámetro poblacional de posición, si se desea estimar la diferencia entre las dos medias poblacionales estimadas de las dos medias muestrales calculadas con los datos de la variable cuantitativa (dependiente) a partir de las dos nominaciones de la variable cualitativa (independiente), empleando el estadístico de prueba o contraste: $Z \sim N(0, 1)$ si ($n \geq 30$) o t-student si $n < 30$, en la evaluación de la significación estadística del resultado con un nivel de confianza (según el valor de alfa prefijado) y grados de libertad correspondientes para el caso que emplee la t-student. El posible planteamiento de hipótesis estadísticas para el contraste sería:

- Contraste bilateral:
  $H_0: \mu_1 - \mu_2 = 0$ *vs.* $H_1: \mu_1 - \mu_2 \neq 0$
- Contraste unilateral:
  - $H_0: \mu_1 - \mu_2 = 0$ *vs.* $H_1: \mu_1 - \mu_2 > 0$ o $H_0: \mu_1 - \mu_2 = 0$ *vs.* $H_1: \mu_1 - \mu_2 < 0$.
  - $H_0: \mu_1 - \mu_2 \geq 0$ *vs.* $H_1: \mu_1 - \mu_2 < 0$ o $H_0: \mu_1 - \mu_2 \leq 0$ *vs.* $H_1: \mu_1 - \mu_2 > 0$.
- Se emplea el estadístico de contraste $Z \sim N(0, 1)$ por la expresión:

  $$Z\ (\text{observado-calculado}) = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$; para el caso en que asumimos las medias poblacionales estimadas de las muestrales en la fórmula anterior y si se conocen las desviaciones estándar poblacionales (σ^1 y σ^2) (que se asumen como igual para las dos poblaciones estimadas).
- Se emplea el estadístico de contraste t-student adoptando la expresión:

  $$\text{t-student (observado-calculado)} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$, que sigue una distribución teórica de probabilidades t-student con n1 + n2 – 2 grados de libertad para el valor de alfa según el nivel de confianza preseleccionado.

  Esto es para el caso en que se asume las medias poblacionales estimadas de las muestrales en la fórmula y cuando se desconocen las varianzas poblacionales (σ^2_1 y σ^2_2) aunque se suponen no significativamente diferentes, siendo estimadas estas últimas por la expresión que sigue a partir de:

  $$S^2 = \frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{(n_1 + n_2 - 2)}$$, que es la cuasivarianza muestral ponderada.

– Se emplea el estadístico de contraste t-student adoptando la expresión:

$$\text{t-student (observado-calculado)} = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} \sim t_f$$, que sigue una distribución teórica de probabilidades t-student con f grados de libertad para el valor de alfa según el nivel de confianza preseleccionado.

En este caso es que asumimos las medias poblacionales estimadas de las muestrales en la fórmula; pero se desconocen las varianzas poblacionales (σ^2_1 y σ^2_2) y se suponen estadísticamente diferentes, siendo estimadas estas últimas a partir de cada cuasivarianza muestral típica respectivamente ($\hat{S}_1^2$ y $\hat{S}_2^2$).

donde:

f: es el número de grados de libertad estimados para la t-student y se calcula mediante la fórmula de Welch:

$$f = \frac{\left(\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}\right)^2}{\frac{1}{n_1 + 1}\left(\frac{\hat{S}_1^2}{n_1}\right)^2 + \frac{1}{n_2 + 1}\left(\frac{\hat{S}_2^2}{n_2}\right)^2} - 2$$

$\hat{S}_1^2 = \sqrt{\frac{n_1}{(n_1 - 1)S_1}}$: cuasivarianza típica de la muestra 1, que corresponde al grupo de datos para la primera categoría de la variable cualitativa dicotómica empleada;

$\hat{S}_2^2 = \sqrt{\frac{n_2}{(n_2 - 1)S_2}}$: cuasivarianza típica de la muestra 2, que corresponde para el grupo de datos de la segunda categoría de la variable cualitativa dicotómica empleada;

S_1: desviación estándar típica de la muestra 1;

S_2: deviación estándar típica de la muestra 2.

Para recordar

A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo T_i el valor inferior del estadístico de prueba o contraste y T_s el valor superior de dicho estadístico empleado para el contraste de hipótesis, entonces:

– Para contraste bilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, T_s)$
    Región crítica = $(-\infty, T_i] \cup [T_S, +\infty)$ donde se rechaza H_0.

– Para contraste unilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, +\infty)$:

Región crítica = $(-\infty, T_i)$ donde se rechaza H_0.

- Intervalo donde no se rechaza $H_0 = (-\infty, T_S)$:
  Región crítica = $(T_{S,} +\infty)$ donde se rechaza H_0.

– Valor de p o p-valor bajo H_0:
  - p-valor = Pr[Z (teórico-esperado) según alfa ≥ Z (observado-calculado)].
  - p-valor = Pr [t-student (teórico-esperado) con n – 2 grados de libertad y según alfa ≥ t-student (observado-calculado)].

Regla de decisión

– Para Z (observado-calculado) < 0:
  - Si Z (calculado-observado) < Z (teórico-esperado) según el valor de alfa prefijado, esto se corresponde con la ubicación de Z calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que disminuye la probabilidad de cometer el error tipo 1, entonces hay evidencias suficientes para rechazar H_0 y, por ende, se acepta H_1 (la hipótesis alternativa), siendo estadísticamente significativo el resultado obtenido al ser la probabilidad relacionada con el azar de los mismos inferior a la esperada del modelo teórico de referencia ($p < \alpha$).
  - Si Z (calculado-observado) ≥ Z (teórico-esperado) según el valor de alfa prefijado y grados de libertad respectivos, no se puede rechazar H_0.

– Para Z (observado-calculado) ≥ 0:
  - Si Z (calculado-observado) > Z (teórico-esperado) según el valor de alfa prefijado, esto se corresponde con la ubicación de Z (calculado) en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que disminuye la probabilidad de cometer el error tipo 1, entonces hay evidencias suficientes para rechazar H_0 y, por ende, se acepta H_1 (la hipótesis alternativa), siendo estadísticamente significativo el resultado obtenido al ser la probabilidad relacionada con el azar de los mismos menor a la esperada del modelo teórico de referencia ($p < \alpha$).
  - Si Z (calculado-observado) ≤ Z (teórico-esperado) según valor de alfa prefijado, no se puede rechazar H_0.

En el caso de emplear la distribución teórica del estadístico t-student:

– Para t-student (observado-calculado) < 0:
  - Si t-student (calculado-observado) < t-student (teórico-esperado) según valor de alfa y grados de libertad correspondientes, esto se corresponde con la ubicación de t-student calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que disminuye la probabilidad de cometer error tipo 1, entonces hay evidencias suficientes para rechazar H_0 y, por ende, se acepta H_1 (la hipótesis alternativa), siendo estadísticamente significativo el resultado obtenido al

ser la probabilidad relacionada con el azar de los mismos, inferior a la esperada del modelo teórico de referencia ($p < \alpha$).
    - Si t-student (calculado-observado) ≥ t-student (teórico-esperado) según valor de alfa y grados de libertad correspondientes, no se puede rechazar H_0.
  - Para T (t-student) observado-calculado ≥ 0:
    - Si t-student (calculado-observado) > t-student (teórico-esperado) según valor de alfa y grados de libertad correspondientes, esto se corresponde con la ubicación de t-student calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que disminuye la probabilidad de cometer error tipo 1, entonces hay evidencias suficientes para rechazar H_0 y, por ende, se acepta H_1 (la hipótesis alternativa), siendo estadísticamente significativo el resultado obtenido al ser la probabilidad relacionada con el azar de los mismos, inferior a la esperada del modelo teórico de referencia ($p < \alpha$).
    - Si t-student (calculado-observado) ≤ t-student teórico-esperado según valor de alfa y grados de libertad correspondientes, no se puede rechazar H_0.

k) Una muestra, una variable cuantitativa aleatoria dependiente que sigue o se aproxima a una distribución normal estándar y otra variable cualitativa politómica u ordinal (independiente), con interés en el parámetro poblacional de posición, si se desea comparar más de dos medias de una variable dependiente cuantitativa según las más de dos nominaciones o categorías en la independiente cualitativa o por los más de dos niveles u ordinalidad, empleando el estadístico de prueba o contraste F de Fisher-Snedecor por análisis de varianza (ANOVA) de una vía o dirección, en la evaluación de la significación estadística del resultado obtenido con un nivel de confianza según el valor de alfa prefijado y grados de libertad correspondientes, el posible planteamiento de hipótesis estadísticas para contrastar sería:
  - Contraste bilateral:

    $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$ *vs.* $H_1: \mu_1 \neq \mu_2 \neq \ldots \neq \mu_k$.

    El estadístico de prueba o contraste a emplear es F de Fisher-Snedecor (por ANOVA de una vía o dirección) para grados de libertad t o k – 1, N – t o k y para un valor alfa de significación que adopta la expresión:

    $$\text{F de Fisher-Snedecor (observado-calculado)} = \frac{\hat{S}_E^2}{\hat{S}_D^2} \rightarrow F_{t-1,N-t}$$ y sigue una distribución teórica de probabilidades F- Snedecor con t – 1 y N – t grados de libertad en el numerador y el denominador, respectivamente, para el valor de alfa según el nivel de confianza preseleccionado;

    donde:

    $\hat{S}_D^2 = \frac{SCD}{N-t}$: intravarianza (cuasivarianza intraniveles, intragrupos o intracategorías);

$SCD = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i\cdot})^2$: suma de las diferencias de cada valor de cada nivel, grupo o categoría respecto a la media intranivel i, intragrupo o intracategoría al cuadrado;

$\hat{S}_E^2 = \frac{SCE}{t-1}$: intervarianza (cuasivarianza entre los niveles, grupos o categorías);

$SCE = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2$: suma de las diferencias de las medias intranivel i respecto a la media muestral, al cuadrado;

$\hat{S}_T^2 = \frac{SCT}{N-1}$: cuasivarianza total (toda la varianza muestral);

$SCT = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{\cdot\cdot})^2$: suma de las diferencias de todos los valores de los diferentes niveles, grupos o categorías respecto a la media muestral, al cuadrado;

$N = \sum_{i=1}^{t} n_i$: número total de observaciones (entre todos los niveles);

$x_{i\cdot} = \sum_{j=1}^{n_i} x_{ij}$: suma de las observaciones del nivel i;

$\bar{x}_{i\cdot} = \frac{x_{i\cdot}}{n_i}$: media muestral del nivel i;

$x_{\cdot\cdot} = \sum_{i=1}^{t}\sum_{j=1}^{n_i} x_{ij} \sum_{i=1}^{t} n_i \bar{x}_{i\cdot}$: suma de todas las observaciones;

$\bar{x}_{\cdot\cdot} = \frac{x_{\cdot\cdot}}{N}$: media muestral de todas las observaciones.

Para recordar

A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo F exp el valor observado-calculado del estadístico de prueba o contraste y F teo el valor teórico-esperado de ese estadístico con grados de libertad t – 1, N – t para un nivel de significación alfa prefijado entonces:

- Intervalo donde no se rechaza H_0, para cualquier valor de F exp ≤ F teo $_{(N-t,\, t-1;\, 1-\alpha)}$: Región crítica para cualquier valor de Fexp > F teo $_{(N-t,\, t-1;\, 1-\alpha)}$, donde se rechaza H_0.
- Valor de p o p-valor bajo H_0:
  p-valor = Pr[F (teórico-esperado) ≥ F (observado-calculado)].

Regla de decisión

La regla de decisión respecto a la significación estadística al contrastar las hipótesis (nula y alternativa) estará determinada por:

– Si F (calculado-observado) > F de Snedecor (teórico-esperado) según los grados de libertad respectivos y el valor de alfa prefijado, esto se corresponde con la ubicación de F calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que disminuye la probabilidad de cometer el error tipo 1, entonces hay evidencias suficientes para rechazar H_0 y, por ende, se acepta H_1 (la hipótesis alternativa), siendo estadísticamente significativa la diferencia entre las más de dos medias utilizadas a partir de los grupos formados por ANOVA de una vía al ser la probabilidad relacionada con el azar de los mismos, inferior a la esperada del modelo teórico de referencia ($p < \alpha$).
– Si F (calculado-observado) ≤ F (teórico-esperado) según los grados de libertad respectivos y el valor de alfa prefijado, esto no se corresponde con la ubicación de F calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que aumenta la probabilidad de cometer el error tipo 1, entonces no hay evidencias suficientes para rechazar H_0 y, por ende, no se acepta H_1 (la hipótesis alternativa), no siendo estadísticamente significativa la diferencia entre las más de dos medias utilizadas a partir de los grupos formados por ANOVA de una vía, al ser la probabilidad relacionada con el azar de los mismos, igual o superior a la esperada del modelo teórico de referencia ($p \geq \alpha$).

l) Una muestra, una variable cuantitativa aleatoria dependiente que sigue o se aproxima a una distribución normal estándar y otra variable cualitativa dicotómica, con interés en el parámetro poblacional de dispersión, si se desea determinar la diferencia a partir de la prueba F del cociente entre dos varianzas poblacionales estimadas de las dos varianzas muestrales calculadas con los datos de la variable dependiente cuantitativa continua, a partir de dos nominaciones de la variable cualitativa dicotómica (independiente), empleando el estadístico de prueba o contraste F de Snedecor-Fisher en la evaluación de la significación estadística del resultado obtenido para un nivel de confianza según el valor de alfa prefijado y los grados de libertad correspondientes, el posible planteamiento de hipótesis estadística para contrastar sería:
– Contraste bilateral:
  $H_0: \sigma_1^2/\sigma_2^2 = 1$ *vs.* $H_1: \sigma_1^2/\sigma_2^2 \neq 1$
– Contraste unilateral:
  • $H_0: \sigma_1^2/\sigma_2^2 = 1$ *vs.* $H_1: \sigma_1^2/\sigma_2^2 > 1$ o $H_0: \sigma_1^2/\sigma_2^2 = 1$ *vs.* $H_1: \sigma_1^2/\sigma_2^2 < 1$.
  • H0: $\sigma_1^2/\sigma_2^2 \geq 1$ *vs.* $H_1: \sigma_1^2/\sigma_2^2 < 1$ o $H_0: \sigma_1^2/\sigma_2^2 \leq 1$ *vs.* $H_1: \sigma_1^2/\sigma_2^2 > 1$.
  El estadístico de prueba o contraste a emplear es la F de Fisher-Snedecor:

$$\text{F (observado-calculado)} = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{S_1^2 \cdot \sigma_2^2}{S_2^2 \cdot \sigma_1^2} = \left(\frac{S_1^2}{S_2^2}\right) \cdot \left(\frac{\sigma_2^2}{\sigma_1^2}\right),$$

donde bajo la H_0 (σ_2^2/σ_1^2) es 1, quedando, por tanto:
F (observado-calculado) = (S_1^2/S_2^2)
que sigue una distribución teórica de probabilidades F-Snedecor con $n_1 - 1$ y $n_2 - 1$ grados de libertad en el numerador y el denominador, respectivamente, para el valor de alfa según el nivel de confianza preseleccionado, donde bajo H0: $\sigma_2^2/\sigma_1^2 = 1$.
También puede ser empleado en este contraste el estadístico de Levene (W) según Cleves (1995):

$$W = \frac{n_1(\bar{Z}_1 - \bar{Z})^2 + n_2(\bar{Z}_2 - \bar{Z})^2}{\frac{\sum_{j=1}^{n_1}(Z_{1j} - \bar{Z}_1)^2 + \sum_{j=1}^{n_1}(Z_{2j} - \bar{Z}_2)^2}{n_1 + n_2 - 2}}$$

, que sigue una distribución teórica de probabilidades F-Snedecor con 1 grado de libertad en el numerador y con $n_1 + n_2 - 2$ para el denominador.
donde:

$X_{i1}, \ldots X_{in_i}$: son las observaciones de la muestra i, con i = 1, 2;

n_i: es el tamaño de la muestra i, con i = 1, 2;

$Zij = |xij - \bar{x}_i|$, con i = 1, 2 y j = 1, 2,..., ni;

$\bar{Z}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij}$: media de Z en el grupo i, con i = 1, 2;

$$\bar{Z} = \frac{1}{n_1 + n_2}\sum_{i=1}^{2}\sum_{j=1}^{n_i} Z_{ij}$$

Para recordar:
Los grados de libertad son el número de unidades independientes de información en una muestra, que son relevantes para la estimación de un parámetro o para el cálculo de un estadístico de contraste. Si se tienen dos muestras de tamaños n_1 y n_2 de una población que se distribuye de forma normal, el cociente S^2_1/S^2_2 sigue una distribución F de Snedecor con $n_1 - 1$ (los de la varianza del numerador) y $n_2 - 1$ (los de la varianza del denominador) grados de libertad, siendo T_i el valor inferior del estadístico de prueba o contraste y T_s el valor superior de dicho estadístico empleado para el contraste de hipótesis, entonces:

- Para contraste bilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, T_s)$
    Región crítica = $(-\infty, T_i] \cup [T_s, +\infty)$ donde se rechaza H_0.
- Para contraste unilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, +\infty)$:
    Región crítica = $(-\infty, T_i)$ donde se rechaza H_0.
  - Intervalo donde no se rechaza $H_0 = (-\infty, T_S)$:

Región crítica = (T_S, $+\infty$) donde se rechaza H_0.

- Valor de p o p-valor bajo H_0:
  p-valor = Pr[F (teórico-esperado) ≥ F (observado-calculado)].

Regla de decisión

La regla de decisión respecto a la significación estadística al contrastar las hipótesis (nula y alternativa) estará determinada por:

- Si F (calculado-observado) > F (teórico-esperado) según el valor de alfa prefijado y los grados de libertad respectivos, esto se corresponde con la ubicación de F calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teóricode referencia, lo que disminuye la probabilidad de cometer el error tipo 1, entonces hay evidencias suficientes para rechazar H_0 y, por ende, se acepta H_1 (la hipótesis alternativa), siendo estadísticamente significativo el resultado obtenido para el análisis de las diferencias entre dos varianzas de grupos de datos muestrales, al ser la probabilidad relacionada con el azar de los mismos o el p-valor bajo H_0 para el valor observado de F, inferior a la esperada del modelo teórico de referencia ($p < \alpha$).
- Si F (calculado-observado) ≤ F (teórico-esperado) según el valor de alfa prefijado y los grados de libertad respectivos, esto no se corresponde con la ubicación de F calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que aumenta la probabilidad de cometer el error tipo 1, entonces no hay evidencias suficientes para rechazar H_0 y, por ende, no se acepta H_1 (la hipótesis alternativa), siendo no estadísticamente significativo el resultado obtenido para el análisis de las diferencias entre dos varianzas de grupos de datos muestrales, al ser la probabilidad relacionada con el azar de los mismos o el p-valor bajo H_0 para el valor observado de F, igual o superior a la esperada del modelo teórico de referencia ($p \geq \alpha$).

m) Una muestra, dos variables cualitativas (una dependiente y la otra independiente) con distribución binomial que se puede aproximar con la normal estándar, si se desea comparar dos proporciones (con solo dos categorías o nominaciones cada variable cualitativa en la muestra), empleando el estadístico de prueba o contraste Z ~ N (0, 1) en la evaluación de la significación del resultado para un nivel de significación alfa predeterminado para muestras grandes (npq > 5), el posible planteamiento de hipótesis estadísticas para contrastar sería:

- Para contraste bilateral:
  $H_0: P_1 = P_2$ *vs.* $H_1: P_1 \neq P_2$
- Para contraste unilateral:
  - $H_0: P_1 = P_2$ *vs.* $H_1: P_1 > P_2$ o $H_0: P_1 = P_2$ *vs.* $H_1: P_1 < P_2$
  - $H_0: P_1 \geq P_2$ *vs.* $H_1: P_1 < P_2$ o $H_0: P_1 \leq P_2$ *vs.* $H_1: P_1 > P_2$

    Los pasos a seguir para realizar el contraste incluyen comenzar con el auxilio gráfico de una tabla cruzada 2x2 simple (Tabla 4.9), mediante la

cual se pueden ubicar los distintos valores correspondientes a los datos de las variables cualitativas dicotómicas empleadas en la misma, con vistas a facilitar la mejor comprensión del procedimiento a realizar:

Tabla 4.9. Representación tabular de los elementos a emplear para el análisis de la diferencia de dos proporciones formadas al asociar dos variables cualitativas dicotómicas

Variable independiente	Variable dependiente evento presente	Variable dependiente evento ausente	
Categoría 1	X_1		n_1
Categoría 2	X_2		n_2
Totales	X_i		n
$p = (X_1 + X_2) / (n_1 + n_2)$; $p_1 = X_1/n_1$; $p_2 = X_2/n_2$; $q = 1 - p$; $Var_c = [pq\ (1/n_1 + 1/n_2)\ ]$			

Como estadístico de prueba se emplea Z ~ N(0, 1), asumiendo que las variables con distribución de frecuencias binomial por la aproximación de la distribución normal-estándar, adopta la expresión que les relaciona:

$$Z\ (\text{observado-calculado}) = \frac{(p_2 - p_1)}{\sqrt{Var_c}}$$, también se puede expresar como:

$$Z\ (\text{observado-calculado}) = \frac{p_2 - p_1}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

donde:

$p = \frac{x_1 + x_2}{n_1 + n_2}$: proporción para el total de enfermos para ambas categorías de la variable independiente dicotómica;

Var_c: varianza combinada de p según n_1 y n_2, se estima como:

$$Var_c = pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$, donde q = 1 − p;

También puede ser expresada la varianza combinada como:

$$Var_c = \left[\frac{p_1(1-p_1)}{n_1} + \frac{p_1(1-p_2)}{n_1}\right]$$

X_1: número de sujetos con la categoría 1 en la variable independiente y con el evento 1 en la dependiente;

X_2: número de sujetos con la categoría 2 en la variable independiente y el evento 1 en la dependiente;

n_1: total de sujetos con la categoría 1 para la variable independiente en la muestra;
n_2: total de sujetos con la categoría 2 para la variable independiente en la muestra;
p_1: es la proporción formada a partir del: valor numérico de X_1 / valor numérico de n_1;
p_2: es la proporción formada a partir del: valor numérico de X_2 / valor numérico de n_2.

Para recordar

A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo T_i el valor inferior del estadístico de prueba o contraste y T_s el valor superior de dicho estadístico empleado para el contraste de hipótesis, entonces:

- Para contraste bilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, T_s)$:
    Región crítica = $(-\infty, T_i] \cup [T_s, +\infty)$ donde se rechaza H_0.
- Para contraste unilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, +\infty)$:
    Región crítica = $(-\infty, T_i)$ donde se rechaza H_0.
  - Intervalo donde no se rechaza $H_0 = (-\infty, T_S)$:
    Región crítica = $(T_S, +\infty)$ donde se rechaza H_0.
- Valor de p o p-valor bajo H_0:
  p-valor = Pr[Z (teórico-esperado) según alfa ≥ Z (observado-calculado)].

Regla de decisión

La regla de decisión respecto a la significación estadística al contrastar las hipótesis (nula y alternativa) está determinada por:

- Para Z (observado-calculado) < 0:
  - Si Z (observado-calculado) < Z (teórico-esperado), según el valor de alfa prefijado, esto se corresponde con la ubicación de Z calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que disminuye la probabilidad de cometer el error tipo 1, entonces hay evidencias suficientes para rechazar H_0 y, por ende, se acepta H_1 (la hipótesis alternativa), siendo estadísticamente significativo el resultado obtenido al ser el valor de probabilidad (p-valor) de cumplirse H_0 para los datos empíricos muestrales empleados o la probabilidad relacionada con el azar de los mismos menor a la esperada del modelo teórico de referencia ($p < \alpha$).
  - Si Z (observado-calculado) ≥ Z (teórico-esperado) según el valor de alfa prefijado, no se puede rechazar H_0.
- Para Z (observado-calculado) ≥ 0:
  - Si Z (observado-calculado) > Z (teórico-esperado) según el valor de alfa prefijado, esto se corresponde con la ubicación de Z calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que disminuye la probabilidad de cometer el error

tipo 1, entonces hay evidencias suficientes para rechazar H_0 y, por ende, se acepta H_1 (la hipótesis alternativa), siendo estadísticamente significativo el resultado obtenido al ser el valor de probabilidad (p-valor) de cumplirse H_0 para los datos empíricos muestrales empleados o la probabilidad relacionada con el azar de los mismos menor a la esperada del modelo teórico de referencia ($p < \alpha$).

- Si Z (observado-calculado) ≤ Z (teórico-esperado) según el valor de alfa prefijado, no se puede rechazar H_0.

n) Una muestra, con dos variables cualitativas dicotómicas, por ejemplo: exposición-riesgo, para el diseño de caso-control pareado, si se desea determinar la asociación entre las dos variables cualitativas dicotómicas utilizando medidas de asociación como razón de productos cruzados (OR) y realizar prueba de independencia entre las dos variables cualitativas empleando el estadístico de prueba o contraste chi-cuadrado especial de McNemar, en la evaluación de la significación estadística del resultado obtenido para un valor de alfa prefijado y grados de libertad correspondientes.

El posible planteamiento de hipótesis estadísticas para contrastar es:

- H_0: existe (poblacionalmente) independencia entre las 2 variables estudiadas.
- H_1: no existe (poblacionalmente) independencia entre las 2 variables estudiadas.

El estadístico de prueba a emplear es chi-cuadrado de McNemar que en estudios de caso y control pareado para tablas de contingencia 2x2 simple a partir de la nomenclatura de las respectivas celdas se expresa como:

chi-cuadrado $(\chi^2)_{(Mc\ Nemar)}$ (observado-calculado) = $(n_{10} - n_{01})^2/(n_{10} + n_{01})$ o también si se utiliza otra nomenclatura en las tablas para representar los datos de este tipo de estudio como:

chi-cuadrado $(\chi_2)_{(Mc\ Nemar)}$ (observado-calculado) = $(b - c)^2/(b + c)$, que sigue una distribución teórica chi-cuadrado con 1 grado de libertad para el valor de alfa según el nivel de confianza preseleccionado.

Para recordar

OR es la razón de momios, de disparidad, de desventaja, de productos cruzados, de chances, de oportunidades relativas de enfermedad, de oportunidades relativas de exposición u *odds ratio* y expresa la relación por un cociente entre cierta oportunidad para un efecto y la oportunidad opuesta para otro efecto. Cada oportunidad viene determinada por una razón de riesgos. Se interpreta como:

- Si OR > 1 o asociación positiva: la exposición o presencia del probable factor causal (variable independiente) está asociado, guarda relación o puede comportarse como una causa o factor de riesgo para el efecto relacionado (variable dependiente).
- Si OR < 1 o asociación negativa: la exposición o presencia del probable factor causal (variable independiente) está asociado, guarda relación o puede com-

portarse como una causa o factor negativo o protector para el efecto relacionado (variable dependiente).

- Si OR = 1 o no asociación: la exposición o presencia del probable factor causal (variable independiente) no está asociado, no guarda relación o no puede comportarse como una causa o factor para el efecto relacionado (variable dependiente).

  Respecto a la significación estadística de la relación entre variables a partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo χ^2 exp (McNemar) el valor observado-calculado del estadístico de prueba o contraste y χ^2 teo el valor teórico-sperado de ese estadístico con grados de libertad (c – 1) (f – 1) y nivel de confianza 1 – α, entonces:
- Intervalo donde no se rechaza H_0 = (para cualquier valor de χ^2 exp < 0 = χ^2 teo $^{(c-1)(f-1),1-\alpha}$.

  Región crítica (donde se rechaza H_0) = (para cualquier valor de χ^2 exp > χ^2 teo $^{(c-1)(f-1),1-\alpha}$.

  donde:

  χ^2 exp: chi-cuadrado experimental o calculado-observado;

  χ^2 teo: chi-cuadrado teórico, esperado, tabulado o de referencia.
- Valor de p o p-valor bajo H_0:

  p-valor = Pr [χ^2 (teórico-esperado) ≥ χ^2 (observado-calculado)].

Regla de decisión

La regla de decisión respecto a la significación estadística al contrastar las hipótesis (nula y alternativa) está determinada por:

- Si χ^2 (observado-calculado) de McNemar > χ^2 (esperado-tabulado) con el valor de alfa y grados de libertad correspondientes para el caso, esto se corresponde con la ubicación del valor de chi-cuadrado McNemar observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que disminuye la probabilidad de cometer el error tipo 1, entonces hay evidencias suficientes para rechazar H_0, es decir, que hay asociación o dependencia significativa entre las variables dicotómicas independiente y dependiente tomada en forma pareada (casos y controles) para la presencia o no del evento, siendo las probabilidades relativas al azar o el p-valor bajo H_0 para estos resultados, inferiores a las esperadas por el modelo teórico de referencia (p < α).
- Si χ^2 (observado-calculado) de McNemar ≤ χ^2 (esperado-tabulado) con el valor de alfa y grados de libertad correspondientes para el caso, esto no se corresponde con la ubicación del valor de chi-cuadrado McNemar observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que aumenta la probabilidad de cometer el error

tipo 1, entonces no hay evidencias suficientes para rechazar H_0, es decir, que hay independencia significativa entre las variables dicotómicas independiente y dependiente tomada en forma pareada (casos y controles) para la presencia o no del evento, siendo las probabilidades relativas al azar o el p-valor bajo H0 para estos resultados, iguales o superiores a las esperadas por el modelo teórico de referencia ($p \geq \alpha$).

ñ) Una muestra con dos variables cualitativas dicotómicas (exposición-riesgo) para el diseño de estudio (transversal, caso-control no pareado y de cohorte por incidencia acumulada), si se desea determinar la existencia de asociación o no entre las dos variables cualitativas dicotómicas, realizar prueba de independencia entre las dos variables empleando el estadístico de prueba o contraste chi-cuadrado (Pearson o corregido de Yates) o prueba de la probabilidad exacta de Fisher, en la evaluación de la significación estadística del resultado obtenido para un valor de alfa prefijado y grados de libertad correspondientes.

El posible planteamiento de hipótesis estadísticas para contrastar sería:

- H_0: existe (poblacionalmente) independencia entre las dos variables estudiadas.
- H_1: no existe (poblacionalmente) independencia entre las dos variables estudiadas.

  El estadístico de prueba a emplear es chi-cuadrado de Pearson de asociación (sin corrección), de Yates (con corrección) o la probabilidad exacta de Fisher para tablas de contingencia 2x2 simple según la frecuencia esperada mayor o igual que 5 en cada celda o no, respectivamente. A partir de la nomenclatura de las respectivas celdas (véase en el capítulo 3 las tablas de contingencia) se expresa como:
- Fórmula para chi-cuadrado de Pearson (χ^2) según la expresión:

$$\chi^2 = \frac{n(ad-bc)^2}{m_1 m_0 n_1 n_0}$$

- Fórmula para chi-cuadrado con corrección de Yates ($\chi^2 C$) según la expresión:

$$\chi^2 = \frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{m_1 m_0 n_1 n_0}$$

- Procedimiento para la prueba exacta de Fisher según Rosner (2000):
  - Siendo el p-valor (valor p) unilateral:

$$\text{p-valor} = \begin{cases} P(+) & \text{si } \dfrac{ad}{(b+1)(c+1)} > \dfrac{bc}{(a+1)(d+1)} \\ P(-) & \text{en caso contrario} \end{cases}$$

- Siendo el p-valor (valor p) bilateral:
  p -valor = P(+) + P(–) – (a)
  donde:

$$P(+) = \sum_{P(a+k)\leq P(a)} P(a+k)$$

$$P(-) = \sum_{P(a-k)\leq P(a)} P(a-k)$$

  P (a + k) y P (a – k) se calculan a partir de la función de probabilidad de una distribución hipergeométrica de parámetros N = a + b + c + d, R = a + b y n = a + c.

A continuación se presenta la explicación detallada de la prueba para la probabilidad exacta de Fisher:

– Puede ser empleado en la comparación de dos proporciones binomiales procedentes de dos muestras independientes en tablas de 2x2 simple con valores de frecuencias esperada menores de 5 en alguna celda.
– Se puede emplear tanto para pruebas de homogeneidad en dos muestras independientes como de Independencia de dos variables cualitativas dicotómicas.
  - Permite determinar la probabilidad (p-valor) exacta de obtener valores de la celda a, fuera de las a + b observaciones para un número N (a + b + c + d = total de observaciones) de las cuales a + c se consideran con éxito dentro del evento estudiado, al comparar dos proporciones binomiales en tablas de contingencia 2x2 simples, esto se puede realizar empleando la probabilidades de la distribución hipergeométrica y estimar dicho valor p (p-valor) como sigue según el procedimiento de Rosner (2000).
– Valor p unilateral:
  - p-valor para la cola derecha se estimacomo probabilidades (+) a partir de la condición:

$$\text{p- valor} = \begin{cases} P(+) \text{ si } \dfrac{ad}{(b+1)(c+1)} > \dfrac{bc}{(a+1)(d+1)} \\ P(-) \text{ en caso contrario} \end{cases}$$

$$P(+) = \sum_{P(a+k)\leq P(a)} P(a+k) \quad \text{cuando se quiere contrastar } p_{1\,>\,p2.}$$

  - p-valor para la cola izquierda, se estima como probabilidades (–) a partir de la condición:

$$\text{p-valor} = \begin{cases} P(-) \text{ si } \dfrac{ad}{(b+1)(c+1)} < \dfrac{bc}{(a+1)(d+1)} \\ P(-) \text{ en caso contrario} \end{cases}$$

$P(-) = \sum_{P(a-k) \leq P(a)} P(a-k)$ cuando se quiere contrastar $p_1 < p_2$.

– Valor p bilateral:
p-valor = P(+) + P(–) – P(a)

donde:

$P(+) = \sum_{P(a+k) \leq P(a)} P(a+k)$, las probabilidades de las tablas estructuradas diferentes a la tabla inicial observada, empleando valores mayores o iguales de los correspondientes a la celda a (de la tabla inicial);

$P(-) = \sum_{P(a-k) \leq P(a)} P(a-k)$, las probabilidades de las tablas estructuradas diferentes a la tabla inicial observada, empleando valores menores o iguales de los correspondientes a la celda a (de la tabla inicial);

P (a + k) y P (a – k) se calculan a partir de la función de probabilidad de una distribución hipergeométrica de parámetros N = a + b + c + d, R = a + b y n = a + c.
La explicación del método es como sigue:

- Se tiene una tabla de contingencia 2x2 simple (Tabla 4.10).

Tabla 4.10. De contingencia 2x2 simple que relaciona el hábito de fumar y la presencia de enfermedad cardiovascular

Hábito de fumar	Presencia de enfermedad cardiovascular	No enfermedad cardiovascular	Totales
Fumador	a	b	a + b
No fumador	c	d	c + d
Totales	a + c	b + d	N = a + b + c + d

La probabilidad de obtener los valores para una tabla 2x2 simple observada se pueden calcular con la siguiente expresión matemática denominada *probabilidad exacta* (Pr) para los valores observados de las celdas a, b, c, d en una tabla 2x2 simple:

$\Pr(a, b, c, d) = \dfrac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$, donde el signo ! significa la operación matemática del factorial para el valor o expresión que le antecede.

- Se asignan los valores a las celdas de la tabla 4.9 (Tabla 4.11).

Tabla 4.11. De contingencia 2x2 simple que relaciona el hábito de fumar y la presencia de enfermedad cardiovascular con los valores asignados

Hábito de fumar	Presencia de enfermedad cardiovascular	No enfermedad cardiovascular	Totales marginales
Fumador	a = 2	b = 5	a + b = 7
No fumador	c = 3	d = 1	c + d = 4
Totales	a + c = 5	b + d = 6	N = a + b + c + d = 11

La probabilidad exacta a obtener para los valores de las celdas empleados para esta tabla 2x2 simple:

$$\Pr(2,\ 5,\ 3,\ 1) = \frac{7!4!5!6!}{11!2!5!3!1!} = \frac{5\ 040(24)(120)(720)}{39\ 916\ 800(2)(120)(6)} = \frac{1{,}0450944 \cdot 10^{10}}{5{,}7480192 \cdot 10^{10}} = 0{,}182$$

Los pasos que se pueden seguir para el cálculo de la probabilidad exacta de Fisher (que incluye la sumatoria de las probabilidades para los distintos valores en las celdas para las tablas formadas, manteniendo sus valores totales marginales), por ejemplo, para la prueba de Independencia de dos variables:

- Teniendo en cuenta que su estrategia básica en el contraste de hipótesis (H_0: $p_1 = p_2$ *vs.* H_1: $p_1 \neq p_2$), siendo p_1 = (individuos con efecto positivo y factor de riesgo positivo/total de individuos de la muestra) y p_2 = (individuos con efecto positivo y factor de riesgo negativo/total de individuos de la muestra) se enumeran todas las posibles tablas a formar manteniendo constantes los totales marginales en ellas, tal y como se encuentran en la tabla observada con los valores iniciales y calcular la probabilidad exacta para cada una de las tablas formadas según la expresión matemática anterior (basada en las probabilidades de la distribución hipergeométrica), cuyo valor esperado E (X) y su varianza (Var (X) para una variable aleatoria X (para la celda a o C1,1), o sea, fila 1, columna 1, se expresa como:

$$E(x) = \frac{M_1 N_1}{N}$$

$$\mathrm{Var}(x) = \frac{M_1 M_2 N_1 N_2}{N^2 (N-1)}$$

- Se reestructuran las filas y columnas de la tabla observada con los valores iniciales de modo que el total marginal inferior por las filas quede en la primera fila y el total marginal por columna inferior en la primera columna.
- Se comienza asignando valor de 0 a la primara celda ($C_{1,1}$) o celda a, tomando en lo siguiente cada celda restante los valores correspondientes que permitan sin variar los valores totales marginales por filas y columnas que presentó la

tabla inicial observada y calcular la probabilidad para estos nuevos valores a partir de la distribución de probabilidades hipergeométrica.

– Construir nuevas tablas incrementando tabla a tabla los valores de la celda a, obteniéndose tablas hasta que alguno de los valores de las celdas restantes adquiera el valor 0 y entonces calcular igualmente la probabilidad para los valores de cada nueva tabla estructurada.
– Si se toma como ejemplo la tabla observada inicial con los valores de a, b, c y d con sus respectivos valores totales marginales (a + b), (c + d), (a + c), (b + d) y el número total de observaciones (N = a + b + c + d) y denominamos tabla (a) como la observada (inicial) y la tabla (k) la última posible obtenida (cuando alguna otra celda toma un valor de 0 debido al aumento de la celda a y que debemos mantener los totales marginales) tenemos que:
  - Para contrastar la hipótesis bilateral (a partir de la estimación de la probabilidad exacta de Fisher) se tiene:

    H_0: $p_1 = p_2$ *vs.* H_1: $p_1 \neq p_2$

    p-valor Fisher (dos colas) = 2mín[Pr(0) + Pr(1) +... + Pr(a), Pr(a) + Pr(a + 1) + …+ Pr(k)], donde min significa el menor valor de las probabilidades y el p-valor (corresponde al duplicado de la menor área bajo la curva, bien de la cola derecha o de la izquierda).

    Otra forma que puede ser aplicada en su estimación es:

    $$\text{p-valor Fisher (dos colas)} = \sum_{\{i:\ Pr(i)\leq Pr(a)\}} Pr(i)$$, donde Pr(i) son las probabilidades obtenidas para los valores de cada tabla y se suman todas las Pr(i) menores o iguales a la Pr(a) de la tabla observada con los valores iniciales en cada celda.
  - Para contrastar la hipótesis unilateral:

    H_0: $p_1 = p_2$ *vs.* H_1: $p_1 < p_2$

    p-valor Fisher = Pr (0) + Pr (1) + … + Pr (a) corresponde al área bajo la curva para la cola izquierda.
  - Para contrastar la hipótesis unilateral:

    H_0: $p_1 = p_2$ *vs.* H_1: $p_1 > p_2$

    p-valor Fisher = Pr (a) + Pr (a + 1) + … Pr (k) corresponde al área bajo la curva para la cola derecha.

    Nota: en este caso el valor de p (p-valor) expresa la probabilidad de obtener una tabla con valores tan extremos o más que los obtenidos con la tabla observada inicial.

Regla de interpretación

– Si p –valor (observado-calculado) exacto de Fisher< nivel de significación alfa de acuerdo con el tipo de contraste de hipótesis a realizar (bilateral o unilateral), entonces las probabilidades estimadas exactas de poder cumplirse

H_0 con los datos observados o más extremos que los mismos son mínimas por lo que existen evidencias suficientes para el rechazo de H_0, por ende se concluye que existen diferencias significativas en las proporciones formadas a partir de las dos variables cualitativas dicotómicas, por lo que tienen asociación estadísticamente significativa ambas variables para el nivel de significación alfa preseleccionado.

- Si p-valor (observado-calculado) exacto de Fisher ≥ nivel de significación alfa de acuerdo con el tipo de contraste de hipótesis a realizar (bilateral o unilateral), entonces las probabilidades estimadas exactas de poder cumplirse H_0 con los datos observados o más extremos que los mismos no son mínimas, por lo que no existen evidencias suficientes para el rechazo de H_0, por ende se concluye que no existen diferencias significativas en las proporciones formadas a partir de las dos variables cualitativas dicotómicas por lo que tienen independencia ambas variables para el nivel de significación alfa preseleccionado.

Para recordar:

La OR o razón de momios, de disparidad, de desventaja, de productos cruzados, de chances, de oportunidades relativas de enfermedad, de oportunidades relativas de exposición u *odds ratio* expresa la relación por un cociente entre cierta oportunidad para un efecto y la oportunidad opuesta para otro efecto. Se interpreta como:

- Si OR > 1 o asociación positiva: la exposición o presencia del probable factor causal (variable independiente) está asociado, guarda relación o puede comportarse como una causa o factor de riesgo para el efecto relacionado (variable dependiente).
- Si OR <1 o asociación negativa: la exposición o presencia del probable factor causal (variable independiente) está asociado, guarda relación o puede comportarse como una causa o factor negativo o protector para el efecto relacionado (variable dependiente).
- Si OR = 1 o no asociación: la exposición o presencia del probable factor causal (variable independiente) no está asociado, no guarda relación o no puede comportarse como una causa o factor para el efecto relacionado (variable dependiente).

  Respecto a la significación estadística de la relación entre las dos variables a partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo χ^2 exp (Pearson o Yates) el valor observado-calculado (exp) del estadístico de prueba o contraste y χ^2 teo el valor teórico-esperado (teo) de dicho estadístico con 1 grado de libertad (c – 1) (f – 1) y nivel de confianza 1 – α, entonces:
- Intervalo donde no se rechaza H_0 = (para cualquier valor de χ^2 exp ≤ χ^2 teo $_{(c-1)\,(f-1);\,1-\alpha.}$

– Región crítica (donde se rechaza H_0) = (para cualquier valor de χ^2 exp > χ^2 teo $_{(c-1)(f-1);1-\alpha}$.
– Valor de p o p-valor bajo H_0:
p-valor = Pr [χ^2 (teórico-esperado) ≥ χ^2 (observado-calculado)].

Regla de decisión

La regla de decisión respecto a la significación estadística al contrastar las hipótesis (nula y alternativa) está determinada por:

– Si χ^2 (observado-calculado) > χ^2 (esperado-teórico-tabulado), esto se corresponde con la ubicación de χ^2 observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que disminuye la probabilidad de cometer el error tipo 1, entonces hay evidencias suficientes para rechazar H_0, por ende, se acepta la alternativa, es decir, hay una relación de dependencia estadísticamente significativa entre las dos variables, siendo las probabilidades relativas al azar o el p-valor bajo H_0 para dichos resultados, inferiores a las esperadas por el modelo teórico de referencia ($p < \alpha$).
– Si χ^2 (observado-calculado) ≤ χ^2 (esperado-teórico-tabulado), esto se corresponde con la no ubicación de χ^2 observado en el intervalo de los valores del área crítica o de rechazo de H0 según el modelo teórico de referencia, lo que aumenta la probabilidad de cometer el error tipo 1, entonces no hay evidencias suficientes para rechazar H_0, es decir, hay una relación de independencia estadísticamente significativa entre las dos variables, siendo las probabilidades relativas al azar o el p-valor bajo H_0 para estos resultados, iguales o superiores a las esperadas por el modelo teórico de referencia ($p \geq \alpha$).

o) Una muestra con dos variables cualitativas para el diseño de estudio con datos no pareados, si se desea determinar independencia entre las dos variables cualitativas nominales al menos con una no dicotómica u ordinal por niveles (k > 2) mediante la prueba de independencia *versus* asociación entre las dos variables empleando el estadístico de prueba o contraste chi-cuadrado (que compara frecuencia observada *versus* esperada), en la evaluación de la significación estadística del resultado obtenido para un valor de alfa prefijado y grados de libertad correspondientes. El posible planteamiento de hipótesis estadísticas para contrastar sería:
   – H_0: existe (poblacionalmente) independencia entre las variables estudiadas.
   – H_1: no existe (poblacionalmente) independencia entre las variables estudiadas.
   Como estadístico de prueba se emplea chi-cuadrado de Pearson, que permite comparar la frecuencia observada y esperada de las celdas tabuladas, adoptando la siguiente expresión en tablas de CxF con una variable al menos politómica u ordinal por niveles:

χ^2 de Pearson (observado-calculado) $= \sum_{i,j} \frac{(x_{ij} - e_{ij})^2}{e_{ij}}$; que sigue una distribución teórica de probabilidades chi-cuadrado ($\chi^2_{(c-1)(f-1);\,1-\alpha}$ con (c – 1) (f – 1) grados de libertad para el valor alfa preseleccionado por el investigador.

También se puede emplear el chi-cuadrado con corrección de Yates (χ^2_C) según la fórmula:

$$\chi^2_C = \sum_{i,j} \frac{(X_{ij} - e_{ij} - 0{,}5)^2}{e_{ij}}$$

donde:

$X_{i,j}$: valor observado para cada celda;

$e_{i,j}$: valor esperado para cada celda = (total de la fila · total de columna correspondiente para cada celda) / tamaño de muestra.

Para recordar:

A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo χ^2 exp el valor observado-calculado del estadístico de prueba o contraste y χ^2 teo el valor teórico-esperado de este estadístico con grados de libertad (c – 1) (f – 1) y nivel de confianza 1 – α, entonces:

- Intervalo donde no se rechaza H_0 = (para cualquier valor de χ^2 exp ≤ χ^2 teo $_{(c-1)(f-1);\,1-\alpha.}$
- Región crítica (donde se rechaza H_0) = (para cualquier valor de χ^2 exp > χ^2 teo $_{(c-1)(f-1);\,1-\alpha.}$
- Valor de p o p-valor bajo H_0:
  p-valor = Pr[χ^2 (teórico-esperado) ≥ t-student (observado-calculado)].

Regla de decisión

- Si χ^2 (observado-calculado) > χ^2 (esperado-teórico-tabulado), esto se corresponde con la ubicación de χ^2 observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que disminuye la probabilidad de cometer el error tipo 1, entonces hay evidencias suficientes para rechazar H_0, por ende, se acepta la alternativa, es decir, hay una relación de dependencia estadísticamente significativa entre las dos variables, siendo las probabilidades relativas al azar o el p-valor bajo H_0 para estos resultados inferiores a las esperadas por el modelo teórico de referencia (p < α).
- Si χ^2 (observado-calculado) ≤ χ^2 (esperado-teórico-tabulado), esto se corresponde con la no ubicación de χ^2 observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que aumenta la probabilidad de cometer el error tipo 1, entonces no hay evidencias suficientes para rechazar H_0, es decir, hay una relación de independencia estadísticamente significativa entre las dos variables, siendo las probabilidades relativas al azar o el p-valor bajo H_0 para estos resultados iguales o superiores inferiores a las esperadas por el modelo teórico de referencia (p ≥ α).

Otra medida de asociación entre variables cualitativas a emplear, derivada del cálculo de χ^2:

- Coeficiente de contingencia (C):

  $C = \left(\frac{\chi^2}{\chi^2 + T} \right)^{\frac{1}{2}}$ cuanto más cercano al valor 1, mayor es el grado de asociación entre las variables estudiadas;

  donde:

  T: suma de los totales por filas y columnas en la tabla de contingencia empleada;

  En caso de coincidir el número de filas y columnas se puede calcular el coeficiente de correlación de atributos.

- Coeficiente de correlación de atributos (r_A):

  $r_A = \left[\frac{C^2}{T(k-1)} \right]^{\frac{1}{2}}$

  donde:

  k: número de filas o columnas;

  C: coeficiente de contingencia;

  T: suma de totales de columnas y filas.

Si se trata de un análisis estratificado que ajustara la asociación entre las dos variables, según "la tercera variable" o factor de estratificación, puede utilizarse el correspondiente estadístico de prueba de Mantell-Haenzsel (promedio del chi-cuadrado individual de cada tabla, o sea, cada parte del estrato).

Algunos aspectos para recordar de situaciones a enfrentar en este análisis:

- La asociación en estudio puede ser significante antes y después de ajustar para una tercera variable sin cambios en la significación estadística de la asociación entre variables (no hay probabilidad de un confusor).
- La asociación puede permanecer significante después de ajustar para la tercera variable; pero con una menor significación del valor de p (no es un buen confusor).
- La asociación en estudio puede ser significante antes de ajustar para una tercera variable y no significante después (hay probablemente un confusor).
- La asociación en estudio al introducir el probable confusor muestra resultados opuestos o aparece mayor asociación entre las variables (hay interacción del probable confusor con la variable independiente para el efecto estudiado).

Para recordar

El empleo de tablas de contingencia es el método graficado para representar los datos de las variables cualitativas y guardan relación con el tipo de prueba estadística o estadístico a emplear para el análisis de asociaciones significativas entre estas mismas:

- Para tablas de 2x2 simples (dos variables cualitativas nominales dicotómicas):
  - Prueba de asociación *versus* independencia por chi-cuadrado sin corrección (Pearson).
  - Prueba de asociación *versus* independencia por chi-cuadrado con corrección (Yates).
  - Prueba de la probabilidad exacta de Fisher.
- Para tablas de 2x2 estratificadas (2 variables cualitativas nominales dicotómicas estratificadas por una tercera:
  - Prueba de asociación de Mantel-Haenszel.
  - Prueba de nomogeneidad entre estratos.
- Para tablas de 2xN simples (1 variable dependiente cualitativa nominal dicotómica y otra independiente cualitativa con ordinalidad por niveles (N):
  - Prueba de tendencia lineal entre los niveles de la variable independiente para el evento en la dependiente.
  - Prueba de homogeneidad entre niveles.
- Para tablas de 2xN estratificadas (lo mismo al anterior, pero estratificado por una tercera variable nominal o factor de estratificación):
  - Prueba de tendencia lineal entre los niveles de la variable independiente para el evento en la dependiente para cada estrato.
  - Prueba de homogeneidad entre niveles para cada estrato.
- Para tablas de NxM con N (número de filas) > 2 nominaciones; M (número de columnas) > 2 nominaciones para variables dependiente e independiente: nominales politómicas:
  - Prueba de independencia empleando dos variables cualitativas con dos o más categorías o nominaciones.
  - Prueba de la razón de verosimilitudes para datos nominales.
  - Prueba de tendencia lineal para datos ordinales (por niveles).

p) Una muestra, dos variables cualitativas dicotómicas en un diseño de estudio de datos no pareados, si se desea determinar reproducibilidad, concordancia o acuerdo en el comportamiento de esas variables cualitativas dicotómicas, por ejemplo, dos pruebas con resultado dicotómico, empleando el índice de acuerdo o concordancia: kappa (por prueba de Cohen) y empleo del estadístico de contraste o de prueba Z ~ N (0, 1) en la evaluación de la significación estadística del resultado obtenido para un valor de alfa prefijado por el nivel de confianza previamente seleccionado.

   El posible planteamiento de hipótesis estadísticas para contrastar es:
   - H_0: no existe concordancia o acuerdo entre los resultados de las dos pruebas a comparar o dos variables cualitativas dicotómicas.
   - H_1: existe concordancia o acuerdo entre los resultados de las dos pruebas a comparar o dos variables cualitativas dicotómicas.

Pasos a seguir:

- Se construye una tabla de contingencia 2x2 para cada variable cualitativa dicotómica o prueba con resultado dicotómico: cantidad de sujetos con respuesta: si = (celda a) y cantidad de sujetos con respuesta: no = (celda d, se suman ambos valores y se obtiene la concordancia simple (Cs).
- Se dividen todos los valores de las celdas de la tabla 2x2 incluyendo los totales marginales entre el número total de sujetos (a + b + c + d).
- Se calcula la suma de la diagonal en que concuerdan los resultados para celdas a y d, su valor es la concordancia observada (p_o).
- Se calcula la concordancia esperada total por el azar (p_e) a partir de: la multiplicación de los totales marginales para la celda (a) más la multiplicación de los totales marginales de la celda (d).
- Se calcula el índice de acuerdo o concordancia kappa como:

$$\text{Índice kappa (k)} = \frac{p_o - p_e}{1 - p_e}$$

Para recordar

- Si: índice kappa > 0,8: excelente la concordancia.
- Si: índice kappa > 0,6: buena la concordancia.
- Si: índice kappa > 0,4: aceptable la concordancia.

  Si se desea realizar el contraste de hipótesis de forma unilateral (un solo lado) en la evaluación de la significación del índice kappa, el planteamiento posible de hipótesis es:

  H_0: K = 0 *vs.* H_1: K > 0.

  La significación estadística para la concordancia a partir del índice kappa es evaluada a partir de la comparación del estadístico de contraste Z ~ N (0, 1) bajo H_0:

$$\text{Z (observado-calculado)} = \frac{K}{se(K)}$$

  donde:

$$se(K) = \sqrt{\frac{1}{n(1-p_e)^2} \cdot \left\{ p_e + p_e^{\,2} - \sum_{i=1}^{c} a_i b_i (a_i + b_i) \right\}}$$

- Valor de p (p-valor), bajo H_0:

  p-valor = Pr [$Z_{1-\alpha}$ (teórico-esperado) ≥ Z (observado-calculado)].

  El p-valor exacto estará dado por:

  $p = 1 - \Phi(z)$

  donde:

Φ (z): valor de probabilidad que deja en una cola a la izquierda el percentil Z de la distribución teórica normal estándar N (0, 1) correspondiente.

Regla de decisión

- Para Z (observado-calculado) < 0:
  - Si Z (observado-calculado) < Z (teórico-esperado) según el valor de alfa prefijado, esto se corresponde con la ubicación de Z calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teóricode referencia, lo que disminuiye la probabilidad de cometer el error tipo 1, entonces hay evidencias suficientes para rechazar H_0 y, por ende, se acepta H_1 (la hipótesis alternativa), o sea, es estadísticamente significativa la concordancia o acuerdo entre las variables cualitativas dicotómicas o entre las pruebas empleadas a partir del valor del índice kappa, siendo la probabilidad relacionada con el azar o el p-valor bajo H_0 para estos resultados inferior a la esperada del modelo teórico de referencia ($p < \alpha$).
  - Si Z (observado-calculado) ≥ Z (teórico-esperado) según el valor de alfa prefijado, no se puede rechazar H_0.
- Para Z (observado-calculado) ≥ 0:
  - Si Z (observado-calculado) > Z (teórico-esperado) según el valor de alfa prefijado, esto se corresponde con la ubicación de Z calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que disminuye la probabilidad de cometer el error tipo 1, entonces hay evidencias suficientes para rechazar H_0 y, por ende, se acepta H_1 (la hipótesis alternativa), o sea, es estadísticamente significativa la concordancia o acuerdo entre las variables cualitativas dicotómicas o entre las pruebas empleadas a partir del valor del índice kappa, siendo la probabilidad relacionada con el azar o el p-valor bajo H_0 para estos resultados inferior a la esperada del modelo teórico de referencia ($p < \alpha$).
  - Si Z (observado-calculado) ≤ Z (teórico-esperado) según el valor de alfa prefijado, no se puede rechazar H_0.

q) Una muestra y dos variables cualitativas, la variable independiente expresa la presencia de exposición o no y la dependiente como tasa de incidencia (TI) que es el número de casos nuevos o eventos según el número de personas-tiempo de seguimiento, si se desea evaluar la significación de la asociación de la exposición con la tasa de incidencia, se emplea la prueba de Z por la aproximación de la normal estándar Z ~ N (0, 1) para un nivel de significación alfa predeterminado, el posible planteamiento de hipótesis estadísticas a contrastar es:

H_0: TI exp = TI no exp *vs.* H_1: TI exp ≠ TI no exp

El estadístico de contraste o prueba a emplear (se emplea solo si V ≥ 5) se expresa como:

$$Z\ (\text{observado-calculado}) = \begin{cases} \dfrac{a - E - \dfrac{1}{2}}{\sqrt{V}} & \text{si } a > E \\ \dfrac{a - E + \dfrac{1}{2}}{\sqrt{V}} & \text{si } a \leq E \end{cases}$$

donde:

$E = \dfrac{nT_1}{T}$;

$V = \dfrac{nT_1T_2}{T^2}$;

$p = \dfrac{T_1}{T}$;

a: número de casos expuestos;

n: total de casos (expuestos más no expuestos);

T_1: número de personas expuestas por el tiempo de seguimiento;

T_2: número de personas no expuestas por el tiempo de seguimiento;

T: es el total de personas por el tiempo de seguimiento.

- Valor de p (p-valor) bajo H_0:

$$\text{p-valor} = \begin{cases} 2\Phi(z) & \text{si} \quad z < 0 \\ 2\left[1 - \Phi(z)\right] & \text{si} \quad z \geq 0 \end{cases}$$

donde:

Φ (Z): se refiere al valor de probabilidad que deja en una cola a la izquierda el percentil Z de la distribución teórica normal estándar N (0, 1) correspondiente.

Regla de decisión

- Para Z (observado-calculado) < 0:
  - Si Z (observado-calculado) < Z (teórico-esperado) según valor de alfa prefijado, esto se corresponde con la ubicación de Z calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que disminuye la probabilidad de cometer el error tipo 1, entonces hay evidencias suficientes para rechazar H_0 y, por ende, se acepta H_1 (la hipótesis alternativa), o sea, es estadísticamente significativa la asociación entre la exposición y la tasa de incidencia, siendo la probabilidad relacionada con el azar o el p-valor bajo H_0 para estos resultados inferior a la esperada del modelo teórico de referencia ($p < \alpha$).
  - Si Z (observado-calculado) ≥ Z (teórico-esperado) según el valor de alfa prefijado no se puede rechazar H_0.

– Para Z (observado-calculado) ≥ 0:
  - Si Z (observado-calculado) > Z (teórico-esperado) según el valor de alfa prefijado, esto se corresponde con la ubicación de Z calculado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que disminuye la probabilidad de cometer el error tipo 1, entonces hay evidencias suficientes para rechazar H_0 y, por ende, se acepta H_1 (la hipótesis alternativa), o sea, es estadísticamente significativa la asociación entre la exposición y la tasa de incidencia, siendo la probabilidad relacionada con el azar o el p-valor bajo H_0 para estos resultados inferior a la esperada del modelo teórico de referencia ($p < \alpha$).
  - Si Z (observado-calculado) ≤ Z (teórico-esperado) según el valor de alfa prefijado, no se puede rechazar H_0.

r) Una muestra con dos tomas de valores de una variable cuantitativa en diferentes momentos (se comporta como dos muestras emparejadas) cuya diferencia de los valores de las observaciones por pares (d_i) constituye una variable aleatoria que sigue o se aproxima su distribución de frecuencia a una normal estándar y se desea comprobar si la media muestral de las diferencias de los valores de la variable cuantitativa para las observaciones en parejas ($\bar{d}$) que estima una media de las diferencias pareadas poblacional μd significativa, se utiliza una prueba de t-student pareada en la evaluación de la significación estadística del resultado para n – 1 grados de libertad y nivel de significación alfa predeterminado; siendo el posible planteamiento de hipótesis estadísticas para el contraste:

– Para contraste bilateral:

$H_0: \mu_d = \Delta$ *vs.* $H_1: \mu_d \neq \Delta$

$H_0: \mu_d = 0$ *vs.* $H_1: \mu_d \neq 0$.

– Para contraste unilateral:

- $\begin{cases} H_0: \mu_d = \Delta \\ H_1: \mu_d < \Delta \end{cases}$ o $\begin{cases} H_0: \mu_d \geq \Delta \\ H_1: \mu_d < \Delta \end{cases}$

  $H_0: \mu_d = 0$ *vs.* $H_1: \mu_d < 0$ o $H_0: \mu_d \geq 0$ *vs.* $H_1: \mu_d < 0$.

- $\begin{cases} H_0: \mu_d = \Delta \\ H_1: \mu_d > \Delta \end{cases}$ o $\begin{cases} H_0: \mu_d \leq \Delta \\ H_1: \mu_d > \Delta \end{cases}$

  $H_0: \mu_d = 0$ *vs.* $H_1: \mu_d > 0$ o $H_0: \mu_d \leq 0$ *vs.* $H_1: \mu_d > 0$.

– Como estadístico de prueba o contraste se emplea t-student pareada adoptando esta relación:

$$\text{t-student pareada (observada-calculada)} = \frac{\bar{d} - \mu_d}{\frac{S_{\bar{d}}}{\sqrt{n}}} \sim t_{n-1}$$ sigue una distribución teórica t-student con n – 1 grados de libertad para el nivel de significación alfa según el nivel de confianza preseleccionado.

donde:

$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{\sum_{i=1}^{n} (x_i - x_2)}{n}$: media muestral de todas las diferencias (d_i) entre las dos observaciones o las parejas de datos de las dos muestras pareadas;

n: número de parejas de datos u observaciones por parejas de datos;

μ_d: valor de la media poblacionalque bajo la hipótesis nula al realizar el contraste adquiere el valor de 0 en la expresión antes mencionada para el cálculo de estadístico de prueba (t-student);

$S_d^2 == \frac{\sum_{i=1}^{n} (d_i - \bar{d})}{n-1}$: cuasivarianza para las diferencias de todas las observaciones pareadas correspondientes a las muestras pareadas que estimaría la varianza poblacional para los datos pareados.

Otra manera de plantear la expresión de la t-student pareada es:

$$\frac{\bar{d} - \Delta}{\frac{1}{\sqrt{n}} \hat{S}_d} \sim t_{n-1}$$

donde:

$\mu_d = \Delta$: media poblacional estimada de la diferencia de las parejas de observaciones con los datos de las muestras pareadas ($\bar{d}$).

Para recordar

A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo Ti el valor inferior del estadístico de prueba o contraste y Ts el valor superior de dicho estadístico empleado para el contraste de hipótesis, entonces:

- Para contraste bilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, T_s)$:
    Región crítica = $(-\infty, Ti] \cup [T_s, +\infty)$ donde se rechaza H_0.
- Para contraste unilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, +\infty)$:
    Región crítica = $(-\infty, T_i)$ donde se rechaza H_0.
  - Intervalo donde no se rechaza $H_0 = (-\infty, T_S)$:
    Región crítica = $(T_S, +\infty)$ donde se rechaza H_0.
  - Valor de p o p-valor bajo H_0:
    p-valor = Pr[t-student (teórico-esperado) ≥ t-student (observado-calculado)].

Regla de decisión

La regla de decisión respecto a la significación estadística al contrastar las hipótesis (nula y alternativa) estará determinada por:

– Para t-student (observado-calculado) < 0:
  - Si t-student (observado-calculado) < t-student (teórico-esperado) según los grados de libertad y el valor de alfa correspondientes, entonces t-student observado se ubica en la región de rechazo para H_0, por lo que existe evidencia suficiente para rechazar hipótesis nula y se acepta la alternativa, por lo que las diferencias entre las medias poblacionales estimadas para pareadas son estadísticamente significativas, al determinarse una probabilidad relativa al azar (alietoriedad de los datos) y una probabilidad de cometer el error tipo 1 menor que la probabilidad alfa para el nivel de confianza seleccionado ($p < \alpha$).
  - Si t-student (observado-calculado) ≥ t-student (teórico-esperado) según los grados de libertad y el valor de alfa correspondientes, entonces t-student observado no se ubica en la región de rechazo para H_0, por lo que no existe evidencia suficiente para rechazar hipótesis nula y no se acepta la alternativa, por lo que las diferencias entre las medias poblacionales estimadas pareadas no son estadísticamente significativas, al determinarse una probabilidad relativa al azar (alietoriedad de los datos) y una probabilidad de cometer el error tipo 1 igual o mayor que la probabilidad de alfa para el nivel de confianza seleccionado ($p \geq \alpha$).

– Para t-student (observado-calculado) ≥ 0:
  - Si t-student (observado-calculado) > t-student (teórico-esperado) según los grados de libertad y el valor de alfa correspondientes, entonces t-student observado se ubica en la región de rechazo para H_0, por lo que existe evidencia suficiente para rechazar hipótesis nula y se acepta la alternativa, por lo que las diferencias entre las medias poblacionales estimadas pareadas son estadísticamente significativas, al determinarse una probabilidad relativa al azar (alietoriedad de los datos) y una probabilidad de cometer el error tipo 1 menor que la probabilidad alfa para el nivel de confianza seleccionado ($p < \alpha$).
  - Si t-student (observado-calculado) ≤ t-student (teórico-esperado) según los grados de libertad y el valor de alfa correspondientes, entonces t-student observado no se ubica en la región de rechazo para H_0, por lo que no existe evidencia suficiente para rechazar hipótesis nula y no se acepta la alternativa, por lo que las diferencias entre las medias poblacionales estimadas pareadas no son estadísticamente significativas, al determinarse una probabilidad relativa al azar (alietoriedad de los datos) y una probabilidad de cometer el error tipo 1 igual o mayores que la probabilidad de alfa para el nivel de confianza seleccionado ($p \geq \alpha$).

s) Dos muestras pareadas, una variable cualitativa dicotómica que sigue una distribución de frecuencias binomial, si se desea determinar la homogeneidad en las poblaciones estimadas de dos muestras pareadas, por ejemplo, de un caso por un control, para las proporciones formadas se emplea el estadístico de prueba o contraste chi-cuadrado deMcNemar en la evaluación de la significación estadística del resultado obtenido para un valor de alfa prefijado y grados de libertad correspondientes. El posible planteamiento de hipótesis estadísticas para contrastar sería:

- H_0: existe homogeneidad en ambas poblaciones pareadas para las proporciones estimadas según la variable cualitativa dicotómica.
- H_1: no existe homogeneidad en ambas poblaciones pareadas para las proporciones estimadas según la variable cualitativa dicotómica.

  El estadístico de prueba a emplear es chi-cuadrado de McNemar que en estudios de casos y controles (muestras pareadas) para tablas de contingencia 2x2 simple a partir de la nomenclatura de las respectivas celdas (acordada entre los diferentes autores) se expresa como:

$$\chi c^2_{(McNemar)}(\text{observado-calculado}) = \frac{n_{10} - n_{01}}{n_{10} + n_{01}}$$

  o también si se utiliza otra nomenclatura para las celdas en las tablas para representar los datos de este tipo de estudio como:

$$\chi c^2_{(McNemar)}(\text{observado-calculado}) = \frac{(b - c)^2}{b + c}$$

  que sigue una distribución teórica de probabilidades chi-cuadrado $(\chi^2_{(f-1)(c-1);\,1-\alpha})$ con (f – 1) (c – 1) grados de libertad para el valor de alfa prefijado.

Para recordar

Respecto a la significación estadística de la relación entre variables a partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo χ^2 exp (McNemar) el valor observado-calculado del estadístico de prueba o contraste y χ^2 teo el valor teórico-esperado de este estadístico con grados de libertad (c – 1) (f – 1) y nivel de confianza 1 – α, entonces:

- Intervalo donde no se rechaza H_0 = (para cualquier valor de χ^2 exp ≤ χ^2 teo $_{(c-1)(f-1);\,1-\alpha}$.
  Región crítica (donde se rechaza H_0) = (para cualquier valor de χ^2 exp > χ^2 teo $_{(c-1)(f-1);\,1-\alpha}$.
  donde:
  χ^2 exp: chi-cuadrado experimental o calculado-observado;

χ^2 teo: chi-cuadrado teórico, esperado, tabulado o de referencia.

– Valor de p o p-valor bajo H_0:
p-valor = Pr [χ^2 (teórico-esperado) ≥ χ^2 de McNemar (observado-calculado)].

Regla de decisión

La regla de decisión respecto a la significación estadística al contrastar las hipótesis (nula y alternativa) estará determinada por:

– Si χ^2 (observado-calculado) de McNemar> χ^2 (esperado-tabulado) con el valor de alfa y grados de libertad correspondientes para el caso, esto se corresponde con la ubicación del valor de chi-cuadrado McNemar observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teóricode referencia, lo que disminuye la probabilidad de cometer el error tipo 1, entonces hay evidencias suficientes para rechazar H_0, es decir, que no hay homogeneidad en las poblaciones estimadas de las muestrales pareadas (de casos y controles) para las proporciones estimadas según la variable cualitativa (dicotómica) empleada, siendo las probabilidades relativas al azar o el p-valor bajo H_0 para dichos resultados, inferiores a las esperadas para el valor crítico por el modelo teórico de referencia ($p < \alpha$).

– Si χ^2 (observado-calculado) de McNemar ≤ χ^2 (esperado-tabulado) con el valor de alfa y grados de libertad correspondientes para el caso; esto no se corresponde con la ubicación del valor de chi-cuadrado McNemar observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que no disminuye la probabilidad de cometer el error tipo 1, entonces no hay evidencias suficientes para rechazar H_0, es decir, que hay homogeneidad en las poblaciones estimadas de las muestrales pareadas (de casos y controles) para las proporciones estimadas según la variable cualitativa (dicotómica) empleada, siendo las probabilidades relativas al azar o el p-valor bajo H_0 para estos resultados superiores o iguales a las esperadas para el valor crítico por el modelo teórico de referencia ($p \geq \alpha$).

t) Dos muestras pareadas (con un caso y más de un control) y una variable cualitativa dicotómica se desea determinar la homogeneidad en las poblaciones estimadas de las muestras pareadas para las proporciones estimadas, respecto a la categoría presente (de alguna variable o factor de riesgo como la exposición) se puede emplear la prueba de Mantel-Haenszel. El planteamiento de hipótesis estadísticas para contrastar sería:

– H_0: existe homogeneidad en las poblaciones pareadas (más de un control por caso) para las proporciones estimadas según la variable cualitativa dicotómica.

– H_1: no existe homogeneidad en las poblaciones pareadas (más de un control por caso) para las proporciones estimadas según la variable cualitativa dicotómica.

Se tiene la representación de los datos (Tabla 4.12).

Tabla 4.12. Representación tabulada de los elementos a emplear en la prueba de Mantell-Haenszel (para probar la homogeneidad en las muestras pareadas con más de un control por caso)

Combinaciones de caso-controles	Categorías de la variable dicotómica para controles			
	Número de casos: con factor = 1 sin factor = 0	Número de Controles con factor presente (X_i) (desde 0 hasta m – 1)	Total de sujetos que tienen el factor presente (n_i) (desde 0 hasta m)	$(n_i)^2$
1				
2				
.....	.....	.....	.....	.....
N				
Total por columnas	∑B – A	∑A	∑B	$\sum (n_i)^2$

El estadístico de contraste (de prueba) es:

$$\chi^2_{\text{(Mantel-Haenzel)}}(\text{observado-calculado}) = \left[\frac{p_2 - p_1}{\text{s.e.}(p_2 - p_1)}\right]^2 = \frac{\left[(m-1)B - mA\right]^2}{mB - \sum_{i=1}^{N} n_i^2}$$

que sigue una distribución teórica de chi-cuadrado con un grado de libertad para el nivel de significación alfa prefijado.

donde:

$A = \sum_{i=1}^{N} x_i$: número total de sujetos (controles) que tienen el factor presente o categoría positiva de la variable;

$B = \sum_{i=1}^{N} n_i$: número total de sujetos (casos y controles) que tiene el factor presente o categoría positiva de la variable;

B – A: número de casos que tiene el factor presente;

$p_1 = \frac{A}{N(m-1)}$: proporción para la cual el factor está presente entre los controles;

$p_2 = \frac{B-A}{N}$: proporción en la cual el factor está presente entre los casos;

m: número de constituyentes de la combinación caso-controles;

m – 1: número de controles por caso a emplear;

N: número total de combinaciones de caso-controles (el número total de observaciones);

s.e.: error estándar para la diferencia de $p_2 - p_1$.

– Valor p (p-valor), bajo H_0:
p-valor = $Pr[\chi^2_{1,\ 1-\alpha}$ (teórico – esperado) ≥ $\chi^2_{\text{(Mantel-Haenszel)}}$ (observado-calculado)].

Regla de decisión

– Si $\chi^2_{\text{(Mantel-Haenszel)}}$ (observado-calculado) > χ^2 (esperado-tabulado) con 1 grado de libertad y el valor de alfa prefijado, entonces el valor de chi-cuadrado observado de Mantel-Haenszel se ubica en el área de rechazo de H_0, por lo que existen evidencias suficientes para rechazar la hipótesis nula (H_0) y se acepta la alternativa (H_1), es decir, no existe homogeneidad significativa en las poblaciones estimadas de las muestras pareadas para las proporciones calculadas según la variable dicotómica, siendo las probabilidades del azar o el p-valor bajo H0 para estos resultados menores que el valor de alfa prefijado según el nivel de confianza escogido ($p < \alpha$).

– Si $\chi^2_{\text{(Mantel-Haenszel)}}$ (observado-calculado) ≤ χ^2 (esperado-tabulado) con 1 grado de libertad y el valor de alfa prefijado, entonces el valor de chi-cuadrado observado de Mantel-Haenszel no se ubica en el área de rechazo de H_0, por lo que no existen evidencias suficientes para rechazar la hipótesis nula (H_0) y no se acepta la alternativa (H_1), es decir, existe homogeneidad significativa en las poblaciones estimadas de las muestrales pareadas para las proporciones calculadas en caso y controles (más de 1) según la variable dicotómica la variable dicotómica, siendo las probabilidades del azar o el p-valor bajo H_0 para estos resultados, mayores o iguales que el valor de alfa prefijado según el nivel de confianza escogido ($p \geq \alpha$).

u) Dos muestras pareadas y una variable cualitativa politómica, si se desea determinar la homogeneidad en las poblaciones estimadas de las muestrales de las proporciones estimadas a partir de una variable politómica según sus categorías (en este caso 3 de ellas) se puede emplear la prueba de Stuart-Maxwell. El posible planteamiento de hipótesis estadísticas para el contraste sería:

– H_0: existe homogeneidad en las poblaciones estimadas de las muestrales pareadas para las proporciones estimadas según la variable cualitativa politómica empleada.

– H_1: no existe homogeneidad en las poblaciones estimadas de las muestrales pareadas para las proporciones estimadas según la variable cualitativa politómica empleada.

Para realizar el contraste de hipótesis se requiere contar con los elementos teóricos necesarios presentados en forma tabulada y se procede al llenado de sus celdas con los datos respectivos procedentes de la información recogida (Tabla 4.13).

Tabla 4.13. Representación tabulada de los elementos teóricos necesarios a emplear en la prueba de Stuart-Maxwell

Categorías de la variable para casos	Categorías de la variable politómica para controles			
	Categoría 1	Categoría 2	Categoría 3	Total
Categoría 1	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$	
Categoría 2	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$	
Categoría 3	$n_{3,1}$	$n_{3,2}$	$n_{3,3}$	
Total				

El estadístico para el contraste (de prueba) sería:

$$\chi^2_{\text{(Stuart-Maxwell)}}(\text{observado-calculado}) = \frac{\bar{n}_{23}d_1^2 + \bar{n}_{13}d_2^2 + \bar{n}_{12}d_3^2}{2(\bar{n}_{12}\bar{n}_{13} + \bar{n}_{12}\bar{n}_{23} + \bar{n}_{13}\bar{n}_{23})}$$

que sigue una distribución teórica de chi-cuadrado con 2 grados de libertad para el nivel de significación alfa prefijado.

donde:

$\bar{n}_{ij} = \frac{n_{ij} + n_{ji}}{2}$: mitad de la suma de los valores para la misma categoría de la variable politómica en ambas muestras (casos y controles), representados por las filas (i) y las columnas (j);

d_n: diferencia entre el total de fila (i) y el total de columna (j), para el mismo orden numérico de i y de j.

– Valor p (p-valor), bajo H_0:
p-valor = Pr [($\chi^2_{2,1-\alpha}$ (teórico-esperado) ≥ $\chi^2_{\text{(Stuart-Maxwell)}}$ (observado-calculado)].

Regla de decisión

– Si $\chi^2_{\text{(Stuart-Maxwell)}}$ (observado-calculado) > $\chi^2_{2;\,1-\alpha}$ (esperado-tabulado), con dos grados de libertad y el valor alfa prefijado, entonces el valor de chi-cuadrado observado de Stuart-Maxwell se ubica en el área de rechazo de H_0, por lo que existen evidencias suficientes para rechazar la hipótesis nula (H_0) y se acepta la alternativa (H_1), es decir, no existe homogeneidad en las poblaciones estimadas pareadas para las proporciones calculadas según la variable politómica, estimándose un p-valor para el chi-cuadrado observado bajo H_0 o siendo las probabilidades del azar menores que el nivel de significación alfa o de cometer el error de tipo 1, según el nivel de confianza escogido ($p < \alpha$).

– Si $\chi^2_{\text{(Stuart-Maxwell)}}$ (observado-calculado) ≤ $\chi^2_{2,\,1-\alpha}$ (esperado-tabulado), con dos grados de libertad y valor alfa prefijado, entonces el valor de chi-cuadrado observado de Stuart-Maxwell no se ubica en el área de rechazo de H_0, por lo

que no existen evidencias suficientes para rechazar la hipótesis nula (H_0) y no se acepta la alternativa (H_1), es decir, existe homogeneidad significativa en las poblaciones estimadas pareadas para las proporciones calculadas según la variable politómica, estimándose un p-valor para el chi-cuadrado observado bajo H_0 o siendo las probabilidades del azar mayores o iguales que el nivel de significación alfa o de cometer el error de tipo 1, según el nivel de confianza escogido ($p \geq \alpha$).

v) Dos muestras no pareadas y una variable cuantitativa aleatoria que sigue o se aproxima a la distribución normal estándar, si se desea determinar la diferencia de dos medias poblacionales estimadas no pareadas con desviaciones estándar poblacionales conocidas para muestras grandes ($n \geq 30$), se emplea el estadístico de prueba $Z \sim N(0, 1)$. El posible planteamiento de hipótesis estadísticas para el contraste sería:

- Para contraste bilateral:

  $H_0: \mu_1 - \mu_2 = 0$ *vs.* $\mu_1 - \mu_2 \neq 0$.

- Para contraste unilateral:
  - $H_0: \mu_1 - \mu_2 = 0$ *vs.* $H_1: \mu_1 - \mu_2 > 0$.

    $H_0: \mu_1 - \mu_2 = 0$ *vs.* $H_1: \mu_1 - \mu_2 > 0$.
  - $H_0: \mu_1 - \mu_2 \geq 0$ *vs.* $H_1: \mu_1 - \mu_2 < 0$

    $H_0: \mu_1 - \mu_2 \leq 0$ *vs.* $H_1: \mu_1 - \mu_2 > 0$

- Como estadístico de prueba se emplea Z que adopta esta relación:

$$Z\ (\text{observado-calculado}) = \frac{\bar{x} - \bar{y} - d_0}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

donde:

d_0: estima la diferencia entre las medias poblacionales empleadas de referencia para contrastar. Si $d_0 = 0$ se asume que la diferencia entre las medias poblacionales de referencia para el contraste (bajo H_0) es igual a 0, procedimiento que debemos asumir a la hora de realizar el cálculo del estadístico de prueba:

$\bar{x}$: media de la muestra 1;

$\bar{y}$: media de la muestra 2;

σ_x^2 : varianza de la muestra 1 que estima una poblacional;

σ_y^2 : varianza de la muestra 2 que estima una poblacional;

n_x: tamaño de la muestra 1;

n_y: tamaño de la muestra 2.

Para recordar

A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo Ti el valor inferior del estadístico de

prueba o contraste y Ts el valor superior de dicho estadístico empleado para el contraste de hipótesis, entonces:

- Para contraste bilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, T_s)$:
    Región crítica = $(-\infty, T_i] \cup [T_s, +\infty)$ donde se rechaza H_0.
- Para contraste unilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, +\infty)$:
    Región crítica = $(-\infty, Ti)$ donde se rechaza H_0.
  - Intervalo donde no se rechaza $H_0 = (-\infty, TS)$:
    Región crítica = $(T_S, +\infty)$ donde se rechaza H_0.
- Valor de p o p-valor bajo H_0:
  p-valor = Pr [Z (teórico-esperado) según alfa ≥ Z (observado-calculado)].

Regla de decisión

La regla de decisión respecto a la significación estadística al contrastar las hipótesis (nula y alternativa) está determinada por:

- Para Z (observado-calculado) < 0:
  - Si Z (observado-calculado) < Z (teórico-esperado) según el valor de alfa prefijado, entonces el valor de Z observado es ubicado en el área de rechazo para H_0 por lo que hay evidencia suficiente para rechazar H_0 y aceptar la alternativa H_1, siendo significativas las diferencias entre las medias de la variable para las dos poblaciones independientes al estimarse las probabilidades (p-valor) bajo H_0 o del azar para el valor del Z observado, menores que el valor alfa prefijado según el nivel de confianza escogido ($p < \alpha$).
  - Si Z (observado-calculado) ≥ Z (teórico-esperado) según el valor de alfa prefijado, entonces el valor de Z observado no es ubicado en el área de rechazo para H0 por lo que no hay evidencia suficiente para rechazar H_0 y aceptar la alternativa H_1, siendo no significativas las diferencias entre las medias de la variable para las dos poblaciones independientes al estimarse las probabilidades (p-valor) bajo H_0 o del azar para el valor del Z observado iguales o mayores que el valor alfa prefijado según el nivel de confianza escogido ($p \geq \alpha$).
- Para Z (observado-calculado) ≥ 0:
  - Si Z (observado-calculado) > Z (teórico-esperado) según el valor de alfa prefijado, entonces el valor de Z observado es ubicado en el área de rechazo para H_0 por lo que hay evidencia suficiente para rechazar H_0 y aceptar la alternativa H_1, siendo significativas las diferencias entre las medias de la variable para las dos poblaciones independientes al estimarse las probabilidades (p-valor) bajo H_0 o del azar para el valor del Z observado menores que el valor alfa prefijado según el nivel de confianza escogido ($p < \alpha$).

- Si Z (observado-calculado) ≤ Z (teórico-esperado) según el valor de alfa prefijado, entonces el valor de Z observado no es ubicado en el área de rechazo para H_0 por lo que no hay evidencia suficiente para rechazar H_0 y aceptar la alternativa H_1, siendo no significativas las diferencias entre las medias de la variable para las dos poblaciones independientes al estimarse las probabilidades (p-valor) bajo H_0 o del azar para el valor del Z observado iguales o mayores que el valor alfa prefijado según el nivel de confianza escogido ($p \geq \alpha$).

w) Dos muestras no pareadas y una variable cuantitativa aleatoria que sigue o se aproxima a la distribución normal estándar, si se desea comparar dos medias poblacionales estimadas a partir de las medias muestrales calculadas con los datos de la variable cuantitativa para tamaños de muestras pequeños (n < 30) y varianzas poblacionales desconocidas, pero supuestamente iguales. El posible planteamiento de hipótesis estadísticas para el contraste sería:

– Para contraste bilateral:
H_0: $\mu_1 - \mu_2 = 0$ *vs.* $\mu_1 - \mu_2 \neq 0$.

– Para contraste unilateral:
- H_0: $\mu_1 - \mu_2 = 0$ *vs.* H_1: $\mu_1 - \mu_2 > 0$.
  H_0: $\mu_1 - \mu_2 = 0$ *vs.* H_1: $\mu_1 - \mu_2 > 0$.
- H_0: $\mu_1 - \mu_2 \geq 0$ *vs.* H_1: $\mu_1 - \mu_2 < 0$.
  H_0: $\mu_1 - \mu_2 \leq 0$ *vs.* H_1: $\mu_1 - \mu_2 > 0$.

– Como estadístico de prueba se emplea t-student adoptando la expresión que les relaciona:

$$\text{t-student (observado-calculado)} = \frac{(x_1 - x_2)(\mu_1 - \mu)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

que sigue una distribución teórica de probabilidades t-student con $(n_1 + n_2) - 2$ grados de libertad y para un valor de alfa predeterminado.
donde:
$\bar{x}_1$: media de la muestra 1;
$\bar{x}_2$: media de la muestra 2;
$(\mu_1 - \mu)$: equivale a plantear $d_0 = 0$, si se asume que la diferencia entre las medias poblacionales de referencia para el contraste (bajo H_0) es igual a 0, procedimiento que se debe asumir al realizar el cálculo del estadístico de prueba;
n_1: tamaño de la muestra 1;
n_2: tamaño de la muestra 2;

$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$: cuasivarianza muestral ponderada de ambas varianzas muestrales (s_1^2 y s_2^2).

Para recordar

A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo T_i el valor inferior del estadístico de prueba o contraste y T_s el valor superior de este estadístico empleado para el contraste de hipótesis, entonces:

- Para contraste bilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, T_s)$:
    Región crítica = $(-\infty, T_i] \cup [T_{s,} +\infty)$ donde se rechaza H_0.
- Para contraste unilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, +\infty)$:
    Región crítica = $(-\infty, T_{i)}$ donde se rechaza H_0.
  - Intervalo donde no se rechaza $H_0 = (-\infty, T_S)$:
    Región crítica = $(T_S, +\infty)$ donde se rechaza H_0.
- Valor de p o p-valor bajo H_0:
  p-valor = Pr [t-student (teórico-esperado) ≥ t-student (observado-calculado)].

Regla de decisión

La regla de decisión respecto a la significación estadística al contrastar las hipótesis (nula y alternativa) está determinada por:

- Para t-student (observado-calculado) < 0:
  - Si t-student (observado-calculado) < t-student (teórico-esperado) según el valor de alfa prefijado y los grados de libertad respectivos, entonces el valor de t-student observado es ubicado en el área de rechazo para H_0 por lo que hay evidencia suficiente para rechazar H_0 y aceptar la alternativa H_1, siendo significativas las diferencias entre las medias de la variable para las dos poblaciones independientes, al estimarse un valor de probabilidades (p-valor) correspondiente con el valor observado de t-student si se cumple H_0 o las probabilidades del azar para el resultado obtenido menores que el valor alfa prefijado o de cometer el error tipo 1 para el modelo teórico según el nivel de confianza escogido ($p < \alpha$).
  - Si t-student (observado-calculado) ≥ t-student (teórico-esperado) según el valor de alfa prefijado y los grados de libertad respectivos, entonces el valor de t-student observado no es ubicado en el área de rechazo para H_0 por lo que no hay evidencia suficiente para rechazar H_0 y aceptar la alternativa H_1, siendo no significativas las diferencias entre las medias de la variable para las dos poblaciones independientes, al estimarse un valor de probabilidades (p-valor) correspondiente con el valor observado de t-student si se cumple H_0 o las probabilidades del azar para el resultado obtenido iguales o mayores que el valor alfa prefijado o de cometer el error tipo 1 para el modelo teórico según el nivel de confianza escogido ($p \geq \alpha$).

– Para t-student (observado-calculado) ≥ 0:
  - Si t-student (observado-calculado) > t-student (teórico-esperado) según el valor de alfa prefijado y los grados de libertad respectivos, entonces el valor de t-student observado es ubicado en el área de rechazo para H_0 por lo que hay evidencia suficiente para rechazar H_0 y aceptar la alternativa H_1, siendo significativas las diferencias entre las medias de la variable para las dos poblaciones independientes, al estimarse un valor de probabilidades (p-valor) para el valor observado de t-student si se cumple H_0 o las probabilidades del azar para el resultado obtenido menores que el valor alfa prefijado o de cometer el error tipo 1 para el modelo teórico según el nivel de confianza escogido ($p < \alpha$).
  - Si t-student (observado-calculado) ≤ t-student (teórico-esperado) según el valor alfa prefijado y los grados de libertad respectivos, entonces el valor de t-student observado no se ubica en el área de rechazo para H_0 por lo que no hay evidencia suficiente para rechazar H_0 y aceptar la alternativa H_1, siendo no significativas las diferencias entre las medias de la variable para las dos poblaciones independientes, al estimarse un valor de probabilidades (p-valor) para el valor observado de t-student si se cumpliera H_0 o las probabilidades del azar para el resultado obtenido iguales o mayores que el valor alfa prefijado o de cometer el error tipo 1 para el modelo teórico según el nivel de confianza escogido ($p \geq \alpha$).

x) Dos muestras no pareadas y una variable cuantitativa aleatoria que sigue o se aproxima a la distribución normal estándar, si se desea comparar dos medias poblacionales estimadas a partir de las medias muestrales calculadas con los datos de la variable cuantitativa para tamaños de muestras pequeños ($n < 30$) y varianzas poblacionales desconocidas y significativamente diferentes. El posible planteamiento de hipótesis estadísticas para el contraste sería:
  – Para contraste bilateral:

    $H_0: \mu_1 - \mu_2 = 0$ *vs.* $H_1: \mu_1 - \mu_2 \neq 0$.
  – Para contraste unilateral:
    - $H_0: \mu_1 - \mu_2 = 0$ *vs.* $H_1: \mu_1 - \mu_2 > 0$.

      $H_0: \mu_1 - \mu_2 = 0$ *vs.* $H_1: \mu_1 - \mu_2 > 0$.
    - $H_0: \mu_1 - \mu_2 \geq 0$ *vs.* $H_1: \mu_1 - \mu_2 < 0$.

      $H_0: \mu_1 - \mu_2 \leq 0$ *vs.* $H_1: \mu_1 - \mu_2 > 0$.
  – Como estadístico de prueba se emplea t-student adoptando la siguiente expresión que les relaciona:

$$\text{t-student (observado-calculado)} = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

que sigue una distribución teórica de probabilidades t-student con f grados de libertad

donde:

$$f = \frac{\left(\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}\right)^2}{\frac{1}{n_1+1}\left(\frac{\hat{S}_1^2}{n_1}\right) + \frac{1}{n_2+1}\left(\frac{\hat{S}_2^2}{n_2}\right)} - 2 : \text{grados de libertad por Fórmula de Welch;}$$

Nota: hay otros autores como Triola (2009) que plantean el cálculo de los grados de libertad el menor de n1 − 1 y de n2 − 1.

$\bar{x}_1$: media de la muestra 1;

$\bar{x}_2$: media de la muestra 2;

d_0: expresa la diferencia entre las medias poblacionales de referencia ($\mu_1 - \mu$) para el contraste que bajo H_0 es igual a 0, procedimiento que se debe asumir al realizar el cálculo del estadístico de prueba;

n_1: tamaño de la muestra 1;

n_2: tamaño de la muestra 2.

Para recordar

A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo Ti el valor inferior del estadístico de prueba o contraste y Ts el valor superior de este estadístico empleado para el contraste de hipótesis, entonces:

- Para contraste bilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, T_s)$:
    Región crítica = $(-\infty, T_{i]} \cup [T_s, +\infty)$ donde se rechaza H_0.
- Para contraste unilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, +\infty)$:
    Región crítica = $(-\infty, T_i)$ donde se rechaza H_0.
  - Intervalo donde no se rechaza $H_0 = (-\infty, T_S)$:
    Región crítica = $(T_S, +\infty)$ donde se rechaza H_0.
- Valor de p o p-valor bajo H_0:
  p-valor = Pr [t-student (teórico-esperado) ≥ t-student (observado-calculado)].

Regla de decisión

La regla de decisión respecto a la significación estadística al contrastar las hipótesis (nula y alternativa) estará determinada por:

- Para t-student (observado-calculado) < 0:
  - Si t-student (observado-calculado) < t-student (teórico-esperado) según el valor de alfa prefijado y los grados de libertad respectivos, entonces el valor de t-student observado es ubicado en el área de rechazo para H_0 por lo

que hay evidencia suficiente para rechazar H_0 y aceptar la alternativa H_1, siendo significativas las diferencias entre las medias de la variable para las dos poblaciones independientes al estimarse las probabilidades (p-valor) bajo H_0 o del azar para el t-student observado menores que el valor alfa prefijado según el nivel de confianza escogido ($p < \alpha$).

- Si t-student (observado-calculado) ≥ t-student (teórico-esperado) según el valor de alfa prefijado y los grados de libertad respectivos, entonces el valor de t-student observado no es ubicado en el área de rechazo para H_0 por lo que no hay evidencia suficiente para rechazar H_0 y aceptar la alternativa H_1, siendo no significativas las diferencias entre las medias de la variable para las dos poblaciones independientes al estimarse las probabilidades (p-valor) bajo H_0 o del azar para el t-student observado iguales o mayores que el valor alfa prefijado según el nivel de confianza escogido ($p \geq \alpha$).

– Para t-student (observado-calculado) ≥ 0:

- Si t-student (observado-calculado) > t-student (teórico-esperado) según el valor de alfa prefijado y los grados de libertad respectivos, entonces el valor de t-student observado es ubicado en el área de rechazo para H_0 por lo que hay evidencia suficiente para rechazar H_0 y aceptar la alternativa H_1, siendo significativas las diferencias entre las medias de la variable para las dos poblaciones independientes al estimarse las probabilidades (p-valor) bajo H_0 o del azar para el t-student observado menores que el valor alfa prefijado según el nivel de confianza escogido ($p < \alpha$).
- Si t-student (observado-calculado) ≤ t-student (teórico-esperado) según el valor de alfa prefijado y los grados de libertad respectivos, entonces el valor de t-student observado no es ubicado en el área de rechazo para H_0 por lo que no hay evidencia suficiente para rechazar H_0 y aceptar la alternativa H_1, siendo no significativas las diferencias entre las medias de la variable para las dos poblaciones independientes al estimarse las probabilidades (p-valor) bajo H_0 o del azar para el t-student observado iguales o mayores que el valor alfa prefijado según el nivel de confianza escogido ($p \geq \alpha$).

y) Dos muestras no pareadas, una variable cualitativa dicotómica que sigue distribución binomial, si se desea comparar dos proporciones poblacionales estimadas a partir de las proporciones muestrales calculadas para una categoría de la variable cualitativa, muestras grandes ($npq > 5$), empleando la prueba de Z por la aproximación a la normal estándar, el posible planteamiento de hipótesis estadísticas para el contraste sería:

– Para contraste bilateral:
$H_0: P_1 = P_2$ *vs.* $H_1: P_1 \neq P_2$

– Para contraste unilateral:

- $H_0: P_1 = P_2$ *vs.* $H_1: P_1 > P_2$ o $H_0: P_1 = P_2$ *vs.* $H_1: P_1 < P_2$.

- H_0: $P_1 \geq P_2$ *vs.* H_1: $P_1 < P_2$ o H_0: $P_1 < o = P_2$ *vs.* H_1: $P_1 > P_2$.

Los pasos a seguir para realizar el contraste de hipótesis incluyen comenzar auxiliándonos gráficamente de la siguiente tabla cruzada 2x2 simple (Tabla 4.14) con los elementos teóricos necesarios y ubicando los distintos valores correspondientes a los datos de la variable cualitativa dicotómica y las dos muestras empleadas en el análisis, lo que puede facilitar la mejor comprensión del procedimiento a realizar:

Tabla 4.14. Representación tabular de los elementos teóricos necesarios en el análisis de la diferencia de proporciones formadas para analizar la homogeneidad entre dos muestras independientes (muestra 1 y 2, que estiman dos poblaciones independientes) utilizando una variable cualitativa dicotómica

Variable cualitativa dicitómica	**Muestra 1**	**Muestra 2**	**Totales**
Categoría 1	$\mathbf{X_1}$	$\mathbf{X_2}$	
Categoría 2			
Totales	$\mathbf{n_1}$	$\mathbf{n_2}$	**n**
$p = (X_1 + X_2) / (n_1 + n_2)$; $p_1 = X_1/n_1$; $p_2 = X_2/n_2$; $q = 1 - p$; $Var_c = [pq\,(1/n_1 + 1/n_2)]$			

– Como estadístico de prueba para el contraste se empleaq Z~N (0, 1) por aproximación a la normal estándar de la binomial, adoptando esta expresión:

$$Z\ (\text{observado-calculado}) = \frac{p_2 - p_1}{\sqrt{Var_c}}$$; también se puede expresar como:

$$Z\ (\text{observado-calculado}) = \frac{p_2 - p_1}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

donde:

Var_c: varianza combinada de p según n_1 y n_2, se estima como:

$$Var_c = pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

donde:

$q = 1 - p$.

Otros autores también expresan la varianza combinada como:

$$Var_c = p_1\left(\frac{1-p_1}{n_1}\right) + p_2\left(\frac{1-p_2}{n_2}\right)$$

donde:

$p = \frac{X_1 - X_2}{n_1 + n_2}$: proporción combinada de la categoría 1 de la variable cualitativa dicotómica para las dos muestras no pareadas;

X_1: sujetos con categoría 1 de la variable en la muestra 1;
X_2: sujetos con categoría 1 de la variable en la muestra 2;
n_1: total de sujetos en la muestra 1;
n_2: total de sujetos en la muestra 2;
p_1 y p_2: dos proporciones (una de cada muestra) a comparar, para la misma categoría de la variable dicotómica (X_1 y X_2) según el total de cada muestra respectivamente (n_1 y n_2 en la tabla anterior), que estiman a dos proporciones poblacionales.

Para recordar

A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo T_i el valor inferior del estadístico de prueba o contraste y T_s el valor superior de dicho estadístico empleado para el contraste de hipótesis, entonces:

- Para contraste bilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, T_s)$:
    Región crítica = $(-\infty, T_{i]} \cup [T_s, +\infty)$ donde se rechaza H_0.
- Para contraste unilateral:
  - Intervalo donde no se rechaza $H_0 = (T_i, +\infty)$:
    Región crítica = $(-\infty, T_{i)}$ donde se rechaza H_0.
  - Intervalo donde no se rechaza $H_0 = (-\infty, T_S)$:
    Región crítica = $(T_S, +\infty)$ donde se rechaza H_0.
- Valor de p o p-valor bajo H_0:
  p-valor = Pr [Z (teórico-esperado) ≥ Z (observado-calculado)].

Regla de decisión

La regla de decisión respecto a la significación estadística al contrastar las hipótesis (nula y alternativa) estará determinada por:

- Para Z (observado-calculado) < 0:
  - Si Z (calculado-observado) < Z (teórico-esperado) según el valor de alfa prefijado, entonces el valor de Z observado es ubicado en el área de rechazo para H_0 por lo que hay evidencia suficiente para rechazar H_0 y aceptar la alternativa H_1, siendo significativas las diferencias entre las medias de la variable para las dos poblaciones independientes, al estimarse un valor de probabilidades (p-valor) para el valor observado de Z si se cumpliera H_0 o las probabilidades del azar para el resultado obtenido menores que el valor alfa prefijado o de cometer el error tipo 1 para el modelo teórico según el nivel de confianza escogido ($p < \alpha$).
  - Si Z (calculado-observado) ≥ Z (teórico-esperado) según el valor de alfa prefijado, entonces el valor de Z observado no es ubicado en el área de

rechazo para H_0 por lo que no hay evidencia suficiente para rechazar H_0 y aceptar la alternativa H_1, siendo no significativas las diferencias entre las medias de la variable para las dos poblaciones independientes, al estimarse un valor de probabilidades (p-valor) para el valor observado de Z si se cumpliera H_0 o las probabilidades del azar para el resultado obtenido iguales o mayores que el valor alfa prefijado o de cometer el error tipo 1 para el modelo teórico según el nivel de confianza escogido ($p \geq \alpha$).

- Para Z (observado-calculado) ≥ 0:
  - Si Z (calculado-observado) > Z (teórico-esperado) según el valor de alfa prefijado, entonces el valor de Z observado es ubicado en el área de rechazo para H_0 por lo que hay evidencia suficiente para rechazar H_0 y aceptar la alternativa H_1, siendo significativas las diferencias entre las medias de la variable para las dos poblaciones independientes, al estimarse un valor de probabilidades (p-valor) para el valor observado de Z si se cumpliera H_0 o las probabilidades del azar para el resultado obtenido menores que el valor alfa prefijado o de cometer el error tipo 1 para el modelo teórico según el nivel de confianza escogido ($p < \alpha$).
  - Si Z (calculado-observado) ≤ Z (teórico-esperado) según el valor de alfa prefijado, entonces el valor de Z observado no es ubicado en el área de rechazo para H_0 por lo que no hay evidencia suficiente para rechazar H_0 y aceptar la alternativa H_1, siendo no significativas las diferencias entre las medias de la variable para las dos poblaciones independientes al estimarse un valor de probabilidades (p-valor) para el valor observado de Z si se cumpliera H_0 o las probabilidades del azar para el resultado obtenido iguales o mayores que el valor alfa prefijado o de cometer el error tipo 1 para el modelo teórico según el nivel de confianza escogido ($p \geq \alpha$).

z) Dos muestras no pareadas, una variable cualitativa dicotómica que sigue distribución de frecuencias binomial, si se desea determinar la homogeneidad en las poblaciones estimadas de las muestrales para las proporciones calculadas, empleando la prueba de Pearson, de Yates o probabilidad exacta de Fisher, tomando en cuenta los grados de libertad correspondientes para el nivel alfa predeterminado por el investigador, el posible planteamiento de hipótesis estadísticas para el contraste sería:

- H_0: existe (poblacionalmente) homogeneidad para las proporciones calculadas a partir de la variable cualitativa dicotómica.
- H_1: no existe (poblacionalmente) homogeneidad para las proporciones calculadas a partir de la variable cualitativa dicotómica.

  Los pasos a seguir para el contraste incluyen la presentación de forma tabulada de los diferentes elementos teóricos (Tabla 4.15).

Tabla 4.15. Representación tabular de los elementos teóricos necesarios a emplear para el análisis de la homogeneidad de dos muestras independientes según el comportamiento de una variable cualitativa dicotómica:

Variable cualitativa dicitómica	Muestra 1	Muestra 2	Totales
Categoría 1	a	b	a + b
Categoría 2	c	d	c + d
Totales	**a + c**	**b + d**	**n = a + b + c + d**

- El estadístico de prueba para el contraste a emplear sería chi-cuadrado de Pearson (sin corrección), de Yates (con corrección) o la prueba para la probabilidad exacta de Fisher.
  - Fórmula para chi-cuadrado de Pearson (χ^2) según la expresión:

$$\chi^2 \text{ Pearson} = \frac{n(ad-bc)^2}{m_1 m_0 n_1 n_0} \text{ o } \chi^2 \text{ Pearson} = \frac{n(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

  - Fórmula para chi-cuadrado con corrección de Yates (χ^2_C) según la expresión:

$$\chi^2_c = \frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{(a+c)(b+d)(a+b)(c+d)}$$

    Donde las letras incluidas en su expresión matemática se corresponden con las celdas de la tabla 4.14 y siguen una distribución teórica de probabilidades chi-cuadrado con (f – 1) (c – 1) = 1 grado de libertad para el nivel alfa prefijado.

Para recordar

Respecto a la significación estadística para la homogeneidad en las dos muestras independientes para la variable dicotómica según las proporciones formadas a partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo χ^2 exp (Pearson o Yates) el valor observado-calculado del estadístico de prueba o contraste y χ^2 teo el valor teórico-esperado de este estadístico con grados de libertad (c – 1) (f – 1) y nivel de confianza (1 – α)%, entonces:

- Intervalo donde no se rechaza H_0 = (para cualquier valor de χ^2 exp ≤ χ^2 teo).
- Región crítica (donde se rechaza H_0) = (para cualquier valor de χ^2 exp > χ^2 teo).
- Valor de p o p-valor bajo H_0:
  p-valor = Pr [χ^2 teo ≥ χ^2 exp].

Regla de decisión

La regla de decisión respecto a la significación estadística al contrastar las hipótesis (nula *versus* alternativa) está determinada por:

- Si χ^2 (observado-calculado) > χ^2 (esperado-tabulado) con un grado de libertad y alfa prefijado, entonces el valor de chi-cuadrado observado se ubica en el área de rechazo de H_0, por lo que existen evidencias suficientes para rechazar la hipótesis nula (H_0) y se acepta la alternativa (H_1), es decir, no existe homogeneidad en las muestras no pareadas a partir de las proporciones formadas según la variable cualitativa dicotómica empleada, estimándose un p-valor bajo H_0 para el chi-cuadrado observado o las probabilidades del azar para este resultado menores que el nivel de significación alfa o de cometer el error de tipo 1 según el nivel de confianza escogido ($p < \alpha$).
- Si χ^2 (observado-calculado) ≤ (χ^2 (esperado-tabulado) con un grado de libertad y alfa prefijado, entonces el valor de chi-cuadrado observado no se ubica en el área de rechazo de H0; por lo que no existen evidencias suficientes para rechazar la hipótesis nula (H_0) y no se acepta la alternativa (H_1), es decir, existe homogeneidad en las proporciones para las muestras no pareadas según la variable cualitativa dicotómica empleada, estimándose un p-valor bajo H_0 para el chi-cuadrado observado o las probabilidades del azar para este resultado iguales o mayores que el nivel de significación alfa o de cometer el error de tipo 1 según el nivel de confianza escogido ($p \geq \alpha$).

  La prueba de la probabilidad exacta de Fisher se emplea cuando las frecuencias esperadas para las diferentes celdas, al menos una es menor de 5:
- Puede ser empleada en la comparación de dos proporciones binomiales procedentes de dos muestras independientes en tablas de 2x2 con valores de frecuencias esperada menores de 5 en alguna celda a partir de los valores de la variable dicotómica.
- Se puede emplear tanto para pruebas de homogeneidad en dos muestras independientes como de independencia de dos variables dicotómicas.
- Permite determinar la probabilidad (p-valor) exacta de obtener valores de la celda a, fuera de las a + b observaciones para un número N (a + b + c + d = total de observaciones) de las cuales a + c se consideran con éxito dentro del evento estudiado al comparar dos proporciones binomiales en tablas de contingencia 2x2 simples, esto se puede realizar empleando las probabilidades de la distribución hipergeométrica y estimar este valor p (p-valor) como sigue según el procedimiento de Rosner (2000):

• Valor p unilateral:

  p-valor para la cola derecha, se estima como probabilidades (+) a partir de la condición:

$$\text{p-valor} = \begin{cases} P(+) \text{ si } \dfrac{ad}{(b+1)(c+1)} > \dfrac{bc}{(a+1)(d+1)} \\ \\ P(-) \text{ en caso contrario} \end{cases}$$

$P(+) = \sum_{P(a+k) \le P(a)} P(a+k)$ cuando se quiere contrastar $p_1 > p_2$

p-valor para la cola izquierda, se estima como probabilidades (–) a partir de la condición:

$$\text{p-valor} = \begin{cases} P(-) \quad \text{si} \dfrac{ad}{(b+1)(c+1)} < \dfrac{bc}{(a+1)(d+1)} \\ P(-) \quad \text{en caso contrario} \end{cases}$$

$P(-) = \sum_{P(a-k) \le P(a)} P(a-k)$ cuando se quiere contrastar $p_1 < p_2$

- Valor p bilateral:
  p-valor = P (+) + P (–) –P (a)
  donde:

  $P(+) = \sum_{P(a+k) \le P(a)} P(a+k)$: quiere decir las probabilidades de las tablas estructuradas diferentes a la tabla observada inicial, empleando valores mayores o iguales de los de la celda a (de la tabla inicial);

  $P(-) = \sum_{P(a-k) \le P(a)} P(a-k)$: quiere decir las probabilidades de las tablas estructuradas diferentes a la tabla observada inicial, empleando valores menores o iguales de los de la celda a (de la tabla inicial);
  P (a + k) y P (a – k): se calculan a partir de la función de probabilidad de una distribución hipergeométrica de parámetros;
  N = a + b + c + d, R = a + b y n = a + c.
  La explicación del método es como sigue:
  Se retoma la tabla 4.14 y se presenta como tabla 4.16.

Tabla 4.16. De contingencia 2x2 simple para determinar el comportamiento del hábito de fumar en dos muestras independientes con y sin enfermedad cardiovascular sin valores asignados en las celdas

Hábito de fumar	Muestra 1 (pacientes con enfermedad cardiovascular)	Muestra 2 (pacientes sin enfermedad cardiovascular)	Totales
Fumador	a	b	a + b
No fumador	c	d	c + d
Totales	**a + c**	**b + d**	**N = a + b + c + d**

La probabilidad de obtener los valores para una tabla 2x2 simple observada se pueden calcular con la siguiente expresión matemática denominada probabilidad exacta (Pr) para los valores observados de las celdas a, b, c, d en una tabla 2x2 simple:

$$\Pr(a,\ b,\ c,\ d) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$, donde el signo (!) significa la operación matemática factorial del valor o expresión que le antecede.

Si se asignan los siguientes valores a las celdas de la tabla 4.16 y se presenta la tabla 4.17.

Tabla 4.17. De contingencia 2x2 simple, para determinar el comportamiento del hábito de fumar en dos muestras independientes con y sin enfermedad cardiovascular con valores asignados en las celdas

Hábito de fumar	Muestra 1 (pacientes con enfermedad cardiovascular)	Muestra 2 (pacientes sin enfermedad cardiovascular)	Totales marginales
Fumador	a = 2	b = 5	a + b = 7
No fumador	c = 3	d = 1	c + d = 4
Totales	**a + c = 5**	**b + d = 6**	**N = a + b + c + d = 11**

La probabilidad exacta a obtener para los valores de las celdas empleados para esta tabla 2x2 simple sería:

$$\Pr(2,\ 5,\ 3,\ 1) = \frac{7!4!5!6!}{11!2!5!3!1!} = \frac{5040(24)(120)(720)}{39\,916\,800(2)(120)(6)} = \frac{1{,}0450944 \cdot 10^{10}}{5{,}7480192 \cdot 10^{10}} = 0{,}182$$

Los pasos que se pueden seguir para el cálculo de la probabilidad exacta de Fisher, que incluye la sumatoria de las probabilidades para los distintos valores en las celdas para las tablas formadas, manteniendo sus valores totales marginales, por ejemplo para la prueba de homogeneidad de proporciones en dos muestras independientes sería:

– Teniendo en cuenta que su estrategia básica en el contraste de hipótesis H_0: $p_1 = p_2$ *vs.* H_1: $p_1 \neq p_2$, siendo p_1 = (individuos de la muestra 1 y factor de riesgo positivo/total de individuos observados) y p_2 = (individuos de la muestra 2 con factor de riesgo positivo/total de individuos observados) será enumerar todas las posibles tablas a formar manteniendo constantes los totales marginales, tal y como se encuentran en la tabla observada con los valores iniciales y calcular la probabilidad exacta para cada una de las tablas formadas según la expresión matemática anterior (basada en las probabilidades de la distribución hipergeométrica), cuyo valor esperado E (X) y su varianza [Var (X)] para una variable aleatoria X (para la celda a o $C_{1,1}$), o sea, fila 1, columna 1, se expresa como:

$$E(x) = \frac{M_1 N_1}{N}$$

$$Var(x) = \frac{M_1 M_2 N_1 N_2}{N^2 (N-1)}$$

- Se reestructuran las filas y columnas de la tabla observada con los valores iniciales de modo que el total marginal inferior por las filas quede en la primera fila y el total marginal por columna inferior en la primera columna.
- Comenzar asignando valor de 0 a la primera celda ($C_{1,1}$) o celda a, tomando en la siguiente celda restante los valores correspondientes que permitan sin variar los valores totales marginales por filas y columnas que presentó la tabla inicial observada y calcular la probabilidad para estos nuevos valores a partir de la distribución de probabilidades hipergeométrica.
- Construir nuevas tablas incrementando tabla a tabla los valores de la celda a, obteniéndose tablas hasta que alguno de los valores de las celdas restantes adquiera el valor 0 y entonces calcular igualmente la probabilidad para los valores de cada nueva tabla estructurada.

  Si se toma como ejemplo la tabla observada inicial con los valores de a, b, c,d con sus respectivos valores totales marginales (a + b), (c + d), (a + c), (b + d) y el número total de observaciones (N = a + b + c + d) y se denomina tabla (a) como la observada (inicial) y la tabla (k) la última posible obtenida (cuando alguna otra celda toma un valor de 0 debido al aumento de la celda a y manteniendo los valores de los totales marginales) tenemos que:
- Para contrastar la hipótesis bilateral a partir de la estimación de la probabilidad exacta de Fisher se tiene:

  H_0: $p_1 = p_2$ *vs.* H_1: $p_1 \neq p_2$

  p-valor Fisher (2 colas) = 2mín [Pr(0) + Pr(1) + … + Pr(a) , Pr(a) + Pr(a + 1) + … + Pr(k)]

  donde:

  mín: el menor valor de las probabilidades y el p-valor, que corresponde al duplicado de la menor área bajo la curva, bien de la cola derecha o de la izquierda.

  Otra forma que puede ser aplicada en su estimación es:

$$\text{p-valor Fisher (2 colas)} = \sum_{\{i:\ Pr(i) \leq Pr(a)\}} Pr(i)$$

  donde:

  Pr (i): probabilidades obtenidas para los valores de cada tabla, y se sumarían todas las Pr (i) menores o iguales a la Pr (a) de la tabla observada con los valores iniciales en cada celda.

– Para contrastar la hipótesis unilateral:
  H_0: $p_1 = p_2$ *vs.* H_1: $p_1 < p_2$
  p-valor Fisher = Pr (0) + Pr (1) + … + Pr (a) corresponde al área bajo la curva para la cola izquierda).
– Para contrastar la hipótesis unilateral:
  H_0: $p_1 = p_2$ *vs.* H_1: $p_1 > p_2$
  p-valor Fisher = Pr (a) + Pr (a + 1) + … Pr (k) corresponde al área bajo la curva para la cola derecha.
  Nota: En este caso el valor de p (p-valor) expresa la probabilidad de obtener una tabla con valores tan extremos o mas que los obtenidos con la tabla observada inicial bajo H_0.

Regla de interpretación

– Si p-valor (observado-calculado) exacto de Fisher < nivel de significación (alfa) de acuerdo con el tipo de contraste de hipótesis a realizar (bilateral o unilateral), las probabilidades estimadas exactas de poder cumplirse H_0 con los datos observados o más extremos que los mismos son mínimas por lo que existen evidencias suficientes para el rechazo de H_0, por ende se concluye que existen diferencias significativas en las proporciones formadas a partir de la variable cualitativa dicotómica para ambas muestras independientes, por lo tanto no hay homogeneidad en las dos muestras comparadas en el nivel de confianza preseleccionado.
– Si p-valor (observado-calculado) exacto de Fisher ≥ nivel de significación (alfa) de acuerdo al tipo de contraste de hipótesis a realizar (bilateral o unilateral), las probabilidades estimadas exactas de poder cumplirse H_0 con los datos observados o más extremos que los mismos no son mínimas por lo que no existen evidencias suficientes para el rechazo de H_0, por ende, se concluye que no existen diferencias significativas en las proporciones formadas a partir de la variable cualitativa dicotómica en ambas muestras independientes, por lo tanto, hay homogeneidad en las dos muestras comparadas en el nivel de confianza preseleccionado.

Z1) Dos muestras no pareadas, una variable cualitativa politómica u ordinal por niveles asumidos como categorías nominales, si se desea determinar la homogeneidad en dos poblaciones estimadas para las proporciones calculadas para la variable cualitativa politómica o cualitativa ordinal, empleando la prueba de Pearson, según los grados de libertad correspondientes y valor de alfa predeterminado por el investigador, el posible planteamiento de hipótesis estadísticas para el contraste sería:

– H_0: existe (poblacionalmente) homogeneidad en el comportamiento de las proporciones en ambas muestras independientes según la variable cualitativa politómica.

– H_1: no existe (poblacionalmente) homogeneidad en el comportamiento de las proporciones en ambas muestras independientes según la variable cualitativa politómica.

  Los pasos a seguir para el contraste incluyen la presentación de forma tabulada de los diferentes elementos teóricos necesarios y la introducción de los respectivos valores en las celdas a partir de la información recogida previamente (Tabla 4.18).

Tabla 4.18. Representación tabular de los elementos teóricos necesarios a emplear para el análisis de la homogeneidad de dos muestras independientes según el comportamiento de una variable cualitativa politómica u ordinal por nominaciones

Variable cualitativa politómica u ordinal	Muestra 1 (pacientes con enfermedad cardiovascular)	Muestra 2 (pacientes sin enfermedad cardiovascular)	Totales marginales
Categoría 1	$X_{1,1}$	$X_{1,2}$	$X_{1,1} + X_{1,2}$
Categoría 2	$X_{2,1}$	$X_{2,2}$	$X_{2,1} + X_{2,2}$
Categoría 3	$X_{3,1}$	$X_{3,2}$	$X_{3,1} + X_{3,2}$
Totales	$X_{1,1} + X_{2,1} + X_{3,1}$	$X_{1,2} + X_{2,2} + X_{3,2}$	$n = X_{1,1} + X_{1,2} + X_{2,1} + X_{2,2} + X_{3,1} + X_{3,2}$

Nota: los subíndices numéricos expresan el número de fila y de columna respectivo de cada celda.

– Como estadístico de prueba en el contraste se emplearía chi-cuadrado de Pearson (χ^2), que compara la frecuencia observada y la esperada en todas las celdas) adoptando la siguiente expresión en tablas de 2xN simple con un variable cualitativa nominal politómica u ordinal por niveles:

$$\chi^2 = \sum_{i,j} \frac{(x_{ij} - e_{ij})^2}{e_{ij}}$$, que sigue una distribución teórica de probabilidades

$\chi^2_{(c-1)(f-1);\,1-\alpha}$ con (f – 1) (c – 1) grados de libertad para el nivel alfa prefijado.

– También se puede emplear el chi-cuadrado con corrección de Yates (χ^2_C) según la fórmula:

$$\chi^2_c = \sum_{i,j} \frac{(x_{i,j} - e_{i,j} - 0{,}5)^2}{e_{i,j}}$$

donde:

$X_{i,j}$: valor observado para cada celda;

$e_{i,j}$: valor esperado para cada celda (total de la fila x total de columna correspondiente para cada celda) /tamaño de muestra;

Grados de libertad χ^2 (esperado-teórico-tabulado) = (Número de columnas – 1) x (Número de filas – 1).

i: número de orden de las filas;

j: número de orden de las columnas.

Para recordar

A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo χ^2 exp el valor observado-calculado del estadístico de prueba o contraste y χ^2 teo el valor teórico-esperado de este estadístico con grados de libertad (c – 1) (f – 1) y nivel de confianza (1 – α)%, entonces:

- Intervalo donde no se rechaza H_0 = (para cualquier valor de χ^2 exp ≤ χ^2 teo $_{(c-1)\,(f-1);1-\alpha.}$
- Región crítica (donde se rechaza H_0) = (para cualquier valor de χ^2 exp> χ^2 teo $_{(c-1)\,(f-1);1-\alpha.}$
- Valor de p o p-valor bajo H_0: p-valor = Pr [χ^2 teo ≥ χ^2 exp].

Regla de decisión

La regla de decisión respecto a la significación estadística al contrastar las hipótesis (nula y alternativa) estará determinada por:

- Si χ^2 (observado-calculado) de Pearson> χ^2 (c –1) (f –1); 1 – α (teórico-esperado), entonces el valor de chi-cuadrado observado de Pearson se ubica en el área de rechazo de H_0, por lo que existen evidencias suficientes para rechazar la hipótesis nula (H_0) y se acepta la alternativa (H_1), es decir, no existe homogeneidad en las proporciones para las dos muestras independientes según la variable politómica, siendo las probabilidades del azar o el p-valor si se cumpliera H_0 para el valor de chi cuadrado observado menores que el valor de alfa prefijado según el nivel de confianza escogido (p < α).
- Si χ^2 (observado-calculado) de Pearson ≤ χ^2 $_{(c-1)\,(f-1);\,1-\alpha}$ (teórico-esperado), entonces el valor de chi-cuadrado observado de Pearson no se ubica en el área de rechazo de H_0, por lo que no existen evidencias suficientes para rechazar la hipótesis nula (H_0) y no se acepta la alternativa (H_1), es decir, existe homogeneidad en las proporciones para las dos muestras independientes según la variable politómica, siendo las probabilidades del azar o el p-valor si se cumpliera H_0 para el valor de chi-cuadrado observado mayores o iguales que el valor de alfa prefijado según el nivel de confianza escogido (p ≥ α).

z2) Dos muestras no pareadas, una variable cualitativa expresada en forma de nuevos eventos según personas-tiempo de seguimiento, si se desea comparar la densidad de incidencia (DI) o tasa de incidencia (TI) de una población estimada respecto a la densidad de incidencia o tasa de incidencia de otra población estimada (ambas a partir de las dos muestras no relacionadas, respectivamente), a partir de los valores observados de un nuevo evento (incidencia) para un número de personas-tiempo de seguimiento, con un nivel de significación alfa prefijado para muestras grandes, empleando la aproximación de la distribución normal

estándar a la binomial, el posible planteamiento de hipótesis estadísticas para contrastar sería:

- Contraste bilateral:
  H_0: $DI_1 = DI_2$ *vs.* H_1: $DI_1 \neq DI_2$ o H_0: $TI_1 = TI_2$ *vs.* H_1: $TI_1 \neq TI_2$
  donde:
  DI_1: densidad de incidencia o tasa de incidencia verdadera en población 1 estimada = valor observado del evento x personas –tiempo = (a 1 x t 1);
  DI_2: densidad de incidencia o tasa de incidencia verdadera en población 2 estimada = valor observado del evento x persona –tiempo = (a 2 x t 2).
- Método de inferencia para datos de dos tasas de incidencia comparadas por aproximación a la normal, con número de eventos grandes:
  H_0: $TI_1 = TI_2$ *vs.* H_1: $TI_1 \neq TI_2$. o H_0: $DI_1 = DI_2$ *vs.* H_1: $DI_1 \neq DI_2$ se emplea solo si V1 ≥ 5.
  Se auxilia primeramente de la representación gráfica (Tabla 4.18) para ubicar los elementos teóricos necesarios a incluir en este contraste:
  Bajo H_0 el número de nuevos eventos de la variable aleatoria empleada para el grupo 1 (muestra de expuestos que estima población 1) es asumido como una variable aleatoria binomial con parámetros:
  $n = a_1 + a_2$
  $p0 = t_1/(t_1 + t_2)$
  $q_0 = 1 - p_0$
  Empleando la aproximación a la distribución normal de la binomial de la variable aleatoria (para el número de eventos en el grupo 1 (a_1), su media o valor esperado:
  $E_1 = np_0 = (a_1 + a_2)\ t_1/(t_1 + t_2)$;
  así como su varianza:
  $V_1 = np_0q_0 = (a_1 + a_2)\ t_1t_2/(t_1 + t_2)^2$;

Tabla 4.19. Representación tabular de los elementos a emplear en el análisis comparativo de dos tasas de incidencias (para dos grupos o muestras según exposición)

Grupo de exposición (equivale a cada muestra)	Número de nuevos eventos o casos	Número de personas por tiempo de seguimiento
1	a_1	t_1
2	a_2	t_2
	n	$t_1 + t_2$

Pasos a seguir:

- Determinar el número de nuevos eventos observados en ambos grupos de exposición (a_1 y a_2).
- Determinar el número de personas-tiempo de seguimiento observados en ambos grupos de exposición o muestras (t_1 y t_2).

– Determinar el número de eventos esperados (parámetro) en la población estimada para el grupo 1 de expuestos (E_1).
– Determinarla varianza (V_1) en la población estimada para el grupo 1 de expuestos.
– Sustituir los valores de correspondientes en la fórmula para el estadístico de prueba o contraste Z ~ N (0, 1) por la aproximación de la distribución normal estándar bajo H_0, como sigue, solo si V ≥ 5:

$$Z\ (\text{observado-calculado}) = \frac{a_1 - E_1 - 0{,}5}{\sqrt{V_1}}\text{, si a1 > E1 o}$$

$$Z\ (\text{observado-calculado}) = \frac{a_1 - E_1 + 0{,}5}{\sqrt{V_1}}\text{, si a1 ≤ E1}$$

donde:

$E_1 = (a_1 + a_2)\ t_1/\ (t_1 + t_2)$;

$V_1 = (a_1 + a_2)\ t_1t_2/\ (t_1 + t_2)^2$;

a_1 y a_2: número de nuevos eventos observados en el grupo 1 y grupo 2, respectivamente;

t_1 y t_2: cantidad de personas por tiempo en el grupo 1 y el 2, respectivamente.

– Plantear p-valor, bajo la H_0: Z ~ N (0, 1) como sigue:
  • Si Z ≥ 0, entonces p-valor = 2[1 – Φ (z)].
  • S: Z < 0, entonces p-valor =2Φ (z)].
– Interpretar resultado del estadístico de contraste Z (observado-calculado) *versus* $Z_{1-\alpha/2}$ o $Z_{\alpha/2}$ (teórico-esperado) si el valor observado de Z fuese mayor o menor que 0 respectivamente, bajo la H0 se sigue la siguiente Regla de decisión

Para prueba de dos lados (bilateral) según el valor de alfa prefijado por el investigador:

– Para Z (observado-calculado) < 0:
  • Si Z (observado-calculado) < Z (teórico-esperado) según el valor de alfa, entonces Z observado se ubica en la región de rechazo para H_0, por lo que existe evidencia suficiente para rechazar la hipótesis nula y se acepta la alternativa, es decir, existen diferencias significativas entre las tasas de incidencias para ambos grupos según exposición o de ambas poblaciones estimadas o no hay homogeneidad en los valores para las tasas de incidencia según exposición, estimándose que el p-valor bajo H_0 según el valor de Z observado o relacionado con el azar para este resultado es menor que el valor alfa prefijado ($p < \alpha$).
  • Si Z (observado-calculado) ≥ Z (teórico-esperado) según el valor de alfa, entonces Z observado no se ubica en la región de rechazo para H_0, por lo que no existe evidencia suficiente para rechazar la hipótesis nula y no se

acepta la alternativa, es decir, no existen diferencias significativas entre las tasas de incidencias para ambos grupos según exposición o de ambas poblaciones estimadas o hay homogeneidad en los valores para las tasas de incidencia según exposición, estimándose que el p-valor bajo H_0 según el valor de Z observado o relacionado con el azar para este resultado es igual o mayor que el valor alfa prefijado ($p \geq \alpha$).

– Para Z (observado-calculado) ≥ 0:
  - Si Z (observado-calculado) > Z (teórico-esperado) según el valor de alfa, entonces Z observado se ubica en la región de rechazo para H_0, por lo que existe evidencia suficiente para rechazar la hipótesis nula y se acepta la alternativa, es decir, existen diferencias significativas entre las tasas de incidencias para ambos grupos según exposición o de ambas poblaciones estimadas o no hay homogeneidad en los valores para las tasas de incidencia según exposición, estimándose que el p-valor bajo H_0 según el valor de Z observado o relacionado con el azar para este resultado es menor que el valor alfa prefijado ($p < \alpha$).
  - Si Z (observado-calculado) ≤ Z (teórico-esperado) según el valor de alfa, entonces Z observado no se ubica en la región de rechazo para H_0, por lo que no existe evidencia suficiente para rechazar la hipótesis nula y no se acepta la alternativa, es decir, no existen diferencias significativas entre las tasas de incidencias para ambos grupos según exposición o de ambas poblaciones estimadas o hay homogeneidad en los valores para las tasas de incidencia según exposición, estimándose que el p-valor bajo H_0 según el valor de Z observado o relacionado con el azar para este resultado es igual o mayor que el valor alfa prefijado ($p \geq \alpha$).

    Método de inferencia para datos de tasas de incidencia, prueba exacta para muestras con número de nuevos eventos pequeños:

    H_0: $TI_1 = TI_2$ *vs.* H_1: $TI_1 \neq TI_2$ o H_0: $DI_1 = DI_2$ *vs.* H_1: $DI_1 \neq DI_2$.

    Se emplea este método solo si $V_1 < 5$.

Pasos a seguir:

– Sean a_1 y a_2 el número de nuevos eventos observados en el grupo 1 y el 2, respectivamente y t_1 y t_2 la cantidad de personas por tiempo en el grupo 1 y el 2, respectivamente y p la proporción verdadera de eventos en el grupo 1.

  Para probar la hipótesis H_0: $DI_1 = DI_2$ (que equivale a $p = p_0$) *vs.* H_1: $DI_1 \neq DI_2$ (que equivale a $p \neq p_0$), donde:

  DI_1: densidad de incidencia verdadera en el grupo 1;

  DI_2: densidad de incidencia en el grupo 2.

  Y si se asume que:

$p_0 = t_1/ (t_1 + t_2)$;

$q_0 = 1 - p_0$

Bajo H_0 el número de nuevos eventos de la variable aleatoria para el grupo de exposición 1 (muestra de expuestos que estima la población 1) es asumido como una variable aleatoria binomial con parámetros:

$n = a_1 + a_2$

$p_0 = t_1/ (t_1 + t_2)$

$q_0 = 1 - p_0$

$V_1 = np_0q_0 = (a_1 + a_2)\, t_1t_2/ (t_1 + t_2)^2$

- El cálculo del p-valor exacto (valor de probabilidad relacionado con el azar o donde se cumple H_0) para ambos lados de la distribución de probabilidades se realiza como sigue:
  - Si $a_1 < (a_1 + a_2)\, p_0$, entonces:

    $$\text{p-valor} = 2\sum_{k=0}^{a_1}\binom{a_1 + a_2}{k} p_0^k q_0^{a_1 + a_2 - k}$$

  - Si $a_1 \geq (a_1 + a_2)\, p_0$, entonces:

    $$\text{p-valor} = 2\sum_{k=a_1}^{a_1 + a_2}\binom{a_1 + a_2}{k} p_0^k q_0^{a_1 + a_2 - k}$$

- Posible interpretación del resultado del p-valor:
  - Si p-valor estimado (observado-calculado) < valor de alfa según nivel de confianza predeterminado, las probabilidades con los datos empleados o más extremas que las mismas de que pudiera cumplirse H_0 o del azar en este resultado son muy bajas (menor al valor de incertidumbre para H_1 o de cometer el error de tipo 1 o alfa), por lo que son significativas las diferencias y se puede rechazar la hipótesis nula H_0.
  - Si p-valor estimado (observado-calculado) ≥ valor de alfa según nivel de confianza predeterminado, las probabilidades con los datos empleados o más extremas que las mismas de que pudiera cumplirse H_0 o del azar en este resultado no son muy bajas (mayor o igual al valor de incertidumbre para H_1 o de cometer el error tipo 1 o alfa), por lo que no son significativas las diferencias y no se puede rechazar la hipótesis nula H_0.

z3) Dos muestras no pareadas, una variable cualitativa expresada como tasa de incidencia (en forma de nuevos eventos según personas-tiempo de seguimiento en expuestos), si se desea determinar la homogeneidad de la razón de tasas de incidencia (de dos grupos expuestos distintos) a través de diferentes estratos, empleando prueba de chi-cuadrado según los grados de libertad correspondientes y para un nivel de significación alfa prefijado, el posible planteamiento de hipótesis estadísticas para su contraste sería:

H_0: IRR_1 = ... = IRR_K *vs.* H_1: al menos dos de las IRR_i son diferentes.
donde:
IRR_1: razón de tasas de incidencia de un estrato;
K: número máximo de estratos.

– El estadístico de contraste a emplear es chi-cuadrado (χ^2_{het}) de heterogeneidad adoptando la expresión:

$$\chi^2_{het} = \sum_{i=1}^{k} w_i \left[\ln(\hat{IRR}_i) - \ln(\hat{IRR}) \right]^2 \text{ o}$$

$$\chi^2_{het} = \sum_{i=1}^{k} w_i \left[\ln(\hat{IRR}_i) \right]^2 - \left(\sum_{i=1}^{k} w_i \right) \left[\ln(\hat{IRR}_i) \right]^2$$, que sigue una distribución chi-cuadrado ($\chi^2_{k-1;1-\alpha}$) con k – 1 grados de libertad para el valor de alfa prefijado.
donde:
wi: promedio ponderado estimado para el número de eventos en cada grupo según exposición para cada estrato, se estima por:

$$w_i = \left(\frac{1}{a_{1i}} + \frac{1}{a_{2i}} \right)^{-1} \text{ o}$$

$$w_i = \frac{1}{Var\left[\ln(\hat{IRR}_i) \right]} \text{ o}$$

$$Var\left[\ln(\hat{IRR}_i) \right] = \frac{1}{a_{1i}} + \frac{1}{a_{2i}}$$

donde:
t: número del estrato correspondiente.
$\hat{IRR}$: razón de tasas de incidencia estimada en el i-ésimo estrato, se estima por:

$$\hat{IRR}_i = \left(\frac{a_{1i}}{t_{1i}} \right) / \left(\frac{a_{2i}}{t_{2i}} \right)$$

$\hat{IRR}$: razón de tasas de incidencia estimada en conjunto se estima por:

$$\hat{IRR} = e^c$$

$$c = \frac{\sum_{i=1}^{k} w_i \ln(\hat{IRR}_i)}{\sum_{i=1}^{k} w_i}$$

– Valor de p (p-valor) bajo H_0:
p-valor = Pr [χ^2_{k-1} (teórico-esperado) ≥ χ^2_{het} (observado-calculado)].

Regla de decisión

- Si $\chi^2_{\text{de heterogeneidad}}$ (observado-calculado) > $\chi^2_{k-1,\,1-\alpha}$ (teórico-esperado, entonces $\chi^2_{\text{de heterogeneidad}}$ observado se ubica en la región de rechazo para H_0, por lo que existe evidencia suficiente para rechazar la hipótesis nula y se acepta la alternativa, estimándose diferencias significativas entre la razón de tasas de incidencia en los diferentes estratos estudiados para ambos grupos, por lo que no hay homogeneidad en el comportamiento de la razón de tasas de incidencia a través de los estratos, estimándose un p-valor correspondiente al chi-cuadrado observado o una probabilidad relacionada con el azar para este resultado menor que el valor alfa para el nivel de confianza seleccionado ($p < \alpha$).
- Si χ^2 de heterogeneidad (observado-calculado) ≤ $\chi^2_{k-1,\,1-\alpha}$ (teórico-esperado), entonces χ^2 de heterogeneidad observado no se ubica en la región de rechazo para H_0, por lo que no existe evidencia suficiente para rechazar la hipótesis nula y se acepta la alternativa, estimándose diferencias no significativas entre la razón de tasas de incidencia en los diferentes estratos estudiados para ambos grupos, por lo que hay homogeneidad en el comportamiento de la razón de tasas de incidencias a través de los estratos, estimándose un p-valor correspondiente al chi-cuadrado observado o una probabilidad relacionada con el azar para este resultado igual o mayor que el valor alfa para el nivel de confianza seleccionado ($p \geq \alpha$).

z4) Más de dos muestras relacionadas, una variable cuantitativa aleatoria que sigue o se aproxima a la distribución de frecuencias normal estándar, si se desea comparar más de dos medias poblacionales estimadas a partir de más de dos medias muestrales calculadas con datos repetidos de la variable dependiente cuantitativa (se comporta como más de dos poblaciones relacionadas) para una caraterística o categoría dada, con empleo del estadístico de prueba F por ANOVA de una vía para medidas repetidas, en la evaluación de la significación estadística del resultado obtenido con los grados de libertad correspondientes y el valor de alfa prefijado el posible planteamiento de hipótesis estadística para el contraste sería:

- Contraste bilateral:

  H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ *vs.* H_1: $\mu_1 \neq \mu_2 \neq \ldots \neq \mu_k$.

  Los pasos a seguir para realizar el contraste comienzan con la ayuda primeramente de la representación gráfica (Tabla 4.20) para ubicar los elementos teóricos necesarios a incluir en este contraste e introducir en cada celda los valores respectivos a partir de la información recogida previamente con los datos de la variable cuantitativa para cierta característica o categoría dada en las más de dos muestras relacionadas.

Tabla 4.20. Representación tabulada de los elementos teóricos necesarios para el análisis comparativo de más de dos medias de una variable cuantitativa para una característica o categoría dada, correspondientes con más de dos muestras relacionadas

Número de observaciones en las muestras	Primera medición (1)	Segunda medición (2)	Última medición (t)
1	$X_{1,1}$	$X_{1,2}$	$X_{1,t}$
2	$X_{2,1}$	$X_{2,2}$	$X_{2,t}$
3	$X_{3,1}$	$X_{3,2}$	$X_{3,t}$
……	…..	…..	…..
N	n_i; donde: i = 1	n_i; donde: i = 2	n_i; donde: i = t
$N = n_1 + n_2 + \ldots + n_t$			

El estadístico de prueba o contraste a emplear es F por ANOVA de una vía o dirección, para grados de libertad k – 1, N – t o k y un nivel de confianza $(1 - \alpha)\%$ que adopta la expresión:

$F\ (\text{observado-calculado}) = \frac{\hat{S}_E^2}{\hat{S}_D^2} \rightarrow F_{t-1,N-t}$, que sigue una distribución teórica de probabilidades F de Fisher-Snedecor con t – 1 y N – t grados de libertad en el numerador y el denominador, respectivamente, para el valor de alfa según el nivel de confianza preseleccionado.

donde:

$\hat{S}_D^2 = \frac{SCD}{N-t}$: intravarianza, cuasivarianza para los valores de la variable en cada medición, tomados en más de dos momentos repetidos diferentes que se comportarían como más de dos grupos o niveles, estas mediciones repetidas realizadas;

$SCD = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (x_{ij} + \overline{x}_{i.})^2$ suma de las diferencias de los valores de la variable cuantitativa para una característica o categoría dada, respecto a la media obtenida en cada medición de esta variable para la misma característica o categoría dada, al cuadrado, donde cada medición se comporta como un nivel, grupo o muestra;

$\hat{S}_E^2 = \frac{SCE}{t-1}$: intervarianza, cuasivarianza entre las distintas mediciones repetidas de la variable para una característica o categoría dada, donde cada medición equivaldría a cada nivel, grupo o muestra relacionada;

$SCE = \sum_{i=1}^{t} n_i(\overline{x}_{i\bullet} + \overline{x}_{\bullet\bullet})^2$: suma de las diferencias de las medias de cada medición para una característica o categoría dada, respecto a la media de las medias para los más de dos momentos de medición repetidos de los valores de la variable cuantitativa (equivale a más de dos niveles, grupos o muestras) para la misma característica o categoría dada, al cuadrado;

$\hat{S}_T^2 = \frac{SCT}{N-t}$: cuasivarianza total, toda la varianza para los más de dos momentos de medición repetidos con los datos de la variable para la misma característica o categoría dada;

$SCT = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(x_{ij} + \overline{x}_{\bullet\bullet})^2$: suma de las diferencias de todos los valores de la variable cuantitativa para una característica o categoría dada de cada momento de medición respecto a la media de las medias de los más de dos momentos de medición repetidos de la variable para la misma característica o categoría dada, al cuadrado;

$N = \sum_{i=1}^{t} n_i$: sumatoria del total de las observaciones de la variable cuantitativa para una característica o categoría dada en la medición i (n_i), en los más de dos momentos de medición repetidos de esta variable (t);

$x_{i\bullet} = \sum_{j=1}^{n_i} x_{ij}$: suma de los valores de la variable cuantitativa para una característica o categoría dada, según el número de observaciones de la medición i (ni) se comporta para las observaciones de la muestra, grupo o nivel i;

$\overline{x}_{i\bullet} = \frac{x_{i\bullet}}{n_i}$: media de los valores de la variable cuantitativa para una característica o categoría dada, según el número de observaciones en la medición i, se comporta para la muestra, grupo o nivel i;

$x_{\bullet\bullet} = \sum_{i=1}^{t}\sum_{j=1}^{n_i} x_{ij} = \sum_{i=1}^{t} n_i\overline{x}_{i\bullet}$: suma de los valores de la variable cuantitativa para una característica o categoría dada según el número de observaciones de la medición i (ni), en los más de dos momentos de medición repetidos realizados (t), se comporta para las observaciones de todas las muestras, grupos o niveles t;

$\overline{x}_{\bullet\bullet} = \frac{x_{\bullet\bullet}}{N}$: media de los valores de la variable cuantitativa para una característica o categoría dada en todos los más de dos momentos de medición repetidos realizados;

n_i: número de observaciones para una de las más de dos mediciones repetidas a realizar;
t: número de momentos de medición repetidos de la variable cuantitativa para una característica o categoría dada.

Para recordar

Los grados de libertad son el número de unidades independientes de información en una muestra, que son relevantes para la estimación de un parámetro o para el cálculo de un estadístico de contraste. A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo Fexp el valor observado-calculado del estadístico de prueba o contraste y Fteo el valor teórico-esperado de este estadístico con grados de libertad t – 1, N – t para un nivel de significación alfa prefijado entonces:

- Intervalo donde no se rechaza H_0, para cualquier valor de Fexp ≤ Fteo $_{(N - t, t - 1; 1 - \alpha}$:
  Región crítica (donde se rechaza H_0), para cualquier valor de Fexp > Fteo $_{(N - t, t - 1; 1 - \alpha)}$.
- Valor de p o p-valor bajo H_0:
  p-valor = Pr (Fteo ≥ Fexp).

Regla de decisión

La regla de decisión respecto a la significación estadística al contrastar las hipótesis (nula y alternativa) está determinada por:

- Si F (calculado-observado) > F (teórico-esperado) según los grados de libertad respectivos y el valor de alfa preseleccionado, esto se corresponde con la ubicación de F observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces hay evidencias suficientes para rechazar H_0 y, por ende, se acepta H_1 (la hipótesis alternativa), siendo estadísticamente significativa la diferencia entre las más de dos medias utilizadas por el método ANOVA de una vía para medidas repetidas, estimándose una probabilidad relacionada con el azar para ese resultado o un p-valor bajo H_0 para el F observado menor a la correspondiente con el error de tipo 1 para el modelo teórico de referencia ($p < \alpha$).
- Si F (calculado-observado) ≤ F (teórico-esperado) según los grados de libertad respectivos y el valor de alfa preseleccionado, esto no se corresponde con la ubicación de F observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces no hay evidencias suficientes para rechazar H_0 y, por ende, no se acepta H_1 (la hipótesis alternativa), no siendo estadísticamente significativa la diferencia entre las más de dos medias utilizadas por el método ANOVA de una vía, estimándose

una probabilidad relacionada con el azar para este resultado o un p-valor bajo H_0 para el F observado igual o superior a la correspondiente con el error de tipo 1 para el modelo teórico de referencia ($p \geq \alpha$).

z5) Más de dos muestras relacionadas, una variable cualitativa dicotómica que sigue una distribución de frecuencias binomial, si se desea determinar la homogeneidad en las poblaciones relacionadas estimadas para las proporciones calculadas a partir dela variable dicotómica, empleando la prueba de Cochran, el posible planteamiento de hipótesis estadísticas para el contraste sería:

- H_0: existe (poblacionalmente) homogeneidad en las más de dos poblaciones relacionadas estimadas para el comportamiento de las proporciones calculadas para la variable dicotómica.
- H_1: no existe (poblacionalmente) homogeneidad en las más de dos poblaciones relacionadas estimadas para el comportamiento de las proporciones calculadas para la variable dicotómica.

  Los pasos a seguir en el contraste incluyen primeramente la ayuda de la presentación tabulada de los elementos teóricos necesarios (Tabla 4.21) y la ubicación en cada celda de los valores respectivos a partir de la información recogida previamente con los datos de la variable cualitativa dicotómica para las más de dos muestras relacionadas.

Tabla 4.21. Representación tabular de los elementos teóricos a emplear en el análisis de la homogeneidad de las más de dos muestras relacionadas según el comportamiento de una variable cualitativa dicotómica

Número de observaciones para el total de las muestras pareadas	**Total de muestras pareadas con el valor (0 o 1) según la categoría de la variable cualitativa dicotómica para cada observación y cada una de las muestras relacionadas**				
	Muestra 1	**Muestra 2**	**Muestra …**	**Muestra M**	**Total**
1	$X_{1,1}$	$X_{1,2}$	…	$X_{1,m}$	S_1
2	$X_{2,1}$	$X_{2,2}$	…	$X_{2,m}$	S_2
…	…	…	…	…	…
N	$X_{N,1}$	$X_{N,2}$	…	$X_{N,m}$	S_N
Total	$\mathbf{T_1}$	$\mathbf{T_2}$	…	$\mathbf{T_m}$	**T**
Proporción	P_1	P_2	…	P_m	P

El estadístico de prueba a emplear para realizar el contraste sería chi-cuadrado (observado-calculado) o Q de Cochran:

$$Q = \frac{N^2(m-1)}{m} \cdot \frac{\sum_{j=1}^{m}(p_j - \bar{p})^2}{N\bar{p}(1-\bar{p}) - \sum_{n=1}^{N}(P_N - \bar{p})^2} = (m-1) \cdot \frac{m\sum_{j=1}^{m}T_j^2 - T^2}{mT - \sum_{n=1}^{N}S_n^2}$$, que sigue una distribución teórica de probabilidades chi-cuadrado ($\chi^2_{m-1,1-\alpha}$) con m – 1 grados de libertad para el nivel de significación alfa prefijado.

donde:

N: total de observaciones realizadas para cada muestra relacionada;

m: número de muestras relacionadas;

cada X: se corresponde a un valor 0 si la respuesta de la variable dicotómica empleada es negativa y valor 1 lo contrario;

S_1: representa el número total de sujetos con respuesta positiva para la variable dicotómica en la primera observación de las m muestras;

T_1: representa el número total de sujetos con respuesta positiva para la variable dicotómica de la primera muestra;

T: representa el número total de sujetos con respuesta positiva de la variable dicotómica en todas las muestras;

$P_1 = T_1/N$: proporción de sujetos con respuesta positiva en la muestra 1 para el total de observaciones de la muestra 1.

$$p_j = \frac{T_j}{N}$$

$$P_n = \frac{S_n}{m}$$

$$\bar{p} = \frac{1}{m}\sum_{j=1}^{m}p_j = \frac{1}{N}\sum_{n=1}^{N}P_N = \frac{T}{Nm}$$

- Valor p (p-valor), bajo H_0:
  p-valor = Pr ($\chi^2_{m-1,1-\alpha}$ (teórico-esperado) ≥ Q de Cochran (observado-calculado)

Regla de decisión

- Si Q de Cochran (calculado-observado) > $\chi^2_{m-1,1-\alpha}$ (teórico-esperado), entonces Q de Cochran se ubica en la región de rechazo para H_0, por lo que existe evidencia suficiente para rechazar la hipótesis nula y se acepta la alternativa, por lo que no existe una homogeneidad entre las proporciones en relación a las categorías de la variable dicotómica entre las más de dos muestras relacionadas, estimándose un p-valor (probabilidades según el valor del estadístico observado de cumplirse H_0 o relacionadas con el azar para el resultado obtenido) menor que el valor de alfa empleado para el nivel de confianza seleccionado ($p < \alpha$).

– Si Q de Cochran (calculado-observado) ≤ $\chi^2_{m-1,1-\alpha}$ (teórico-esperado), entonces Q de Cochran observado no se ubica en la región de rechazo para H_0, por lo que no existe evidencia suficiente para rechazar la hipótesis nula y no se acepta la alternativa, por lo que existe una homogeneidad de las proporciones en relación a las categorías de la variable dicotómica entre las más de dos muestras relacionadas, estimándose un p-valor (probabilidades según el valor del estadístico observado de cumplirse H_0 o relacionadas con el azar para el resultado obtenido) igual o mayor que el valor de alfa empleado para el nivel de confianza seleccionado ($p \geq \alpha$).

z6) Más de dos muestras no relacionadas, una variable cuantitativa aleatoria que sigue o se aproxima su distribución de frecuencias a la distribución de probabilidades normal estándar, si se desea comparar más de dos medias muestrales calculadas con datos no repetidos de una misma variable cuantitativa (se comporta como más de dos poblaciones estimadas no relacionadas) según una característica o categoría dada con el empleo del estadístico de prueba F por ANOVA de una vía para medidas no repetidas, en la evaluación de la significación estadística del resultado obtenido con los grados de libertad correspondientes y el valor de alfa prefijado el posible planteamiento de hipótesis estadísticas para el contraste sería:

– Contraste bilateral:

H_0: $\mu_1 = \mu_2 = ... = \mu_k$ *vs.* H_1: $\mu_1 \neq \mu_2 \neq ... \neq \mu k$.

Los pasos a seguir para realizar el contraste comienzan con la ayuda primeramente de la representación gráfica (Tabla 4.22) para ubicar los elementos teóricos necesarios a incluir en este contraste e introducir en cada celda los valores respectivos a partir de la información recogida previamente con los datos de la variable cuantitativa para cierta característica o categoría dada en las más de dos muestras no relacionadas.

Tabla 4.22. Representación tabulada de los elementos teóricos necesarios para el análisis comparativo de más de dos medias de una variable cuantitativa para una característica o categoría dada, correspondientes con más de dos muestras no relacionadas

Número de observaciones en las muestras	Muestra (1)	Muestra (2)	Muestra (t)
1	$X_{1,1}$	$X_{1,2}$	$X_{1,t}$
2	$X_{2,1}$	$X_{2,2}$	$X_{2,t}$
3	$X_{3,1}$	$X_{3,2}$	$X_{3,t}$
...	...	...	...
n (tamaño muestral)	n_i; donde: i = 1	n_i; donde: i = 2	n_i; donde: i = t
$N = n_1 + n_2 + ... + n_t$			

Asumiendo que las poblaciones estimadas de las más de dos muestras independientes tienen la misma varianza σ^2, el estadístico de prueba a seleccionar F es la razón o el cociente de los siguientes dos estimados de σ^2:

- La variación entre muestras con base en la variación entre medias muestrales.
- La variación dentro de muestras con base en las varianzas muestrales.

  Este contraste de hipótesis sigue el procedimiento de Triola (2009) y tiene dos variantes según sea el tamaño de las muestras o grupos de datos independientes iguales o no como sigue.
- Para tamaños muestrales n iguales:
  - Primer método:

    El estadístico de contraste sería F y se expresa como:

    $$F = \frac{\text{Varianza entre las muestras}}{\text{Varianza dentro de las muestras}}$$

    $$F = \frac{\text{Varianza para las medias muestrales}}{\text{Media de las varianzas muestrales}} = \frac{nS_{\bar{x}}^2}{S_p^2}$$

    donde:

    $S_{\bar{x}}^2$: varianza de las medias muestrales o de los grupos de datos;

    n: tamaño de cada una de las muestras;

    S_p^2: varianza agrupada que se obtiene al calcular la media de las varianzas muestrales.

    Este estadístico sigue la distribución de F de Snedecor con grados de libertad k – 1 (en el numerador); k (n – 1) o N – k, donde N = Σn, para el denominador.
  - Segundo método:

    ANOVA de una vía por el método de la suma de cuadrados (SC) y cuadrados medios (CM):

    El estadístico de contraste sería F y se expresa como:

    $$F = \frac{\text{CM (del tratamiento)}}{\text{CM (del error)}}$$

    donde:

    Grados de libertad en el numerador k – 1 y en el denominador N – k;

    $\text{CM (del tratamiento)} = \frac{\text{SC (del tratamiento)}}{k-1}$: intervarianza entre grupos, entre muestras y entre tratamientos;

    $\text{CM (del error)} = \frac{\text{SC (del error)}}{N-k}$: intravarianza dentro de los grupos, dentro de las muestras y del error;

    $\text{CM (total)} = \frac{\text{SC (total)}}{N-1}$: cuasivarianza total;

SC (del tratamiento): entre los grupos, entre las muestras y de tratamientos:

$$\text{SC (del tratamiento)} = n_1(\overline{x}_1 - \overline{\overline{x}})^2 + n_2(\overline{x}_2 - \overline{\overline{x}})^2 + n_k(\overline{x}_k - \overline{\overline{x}})^2 = \sum n_i(\overline{x}_i - \overline{\overline{x}})^2$$

SC (del error): dentro de los grupos, dentro de las muestras y del error:

$$\text{SC (del error)} = (n_1 - 1)s_1^2 + (n_1 - 1)s_1^2 + (n_k - 1)s_k^2 = \sum (n_i - 1)s_i^2$$

$$\text{SC (total)} = \sum (\overline{x}_i - \overline{\overline{x}})^2$$

SC (total) = SC (del tratamiento) + SC (del error)

En otros textos este contraste de hipótesis se puede encontrar con otra simbología, aunque en síntesis sigue el mismo hilo conductor hasta arribar a las mismas conclusiones para la diferencia de más de dos medias poblacionales estimadas de las de los grupos o muestras para los datos de la variable cuantitativa utilizando la relación entre los parámetros cuasivarianza entre niveles, entre grupos o intervarianza:

$$\hat{S}_E^2 = \frac{SCE}{t-1}$$

y la cuasivarianza intranivel, intragrupo o intravarianza:

$$\hat{S}_D^2 = \frac{SCD}{N-t}$$

El estadístico de prueba o contraste a emplear es F por ANOVA de una vía o dirección y se expresa como:

$$\text{F-Snedecor (observado} - \text{experimental)} = \frac{\hat{S}_E^2}{\hat{S}_D^2} \rightarrow F_{t-1,N-t}$$, que sigue una distribución con los grados de libertad t – 1 o k – 1 (en el numerador) y N – t o N – k (en el denominador) para un nivel de confianza (1 – α) %.

donde:

$\hat{S}_E^2 = \frac{SCE}{t-1}$: intervarianza, cuasivarianza entre las más de dos muestras independientes o grupos de datos para los valores de la variable cuantitativa en relación a una característica o categoría dada;

$SCE = \sum_{i=1}^{t} n_i(\overline{x}_{i\cdot} + \overline{x}_{\cdot\cdot})^2$: suma de las diferencias de las medias de los valores de la variable cuantitativa para una característica o categoría de cada muestra, respecto a la media para todas las muestras en conjunto, al cuadrado;

$\hat{S}_D^2 = \frac{SCD}{N-t}$: intravarianza, cuasivarianza dentro de las más de dos muestras independientes o grupos de datos para los valores de la variable cuantitativa para una categoría o característica dada;

$SCD = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(x_{ij} + \bar{x}_{i\cdot})^2$: suma de las diferencias de los valores de la variable cuantitativa para una característica o categoría dada dentro de cada muestra respecto a la media obtenida de dichos valores, al cuadrado;

$\hat{S}_T^2 = \frac{SCT}{N-t}$: cuasivarianza total, toda la varianza para el total de los valores de la variable cuantitativa para una característica o categoría dada en todas las observaciones de las muestras independientes empleadas;

$SCT = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(x_{ij} + \bar{x}_{\cdot\cdot})^2$: suma de las diferencias de todos los valores de la variable cuantitativa para una característica o categoría dada de todas las muestras independientes o grupos de datos respecto a la media de todos esos valores de las más de dos muestras independientes, al cuadrado;

$N = \sum_{i=1}^{t} n_i$: sumatoria de los tamaños de muestra (ni), en las más de dos muestras independientes (i);

$x_{i\cdot} = \sum_{j=1}^{n_i} x_{ij}$: suma de todos los valores (j) de la variable cuantitativa (X) en relación a una característica o categoría dada, para todas las muestras (i);

$\bar{x}_{i\cdot} = \frac{x_{i\cdot}}{n_i}$: media de los valores de cada muestra (i) de la variable cuantitativa (X) para una característica o categoría dada;

$x_{\cdot\cdot} = \sum_{i=1}^{t}\sum_{j=1}^{n_i} x_{ij} = \sum_{i=1}^{t} n_i \bar{x}_{i\cdot}$: suma de los valores (j) de la variable cuantitativa (X) para una característica o categoría dada para las más de dos muestras independientes empleadas (i);

t: total de muestras independientes o grupos de datos;

– Para tamaños muestrales n desiguales:

$$F = \frac{\text{Varianza entre las muestras}}{\text{Varianza dentro de las muestras}} = \frac{\left[\frac{\sum n_i(\bar{x}_i - \bar{\bar{x}})^2}{k-1}\right]}{\left[\frac{\sum (n_i - 1)s_i^2}{\sum (n_i - 1)}\right]}$$

donde:

$\bar{\bar{x}}$: media de todos los valores muestrales combinados;

k: número de medias poblacionales estimadas (grupos de datos) que se están comparando;

ni: número de valores en la i-ésima muestra;

$\bar{x}$: media de los valores en la i-ésima muestra;

s_i^2 : varianza de los valores en la i-ésima muestra.

Para recordar

El numerador del estadístico de prueba F mide la variación entre medias muestrales. El estimado de la varianza en el denominador depende únicamente de las varianzas muestrales y no se ve afectado por las diferencias entre las medias muestrales. Como consecuencia las medias muestrales que tienen valores cercanos dan como resultado un estadístico de prueba F pequeño y se concluye que no existen diferencias significativas entre las medias muestrales. Pero si el valor de F es excesivamente grande, entonces se rechaza la aseveración de igualdad de medias. Los ambiguos términos "pequeño" y "excesivamente grande" se vuelven objetivos por medio del valor p correspondiente, que indica si el estadístico de prueba F está o no en la región crítica.

A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo Fexp el valor observado-calculado del estadístico de prueba o contraste y Fteo el valor teórico-esperado de este estadístico con grados de libertad t – 1, N – t para un nivel de significación alfa prefijado, entonces:

- Intervalo donde no se rechaza H_0, para cualquier valor de
  $Fexp \leq Fteo_{(N-t,\, t-1;\, 1-\alpha)}$
  Región crítica (donde se rechaza H_0), para cualquier valor de
  $Fexp > Fteo_{(N-t,\, t-1;\, 1-\alpha)}$.
- Valor de p o p-valor bajo H_0:
  p-valor = Pr (Fteo ≥ F exp).

Regla de decisión

La regla de decisión respecto a la significación estadística al contrastar las hipótesis (nula y alternativa) estará determinada por:

- Si F (calculado-observado) > F (teórico-esperado) según los grados de libertad correspondientes y el valor de alfa, esto se corresponde con la ubicación de F observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces hay evidencias suficientes para rechazar H_0 y, por ende, se acepta H_1 (la hipótesis alternativa), siendo estadísticamente significativa la diferencia entre las más de dos medias utilizadas por el método ANOVA de una vía, estimándose una probabilidad relacionada con el azar para esta resultado o un p-valor para el F observado bajo H_0 menor a la esperada del modelo teórico de referencia ($p < \alpha$).
- Si F (calculado-observado) ≤ F (teórico-esperado) según el valor de alfa y los grados de libertad prefijados, esto no se corresponde con la ubicación de F observado en el intervalo de los valores del área crítica o de rechazo de H_0

según el modelo teórico de referencia, lo que aumenta la probabilidad de cometer el error tipo 1, entonces no hay evidencias suficientes para rechazar H_0 y, por ende, no se acepta H_1 (la hipótesis alternativa), no siendo estadísticamente significativa la diferencia entre las más de dos medias utilizadas por el método ANOVA de una vía, estimándose una probabilidad relacionada con el azar para este resultado o un p-valor para el F observado bajo H_0 igual o superior a la esperada del modelo teórico de referencia ($p \geq \alpha$).

z7) Más de dos muestras no relacionadas, una variable cuantitativa aleatoria que sigue o se aproxima su distribución de frecuencias a la distribución de probabilidades normal estándar, si se desea comparar más de dos varianzas muestrales calculadas con datos de una misma variable cuantitativa (se comporta como más de dos poblaciones estimadas) según una característica o categoría dada, el posible planteamiento de hipótesis estadísticas para el contraste sería:

– Contraste bilateral:

H_0: $\sigma^2_1 = \sigma^2_2 = \sigma^2$ k *vs.* H_1: $\sigma^2_1 \neq \sigma^2_2 \neq \sigma^2$k.

Entre las pruebas especificas para este tipo de contraste están:

• Prueba de Barlett:

Para la mayoría de los autores es la prueba más utilizada para contrastar si más de dos muestras aleatorias son homocedásticas (igual varianzas) y brinda la posibilidad de comparar grupos de datos o muestras con diferentes tamaños.

Tiene como probable planteamiento para el contraste de hipótesis:

$$H_0: \ \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

$$H_1: \ \sigma_1^2 \neq \sigma_j^2 \quad \text{para algún i, j}$$

donde: los i son los valores observados dentro de cada grupo de datos o muestras j y no se requiere igual tamaño de cada muestra.

El estadístico de prueba o de contraste es chi-cuadrado, que adopta la expresión:

$$\chi^2(\text{observado-calculado}) \text{ de Barlett } = \frac{1}{k}\left[(N-T)\ln\hat{S}_D^2 - \sum_{i=1}^{t}\ln\hat{S}_i^2\right],$$

que sigue una distribución teórica chi-cuadrado ($\chi^2_{t-1,1-\alpha}$), con t – 1 grados de libertad para el valor de alfa prefijado;

donde:

N: número de observaciones (sujetos) en toda la muestra;

t: número total de muestras o grupos según las categorías del factor (variable cualitativa multinominal);

$\hat{S}$: cuasivarianza de cada muestra.

También se ha utilizado por diferentes autores esta expresión para el estadístico de Barlett:

$$\chi^2(\text{observado-calculado) de Barlett } = \frac{\left[\ln\frac{\sum\sigma^2(n-1)}{\sum(n-1)}\sum(n-1)\right] - \sum\ln\sigma^2(n-1)}{1+\frac{K+1}{3(K-1)(N-K)}}$$

que sigue una distribución teórica de probabilidades chi-cuadrado ($\chi^2_{n-1,1-\alpha}$) con n – 1 grados de libertad para el valor de alfa prefijado donde:

σ^2: varianza de cada muestra que estima una poblacional (se asume como igual a S^2);

n: tamaño de cada muestra;

K: número de muestras comparadas;

N (Σn_i): tamaño total de cada muestra.

Valor de p o p-valor bajo H0:

p-valor = Pr[χ^2 (teórico-esperado) ≥ χ^2 (observado-calculado)].

Regla de decisión

- Si χ^2 (observado-calculado) de Barlett > $\chi^2_{t-1,1-\alpha}$ (teórico-esperado), esto ubica el valor de χ^2 observado en el área crítica o de rechazo en la distribución de valores F teóricos, por ende, hay evidencia para rechazar H_0 y se puede aceptar H_1, por lo cual existen evidencias para estimar que al menos las varianzas de dos de las muestras son significativamente diferentes entre sí, al estimarse un valor de probabilidades (p-valor) bajo H_0 para el valor χ^2 de Barlett observado o que se relacionan con el azar inferiores al valor de probabilidad del modelo teórico de referencia según el valor alfa prefijado ($p < \alpha$).
- Si χ^2 (observado-calculado) de Barlett ≤ $\chi^2_{t-1,1-\alpha}$ (teórico-esperado), esto no ubica el valor de χ^2 de Barlett observado en el área crítica o de rechazo en la distribución de valores F teóricos, por ende, no se puede rechazar H^0, por lo que no existen evidencias para estimar que al menos las varianzas de dos de las muestras son significativamente diferentes entre sí, al estimarse un valor de probabilidades (p-valor) bajo H^0 para el valor χ^2 de Barlett observado o que se relacionan con el azar iguales o superiores al valor de probabilidad del modelo teórico de referencia según el valor alfa prefijado ($p \geq \alpha$).
- Prueba de Levene:

  Tiene como probable planteamiento para el contraste de hipótesis:

  $H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$

  $H_1: \sigma_1^2 \neq \sigma_j^2$ para algún i, j

donde los i son los valores observados dentro de cada grupo de datos o muestras j y no se requiere igual tamaño de cada muestra.

Los pasos a seguir en esta prueba:

- Calcular la diferencia (en valor absoluto) entre cada valor y la media de su grupo de datos:

  $$D_{ij} = \left| x_{ij} - \overline{x}_j \right|$$

  donde:

  x_{ij}: valor de la variable correspondiente al sujeto i, perteneciente al grupo j;

  $\overline{x}_j$: media del grupo j.

- Calcular la media de las diferencias de cada grupo:

  $$\overline{D}_j = \frac{\sum D_{ij}}{n_i}$$

  donde:

  $\sum D_{ij}$: suma de las diferencias;

  n_i: tamaño de cada grupo.

- Calcular la media total de las diferencias:

  $$\overline{D}_t = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} D_{ij}}{N}$$

- Calcular la suma de cuadrados intragrupos:

  $$SC_{intra} = \sum_{i=1}^{n}\sum_{j=1}^{k}(D_{ij} - D_j)^2$$

- Calcular la suma de cuadrados intergrupos:

  $$SC_{inter} = \sum_{i=1}^{k} n_j(\overline{D}_j - \overline{D}_t)^2$$

- Calcular los grados de libertad intergrupos:

  GL intergrupos = k – 1, donde k es el número de grupos de datos.

- Calcular los grados de libertad intragrupos:

  $$\text{GL intragrupos} = \sum_{i=1}^{k}(n_j - 1)^2$$

- Calcular la media cuadrática intergrupos:

  $MC_{inter} = SC_{inter}/GL_{inter}$

- Calcular la media cuadrática intragrupos:

  $MC_{intra} = SC_{intra}/GL_{intra}$

- Calcular el estadístico de prueba o contraste:

  $F = MC_{inter}/MC_{intra}$

- Calcular el valor de F teórico-esperado (crítico) por la tabla para la distribución de F de Snedecor según los grados de libertad en el numerador (k – 1) y en el denominador $\sum_{j=1}^{k}(n_j - 1)^2$ para un nivel de significación alfa seleccionado por el investigador.
- Comparar el F (observado- calculado) *vs.* F (tabulado-crítico).
  Valor de p o p-valor bajo $H_{0:}$
  p-valor = Pr[F (teórico-esperado) ≥ F (observado-calculado)].

Regla de decisión

- Si F (observado-calculado) > F (teórico- esperado) según los grados de libertad correspondientes y para (1 – α)% como nivel de confianza, esto ubica el valor de F observado en el área crítica o de rechazo en la distribución de valores F teóricos, por ende, hay evidencia para rechazar H_0 y se puede aceptar H_1, por lo que existen evidencias para estimar que al menos las varianzas de dos de las muestras son significativamente diferentes entre sí, al estimarse un valor de probabilidades (p-valor) bajo H_0 para el valor de F observado o que se relacionan con el azar inferiores al valor de probabilidad del modelo teórico de referencia según el valor alfa prefijado (p < α).
- Si F (observado-calculado) ≤ F (teórico- esperado) según los grados de libertad correspondientes y para (1 – α)% como nivel de confianza, esto no ubica el valor de F observado, por ende, no se puede rechazar H_0, por lo que no existen evidencias para estimar que al menos las varianzas de dos de las muestras son significativamente diferentes entre sí, al estimarse un valor de probabilidades (p-valor) bajo H_0 para el valor de F observado o que se relacionan con el azar iguales o superiores al valor de probabilidad del modelo teórico de referencia según el valor alfa prefijado (p ≥ α).
- Prueba de Cochran:
  Aunque implica cálculos más sencillos, está algo restringida para el caso de tamaños de muestra o de los grupos de datos que sean iguales y que se cumpla la alietoriedad en el orden de extracción de los datos y la normalidad en la distribución para las observaciones realizadas. Es principalmente útil cuando se desea detectar si una varianza es mucho mayor que las otras.
  Tiene como probable planteamiento para el contraste de hipótesis:

  $H_0: \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$

  $H_1: \sigma_1^2 \neq \sigma_j^2$ para algún i, j

  El estadístico de prueba o de contraste es R (Cochran), que adopta la expresión:

R de Cochran (observado-calculado)$= \frac{\text{máx}\left\{\hat{S}_i^2\right\}_{i=1}^{t}}{\sum_{i=1}^{t}\hat{S}_i^2}$, que sigue una distribución teórica de probabilidades R de Cochran para n – 1, t o k, 1 – a

donde:

$\hat{S}_i^2$: cuasivarianza muestral para la muestra, el nivel o grupo de datos i y se calcula por la expresión:

$$\hat{S}_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} + \overline{x}_{i\bullet})^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}x_{ij}^2 - \frac{n_i}{n_i - 1}\overline{x}_{i\bullet}^2$$

N = n x t: representa el número total de observaciones;

n: tamaño de cada muestra, nivel o grupo de datos (es el mismo en cada uno);

t o k: representa el número total de muestras, niveles o grupos de datos;

j: valores observados u observaciones dentro de cada grupo de datos o muestras;

i: números correspondientes a cada muestra o grupo de datos observados.

Una vez calculado el valor del estadístico de prueba o contraste: R de Cochran, del mismo se comparará su valor observado *versus* R de Cochran $_{n-1,\,t\,o\,k,\,1-\alpha}$ teórico-esperado ubicado en las tablas de referencia para el mismo (véase el anexo 1).

Valor de p o p-valor bajo H_0:

p-valor = Pr [R de Cochran (teórico-esperado) ≥ R de Cochran (observado-calculado)].

Regla de decisión

- Si R de Cochran (observado-calculado) > R de Cochran $_{n-1,\,k,\,1-\alpha}$ (teórico-esperado), esto ubica el valor de R de Cochran observado en el área crítica o de rechazo en la distribución de valores F teóricos, por ende, hay evidencia para rechazar H_0 y aceptar H_1, por lo que existen evidencias para estimar que al menos las varianzas de dos de las muestras son significativamente diferentes entre sí, al estimarse un valor de probabiliadades (p-valor) bajo H_0 para el valor R de Cochran observado o que se relacionan con el azar inferiores al valor de probabilidad del modelo teórico de referencia según el valor alfa prefijado ($p < \alpha$).
- Si R de Cochran (observado-calculado) ≤ R Cochran $_{n-1,\,k,\,1-\alpha}$ (teórico-esperado), esto no ubica el valor de R de Cochran observado en el área crítica o de rechazo en la distribución de valores F teóricos, por ende, hay evidencia para no rechazar H_0, por lo que no existen evidencias para estimar que al menos las varianzas de dos de las muestras son significativamente

diferentes entre sí, al estimarse un valor de probabiliadades (p-valor) bajo H_0 para el valor R de Cochran observado o que se relacionan con el azar iguales o superiores al valor de probabilidad del modelo teórico de referencia según el valor alfa prefijado ($p \geq \alpha$).

z8) Más de dos muestras no relacionadas, una variable cualitativa nominal u ordinal, si se desea determinar la homogeneidad de las poblaciones estimadas de las muestras seleccionadas para las proporciones calculadas a partir de la variable cualitativa empleada, empleando la prueba de Pearson con el uso de chi-cuadrado según los grados de libertad correspondientes y el nivel de significación alfa, auxiliándonos de tablas de contingencia MxN o FxC, el planteamiento de hipótesis estadísticas para contrastar sería:

- H_0: existe (poblacionalmente) homogeneidad en las poblaciones estimadas para las proporciones calculadas según la variable cualitativa nominal u ordinal por niveles.
- H_1: no existe (poblacionalmente) homogeneidad en las poblaciones estimadas para las proporciones calculadas según la variable cualitativa nominal u ordinal por niveles.

  Los pasos a seguir para el contraste incluyen la presentación de forma tabulada de los diferentes elementos teóricos necesarios y la introducción de los respectivos valores en las celdas a partir de la información recogida previamente (Tabla 4.23).

Tabla 4.23. Representación tabular de los elementos teóricos necesarios a emplear para el análisis de la homogeneidad de más de dos muestras independientes según el comportamiento de una variable cualitativa nominal u ordinal por niveles (se comporta como una nominal también)

Variable cualitativa politómica u ordinal	**Muestra 1 (obreros industriales)**	**Muestra 2 (constructores)**	**Muestra 3 (campesinos)**	**Totales marginales**
Categoría 1	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,1} + X_{1,2} + X_{1,3}$
Categoría 2	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	$X_{2,1} + X_{2,2} + X_{2,3}$
Categoría 3	$X_{3,1}$	$X_{3,2}$	$X_{3,3}$	$X_{3,1} + X_{3,2} + X_{3,3}$
Totales	$\mathbf{X_{1,1} + X_{2,1} + X_{3,1}}$	$\mathbf{X_{1,2} + X_{2,2} + X_{3,2}}$	$\mathbf{X_{1,3} + X_{2,3} + X_{3,3}}$	$\mathbf{n = X_{1,1} + X_{1,2} + X_{1,3} + X_{2,1} + X_{2,2} + X_{2,3} + X_{3,1} + X_{3,2} + X_{3,3}}$

– El estadístico de prueba a emplear para el contraste es:

χ^2de Pearson $= \sum_{i,j} \frac{(x_{ij} - e_{ij})^2}{e_{ij}}$, que sigue una distribución teórica de probabilidades chi-cuadrado ($\chi^2_{(c-1)(f-1);1-\alpha}$) con (c –1) (f – 1) grados de libertad para el valor de alfa prefijado.

También se puede emplear el chi-cuadrado con corrección de Yates (χ^2_C) según la fórmula:

$$\chi^2_C = \sum_{i,j} \frac{(x_{ij} - e_{ij} - 0{,}5)^2}{e_{ij}}$$

donde:

$X_{i,j}$: valor observado para cada celda;

$e_{i,j}$: valor esperado para cada celda (total de la fila x total de columna correspondiente para cada celda)/tamaño de muestra;

Grados de libertad: (c – 1) (f– 1);

i: número de orden de las filas;

j: número de orden de las columnas.

Para recordar

A partir de la representación gráfica de la distribución teórica de probabilidades del estadístico de prueba o contraste siendo χ^2 exp el valor observado-calculado del estadístico de prueba o contraste y χ^2 teo el valor teórico-esperado de este estadístico con grados de libertad (c – 1) (f – 1) y nivel de confianza (1 – α)%, entonces:

– Intervalo donde no se rechaza H_0 = (para cualquier valor de χ^2 exp ≤ χ^2 teo $_{(c-1)(f-1);1-\alpha)}$.
Región crítica (donde se rechaza H_0) = (para cualquier valor de χ^2 exp > χ^2 teo $_{(c-1)(f-1);1-\alpha)}$.

– Valor de p o p-valor bajo H_0:
p-valor = Pr[χ^2 (teórico-esperado) ≥ χ^2 (observado-calculado)].

Regla de decisión

La regla de decisión respecto a la significación estadística al contrastar las hipótesis (nula y alternativa) estará determinada por:

– Si χ^2 (observado-calculado) de Pearson> $\chi^2_{(c-1)(f-1);1-\alpha}$ (teórico-esperado), entonces el valor de chi-cuadrado observado de Pearson se ubica en el área de rechazo de H_0, por lo que existen evidencias suficientes para rechazar la hipótesis nula (H_0) y se acepta la alternativa (H_1), es decir, no existe homogeneidad en las poblaciones estimadas de las más de dos muestras independientes para las proporciones calculadas según la variable cualitativa politómica u ordinal, estimándose un p-valor bajo H_0 para el chi-cuadrado observado o siendo las probabilidades del azar para este resultado menores que el valor

de alfa prefijado o de cometer el error de tipo 1 según el nivel de confianza escogido ($p < \alpha$).

– Si χ_2 (observado-calculado) de Pearson $\leq \chi_{2\ (c-1)\ (f-1);\ 1-\alpha}$ (teórico-esperado), entonces el valor de chi-cuadrado observado de Pearson no se ubica en el área de rechazo de H_0, por lo que no existen evidencias suficientes para rechazar la hipótesis nula (H_0) y no se acepta la alternativa (H_1), es decir, no existe homogeneidad en las poblaciones estimadas de las más de dos muestras independientes para las proporciones calculadas según la variable cualitativa politómica u ordinal; estimándose un p-valor bajo H_0 para el chi-cuadrado observado o siendo las probabilidades del azar para este resultado iguales o mayores que el valor de alfa prefijado o de cometer el error de tipo 1 según el nivel de confianza escogido ($p \geq \alpha$).

z9) Más de dos muestras no relacionadas (de grupos según exposición) y una variable cualitativa expresada en forma de nuevos eventos según personas x tiempo de seguimiento (tasas de incidencia), si se desea probar que existe una significativa tendencia lineal al incremento o decrecimiento en las tasas de incidencia para más de dos grupos de expuestos (con diferentes niveles de exposición) empleando la aproximación de la distribución normal estándar: $Z \sim N(0, 1)$, para un nivel de significación alfa predeterminado, el posible planteamiento de hipótesis estadísticas para el contraste sería:

$H_0: \beta = 0$ *vs.* $H_1: \beta \neq 0$

Los pasos a seguir para el contraste incluyen la presentación de forma tabulada de los diferentes elementos teóricos necesarios y la introducción de los respectivos valores en las celdas a partir de la información recogida previamente (Tabla 4.24).

Tabla 4.24. Representación tabular de los elementos teóricos necesarios a emplear para el análisis de tendencia en las tasas de incidencia para grupos con diferentes niveles de exposición.

Niveles de exposición de los grupos estudiados.	Número de nuevos eventos o casos	Número de personas por tiempo de seguimiento
1	a_1	t_1
2	a_2	t_2
…	…	…
N	a_n	t_n

Para la construcción del modelo se emplea la siguiente expresión o fórmula matemática:

$\ln (p_{ij}) = \alpha_i + \beta S_j$

donde:

ln (p_{ij}): representa el logaritmo natural de la incidencia (tasa de incidencia) para cada estrato (i) y nivel de exposición (j);

αi: representa la variación en el logaritmo natural (incidencia) para determinado estrato (i) y el primer nivel de exposición (j = 1);

β: representa la variación (incremento o disminución) en el ln (tasa de incidencia) a partir de la variación del nivel de exposición (en relación al primer nivel de exposición) para los grupos con diferentes niveles de exposición;

Sj: representa el promedio del nivel de exposición dentro de cada grupo expuesto;

a: número de nuevos eventos observados para los distintos niveles de exposición en los grupos estudiados;

t: cantidad de personas por tiempo de seguimiento para los distintos niveles de exposición.

El estadístico de prueba o contraste es Z ~ N (0, 1), que adopta la expresión:

$$Z\ (\text{observado-calculado}) = \frac{\hat{\beta}}{se(\hat{\beta})} \sim N(0,\ 1)$$

donde:

$$\hat{\beta} = \frac{L_{xy}}{L_{xx}}$$

$$L_{xy} = \frac{\sum_{i=1}^{s}\sum_{j=1}^{k} w_{ij}S_j \ln(\hat{p}_{ij}) - \left(\sum_{i=1}^{s}\sum_{j=1}^{k} w_{ij}S_j\right)\left[\sum_{i=1}^{s}\sum_{j=1}^{k} w_{ij}\ln(\hat{p}_{ij})\right]}{\sum_{i=1}^{s}\sum_{j=1}^{k} w_{ij}}$$

$$L_{xx} = \frac{\sum_{i=1}^{s}\sum_{j=1}^{k} w_{ij}S_i^2 - \left(\sum_{i=1}^{s}\sum_{j=1}^{k} w_{ij}S_j\right)^2}{\sum_{i=1}^{s}\sum_{j=1}^{k} w_{ij}}$$

$w_{ij} = a_{ij}$: número de casos en el i-ésimo estrato y el j-ésimo nivel de exposición;

$se(\hat{\beta}) = \frac{1}{\sqrt{L_{xx}}}$: error estándar para $(\hat{\beta})$.

- Estimación del valor de p (p-valor) bajo H_0:
  - Si $z < 0$, p-valor = $2\Phi(z)$.
  - Si $z \geq 0$, p-valor = $2[1 - \Phi(z)]$.

Regla de decisión

- Para Z (observado-calculado) < 0:
  - Si Z (observado–calculado) < Z (teórico-esperado) según alfa, entonces Z observado se ubica en la región de rechazo para H_0, por lo que existe evi-

dencia suficiente para rechazar la hipótesis nula y se acepta la alternativa, por lo que existe una tendencia lineal significativa a la variación en las tasas de incidencia en las más de dos muestras en relación a la variación de los niveles de exposición; estimándose un p-valor bajo H_0 según el valor del estadístico Z observado o de probabilidades relacionadas con el azar para este resultado menor que el valor de alfa empleado para el nivel de confianza seleccionado ($p < \alpha$).

- Si Z (observado –calculado) ≥ Z (teórico-esperado) según alfa, entonces Z observado no se ubica en la región de rechazo para H_0, por lo que no existe evidencia suficiente para rechazar la hipótesis nula, por lo que no existe una tendencia lineal significativa a la variación en las tasas de incidencia en las más de dos muestras en relación a la variación de los niveles de exposición, estimándose un p-valor bajo H_0 según el valor del estadístico Z observado o de probabilidades relacionadas con el azar para este resultado igual o mayor que el valor de alfa empleado para el nivel de confianza seleccionado ($p \geq \alpha$).

– Para Z (observado-calculado) ≥ 0:

- Si Z (observado-calculado) > Z (teórico-esperado) según alfa, entonces Z observado se ubica en la región de rechazo para H_0, por lo que existe evidencia suficiente para rechazar la hipótesis nula y se acepta la alternativa, por lo que existe una tendencia significativa a la variación en las tasas de incidencia en las más de dos muestras en relación a la variación de los niveles de exposición, estimándose un p-valor bajo H_0 según el valor del estadístico Z observado o de probabilidades relacionadas con el azar para este resultado menor que el valor de alfa empleado para el nivel de confianza seleccionado ($p < \alpha$).
- Si Z (observado –calculado) ≤ Z (teórico-esperado) según alfa, entonces Z observado no se ubica en la región de rechazo para H_0, por lo que no existe evidencia suficiente para rechazar la hipótesis nula, por lo que no existe una tendencia significativa a la variación en las tasas de incidencia en las más de dos muestras en relación a la variación de los niveles de exposición, estimándose un p-valor bajo H_0 según el valor del estadístico Z observado o de probabilidades relacionadas con el azar para este resultado igual o mayor que el valor de alfa empleado para el nivel de confianza seleccionado ($p \geq \alpha$).

z10) Más de dos muestras independientes, una variable cualitativa que se emplea de ella una categoría o característica de interés, si se desea determinar la diferencia entre tres proporciones correspondientes a igual número de muestras indepen-

dientes, empleando la prueba de chi-cuadrado para los grados de libertad correspondientes y el nivel de significación alfa predeterminado por el investigador, el posible planteamiento de hipótesis estadísticas para el contraste sería:

H_0: $P_1 = P_2 = P_3 \ldots. = Pn$ *vs.* H_1: $P_1 \neq P_2 \neq P_3 \ldots.\neq Pn$

Los pasos a seguir para el contraste incluyen la presentación de forma tabulada de los diferentes elementos teóricos necesarios y la introducción de los respectivos valores en las celdas a partir de la información recogida previamente (Tabla 4.25).

Tabla 4.25. Representación tabular de los elementos teóricos necesarios a emplear para el análisis comparativo de más de dos proporciones en igual número de muestras no relacionadas

Número de las muestras	Total de sujetos de cada muestra	Número de sujetos con la característica de interés en cada muestra	Número de sujetos sin la característica en cada muestra	Proporción de sujetos con la característica de interés en cada muestra
1	n_1	n_{11}	$n_{1,2}$	P_1
2	n_2	n_{21}	$n_{2,2}$	P_2
……	……..	……..	…….	……..
.				
m	n_m	n_{m1}	n_{m2}	P_m
Total	$n_{..}$	$n_{.1}$	$n_{.2}$	$\overline{p}$

– El estadístico de prueba a emplear en el contraste sería:

$\chi^2(\text{observado-calculado}) = \frac{1}{\overline{p}\,\overline{q}}\sum_{i=1}^{m,} n_i(p_i - \overline{p})^2$, que sigue una distribución teórica de probabilidades $\chi^2_{\,m-1,\,1-\alpha}$ con m – 1 grados de libertad para el valor de alfa prefijado

donde:

m: número de muestras para las proporciones (p_i) empleadas;

$\overline{q} = 1 - \overline{p}$

n_i: tamaño de una muestra (i);

$p_i = \frac{n_{i1}}{n_i}$

$\overline{p} = \frac{\sum n_i \cdot p_i}{\sum n_i}$

– Valor de p (p-valor), bajo H_0:

p-valor = $Pr[\chi^2$ (observado- calculado) $\leq \chi^2_{m-1,\,1-\alpha}$ (teórico-esperado)]

Regla de decisión

La regla de decisión respecto a la significación estadística al contrastar las hipótesis (nula y alternativa) estará determinada por:

- Si: χ^2 (observado-calculado) > $\chi^2_{m-1,1-\alpha}$ (teórico-esperado), entonces chi-cuadrado observado se ubica en la región de rechazo para H_0, por lo que existe evidencia suficiente para rechazar la hipótesis nula y se acepta la alternativa, por lo que existe una diferencia significativa de las proporciones muestrales para la característica seleccionada, al estimarse un p-valor bajo H_0 para el valor observado de chi-cuadrado o una probabilidad relativa al azar para este resultado inferior a la probabilidad alfa para el nivel de confianza seleccionado ($p \geq \alpha$).
- Si χ^2 (observado-calculado) ≤ $\chi^2_{m-1,1-\alpha}$ (teórico-esperado)], entonces chi-cuadrado observado no se ubica en la región de rechazo para H_0, por lo que no existe evidencia suficiente para rechazar la hipótesis nula y no se acepta la alternativa, por lo que no existe una diferencia significativa de las proporciones muestrales para la característica seleccionada, al estimarse un p-valor bajo H_0 para el valor observado de chi-cuadrado o una probabilidad relativa al azar para este resultado igual o superior a la probabilidad alfa para el nivel de confianza seleccionado ($p \geq \alpha$).

z11) Más de dos muestras independientes (determinadas a partir de las clases para una variable cuantitativa) y una variable cualitativa en que se toma una categoría de la misma (característica de interés), si se desea probar que existe una significativa asociación lineal entre las proporciones para cada clase (muestra) y el punto medio de cada clase (número de muestras independientes o clases mayor o igual a tres), para un cierto nivel de significación alfa según el nivel de confianza seleccionado previamente, empleando el procedimiento o prueba de Cochran-Armitagede linealidad en la asociación, el posible planteamiento de hipótesis para el contraste sería:

- H_0: no existe poblacionalmente asociación lineal entre las proporciones para cada clase o muestras independientes y el punto medio de cada clase.
- H_1: existe poblacionalmente asociación lineal entre las proporciones para cada clase o muestras independientes y el punto medio de cada clase.

  Los pasos a seguir para el contraste incluyen la presentación de forma tabulada de los diferentes elementos teóricos necesarios y la introducción de los respectivos valores en las celdas a partir de la información recogida previamente (Tabla 4.26).

Tabla 4.26. Representación tabular de los elementos teóricos necesarios a emplear para el análisis de la asociación lineal entre las proporciones para cada clase (muestra) y el punto medio de cada clase mayor o igual a 3 muestras o grupos de datos independientes

Número de clases para una variable (equivale al número de muestras) (m)	Número de observaciones por sujetos en cada clase o muestra (n_i)	Número de observaciones por sujetos con la característica de interés en cada clase o muestra ($n_{i,1}$)	Proporción de sujetos con la característica en cada clase o muestra (P_i)	Punto o valor medio de cada clase o muestra (X_i)
1	n_1	$n_{1,1}$	P_1	X_1
2	n_2	$n_{2,1}$	P_2	X_2
.	.	.	.	.
.	.	.	.	.
m	n_m	$n_{m,1}$	P_m	X_m
Total	$n_{..}$	$n_{.1}$	$\overline{p}$	$\overline{x}$

– El estadístico de prueba a emplear para el contraste sería:

χ^2(observado-calculado) de Cochran-Armitage $= \dfrac{\sum_{i=1}^{m} n_i(p_i - \hat{p}_i)^2}{\overline{p}\,\overline{q}}$, que sigue una distribución teórica de probabilidades $\chi^2_{m-2,1-\alpha}$ con m – 2 grados de libertad para el valor de alfa prefijado

donde:

n_i: total de sujetos en una clase o muestra (i);

$p_i = \dfrac{n_{i1}}{n_i}$: proporción de sujetos con la característica o categoría de interés en cada clase o muestra;

$\overline{x} = \sum_{i=1}^{m} \dfrac{n_i \cdot x_i}{n_{..}}$: promedio de los puntos medios de las m clases o muestras;

$\overline{p}_i = \overline{p} + b(x_i - \overline{x})$

$$b = \frac{\sum_{i=1}^{m} n_i(p_i - \overline{p})(x_i - \overline{x})}{\sum_{i=1}^{m} n_i(x_i - \overline{x})^2}$$

$\overline{q} = 1 - \overline{p}$

– Valor de p (p-valor), bajo H_0:
p-valor = Pr[χ^2 (observado-calculado) de linealidad (Cochran-Armitage) ≤ $\chi^2_{m-2,1-\alpha}$ (teórico-esperado)].

Regla de decisión

La regla de decisión respecto a la significación estadística al contrastar las hipótesis (nula y alternativa) está determinada por:

– Si χ^2 (observado-calculado) de Cochran-Armitage> $\chi^2_{m-2,1-\alpha}$ (teórico-esperado), entonces chi-cuadrado observado se ubica en la región de rechazo para H_0, por lo que existe evidencia suficiente para rechazar la hipótesis nula y se acepta la alternativa, por lo que existe una significativa linealidad entre los valores de las proporciones y los puntos medios de las clases o muestras, al estimarse un p-valor bajo H_0 para el valor observado de chi-cuadrado o una probabilidad relativa al azar con el resultado obtenido menor al nivel de significación alfa o de cometer el error de tipo 1 para el nivel de confianza seleccionado ($p < \alpha$).

– Si χ^2 (observado-calculado) de Cochran-Armitage ≤ $\chi^2_{m-2,1-\alpha}$ (teórico-esperado), entonces chi-cuadrado observado no se ubica en la región de rechazo para H_0, por lo que no existe evidencia suficiente para rechazar la hipótesis nula y no se acepta la alternativa, por lo que no existe una significativa linealidad entre los valores de las proporciones y los puntos medios de las clases o muestras, al estimarse un p-valor bajo H_0 para el valor observado de chi-cuadrado o una probabilidad relativa al azar para el resultado obtenido igual o mayor al nivel de significación alfa o de cometer el error de tipo 1 para el nivel de confianza seleccionado ($p \geq \alpha$).

z12) Para tres o más muestras independientes y una variable cualitativa ordinal; si se desea determinar la existencia de un gradiente de variación en las tres o más proporciones calculadas según la variable cualitativa ordinal para igual número de muestras independientes, empleando el procedimiento o la prueba de Bartholomew para un nivel de confianza predeterminado por el investigador, el planteamiento de hipótesis sería:

– H_0: no existe poblacionalmente un gradiente de variación de las tres o más proporciones poblacionales estimadas a partir de la variable cualitativa ordinal.

– H_1: existe poblacionalmente un gradiente de variación de las tres o más proporciones poblacionales estimadas a partir de la variable cualitativa ordinal.
Los pasos a seguir para el contraste incluyen la presentación de forma tabulada de los diferentes elementos teóricos necesarios y la introducción de los respectivos valores en las celdas a partir de la información recogida previamente (Tabla 4.27).

Tabla 4.27. Representación tabular de los elementos teóricos necesarios a emplear para el análisis del gradiente de variación de las tres o más proporciones poblacionales estimadas por igual número de muestras o grupos de datos independientes formadas a partir de una variable cualitativa ordinal

Escala ordinal de la variable cualitativa que equivale al número d muestras (m)	Número de observaciones por sujetos en cada nivel u orden de la variable cualitativa al inicio del estudio (n_i)	Número de observaciones por sujetos en cada nivel u orden de la variable cualitativa pasado un tiempo o al final del estudio	Proporción de sujetos en cada nivel u orden de la variable cualitativa (P_i)	Proporción revisada de sujetos en cada nivel u orden de la variable cualitativa ($P'_{i)}$
1	n_1	$n_{1,1}$	P_1	(P'_1)
2	n_2	$n_{2,1}$	P_2	(P'_2)
.	.	.	.	
.	.	.	.	
m	n_m	$n_{m,1}$	P_m	(P'_m)
Totales:	$n_{..}$	$n_{.1}$	$\bar{p}$	$\bar{p}$ '

– El estadístico de prueba para realizar el contraste sería el χ^2 (observado-calculado) de Bartholomew, que adopta la expresión:

$$\chi^2(\text{observado-calculado}) \text{ de Bartholomew} = \frac{1}{\bar{p}\,\bar{q}} \sum_{i=1}^{m} n_i (p'_i - \bar{p})^2$$

Como parte del procedimiento a realizar en caso de que no se cumpla:
$p_1 > p_2 > p_3 > p_4$,
se calcula el valor promedio ponderado para las dos proporciones que no lo cumplan y se coloca este resultado en la última columna de la tabla expuesta anteriormente. La fórmula a emplear sería:

$$P'_i = P_{i,\,i+1} = \frac{[n_i P_i + n_{i+1} P_{i+1}]}{(n_i + n_{i+1})}$$

Por ejemplo entre la P_1 y P_2 la proporción revisada sería:

$$P'_{1,2} = P_{i,\,i+1} = \frac{\left[(n_1 P_1) + (n_2 P_2)\right]}{(n_1 + n_2)}$$

– En caso de cumplirse que $P_1 > P_2 > P_3 \ldots > P_m$, se emplea P'_i coincidiendo con P_i.
– Las proporciones revisadas que son las correspondientes a la última columna son las empleadas en el cálculo de estadístico de prueba de Bartholomew.

Regla de decisión

El valor observado-calculado del estadístico chi-cuadrado de Bartholomew no se compara con el estadístico teórico-esperado-tabulado de chi-cuadrado

según los grados de libertad y nivel alfa . En su lugar el procedimiento a realizar es como sigue:

- Cuando son empleadas tres proporciones, es decir, m = 3, entonces calcular:

$$c = \sqrt{\frac{n_1 n_3}{(n_1 + n_2)(n_2 + n_3)}}$$

el resultado obtenido de conjunto al nivel de significación alfa preseleccionado previamente por el investigador son consultados en la tabla de referencia para esta prueba de Bartholomew (véase el anexo 1) y seguidamente el valor hallado se compara con el valor del estadístico $\overline{\chi}^2$ (observado-calculado) de Bartholomew.

- Cuando son empleadas cuatro proporciones, es decir, m = 4, entonces calcular:

$$c_1 = \sqrt{\frac{n_1 n_3}{(n_1 + n_2)(n_2 + n_3)}}$$

y también:

$$c_2 = \sqrt{\frac{n_2 n_4}{(n_2 + n_3)(n_3 + n_4)}}$$

entonces con estos valores y el nivel de significación alfa seleccionado por el investigador se consulta la tabla de referencia de valores críticos de Bartholomew, el valor obtenido se compara con el del estadístico $\overline{\chi}^2$ (observado-calculado) de Bartholomew.

Regla de decisión

- Si el estadístico de prueba o de contraste $\overline{\chi}^2$ (observado-calculado) de Bartholomew supera el valor crítico-tabulado o de referencia (según c y alfa seleccionado) para los puntos porcentuales de las tablas de referencia, existen evidencias suficientes para rechazar la hipótesis nula y aceptar la alternativa, por lo que el gradiente de las variaciones de las proporciones muestrales según los órdenes o niveles de la variable cualitativa es estadísticamente significativo, al obtenerse un p-valor bajo H_0 menor al nivel de significación alfa) seleccionado a partir de su correspondiente nivel de confianza.
- Si el estadístico de prueba o de contraste $\overline{\chi}^2$ (observado-calculado) de Bartholomew no supera el valor crítico-tabulado o de referencia (según c y alfa seleccionado) para los puntos porcentuales de las tablas de referencia, no existen evidencias suficientes para rechazar la hipótesis nula y aceptar la alternativa, por lo que el gradiente de las variaciones de las proporciones muestrales según los órdenes o niveles de la variable cualitativa no es estadísticamente significativo, al obtenerse un p-valor bajo H_0 igual o mayor al nivel de significación alfa seleccionado a partir de su correspondiente nivel de confianza.

1. Para pruebas no paramétricas:

   Aspectos relacionados con los métodos o técnicas no paramétricas empleadas en el contraste de hipótesis:

   a) Son métodos paralelos a los paramétricos, cuyos procedimientos no precisan la estimación de parámetros ni suponer conocida alguna ley de probabilidad subyacente en la población de la que se extrae la muestra.
   b) Se pueden aplicar en los mismos casos y con mayor facilidad que las técnicas paramétricas.
   c) Se utilizan cuando los datos que sirven para el análisis constan simplemente de categorías o clasificaciones, es decir, los datos pueden no estar basados en una escala de medición lo suficientemente sólida como para permitir las operaciones aritméticas necesarias para llevar a cabo los procedimientos paramétricos.
   d) Desempeñan una función fundamental la ordenación de los datos en una relación de valores de menor a mayor denominada: rangos.
   e) El uso de procedimientos no paramétricos con datos que pueden manejarse con un procedimiento paramétrico produce un desperdicio de información.
   f) La aplicación de algunas de las pruebas no paramétricas puede ser muy laboriosa para muestras grandes.
   g) Tienen como inconveniente la falta de sensibilidad para detectar efectos importantes.

Algunos ejemplos de situaciones investigativas en las cuales como parte del procesamiento estadístico de la información para el manejo de muestras y variables se requiere del uso de estas pruebas no paramétricas y los procedimientos a seguir en las mismas se presentan a continuación:

   a) Una muestra, una variable cuantitativa aleatoria que no sigue la distribución normal estándar (tiene distribución libre), si se desea determinar si los datos de esta variable cuantitativa de una población estimada a partir de la muestral, tienen similar distribución que los de una población de referencia, empleando el método no paramétrico: prueba de la U de Mann- Whitney, un posible planteamiento de hipótesis estadísticas sería:
      - Contraste bilateral:
        - H_0: los valores de la variable cuantitativa estimados de una población y los de la de referencia están equidistribuídos o las medianas de la variable cuantitativa son iguales.
        - H_1: los valores de la variable cuantitativa estimados de una población y los de la población de referencia no están equidistribuídos o las medianas de la variable cuantitativa son diferentes.

      Si tenemos una muestra con los datos de una variable cuantitativa la cual sigue una distribución libre, que estima los datos de una población y se comparan con los de una población de referencia:

– Pasos a seguir para obtener el estadístico U:
  • Determinar el número de observaciones para los grupos de datos independientes que se corresponden con los de la población estimada y la de referencia.
  • Ordenar de menor a mayor los datos de las observaciones de ambas poblaciones: estimada y de referencia y se les asigna un rango (R) en orden ascendente comenzando por el 1 a cada valor de la variable cuantitativa hasta n según el número total de observaciones realizadas.
  • Se calculan los estadísticos:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$, para la población estimada de la muestra de estudio

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$, para la población de referencia

donde:

R_1 y R_2: sumatoria de los respectivos rangos de los valores de la variable cuantitativa aleatoria para ambas poblaciones (estimada a partir de la muestral y la de referencia empleada) respectivamente;

n_1 y n_2: número de observaciones de la población (estimada de la muestra) y el de la población de referencia respectivamente.

– Se selecciona el mayor valor entre U_1 y U_2 y se denota por estadístico de Mann-Whitney como U (observado-calculado), así como se determina el valor de la U (teórico-esperado) en su valor crítico superior según la tabla de referencia para el estadístico U de Mann-Whitney para efectuar la comparación entre ambos valores (véase el anexo 1).

Regla de decisión

Para el contraste bilateral: si $n_1 + n_2 \leq 20$ se emplea el estadístico U de Mann-Whitney como sigue: si U (observado-calculado) > U (teórico-esperado) valor crítico superior según el número de observaciones y el nivel de significación alfa, entonces se rechaza H_0, es decir, las dos poblaciones no tienen equidistribución significativa en los valores de la variable cuantitativa empleada, al observarse diferencias de los rangos para los valores de la variable en ambas poblaciones para el nivel de significación alfa según el nivel de confianza preseleccionado:

– Otra forma de realizar el contraste de hipótesis, según Snedecor y Cochran (1989) es empleando la aproximación de la distribución normal estándar Z ~ N (0, 1), empleando el estadístico de contraste Z, para los dos grupos de datos en conjunto n > 20, lo que conllevaría a dar estos pasos:
  • Ordenar los valores de la variable cuantitativa en orden ascendente a partir del 1 de ambas poblaciones (estimada de la muestral y la de referencia) y dar un rango a estos valores.

- Realizar la suma de rangos observada R_1 y R_2 para el grupo de datos de las poblaciones correspondientes, la R_1 estimada de la muestral y la R_2 que es para la de referencia.
- Calcular el valor de U a partir de R_1 o R_2, se prefiere emplear el menor entre R_1 y R_2 según la expresión:

  $U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$, se escoge R_1 en este caso si fuera el menor
- Determinar el valor esperado [E (R_1)] y la varianza [Var (R_1)] para la sumatoria de rangos observada menor, en la población estimada o la de referencia según la expresión:

  E (R_1) = n_1 (n_1 + n_2 +1)/2

  Var (R_1) = $n_1 n_2$ (n_1 + n_2 +1)/12
- Expresar de forma general el valor esperado [E(U) o mU] y la varianza [Var(U) o σU] para el estadístico U de Mann-Whitney como sigue:

  $E(U) = n_1 n_2 + \frac{n_1(n_1+1)}{2} - E(R_1) = \frac{n_1 n_2}{2}$ o

  $m_U = \frac{n_1 n_2}{2}$

  $Var(U) = Var(R_1) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$ o

  $\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$
- Estimación del valor del estadístico de contraste para la distribución normal estándar Z ~ N (0, 1):

  $Z\ (\text{observado-calculado}) = \frac{U - E(U)}{\sqrt{Var(U)}}$ o

  $Z\ (\text{observado-calculado}) = \frac{U - m_U}{\sigma_U}$ o

  $Z\ (\text{observado-calculado}) = \frac{U - (n_1 n_2)/2}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$

– Valor de p (p-valor), bajo H_0:

$$p - valor = \begin{cases} 2\Phi(z) & \text{si } z \leq 0 \\ 2[1 - \Phi(z)] & \text{si } z > 0 \end{cases}$$

Regla de decisión

– Para Z (observado-calculado) < 0:

- Si Z (observado- calculado) < Z (teórico-esperado), según el valor alfa prefijado, esto se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces hay evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones (estimada vs referencia) los valores de la variable aleatoria no están equidistribuídos, siendo estadísticamente significativo el resultado obtenido al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa a la azar para el resultado obtenido, inferior a la correspondiente con el valor de significación alfa prefijado ($p < \alpha$).
- Si Z (observado- calculado) ≥ Z (teórico-esperado), según el valor de alfa prefijado, esto no se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que aumenta la probabilidad de cometer el error tipo 1, entonces no existen evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones (estimada *vs.* referencia) los valores de la variable aleatoria están equidistribuídos, siendo estadísticamente no significativo el resultado obtenido al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa a la azar para el resultado obtenido igual o superior a la correspondiente con el valor de significación alfa prefijado ($p \geq \alpha$).

– Para Z (observado-calculado) ≥ 0:
  - Si Z (observado-calculado) > Z (teórico-esperado), según el valor alfa prefijado, esto se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que disminuye la probabilidad de cometer el error tipo 1, entonces hay evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones (estimada *vs.* referencia) los valores de la variable aleatoria no están equidistribuídos, siendo estadísticamente significativo el resultado obtenido al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa a la azar para el resultado obtenido inferior a la correspondiente con el valor de significación alfa prefijado ($p < \alpha$).
  - Si Z (observado- calculado) ≤ Z (teórico-esperado), según el valor alfa prefijado, esto no se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que aumenta la probabilidad de cometer el error tipo 1, entonces no hay evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones (estimada *vs.* referencia) los valores de la variable aleatoria están equidistribuídos, siendo estadísticamente no significativo el resultado obtenido al estimarse un p-valor bajo H_0 para Z observado o

una probabilidad relativa a la azar para el resultado obtenido igual o superior a la correspondiente con el valor de significación alfa prefijado ($p \geq \alpha$).

b) Una muestra, una variable cuantitativa que no sigue la distribución normal estándar (tiene distribución libre), si se desea determinar si los datos de esta variable cuantitativa de una población estimada a partir de la muestral, tienen similar distribución que los de una población de referencia, empleando el método no paramétrico prueba de la suma de rangos de Wilcoxon, un posible planteamiento de hipótesis estadísticas sería:
   - Contraste bilateral:
     - H_0: los valores de la variable cuantitativa estimados de una población de estudio estimada y los de una población de referencia están equidistribuidos o tienen el mismo valor para la suma de los rangos de esta variable.
     - H_1: los valores de la variable cuantitativa estimados de una población de estudio estimada y los de una población de referencia no están equidistribuidos o no tienen el mismo valor para la suma de los rangos de esta variable.

     Si se tiene una muestra con datos de una variable cuantitativa con distribución libre que estima una poblacional y se compara con los de una población de referencia.

     Pasos a seguir para obtener el estadístico T (de Wilcoxon):
   - Determinar el número de observaciones en las poblaciones estudiadas (los dos grupos de datos).
   - Ordenar numéricamente en orden ascendente los valores de la variable cuantitativa en ambos grupos de datos de forma conjunta.
   - Asignar un número de orden o rango, de menor a mayor, a los valores de la variable para cada una de las observaciones de las poblaciones (estimada y de referencia) a comparar, tomadas conjuntamente.
   - En caso de que una observación tenga el mismo valor en cada grupo de datos independientes (da las poblaciones comparadas), a cada observación se le adjudica el promedio de ambos rangos. En forma análoga se manejan los "empates" entre más de dos observaciones idénticas, en cambio, si las observaciones idénticas se repiten en un mismo grupo, no es necesario asignarles un rango promedio.
   - Proceder a sumar los números de orden o rangos dentro de cada grupo de datos independientes que son los correspondientes con la población (estimada y de referencia).
   - Comparar los valores de la sumatoria de rangos obtenidos para cada grupo y seleccionar el menor de los dos valores de la sumatoria de rangos R_1 o R_2, el cual se emplea como estadístico de Wilcoxon (T o R) para realizar su comparación con el valor T tabulado-esperado del mismo (con el valor crítico inferior

en este caso de los dos que ofrece la tabla de referencia para esta prueba), según el número de observaciones para ambos grupos de datos y el nivel de significación alfa preseleccionado por el investigador (véase el anexo 1).

Nota: siempre que los grupos sean del mismo tamaño, la suma de los rangos debe ser igual o muy semejante en ambos grupos bajo H_0. En cambio, si las observaciones tienden a ser mayores en un grupo o población estimada, la correspondiente suma de los rangos también tiende a ser mayor en este grupo de datos y esto es una evidencia en contra de la hipótesis nula y a favor de una diferencia entre ambos grupos o poblaciones para la distribución de la variable cuantitativa.

Regla de decisión

Para el contraste bilateral si $n_1 + n_2 \leq 20$ se emplea el estadístico T o R de Wilcoxon como sigue:

- Si T de Wilcoxon (observado-calculado) < T (teórico-esperado-tabulado) valor crítico inferior para el número de observaciones en los dos grupos de datos independientes según el valor de alfa, entonces hay evidencias para rechazar H_0, es decir, las dos poblaciones no están equidistribuidas en los valores de la variable aleatoria cuantitativa con libre distribución al observarse diferencias significativas a partir de la comparación de la sumatoria de los rangos para las mismas, donde los resultados obtenidos que pueden ser atribuidos al azar o donde se cumpliría H_0 son muy poco probables, es decir, es menor que el nivel de significación alfa preseleccionado por el investigador.
- Método por la aproximación de la distribución Normal estándar Z ~N (0, 1), Según el procedimiento descrito por Triola (2009):

  Cuando el número de casos en cada grupo (para cada característica en que se dividió la muestra) excede los tabulados (n_1, $n_2 > 20$) y si ninguno de los grupos contiene menos de cuatro observaciones, puede utilizarse una aproximación a través de la distribución normal estándar empleando la expresión del estadístico Z ~ N (0,1):

$$Z \text{ (observado-calculado)} = \frac{R - \mu_R}{\sigma_R}$$

  donde:

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2}$$

$$\sigma_R = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

  n_1: tamaño de la muestra a partir de la cual se calcula la suma de rangos R;

  n_2: tamaño para la población de referencia según sus datos;

  R: suma de rangos de la muestra con tamaño n_1 (se suele emplear como R_1);

μ_R: media de los valores muestrales R que se espera cuando las dos poblacio nes tienen medianas iguales;
σ_R: desviación estándar de los valores muestrales R cuando las dos poblaciones tienen medianas iguales.

– Valor de p (p-valor), bajo H_0:

$$p = \begin{cases} 2\Phi(z) & \text{si } z \leq 0 \\ 2\left[1-\Phi(z)\right] & \text{si } z > 0 \end{cases}$$

Regla de decisión

– Para Z (observado-calculado) < 0:
  - Si Z (observado- calculado) < Z (teórico-esperado), según el valor alfa prefijado, esto se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que disminuye la probabilidad de cometer el error tipo 1, entonces hay evidencias suficientes para rechazar H_0, es decir, se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa al azar para ese resultado inferior a la correspondiente con el valor de significación alfa prefijado ($p < \alpha$).
  - Si Z (observado- calculado) ≥ Z (teórico-esperado), según el valor alfa prefijado, esto no se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que aumenta la probabilidad de cometer el error tipo 1, entonces no existen evidencias suficientes para rechazar H_0, es decir, no se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa a la azar para ese resutado igual o superior a la correspondiente con el valor de significación alfa prefijado ($p \geq \alpha$).

– Para Z (observado-calculado) ≥ 0:
  - Si Z (observado-calculado) > Z (teórico-esperado), según el valor alfa prefijado, esto se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que aumenta la probabilidad de cometer el error tipo 1, entonces existen evidencias suficientes para rechazar H_0, es decir, se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa a la azar para ese resultado menor a la correspondiente con el valor de significación alfa prefijado ($p < \alpha$).

- Si Z (observado- calculado) ≤ Z (teórico-esperado), según el valor alfa prefijado, esto no se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que aumenta la probabilidad de cometer el error tipo 1, entonces no existen evidencias suficientes para rechazar H_0, es decir, no se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa a la azar para ese resultado igual o superior a la correspondiente con el valor de significación alfa prefijado ($p \geq \alpha$).

c) Una muestra, dos variables cuantitativas que no siguen o no se aproximan a la distribución normal estándar (distribución libre), si se desea evaluar la significación de la correlación por rangos entre estas variables cuantitativas mediante el método no paramétrico para el coeficiente de correlación por rangos de Spearman (ρ_S), empleando el estadístico de prueba o contraste t-student para un nivel de significación alfa y los grados de libertad correspondientes al caso, el planteamiento de hipótesis estadísticas para el contraste sería:
   - Contraste bilateral:
     - H_0: (ρ_s) = 0, no existe una correlación lineal simple entre las dos variables.
     - H_1: (ρ_s) ≠ 0, existe una correlación lineal simple entre las dos variables.

     Los pasos a seguir en el contraste serían:
     - Seleccionar de forma aleatoria los datos muestrales de las variables cuantitativas de libre distribución por pares (no existe el requisito de que los datos muestrales apareados tengan una distribución normal bivariada).
     - Asignar un rango a cada uno de los datos correspondientes de cada variable (representan a los datos por pares de las dos muestras).
     - Hallar las diferencias de rangos para los pares de datos de ambas muestras (para ambas variables o mediciones).

     En el caso del número de observaciones por pares n ≤ 30:
   - Estadístico para el contraste (de prueba):
     - Después de convertir los datos de cada muestra a rangos, si no existen empates entre los rangos para la primera variable y no existen empates entre los rangos para la segunda variable, el valor exacto del estadístico de prueba r_s puede calcularse utilizando esta fórmula:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

     - Después de convertir los datos de cada muestra a rangos, si cualquier variable tiene empates entre sus rangos, el valor exacto del estadístico de prueba r_s puede calcularse utilizando la fórmula con los rangos para cada variable x, y:

$$r_s = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x)^2 - (\sum x)^2}\sqrt{n(\sum y)^2 - (\sum y)^2}}$$

donde:

ρ_s: coeficiente de correlación lineal simple poblacional estimado a partir del muestral de Spearman (r_s);

x, y: valores de las variables cuantitativas de libre distribución;

d o D_i: diferencia entre los rangos de los dos valores de las variables cuantitativas X_i y Y_i para cada par;

n: número de pares de valores entre las variables X_i y Y_i o número de pares de datos muestrales.

Para recordar

– El coeficiente de correlación por rangos de Spearman (r_s) al igual que el coeficiente de correlación lineal momento producto de Pearson, expresa el grado, fuerza o intensidad de la asociación de la variable dependiente con la independiente, ambas cuantitativas, solo que los valores (datos) de cada variable son sustituidos por el rango o lugar que ocupan dentro de la serie ordenada de datos de cada variable en la muestra de estudio, siendo este el punto de partida para su análisis. Sus posibles valores se evalúan por la proximidad a 1, es decir, mientras más cercano a 1 mayor correlación presentan las dos variables cuantitativas estudiadas.

Regla de decisión

Se compara el valor observado del coeficiente de correlación por rangos de Spearman (r_s) con el valor teórico-esperado-tabulado (véase el anexo 1, según el nivel de confianza con valor de alfa seleccionado, siendo rechazada H_0 cuando rs (observado) > r_s (teórico-tabulado) según el nivel de significación alfa preseleccionado por el investigador.

– En el caso del número de observaciones por pares n > 30 el estadístico para el contraste (de prueba):

$$rs = \frac{\pm z}{\sqrt{n-1}}$$

donde:

z: corresponde al valor del percentil de la distribución de probabilidad estándar según el nivel de significación alfa seleccionado, por ejemplo, si alfa = 0,05, Z = 1,96.

– Método por la aproximación de la distribución de probabilidad para la t-student en el caso de de muestras grandes (n > 30), se puede emplear la aproximación de la distribución de probabilidad de t-student según los grados de libertad y valor de alfa preseleccionado correspondientes como sigue:

t-student (observado-calculado) $= \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$, que sigue una distribución teórica de probabilidades t-student con n – 2 grados de libertad y para el valor de alfa prefijado por el investigador.

– Valor de p o p-valor bajo $H_{0:}$
  p-valor = Pr [t-student (teórico-esperado) ≥ t-student (observado-calculado)].

Regla de decisión

– Para t-student (observado-calculado) < 0:
  • Si t-student (calculado-observado) < t-student (teórico-esperado) según los grados de libertad respectivos y el valor alfa preseleccionado, entonces t-student observado se ubica en la región de rechazo para H_0, por lo que existe evidencia suficiente para rechazar la hipótesis nula y se acepta la alternativa, por lo que la correlación por rangos de ambas variables cuantitativas es estadísticamente significativa, al estimarse una probabilidad relativa al azar para este resultado o un p-valor si se cumpliera H_0 para el t-student observado menor que el nivel de significación para el nivel de confianza seleccionado ($p \geq \alpha$).
  • Si t-student (calculado-observado) ≥ t-student (teórico-esperado) según los grados de libertad respectivos y el valor alfa preseleccionado, entonces t-student observado no se ubica en la región de rechazo para H_0, por lo que no existe evidencia suficiente para rechazar la hipótesis nula y no se acepta la alternativa, por lo que la correlación por rangos de ambas variables cuantittativas no es estadísticamente significativa, al estimarse una probabilidad relativa al azar para este resultado o un p-valor si se cumpliera H_0 para el t-student observado igual o mayor que la probabilidad alfa para el nivel de confianza seleccionado ($p \geq \alpha$).

– Para t-student (observado-calculado) ≥ 0:
  • Si t-student (calculado-observado) > t-student (teórico-esperado) según los grados de libertad respectivos y el valor alfa preseleccionado, entonces t-student observado se ubica en la región de rechazo para H_0, por lo que existe evidencia suficiente para rechazar hipótesis nula y se acepta la alternativa, por lo que la correlación por rangos de ambas variables cuantitativas es estadísticamente significativa, al estimarse una probabilidad relativa al azar para este resultado o un p-valor si se cumpliera H_0 para el t-student observado menor que la probabilidad alfa para el nivel de confianza seleccionado ($p < \alpha$).
  • Si t-student (calculado-observado) ≤ t-student (teórico-esperado) según los grados de libertad respectivos y el valor alfa preseleccionado, entonces t-student observado se ubica en la región de rechazo para H_0, por lo que

existe evidencia suficiente para rechazar hipótesis nula y se acepta la alternativa, por lo que las diferencias entre las medias poblacionales estimadas pareadas son estadísticamente significativas, al estimarse una probabilidad relativa al azar para este resultado o un p-valor si se cumpliera H_0 para el t-student observado igual o mayor que el nivel de significación para el nivel de confianza seleccionado ($p \geq \alpha$).

d) Una muestra, una variable cuantitativa aleatoria que no sigue ni se aproxima a la distribución normal estándar (tiene distribución libre) y una variable cualitativa dicotómica, si se desea determinar si los datos de dicha variable cuantitativa aleatoria para los dos grupos de datos formados a partir de las dos categorías de la variable cualitativa dicotómica pueden provenir de la misma población, se emplea el método no paramétrico prueba de la U de Mann-Whitney-Kendall y un posible planteamiento de hipótesis estadísticas sería:
   - Para el contraste bilateral:
     - H_0: las poblaciones de las que provienen las dos muestras (según los dos grupos de datos para las dos categorías de la variable cualitativa dicotómica) están equidistribuídas para los valores de la variable cuantitativa.
     - H_1: las poblaciones de las que provienen las dos muestras (según los dos grupos de datos para las dos categorías de la variable cualitativa dicotómica) no están equidistribuídas para los valores de la variable cuantitativa.

     Si se tienen dos muestras (formadas para las dos categorías de la variable cualitativa dicotómica) con los datos tomados de una variable cuantitativa de libre distribución los pasos a seguir para obtener el estadístico U:
     - Determinar el número de observaciones de los dos grupos de datos.
     - Ordenar de menor a mayor los datos de las observaciones de ambas poblaciones estimadas (para ambos grupos de datos, correspondiendo con las dos categorías de la variable cualitativa dicotómica) y se les asigna un rango (R) en orden ascendente comenzando por el 1 a cada valor de la variable cuantitativa hasta n en los dos grupos o poblaciones estimadas a comparar:
     - Se calculan los estadísticos:

       $U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$, para la muestra que estimaría una población 1 o para la categoría 1 de la variable cualitativa.

       $U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$, para la muestra que estimaría una población 2 o para la categoría 2 de la variable cualitativa.

       donde:

       R_1 y R_2: sumatoria de los respectivos rangos de los valores de la variable cuantitativa para ambas poblaciones estimadas (a partir de las dos muestras para las dos categorías).

n_1 y n_2: número de observaciones de la población estimada de la muestra 1 (para categoría 1 de la variable dicotómica) y el correspondiente con la población estimada de la muestra 2 (para categoría 2 de la misma variable dicotómica).

- Se selecciona el mayor valor entre U_1 y U_2 y se denota por estadístico de Mann-Whitney como U (observado-calculado), así como se determina el valor de la U (teórica-esperada) según la tabla de referencia para la distribución de Mann-Whitney (véase el anexo 1) para efectuar la comparación entre ambos valores.

Regla de decisión

Para el contraste bilateral: si $n_1 + n_2 > 20$ se emplea el estadístico U de Mann-Whitney como sigue:

– Si U (observado-calculado) > U (teórica-esperada) valor crítico superior según el número de observaciones y el nivel de significación alfa, entonces se rechaza H_0, es decir, las dos poblaciones estimadas no están equidistribuidas significativamente en los valores de la variable cuantitativa con libre distribución, al observarse diferencias de los rangos para los valores de la variable en ambas poblaciones que representan una probabilidad de cumplirse H_0 menor que el nivel de significación alfa predeterminado.

– Otra forma de realizar el contraste de hipótesis según Snedecor y Cochran (1989) es empleando la aproximación de la distribución normal con el estadístico de contraste $Z \sim N(0, 1)$ para grupos de datos o muestras mayores (n_1 y $n_2 > 10$) lo que conlleva a seguir estospasos:

- Ordenar los valores de la variable cuantitativa en orden ascendente de ambas poblaciones (estimadas de las dos muestras para los dos grupos de datos de las dos categorías correspondientes a la variable cualitativa) y asignar un rango a dichos valores empezando por 1 hasta n en orden ascendente.
- Realizar la suma de rangos observada R_1 y R_2 para el grupo de datos de las poblaciones correspondientes (la 1 estimada de la muestra para los datos de la categoría 1 de la variable dicotómica y la 2 estimada de la muestra para los datos corresondientes a la categoría 2 de la misma variable dicotómica).
- Calcular el valor de U a partir de R_1 o R_2 (se prefiere emplear el menor entre R_1 y R_2):

  $U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$, se escogería R1 en este caso si fuese el menor.

- Determinar el valor esperado [E (R_1)] y la varianza [Var (R_1)] para la sumatoria de rangos observada en la población estimada de la muestra 1 (para datos de la variable de la categoría 1), según la expresión:

$E(R_1) = n_1 (n_1 + n_2 + 1)/2$

$Var(R_1) = n_1 n_2 (n_1 + n_2 + 1)/12$

- Expresar de forma general el valor esperado [E(U) o mU] y la varianza [Var(U) o σU] para el estadístico U de Mann-Whitney según la expresión:

$$E(U) = n_1 n_2 + \frac{n_1(n_1+1)}{2} - E(R_1) = \frac{n_1 n_2}{2} \quad \text{o}$$

$$m_U = \frac{n_1 n_2}{2}$$

$$Var(U) = Var(R_1) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} \quad \text{o}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

- Estimación del valor del estadístico de contraste para la distribución normal estándar Z ~ N (0, 1):

$$Z \text{ (observado-calculado)} = \frac{U - E(U)}{\sqrt{Var(U)}} \quad \text{o}$$

$$Z \text{ (observado-calculado)} = \frac{U - m_U}{\sigma_U} \quad \text{o}$$

$$Z \text{ (observado-calculado)} = \frac{U - (n_1 n_2)/2}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

– Valor de p (p-valor), bajo H_0:

$$p - valor = \begin{cases} 2\Phi(z) & \text{si } z \leq 0 \\ 2\left[1 - \Phi(z)\right] & \text{si } z > 0 \end{cases}$$

Regla de decisión

– Para Z (observado-calculado) < 0:

- Si Z (observado- calculado) < Z (teórico-esperado), según el valor de alfa prefijado, esto se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces hay evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas (para los dos categorías de la variable dicotómica) se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa al azar para este resultado inferior a la correspondiente con el valor de significación alfa prefijado ($p < \alpha$).

- Si Z (observado-calculado) ≥ Z (teórico-esperado), según el valor de alfa prefijado, esto no se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teóricode referencia, entonces no existen evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas (para las dos categorías de la variable dicotómica) no se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para el valor de Z observado o una probabilidad relativa a la azar para este resultado igual o superior a la correspondiente con el valor de significación alfa prefijado ($p \geq \alpha$).

– Para Z (observado-calculado) ≥ 0:
  - Si Z (observado-calculado) > Z (teórico-esperado), según el valor de alfa prefijado, esto se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces existen evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas (para las dos categorías de la variable dicotómica) se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para el valor de Z observado o una probabilidad relativa a la azar para este resultado menor a la correspondiente con el valor de significación alfa prefijado ($p < \alpha$).
  - Si Z (observado- calculado) ≤ Z (teórico-esperado), según el valor de alfa prefijado, esto no se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces no existen evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas (para las dos categorías de la variable dicotómica) no se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para el valor de Z observado o una probabilidad relativa a la azar para este resultado igual o superior a la correspondiente con el valor de significación alfa prefijado ($p \geq \alpha$).

e) Una muestra, una variable cuantitativa aleatoria que no sigue ni se aproxima a la distribución normal estándar (tiene distribución libre), si se desea determinar si los datos de dicha variable cuantitativa para los dos grupos de datos formados a partir de las dos categorías de la variable cualitativa dicotómica pueden provenir de la misma población**,** empleamos el método no paramétrico prueba de la suma de rangos de Wilcoxon y un posible planteamiento de hipótesis estadísticas sería:
   – Contraste bilateral:
     - H_0: las poblaciones de las que provienen las dos muestras (según los dos grupos de datos de la variable cualitativa dicotómica) están equidistribui-

das para los valores de la variable cuantitativa o las medianas de ambos grupos de datos son iguales.

- H_1: las poblaciones de las que provienen las dos muestras (según los dos grupos de datos de la variable cualitativa dicotómica) no están equidistribuidas para los valores de la variable cuantitativa o las medianas de ambos grupos de datos no son iguales.

  Si se tienen dos muestras que estiman dos poblaciones (formadas para las dos categorías de la variable cualitativa) con datos de una variable cuantitativa con distribución libre los pasos a seguir para obtener el estadístico T (de Wilcoxon):
- Determinar el número de observaciones de cada grupo de datos.
- Ordenar los valores de la variable cuantitativa de ambos grupos de datos independientes en conjunto.
- Asignar un número de ordeno rango, de menor a mayor, a los valores de la variable cuantitativa para cada una de las observaciones de las poblaciones estimadas (en relación a los dos grupos de datos con que contamos según variable cualitativa dicotómica).
- En caso de que una observación tenga el mismo valor en cada grupo o población estimada, a cada una se le adjudica el promedio de ambos rangos. En forma análoga se manejan los "empates" entre más de dos observaciones idénticas, en cambio, si las observaciones idénticas se repiten en un mismo grupo no es necesario asignarles un rango promedio.
- Proceder a sumar los números de orden o rangos dentro de cada población estimada; es decir para cada grupo de datos independientes según las dos categorías de la variable cualitativa.
- Comparar los valores de la sumatoria de rangos obtenidos para cada grupo y seleccionar el menor de los dos valores de rangos sumados, el cual se emplea como estadístico de Wilcoxon (T o R) para realizar su comparación con el valor T o R (tabulado-esperado) crítico inferior del mismo en las tablas correspondientes para este procedimiento, siguiendo el número de observaciones para ambos grupos de datos y el nivel de significación alfa preseleccionado por el investigador (véase el anexo 1).

  Nota: aiempre que los grupos sean del mismo tamaño, la suma de los rangos debe ser igual o muy semejante en ambos grupos bajo H_0. En cambio, si las observaciones tienden a tener valores mayores en un grupo o población, la correspondiente suma de los rangos también tienede a ser mayor en esa población estimada o grupo de datos y es una evidencia en contra de la hipótesis nula y a favor de una diferencia entre ambos grupos o poblaciones.

Regla de decisión

Para el contraste bilateral: si $n_1 + n_2 > 20$ se emplea el estadístico T o R de Wilcoxon como sigue:

- Si T o R (observado-calculado) < T o R (teórico-esperado-tabulado) crítico inferior para el número de observaciones en los dos grupos de datos independientes según el valor de alfa, entonces hay evidencias para rechazar H_0, es decir, las dos poblaciones no están equidistribuidas en los valores de la variable aleatoria cuantitativa con libre distribución al observarse diferencias significativas a partir de la comparación de la sumatoria de sus rangos, donde los resultados obtenidos que pueden ser atribuidos al azar son muy poco probables, es decir, son menor que el nivel de significación alfa preseleccionado por el investigador.
- Método por la aproximación de la distribución Normal estándar $Z \sim N(0, 1)$, según el procedimiento descrito por Triola (2009), cuando el número de casos en cada grupo excede los tabulados (n_1, n_2 >10) y si ninguno de los grupos contiene menos de cuatro observaciones, puede utilizarse una aproximación a través de la distribución normal estándar empleando la expresión del estadístico Z ~ N (0, 1):

$$Z\ (\text{observado-calculado}) = \frac{R - \mu_R}{\sigma_R}$$

donde:

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2}$$

$$\sigma_R = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

n_1: tamaño de la muestra a partir de la cual se calcula la suma de rangos R;
n_2: tamaño de la otra muestra;
R: suma de rangos de la muestra con tamaño n_1 (se suele emplear como R_1);
μ_R: media de los valores muestrales R que se espera cuando las dos poblaciones tienen medianas iguales;
σ_R: desviación estándar de los valores muestrales R cuando las dos poblaciones tienen medianas iguales.

- Valor de p (p-valor), bajo H_0:

$$p = \begin{cases} 2\Phi(z) & \text{si } z \leq 0 \\ 2\left[1 - \Phi(z)\right] & \text{si } z > 0 \end{cases}$$

Regla de decisión

- Para Z (observado-calculado) < 0:

- Si Z (observado-calculado) < Z (teórico-esperado), según el valor de alfa prefijado, esto se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces hay evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas (para los dos categorías de la variable dicotómica) se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para el valor de Z observado o una probabilidad relativa al azar para este resultado inferior a la correspondiente del nivel de significación alfa prefijado ($p < \alpha$).
- Si Z (observado- calculado) ≥ Z (teórico-esperado), según el valor de alfa prefijado, esto no se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces no existen evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas (para las dos categorías de la variable dicotómica) no se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para el valor de Z observado o una probabilidad relativa a la azar para este resultado igual o superior a la correspondiente del nivel de significación alfa prefijado ($p \geq \alpha$).

– Para Z (observado-calculado) ≥ 0:
  - Si Z (observado-calculado) > Z (teórico-esperado), según el valor de alfa prefijado, esto se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces existen evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas (para las dos categorías de la variable dicotómica) se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para el valor de Z observado o una probabilidad relativa a la azar para este resultado menor a la correspondiente con el nivel de significación alfa prefijado ($p < \alpha$).
  - Si Z (observado- calculado) ≤ Z (teórico-esperado), según el valor de alfa prefijado, esto no se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces no existen evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas (para las dos categorías de la variable dicotómica) no se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para el valor de Z observado o

una probabilidad relativa a la azar para este resultado igual o superior a la correspondiente con el nivel de significación alfa prefijado ($p \geq \alpha$).

f) Una muestra, una variable cuantitativa que no sigue ni se aproxima a la distribución normal estándar (tiene distribución libre) y una variable cualitativa dicotómica, si se desea determinar si los datos de esta variable cuantitativa para los dos grupos formados o muestras a partir de las dos categorías de la variable cualitativa dicotómica pueden provenir de la misma población cuando existen valores atípicos (*outliers*) en la distribución de frecuencias de sus valores o sospecha de sesgo en la obtención de los mismos, empleamos el método no paramétrico prueba no paramétrica de la mediana yun posible planteamiento de hipótesis sería:
   - H_0: mediana de la población estimada 1 = mediana de la población estimada 2.
   - H_1: mediana de la población estimada 1 ≠ mediana de la población estimada 2.

   Nota: se considera un caso especial de la prueba de independencia con el estadístico chi-cuadrado (Pearson o corregido de Yates), pues se basa en esta última.

   Pasos a seguir en el contraste de hipótesis
   - Ordenar los valores de la variable cuantitativa continua en cada muestra independiente (grupos para la variable dicotómica) y determinar la mediana en cada una, así como la mediana común para ambas.
   - Calcular la frecuencia absoluta de valores por encima y por debajo de la mediana común en cada muestra independiente.
   - Introducir los valores calculados en una tabla de contingencia 2x2 simple: filas: número de valores mayores y menores que la mediana común; columnas: muestra 1 y muestra 2 (Tabla 4.28).

Tabla 4.28. Representación tabulada de los elementos teóricos necesarios para realizar el contraste de hipótesis empleando la mediana de dos grupos de datos independientes con una variable cuantitativa de libre distribución

Valores de la variable cuantitativa en las obseraciones realizadas.	Muestra o grupo de datos 1 (para categoría 1 de la variable dicotómica)	Muesta o grupo de datos 2 (para categoría 2 de la variable dicotómica)	Total
Con valores mayores que la mediana común	a	b	a + b
Con valores Menores que la mediana común	c	d	c + d
Total	**a + c**	**b + d**	**N = a + b + c + d**

– El estadístico de prueba a emplear en el contraste sería:

χ^2 de Pearson(observado-calculado) $= \dfrac{n(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$ sigue una distribución teórica de probabilidades $\chi^2_{(c-1)(f-1);1-\alpha}$ con 1 grado de libertad para el valor de alfa prefijado.

Nota: se requiere que ninguna de las celdas tenga una frecuencia esperada menor de 5 para su aplicación.

– Valor de p (p-valor), bajo H_0:

p-valor = Pr [$\chi^2_{1,1-\alpha}$ (teórico-esperado) ≥ χ^2 de Pearson (observado-calculado)].

Regla de decisión

– Si χ^2 (observado-calculado) de Pearson > $\chi^2_{1,1-\alpha}$ (teórico-esperado), esto se corresponde con la ubicación de χ^2 observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces hay evidencias suficientes para rechazar H_0, es decir, hay diferencias significativas entre los valores distribuidos en las muestras en relación a la mediana, estimándose un p-valor bajo H_0 para el χ^2 observado o las probabilidades relativas al azar para este resultado, menor que el nivel de significación alfa para el nivel de confianza seleccionado ($p < \alpha$).

– Si χ^2 (observado-calculado) de Pearson ≤ $\chi^2_{1,1-\alpha}$ (teórico-esperado), esto se corresponde con la no ubicación de χ^2 observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces no hay evidencias suficientes para rechazar H_0, es decir, no hay diferencias no significativas entre los valores distribuidos en las muestras en relación a la mediana, estimándose un p-valor bajo H_0 para el χ^2 observado o las probabilidades relativas al azar para este resultado iguales o mayores que el nivel de significación alfa para el nivel de confianza seleccionado ($p \geq \alpha$).

g) Una muestra, una variable cuantitativa que no sigue ni se aproxima a la distribución normal estándar (tiene distribución libre) y una variable cualitativa politómica u ordinal, si se desea determinar si los datos de esta variable cuantitativa aleatoria para los más de dos grupos formados o muestras a partir de las más de dos categorías o niveles de la variable cualitativa politómica u ordinal respectivamente pueden provenir de la misma población**,** empleamos el método no paramétrico prueba no paramétrica de Kruskal-Wallis y un posible planteamiento de hipótesis estadísticas sería:

– H_0: las poblaciones de las que provienen las más de dos muestras (formadas según la variable cualitativa) están equidistribuídas para los valores de la variable cuantitativa o las medianas son iguales para los más de dos grupos de datos formados.

– H_1: las poblaciones de las que provienen las más de dos muestras (formadas según la variable cualitativa) no están equidistribuídas para los valores de la

variable cuantitativa o las medianas no son iguales para los más de dos grupos de datos formados.

Pasos a seguir

- Determinar el número total de observaciones según los más de dos grupos de datos formados, cuya extracción de estas observaciones en las muestras debe cumplir con la alietoriedad.
- Ordenar en forma ascendente los valores de la variable cuantitativa aleatoria correspondiente a todas las observaciones (sujetos) de todas las muestras.
- Asignar a cada valor de las observaciones ordenadas un valor de rango (Ri) correspondiente comenzando por el 1.
- Distribuir los rangos correspondientes en cada muestra y calcular el valor de su sumatoria.
- Calcular el estadístico H de Kruskal-Wallis o de prueba chi-cuadrado según la expresión:

$$\text{H observado-calculado (de Kruskal} - \text{Wallis)} = \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(N+1)$$

sigue una distribución teórica de probabilidades $\chi^2_{k-1,\,1-\alpha}$ con k – 1 grados de libertad para el nivel alfa prefijado en muestras grandes (N ≥ 15 o k > 5).

donde:

k: número de muestras (siempre k > 2);

ni: tamaño de cada muestra i;

N: total de observaciones (sujetos) = $n_1 + n_2 + n_k$;

R_i: rangos de la muestra i.

Otra forma de expresar el estadístico de contraste:

$$\text{H observado-calculado} = (N-1)\frac{\sum_{i=1}^{g} n_i(\overline{r}_i - \overline{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \overline{r})^2}$$

, sigue una distribución teórica de probabilidades $\chi^2_{n-1,\,1-\alpha}$ con n – 1 grados de libertad para el nivel alfa prefijado de igual forma para muestras grandes (N ≥ 15 o k > 5).

donde:

n: número total de muestras o grupos independientes (siempre n > 2);

g: número de observaciones en cada muestra;

j: representa cada muestra;

i: representa cada observación de cada muestra independiente;

N: total de observaciones (sujetos) de todas las muestras independientes;

rij: rango u orden del valor de la variable en cada observación (sujeto) para cada muestra independiente;

$\overline{r}$: rango medio para todas las observaciones (m) y todas las muestras independientes;

$\overline{r}_i$: rango medio de las observaciones de la muestra i.

Regla de decisión

- Se compara el valor observado-calculado del estadístico H *vs.* el valor teórico-esperado-tabulado del mismo, según el nivel de confianza preseleccionado (véase el anexo 1), siendo rechazada H_0 cuando H (observado-calculado) > H tabulado.
- Método que emplea la distribución de chi-cuadrado: en el caso de muestras grandes (N ≥ 15 o k > 5), se puede emplear la distribución de probabilidad de chi-cuadrado según los grados de libertad y valor de alfa correspondientes para obtener el valor observado de H y el criterio de decisión sería como sigue:
  - Valor de p o p-valor bajo H_0:
    p-valor = Pr [χ^2 (teórico-esperado) ≥ χ^2 (observado-calculado)].

Regla de decisión

- Si χ^2 (observado-calculado) > $\chi^2_{k-1,1-\alpha}$ o $\chi^2_{n-1,1-\alpha}$ (teórico-esperado), esto se corresponde con la ubicación de chi-cuadrado observado (para H estimado) en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces hay evidencias suficientes para rechazar H_0, es decir, en las más de dos poblaciones estimadas de los grupos formados para la variable cualitativa independiente, los valores de la variable cuantitativa aleatoria no están equidistribuidos, hay al menos una muestra con mediana diferente a las otras de forma estadísticamente significativa, al estimarse un p-valor bajo H_0 o una probabilidad relativa al azar según el chi-cuadrado observado menor que la esperada según el nivel de significación alfa para el modelo teórico de referencia ($p < \alpha$).
- Si χ^2 (observado- calculado) ≤ $\chi^2_{k-1,1-\alpha}$ o $\chi^2_{n-1,1-\alpha}$ (teórico-esperado), esto no se corresponde con la ubicación de χ^2 observado (para H estimado) en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces no hay evidencias suficientes para rechazar H_0, es decir, en las más de dos poblaciones estimadas de los grupos formados para la variable cualitativa independiente, los valores de la variable dependiente cuantitativa aleatoria están equidistribuidos, no hay al menos una muestra con mediana diferente a las otras de forma estadísticamente significativa, al estimarse un p-valor bajo H_0 o una probabilidad relativa al azar según el χ^2 observado igual o superior que la esperada según el nivel de significación alfa para el modelo teórico de referencia ($p \geq \alpha$).

h) Dos muestras dependientes o pareadas y una variable aleatoria cuantitativa que no sigue ni se aproxima a la distribución de frecuencia normal estándar, si se desea realizar la comparación no paramétrica de medias dependientes (emparejadas) con número de observaciones grandes, empleando prueba de los signos de Wilcoxon, el posible planteamiento para el contraste de hipótesis bilateral sería:
   - Contraste bilateral:

- H_0: existe equidistribución para los valores de la variable en ambas poblaciones estimadas pareadas o existe una igualdad en la cantidad de signos positivos y negativos para las diferencias en los n pares de observaciones o la probabilidad de obtener signos positivo o negativo para las diferencia de los valores de la variable cuantitativa en ambas poblaciones estimadas es de 0,5; es decir, H_0: P = 0,5
- H_1: no existe equidistribución para los valores de la variable en ambas poblaciones estimadas o existe una desigualdad en la cantidad de signos positivos y negativos para las diferencias en los n pares de observaciones o la probabilidad de obtener signos positivo o negativo para las diferencia de los valores de la variable cuantitativa en ambas poblaciones estimadas no es de 0,5; es decir, H_1: P ≠ 0,5.

  El estadístico de contraste es S o X, que expresa el número de observaciones por pares (n) con el respectivo signo (+ o –) a partir de la diferencia entre los dos valores por pares tomados de la variable cuantitativa de libre distribución. Los pasos a seguir:
- Seleccionar los datos muestrales (por pares) en forma aleatoria.
- Determinar el número de observaciones por pares n.
- Determinar el valor y signo de las diferencias entre los valores de la variable cuantitiva de cada observación por pares.
- Sumar el número de observaciones por parejas cuyas diferencias por ejemplo tienen signo positivo (S_+) para el caso en que sea el menos frecuente las de signo +.
- Determinar en la tabla de referencia de la distribución binomial (véase loa anexos), las probabilidades exactas esperadas de obtener para un igual o más extremo resultado que los determinados para el número de signos positivos (S_+) observados si no hubieran diferencias entre las probabilidades para cada signo a estimar, según el número de pares de observaciones realizadas (n) para obtener una proporción (p) = 0,5, mitad de los signos positivos y mitad negativos.

  Nota: cuando ocurren empates al aplicar la prueba del signo, el procedimiento que se recomienda seguir es el de ignorarlos y emplear la prueba solo para aquellos pares en los que no ocurren empates. Este procedimiento puede representar un problema si se tienen numerosos empates y el número original de pares es relativamente pequeño.

Regla de decisión

– Para el caso de que 9 < n < 25 se compara el número de signos para las diferencias en las observaciones por pares menos frecuente (S observado) versus su valor crítico (tabulado-esperado) según n y el nivel de significación alfa

seleccionado (0,01 o 0,05) para esta prueba (véase el anexo 1) y como criterio de decisión en el contraste de hipótesis se utiliza:

- La hipótesis nula (H_0) se rechaza si el número de veces que se repite el signo menos frecuente observado en la muestra de estudio (X o S) es menor o igual que el valor crítico inferior esperado que presenta la tabla de referencia para el nivel de significación preseleccionado por el investigador.
- Método por el empleo de la distribución binomial: para el caso en que la probabilidad exacta estimada empleando la distribución binomial según el valor de S+ observado para el número de observaciones por pares (n) y P = 0,5 sea menor que el valor de alfa prefijado por el investigador, esto significa que la probabilidad de equivocación si se rechaza H_0 es menor que el límite establecido por alfa (para el error tipo 1), es decir, se puede rechazar H_0, por lo que no se debe aceptar que existe una equidistribución para los valores de la variable en ambas poblaciones estimadas pareadas alrededor de un valor central igual a 0 (de los dos grupos de datos emparejados respectivos, uno de cada muestra) a partir de la desigual cantidad de signos positivos y negativos para las diferencias entre los n pares de observaciones.
- Método por la aproximación de la distribución normal estándar Z ~ N (0, 1), según el procedimiento descrito por Triola (2009): para el caso en que el tamaño de los grupos o el número de observaciones por parejas n > 25, se puede realizar la aproximación de la distribución Normal estándar y se calcula Z (observado-calculado) como sigue:

$$Z\ (\text{observado-calculado}) = \frac{(x + 0{,}5) - \left(\frac{n}{2}\right)}{\frac{\sqrt{n}}{2}}$$

donde:

X: número de observaciones n con el signo para la diferencia de los valores por pares de la variable cuantitativa, que menor frecuencia se observó.

– Valor de p o p-valor bajo H_0:
p-valor = Pr [Z (teórico-esperado) ≥ Z (observado-calculado)].

Regla de decisión

– Para Z (observado-calculado) < 0:
  - Si Z (observado-calculado) < Z (teórico-esperado), según el valor alfa prefijado, esto se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces hay evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas se puede rechazar que tengan equidistribución los datos de la variable cuantitativa, al estimarse un

p-valor bajo H_0 para Z observado o una probabilidad relativa al azar para este resultado, inferior a la correspondiente con el valor de significación alfa prefijado ($p < \alpha$).

- Si Z (observado- calculado) ≥ Z (teórico-esperado), según el valor alfa prefijado, esto no se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que aumenta la probabilidad de cometer el error tipo 1, entonces no existen evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas no se puede rechazar que tengan equidistribución los datos de la variable cuantitativa, al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa a la azar para este resultado igual o superior a la correspondiente con el valor de significación alfa prefijado ($p \geq \alpha$).

– Para Z (observado-calculado) ≥ 0:

- Si Z (observado-calculado) > Z (teórico-esperado), según el valor alfa prefijado, esto se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces existen evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas se puede rechazar que tengan equidistribución los datos de la variable cuantitativa, al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa a la azar para este resultado menor a la correspondiente con el valor de significación alfa prefijado ($p < \alpha$).
- Si Z (observado- calculado) ≤ Z (teórico-esperado), según el valor alfa prefijado, esto no se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces no existen evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas no se puede rechazar que tengan equidistribución los datos de la variable cuantitativa, al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa a la azar para este resultado igual o superior a la correspondiente con el valor de significación alfa prefijado ($p \geq \alpha$).

i) Dos muestras dependientes o pareadas y una variable cuantitativa que no sigue la distribución normal estándar, si se desea realizar la comparación no paramétrica de medias dependientes (emparejadas), empleando la prueba de los rangos y de signos de Wilcoxon con la aproximación de la normal $Z \sim N(0, 1)$ para muestras grandes (número de observaciones por pares, $n > 30$), el posible planteamiento para el contraste de hipótesis bilateral sería:

– H_0: los datos apareados tienen diferencias que provienen de una población con una mediana igual a cero o que las dos poblaciones (estimadas de los

dos grupos muestrales emparejados) tienen la misma distribución para los valores de la variable cuantitativa.

– H_1: los datos apareados tienen diferencias que provienen de una población con una mediana igual a cero o que las dos poblaciones (estimadas de los dos grupos muestrales emparejados) no tienen la misma distribución para los valores de la variable cuantitativa.

  Nota: esta prueba de los rangos con signo tiene la ventaja de tomar en cuenta la magnitud de las diferencias entre los pares de observaciones y no solo su signo; por lo que este es el mejor método no paramétríco al utilizar observaciones en parejas.

  El estadístico de contraste es $T_{+,}$ que se estima a partir de la sumatoria de los rangos con signo positivo para la diferencia de las n observaciones por pares.

  Los pasos a seguir:

  - Seleccionar los datos muestrales (por pares) en forma aleatoria.
  - Determinar el número de observaciones por pares n en el estudio.
  - Calcular las diferencias en el total de observaciones de los n pares de valores de la variable cuantitativa e ignore el signo obtenido.
  - Ordenar de menor a mayor (en forma ascendente) las diferencias halladas según su valor absoluto para las n observaciones por pares observadas y otorgarle a cada diferencia un valor de rango u orden.

    Nota: los empates entre las diferencias calculadas al adjudicar un valor de rango se manejan de la misma manera que en la prueba de Mann-Whitney, es decir, se le asigna un valor promedio al rango de acuerdo con el orden que ocupan las observaciones involucradas en su secuencia ordenada, pero si una diferencia es igual a cero, el procedimiento que se sugiere es omitir el par y ajustar n (n ajustado = número de observaciones o pares restándole los pares con diferencia igual a 0).
  - Colocar delante de cada rango adjudicado para el resultado de la diferencia entre los dos valores calculados de la variable cuantitativa en las n observaciones por pares, el signo correspondiente para cada diferencia calculada previamente y agrupar estos valores de rango según el tipo de signo especifico.
  - Realizar la sumatoria de la magnitudes correspondientes a los rangos con signo positivo (T_+) y con signo negativo (T_-).
  - Determinar en la tabla de referencia para esta prueba de rangos y signos de Wilcoxon (véase el anexo 1), el valor crítico (superior e inferior) del estadístico T según el número de pares de observaciones (n) y el nivel de significación alfa predeterminado por el investigador.
  - Comparar el mayor valor absoluto observado del estadístico: sumatoria de rangos (según signos: positivo y negativo) *versus* el respectivo estadístico

en su punto crítico superior tabulado, así como el menor valor absoluto observado del estadístico: sumatoria de rangos versus el punto crítico inferior para el estadístico respectivo tabulado-esperado-teórico, según el número de observaciones por pares empleadas y el nivel de significación alfa seleccionado por el investigador.

Regla de decisión

- Si el menor valor absoluto del estadístico para la sumatoria de rangos T (con signo observado-calculado) < T (con signo teórico-esperado) (valor crítico inferior) según el número de observaciones por pares n y el nivel de significación alfa a dos colas preseleccionado por el investigador, esto significa que la probabilidad de equivoción si se rechaza H_0 es menor que el límite establecido para alfa, es decir, se puede rechazar H_0, por lo que no existe equidistribución significativa para los valores de la variable en ambas poblaciones estimadas pareadas (de los dos grupos de datos emparejados respectivos, uno de cada muestra).
- Método por la aproximación de la distribución normal estándar Z ~N (0, 1) según el procedimiento descrito por Triola M (2009) para el caso que el número de pares de observaciones (n > 30) se puede emplear la aproximación de la distribución normal estándar Z ~ N (0, 1) y emplear el estadístico T_+ en el contraste de hipótesis como sigue:
  - Calcular el valor del estadístico T_+, que se corresponde con la sumatoria de los rangos con signo positivo de las diferencias para los n pares de observaciones (generalmente se escoge el menor valor de los dos para los pasos siguientes).
  - Estimar el valor esperado [E (T_+)] y la varianza [Var (T_+)] para la sumatoria de rangos con signo positivo de las diferencias ordenadas en los n > 30 pares de observaciones, según la expresión:
    $E(T_+) = n\,(n + 1)\,/4$
    $Var(T_+) = n\,(n + 1)\,(2n + 1)/24$
  - Estimación del valor del estadístico de contraste Z (observado-calculado) para este caso que sigue la distribución normal Z ~ N (0, 1):

$$Z\ (\text{observado-calculado}) = \frac{T - E(T_+)}{\sqrt{E(T_+)}} \quad \text{o}$$

$$Z\ (\text{observado-calculado}) = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

Otra forma de realizar el contraste de hipótesis según Snedecor y Cochran (1989) es empleando el método por la aproximación de la distribución normal con el estadístico de contraste Z ~ N (0, 1), asumiendo una distribución nor-

mal estándar para la diferencia entre los valores emparejados de la variable cuantitativa (d_i). Los elementos a utilizar son:

Suma de rangos observada:

$R_+ = \Sigma\{\text{rango}(|d_i|)/d > 0\}$, para el grupo de diferencias positivas.

$R_- = \Sigma\{\text{rango}(|d_i|)/d < 0\}$, para el grupo de diferencias negativas.

$R_0 = \Sigma\{\text{rango}(|d_i|)/d = 0\}$, para el grupo de diferencias negativas.

Suma de rangos esperada:

$RE_+ = \frac{\sum|r_i|}{2}$, para el grupo de diferencias positivas.

$RE_- = \frac{\sum|r_i|}{2}$, para el grupo de diferencias negativas.

$RE_0 = \Sigma\{\text{rango}(|d_i|)/d_i = 0\}$, para el grupo de diferencias negativas.

- Estadístico de prueba o contraste (T) para la prueba de los rangos y signos de Wilcoxon cuando las parejas de datos para ambas poblaciones con $n \leq 30$:
  $T = \Sigma\{r_i/d_i > 0\}$
- Estadístico $Z \sim N(0, 1)$, aproximación de la normal aplicable a la prueba de los rangos y signos de Wilcoxon, cuando las parejas de datos para ambas poblaciones $n > 30$:

$$Z\ (\text{observado-calculado}) = \frac{T - E(T)}{\sqrt{E(T)}} \approx N(0,1)$$

donde:

$T = \Sigma\{r_i/d_i > 0\}$: valor observado de la suma de rangos para las diferencias positivas, entre los valores de la variable por parejas (mayor que cero);

$E(T) = \frac{n(n+1)}{4}$: valor esperado de T;

$Var(T) = \frac{(2n+1)E(T)}{6}$: varianza de T;

n: tamaño de cada muestra;

$X_{11}, \ldots, X_{1n}$: observaciones de la muestra 1;

$X_{21}, \ldots, X_{2n}$: observaciones de la muestra 2;

$d_i = X_{1i} - X_{2i}$, $i = 1, \ldots, n$;

r_i = signo (d_i) rango $(|d_i|)$;

Φ: función de distribución normal estándar.

– Valor de p (p-valor), bajo H_0:

$$p = \begin{cases} 2\Phi(z) & \text{si } z \leq 0 \\ 2[1 - \Phi(z)] & \text{si } z > 0 \end{cases}$$

Regla de decisión

– Para Z (observado-calculado) < 0:
  - Si Z (observado-calculado) < Z (teórico-esperado) según el valor de alfa, entonces Z observado se ubica en la región de rechazo para H_0, por lo que existe evidencia suficiente para rechazar la hipótesis nula y se acepta la alternativa, es decir, en ambas poblaciones estimadas los valores de la variable cuantitativa no están equidistribuidos, al estimarse una probabilidad (p-valor) donde se cumpliría H_0 o relativa al azar para el valor de Z observado menor que el nivel de significación alfa para el nivel de confianza seleccionado ($p < \alpha$).
  - Si Z (observado- calculado) ≥ Z (teórico-esperado) según el valor de alfa, entonces Z observado no se ubica en la región de rechazo para H_0, por lo que no existe evidencia suficiente para rechazar la hipótesis nula y no se acepta la alternativa, es decir, en ambas poblaciones estimadas los valores de la variable cuantitativa están equidistribuidos, al estimarse una probabilidad (p-valor) donde se cumpliría H_0 o relativa al azar para el valor de Z observado igual o mayor que el nivel de significación alfa para el nivel de confianza seleccionado ($p \geq \alpha$).

– Para Z (observado-calculado) ≥ 0:
  - Si Z (observado-calculado) > Z (teórico-esperado) según el valor de alfa, entonces Z calculado se ubica en la región de rechazo para H_0, por lo que existe evidencia suficiente para rechazar la hipótesis nula y se acepta la alternativa, es decir, en ambas poblaciones estimadas los valores de la variable cuantitativa no están equidistribuidos, al estimarse una probabilidad (p-valor) donde se cumpliría H_0 o relativa al azar para el valor de Z observado menor que el nivel de significación alfa para el nivel de confianza seleccionado ($p < \alpha$).
  - Si Z (observado- calculado) ≤ Z (teórico-esperado) según el valor de alfa, entonces Zcalculado no se ubica en la región de rechazo para H_0, por lo que no existe evidencia suficiente para rechazar la hipótesis nula y no se acepta la alternativa, es decir, en ambas poblaciones estimadas los valores de la variable cuantitativa están equidistribuidos, al estimarse una probabilidad (p-valor) donde se cumpliría H_0 o relativa al azar para el valor de Z observado igual o mayor que el nivel de significación alfa para el nivel de confianza seleccionado ($p \geq \alpha$).

j) Dos muestras no pareadas o independientes, una variable cuantitativa aleatoria que no sigue ni se aproxima a la distribución normal estándar (tiene distribución libre) con datos en dos muestras independientes, si se desea determinar si ambas muestras independientes proceden de la misma población, se emplea el método no paramétrico prueba de la U de Mann-Whitney y un posible planteamiento de hipótesis estadísticas sería:

– H_0: las dos muestras no pareadas proceden de la misma población o las poblaciones de las que provienen las dos muestras no pareadas están equidistribuidas en los valores de la variable cuantitativa o las medianas de la variable cuantitativa en los dos grupos de datos son iguales.
– H_1: las dos muestras no pareadas no proceden de la misma población o las poblaciones de las que provienen las dos muestras no pareadas no están equidistribuidas en los valores de la variable cuantitativa o las medianas de la variable cuantitativa en los dos grupos de datos no son iguales.

Si se tienen dos muestras no pareadas o independientes con los datos tomados de una variable cuantitativa de libre distribución los pasos a seguir para obtener el estadístico U:

- Seleccionar los datos muestrales de forma aleatoria.
- Determinar el número de observaciones de los dos grupos de datos.
- Ordenar de menor a mayor los datos de las observaciones de ambas poblaciones estimadas (para ambos grupos de datos, correspondiendo con las dos categorías de la variable cualitativa dicotómica) y se les asigna un rango R en orden ascendente comenzando por el 1 a cada valor de la variable cuantitativa hasta n en los 2 grupos o poblaciones estimadas a comparar:
- Se calculan los estadísticos:

$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$, para la muestra que estimaría una población 1.

$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$, para la muestra que estimaría una población 2.

donde:

R_1 y R_2: sumatoria de los respectivos rangos de los valores de la variable cuantitativa para ambas poblaciones estimadas (a partir de las dos muestras utilizadas);

n_1 y n_2: número de observaciones de la población estimada de la muestra 1 y el correspondiente con la población estimada de la muestra 2;

- Se selecciona el mayor valor entre U_1 y U_2 y se denota por estadístico de Mann-Whitney como U (observado-calculado), así como se determina el valor crítico superior de la U (teórica-esperada) según la tabla de referencia para la distribución de Mann-Whitney (véase el anexo 1) para efectuar la comparación entre ambos valores.

Regla de decisión

Para el contraste bilateral si $n_1 + n_2 \leq 20$ se emplea el estadístico U de Mann-Whitney como sigue:

– Si U (observado-calculado) > U (teórico-esperado) según número de observaciones y el nivel de significación alfa, entonces se rechaza H_0, es decir, las

dos poblaciones estimadas no están equidistribuidas significativamente en los valores de la variable cuantitativa con libre distribución, al observarse diferencias de los rangos para los valores de la variable y un valor de probabilidad donde se cumpliría H_0 menor que el nivel de significación alfa predeterminado.

– Método por la aproximación de la distribución normal estándar Z ~N (0, 1), según el procedimiento descrito por Snedecor y Cochran (1989) y Triola (2009). Otra forma de realizar el contraste de hipótesis es empleando la aproximación de la distribución normal Z ~ N (0, 1), para grupos de datos o muestras mayores (n_1 y n_2 > 10), lo que conlleva a seguir los siguientes pasos:
  - Ordenar los valores de la variable cuantitativa en orden ascendente de ambas poblaciones (estimadas de las dos muestras para los dos grupos de datos de las dos categorías correspondientes a la variable cualitativa) y asignar un rango a estos valores empezando por 1 hasta n en orden ascendente.
  - Realizar la suma de rangos observada R_1 y R_2 para el grupo de datos de las poblaciones correspondientes estimadas de la muestra 1 y la muestra 2.
  - Calcular el valor de U a partir de R_1 o R_2, se prefiere emplear el menor entre R_1 y R_2, según la expresión:

    $U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$ se escoge R_1 en este caso si fuese el menor).
  - Determinar el valor esperado [E (R_1)] y la varianza [Var (R_1)] para la sumatoria de rangos observada en la población estimada de la muestra 1 (para datos de la variable de la categoría 1), según la expresión:

    E (R_1) = n_1 (n_1 + n_2 +1)/2

    Var (R_1) = $n_1 n_2$ (n_1 + n_2 +1)/12
  - Expresar de forma general el valor esperado [E(U) o mU] y la varianza [Var (U) o σU] para el estadístico U de Mann-Whitney según la expresión:

    $$E(U) = n_1 n_2 + \frac{n_1(n_1+1)}{2} - E(R_1) = \frac{n_1 n_2}{2} \quad \text{o}$$

    $$m_U = \frac{n_1 n_2}{2}$$

    $$Var(U) = Var(R_1) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} \quad \text{o}$$

    $$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$
  - Estimar el valor del estadístico de contraste para la distribución normal estándar Z ~ N (0, 1):

    $$Z \text{ (observado-calculado)} = \frac{U - E(U)}{\sqrt{Var(U)}} \quad \text{o}$$

$$Z\ (\text{observado-calculado}) = \frac{U - \frac{(n_1 n_2)}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

– Valor de p (p-valor), bajo H_0:

$$p - \text{valor} = \begin{cases} 2\Phi(z) & \text{si } z \leq 0 \\ 2\left[1 - \Phi(z)\right] & \text{si } z > 0 \end{cases}$$

Regla de decisión

– Para Z (observado-calculado) < 0:
  - Si Z (observado-calculado) < Z (teórico-esperado), según el valor alfa prefijado, esto se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces hay evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa al azar para este resultado inferior a la correspondiente con el nivel de significación alfa prefijado ($p < \alpha$).
  - Si Z (observado- calculado) ≥ Z (teórico-esperado), según el valor alfa prefijado, esto no se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces no existen evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas no se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa al azar para este resultado igual o superior a la correspondiente con el nivel de significación alfa prefijado ($p \geq \alpha$).

– Para Z (observado-calculado) ≥ 0:
  - Si Z (observado- calculado) > Z (teórico-esperado), según el valor alfa prefijado, esto se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces existen evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa al azar para este resultado menor a la correspondiente con el valor de significación alfa prefijado ($p < \alpha$).
  - Si Z observado- calculado ≤ Z (teórico-esperado), según el valor alfa prefijado, esto no se corresponde con la ubicación de Z observado en el intervalo

de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces no existen evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas no se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa a la azar para este resultado igual o superior a la correspondiente con el valor de significación alfa prefijado ($p \geq \alpha$).

k) Dos muestras no pareadas o independientes, una variable cuantitativa aleatoria que no sigue ni se aproxima a la distribución normal estándar (tiene distribución libre) con datos en dos muestras independientes, si se desea determinar si ambas muestras independientes proceden de la misma población, se emplea el método no paramétrico prueba de la suma de rangos de Wilcoxon y un posible planteamiento de hipótesis estadísticas sería:
   - Contraste bilateral:
     - H_0: las dos muestras no pareadas proceden de la misma población o las poblaciones de las que provienen las dos muestras no pareadas están equidistribuidas en los valores de la variable cuantitativa o las medianas de la variable cuantitativa para los dos grupos de datos son iguales.
     - H_1: las dos muestras no pareadas no proceden de la misma población o las poblaciones de las que provienen las dos muestras no pareadas no están equidistribuidas en los valores de la variable cuantitativa o las medianas de la variable cuantitativa para los dos grupos de datos no son iguales.

       Si se tienen dos muestras no pareadas o independientes con los datos tomados de una variable cuantitativa de libre distribución. Los pasos a seguir para obtener el estadístico T o R (de Wilcoxon):
     - Seleccionar los datos muestrales en forma aleatoria.
     - Determinar el número de observaciones en total entre ambos grupos de datos.
     - Ordenar los valores de la variable cuantitativa en ambos grupos de datos independientes en conjunto.
     - Asignar un número de orden o rango, de menor a mayor comenzando por el 1, a los valores de la variable cuantitativa para cada una de las observaciones de las poblaciones estimadas (en relación a los dos grupos de datos con que contamos según ambas muestras no pareadas) tomadas en conjunto.
     - En caso de que una observación tenga el mismo valor en cada grupo o población estimada, a cada una se le adjudica el promedio de ambos rangos. En forma análoga se manejan los "empates" entre más de dos observaciones idénticas, en cambio, si las observaciones idénticas se repiten en una misma muestra no es necesario asignarles un rango promedio.

- Proceder a sumar los números de orden o rangos dentro de cada población estimada de cada muestra observada.
- Comparar los valores de la sumatoria de rangos obtenidos para cada muestra independiente y seleccionar el menor de los dos valores de rangos sumados, el que se emplea como estadístico de Wilcoxon (T o R) para realizar su comparación con el valor T o R tabulado-esperado (crítico inferior del mismo), mediante la tabla correspondiente de referencia para este procedimiento siguiendo el número de observaciones para ambos grupos de datos y el nivel de significación alfa preseleccionado por el investigador al consultar esa tabla de referencia (véase el anexo 1).

  Nota: siempre que las muestras no pareadas sean del mismo tamaño, la suma de los rangos debe ser igual o muy semejante en las mismas bajo H_0. En cambio, si las observaciones tienden a tener valores mayores en una muestra, la correspondiente suma de los rangos también tiende a ser mayor en esa muestra y es una evidencia en contra de la hipótesis nula y a favor de una diferencia entre ambas poblaciones estimadas.

Regla de decisión

Para el contraste bilateral si $n_1 + n_2 \leq 20$ se emplea el estadístico T o R de Wilcoxon como sigue:

– Si T o R de Wilcoxon (observado-calculado) < T o R valor crítico inferior (teórico-esperado-tabulado) para el número de observaciones en los dos grupos de datos independientes según el valor de alfa, entonces hay evidencias suficientes para rechazar H_0, es decir, las dos poblaciones no están equidistribuídas en los valores de la variable aleatoria cuantitativa con libre distribución o las dos muestras no pareadas no pertenecen a la misma población al observarse diferencias significativas para los rangos sumados correspondientes a ambas poblaciones estimada, donde los resultados obtenidos que pueden ser atribuidos al azar son muy poco probables o la probabilidad de cumplirse H_0 con los datos observados son menor que el nivel de significación alfa preseleccionado por el investigador.

– Método por la aproximación de la distribución normal estándar $Z \sim N(0, 1)$ según el procedimiento descrito por Triola F (2009). Cuando el número de casos en cada grupo excede los tabulados (n_1, $n_2 > 10$) y si ninguno de los grupos contiene menos de cuatro observaciones, puede utilizarse una aproximación a través de la distribución normal estándar empleando la expresión del estadístico $Z \sim N(0, 1)$ como sigue:

$$Z \text{ (observado-calculado)} = \frac{R - \mu_R}{\sigma_R}$$

donde:

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2}$$

$$\sigma_R = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

n_1: tamaño de la muestra a partir de la cual se calcula la suma de rangos R;
n_2: tamaño de la otra muestra;
R: suma de rangos de la muestra con tamaño n1 (se suele emplear como R_1);
μ_R: media de los valores muestrales R que se espera cuando las dos poblaciones tienen medianas iguales;
σ_R: desviación estándar de los valores muestrales R cuando las dos poblaciones tienen medianas iguales.

– Valor de p (p-valor), bajo H_0:

$$p = \begin{cases} 2\Phi(z) & \text{si } z \leq 0 \\ 2\left[1 - \Phi(z)\right] & \text{si } z > 0 \end{cases}$$

Regla de decisión

– Para Z (observado-calculado) < 0:
  • Si Z (observado-calculado) < Z (teórico-esperado), según el valor alfa prefijado, esto se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces hay evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa al azar para este resultado inferior a la correspondiente con el nivel de significación alfa prefijado ($p < \alpha$).

• Si Z (observado- calculado) ≥ Z (teórico-esperado), según el valor alfa prefijado, esto no se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces no existen evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas no se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa al azar para este resultado, igual o superior a la correspondiente con el nivel de significación alfa prefijado ($p \geq \alpha$).

– Para Z (observado-calculado) ≥ 0:
  • Si Z (observado-calculado) > Z (teórico-esperado), según el valor alfa prefijado, esto se corresponde con la ubicación de Z observado en el intervalo

de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces existen evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa al azar para este resultado menor a la correspondiente con el nivel de significación alfa prefijado ($p < \alpha$).

- Si Z (observado- calculado) ≤ Z (teórico-esperado), según el valor alfa prefijado, esto no se corresponde con la ubicación de Z observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces no existen evidencias suficientes para rechazar H_0, es decir, en ambas poblaciones estimadas no se puede rechazar que tengan equidistribución los datos de la variable cuantitativa en ambas poblaciones estimadas, al estimarse un p-valor bajo H_0 para Z observado o una probabilidad relativa a la azar para este resultado igual o superior a la correspondiente con el nivel de significación alfa prefijado ($p \geq \alpha$).

l) Dos muestras independientes o no pareadas, una variable cuantitativa que no sigue ni se aproxima a la distribución normal estándar (tiene distribución libre), si se desea determinar si los datos de dicha variable cuantitativa aleatoria para los dos grupos formados pueden provenir de la misma población cuando existen valores atípicos (*outliers*) en la distribución de frecuencias o hay sospecha de sesgo en los mismos, empleamos el método no paramétrico prueba no paramétrica de la mediana y un posible planteamiento de hipótesis sería:

- H_0: mediana de la población estimada 1 = mediana de la población estimada 2.
- H_1: mediana de la población estimada 1 ≠ mediana de la población estimada 2.

Nota: se considera un caso especial de la prueba de independencia con el empleo del estadístico de prueba chi-cuadrado (de Pearson o corregido de Yates), pues se basa en las pruebas por chi-cuadrado. Los pasos a seguir en el contraste de hipótesis:

- Ordenar los valores de la variable cuantitativa en cada muestra independiente y determinar la mediana en cada una; así como, la mediana común para ambas.
- Calcular la frecuencia absoluta de valores (por encima y por debajo) de la mediana común para ambas muestras.
- Introducir los valores calculados en una tabla de contingencia 2x2 simple (filas: número de valores mayores y menores que la mediana común; columnas: muestra 1 y muestra 2) (Tabla 4.29).

Tabla 4.29. Representación tabulada de los elementos teóricos necesarios para realizar el contraste de hipótesis empleando la mediana de dos grupos de datos independientes con una variable cuantitativa de libre distribución

Valores de la variable cuantitativa en las observaciones realizadas respecto a la mediana.	Muestra o grupo de datos 1	Muesta o grupo de datos 2	Total
Mayores que la mediana común	a	b	a + b
Menores que la mediana común	c	d	c + d
Total	a + c	b + d	N = a + b + c + d

– El estadístico de prueba a emplear en el contraste sería:

χ^2 de Pearson(observado-calculado) $= \dfrac{n(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$, que sigue una distribución teórica de probabilidades $\chi^2_{(c-1)(f-1);\,1-\alpha}$ con 1 grado de libertad para el valor de alfa según el nivel de confianza preseleccionado por el investigador

donde:

a, b, c, d: corresponde a los valores de las celdas en la respectiva tabla de contingencia de dos columnas por dos filas simple y n es el total de observaciones.

– Valor de p (p-valor), bajo H_0:

p-valor = Pr [$\chi^2_{1,\,1-\alpha}$ (teórico-esperado) ≥ χ^2 de Pearson (observado-calculado)].

Regla de decisión

– Si χ^2 de Pearson (observado-calculado) > $\chi^2_{1,\,1-\alpha}$ (teórico-esperado), esto se corresponde con la ubicación de chi-cuadrado observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces hay evidencias suficientes para rechazar H_0, es decir, hay diferencias estadísticamente significativas entre los valores distribuidos en las muestras de la variable en relación a la mediana, estimándose un p-valor bajo H_0 para el chi-cuadrado observado o las probabilidades relativas al azar para este resultado inferiores al nivel de significación alfa para el nivel de confianza seleccionado ($p < \alpha$).

– Si χ^2 de Pearson (observado-calculado) ≤ $\chi^2_{1,\,1-\alpha}$ (teórico-esperado), esto se corresponde con la no ubicación de chi-cuadrado observado en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces no hay evidencias suficientes para rechazar H_0, es decir, no hay diferencias estadísticamente significativas entre los valores distribuidos en las muestras de la variable en relación a la mediana, estimándose un p-valor bajo H_0 para el chi-cuadrado observado o las probabilidades relativas

al azar para este resultado iguales o mayores al nivel de significación alfa para el nivel de confianza seleccionado ($p \geq \alpha$).

m) Una muestra con una variable cuantitativa de libre distribución, cuyos valores son utilizados para formar dos grupos o muestras independientes de acuerdo con cierta característica dicotómica, si se desea evaluar la alietoriedad del orden de los datos a partir del número de rachas o tendencias para los valores de dicha variable cuantitativa, empleando la prueba de tendencias o de rachas, tendríamos como posible planteamiento de hipótesis estadística para el contraste:
   - H_0: la secuencia para los valores de la variable de libre distribución son el resultado de algún proceso aleatorio del orden de los datos.
   - H_1: la secuencia para los valores de la variable de libre distribución no son el resultado de algún proceso aleatorio del orden de los datos.

   Nota: la prueba de rachas de Wald-Wolfowitz en su hipótesis alternativa es más amplia que la prueba U Mann-Whitney ya que establece que las distribuciones difieren en algún aspecto como en la tendencia central, en la dispersión o la asimetría.

   Pasos a seguir:
   - Los datos muestrales se acomodan de acuerdo con algún esquema de orden, por ejemplo, el orden en el que se obtuvieron los valores muestrales.
   - Cada valor de los datos de la variable cuantitativa se debe categorizar en una de dos posibilidades separadas (de forma dicotómica).
   - Determinar el número de valores que se corresponden con cada posibilidad al categorizarlos previamente y que formarán cada grupo como n_1 y n_2 respectivamente.
   - Calcular el número de tendencias o rachas presentes observadas (R o G) de cada posibilidad de la categoría para cada grupo individualmente en el total de observaciones ($n_1 + n_2$), siendo R el número total de tendencias observadas en una secuencia ordenada de $n_1 + n_2$ observaciones, donde n_1 y n_2 son los respectivos tamaños de las muestras, entonces: si la hipótesis nula de que la secuencia de los valores de la variable es aleatorio es cierta, las observaciones de las dos muestras o grupos de datos según la categorización realizada en la secuencia ordenada deben encontrarse bien mezcladas sin llegar a producir un número muy grande ni muy pequeño de tendencias (R).

Regla de decisión

Cuando se cumple que $n_1 + n_2 \leq 20$, se compara R (observado calculado) de las dos muestras con el valor R crítico (teórico tabulado) en un nivel de significación alfa prefijado (véase el anexo 1):

- Si R (observado-calculado) ≤ R valor crítico inferior (esperado-tabulado) o si R observado-calculado ≥ R crítico superior (esperado-tabulado) según el nivel alfa prefijado (generalmente alfa igual a 0,05), esto se interpreta como que no

se entremezclan bien los valores de la variable entre las dos muestras independientes (para dos grupos de datos con características o categoría diferentes) y sugiere que son muestras que estiman poblaciones diferentes o que no tienen una alietoriedad en el orden de la extracción de las mismas en un nivel alfa prefijado, siendo por tanto rechazada H_0.

– Método de la aproximación por la distribución normal estándar Z~ N (0, 1) cuando se cumple que $n_1 + n_2 > 20$, la distribución de R se encuentra, en forma adecuada, aproximada por una distribución normal estándar con: media (que se corresponde con el valor esperado de R [E (R)] y la varianza [Var (R)] del número de tendencias o rachas R, según la expresión:

$$E(R) = \frac{2n_1n_2}{n_1 + n_2} + 1$$

$$Var(R) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

– Estimación del valor del estadístico de contraste (de prueba):

$$Z\ (\text{observado-calculado}) = \frac{R\ (\text{observado}) - E(R)}{\sqrt{Var(R)}}\ \text{o}$$

$$Z\ (\text{observado-calculado}) = \frac{G - \mu_G}{\sigma_G}$$

– Valor de p (p-valor), bajo H_0:
Valor p = [P_r |Z (observado-calculado)| ≤ |Z (teórico-esperado)|]

Regla de decisión

– Para Z (observado-calculado) ≤ 0:
  - Si Z (observado- calculado) ≥ Z (teórico-esperado) según el nivel de significación alfa prefijado; entonces no se ubica el valor observadode Z en el área de rechazo para H_0, por lo que no existen evidencias suficientes pararechazar H_0, es decir, los dos grupos de valores de la variable cuantitativa no tienen un número de rachas muy grande o muy pequeño y eso conlleva a que siguen un orden aleatorio, estimándose un valor de p bajo H_0 para el Z observado o de la probabilidad relacionada con el azar mayor o igual al valor de referencia de alfa seleccionado (p-valor > α).
  - Si Z (observado- calculado) < Z (teórico-esperado) según el nivel de significación alfa prefijado, entonces se ubica el valor observadode Z en el área de rechazo para $H_{0,}$ por lo que existen evidencias suficientes pararechazar H_0, es decir, los dos grupos de valores de la variable cuantitativa tienen un número de rachas muy pequeño y eso conlleva a que no siguen un orden aleatorio; estimándose un valor de p bajo H_0 para el Z observadoo de la probabilidad relacionada con el azar menor al valor de referencia de alfa seleccionado (p-valor < α).

- Para Z (observado-calculado) > 0:
  - Si Z (observado- calculado) ≤ Z (teórico-esperado) según el nivel de significación alfa prefijado; entonces no se ubica el valor observadode Z en el área de rechazo para H_0; por lo que no existen evidencias suficientes para rechazar H_0, es decir, los dos grupos de valores de la variable cuantitativa no tienen un número de rachas muy grande y eso conlleva a que siguen un orden aleatorio, estimándose un valor de p bajo H_0 para el Z observadoo de la probabilidad relacionada con el azar menor al valor de referencia de alfa seleccionado (p-valor ≥ α).
  - Si Z (observado- calculado) > Z (teórico-esperado) según el nivel de significación alfa prefijado; entonces se ubica el valor observadode Z en el área de rechazo para H_0, por lo que existen evidencias suficientes para rechazar H_0, es decir, los dos grupos de valores de la variable cuantitativa tienen un número de rachas muy grande y eso conlleva a que no siguen un orden aleatorio; estimándose un valor de p bajo H_0 para el Z observadoo de la probabilidad relacionada con el azar menor al valor de referencia de alfa seleccionado (p-valor < α).

n) Más de dos muestras relacionadas, una variable cuantitativa que no sigue ni se aproxima su distribución de frecuencias a la normal estándar, si se desea realizar la comparación no paramétrica de los valores con su asignación por rangos de una variable cuantitativa aletatoria de distribución libre con más de dos mediciones realizadas (equivale a más de dos muestras relacionadas), asumiendo la alietoriedad en la extracción de la muestra sin requerir normalidad ni homocedasticidad en los valores de la variable para las más de dos muestras relacionadas o grupos relacionados de datos de pequeño tamaño donde no se puede aplicar el ANOVA para medidas repetidas, empleando la prueba de Friedman, un posible planteamiento de hipótesis estadísticas para el contraste sería:
   - H_0: las más de dos muestras relacionadas provienen de una misma población o las mediciones repetidas de datos cuantitativos en una población estimada no guardan diferencias.
   - H_1: las más de dos muestras relacionadas no provienen de una misma población o las mediciones repetidas de datos cuantitativos en una población estimada guardan diferencias.

   Pasos a seguir:
   - Colocar los valores para la variable cuantitativa de cada observación realizada (i) en los más de dos momentos de medición (j) que equivalen a las más de dos muestras relacionadas o tratamientos.
   - Adjudicar un valor de rango o jerarquía (en orden creciente empezando por el 1 hasta el máximo k) a cada valor de la variable cuantitativa empleada para cada observación realizada (R_i) a lo largo de los más de dos momentos de medición (muestras) empleados en el estudio.

– Sumar los rangos correspondientes de cada observación realizada para cada momento de medición o muestra relacionada (ΣR_i) lo que permite obtener el total de rangos de cada muestra, grupo o tratamiento empleado en el estudio (R_j).
– Calcular los rangos medios para cada momento en que se realizó la medición (R_j medio), es decir, para las más de dos mediciones que equivale a más de dos muestras, grupos o tratamiento relacionados).
(R_j medio) = $R_j/2$.
– El procedimiento de Friedman determina cuándo una disparidad observada entre los R_j es suficiente para rechazar la hipótesis nula.
– Emplear el estadístico para el contraste:

$$\text{S o Q de Friedman (observado-calculado)} = \frac{12}{nk(k+1)}\left[\sum_{j=1}^{k} R_j^2\right] - 3nk(k+1)\ ,$$ que sigue una distribución teórica de probabilidades $\chi^2_{k-1,\,1-\alpha}$ con k – 1 grados de libertad para el valor de alfa prefijado para muestras grandes preferentemente

donde:

k: total de momentos o veces en que se realizó mediciones de la variable cuantitativa (equivale al número total de muestras, grupos o tratamientos);

n: número total de observaciones realizadas para el total de momentos empleados para las mismas;

Rj: la sumatoria de rangos para cada momento de medición, muestra, grupo o tratamiento.

Nota aclaratoria: si existen valores repetidos de la variable para las observaciones el resultado del χ^2 (observado-calculado) debe ajustarse dividiendo su valor entre el factor de corrección (FC):

$$FC = 1 - \frac{\sum_i T_i}{nk(k^2-1)}$$

donde:

$$T_i = \sum_h T_{ih}^3 - \sum_h T_{ih}\ ;$$

t: número de observaciones empatadas para un rango determinado en el i-ésimo sujeto u observación.

Regla de decisión

– Si S o Q de Friedman (observado-calculado) > $\chi^2_{k-1,\,1-\alpha}$ o $\chi^2_{n-1,\,1-\alpha}$ (teórico-esperado), entonces S o Q Friedman observado se ubica en el área de rechazo para H_0, por lo que existe evidencia suficiente para rechazar H_0 y aceptar la alternativa, es decir, las más de dos muestras relacionadas con datos cuantitativos no tienen equidistribución de los mismos, por lo que no tienen una misma población de procedencia, estimándose un p-valor bajo H_0 para el valor observado de Friedman (chi-cuadrado) o probabilidades del azar para

este resultado menor que las correspondientes al nivel de significación alfa preseleccionado ($p < \alpha$).

- Si S o Q de Friedman (observado- calculado) $\leq \chi^2_{k-1,1-\alpha}$ o $\chi^2_{n-1,1-\alpha}$ (teórico-esperado), entonces S o Q Friedman observado no se ubica en el área de rechazo para H_0, por lo que no existe evidencia suficiente para rechazar H_0 y aceptar la alternativa, es decir, las más de dos muestras con datos cuantitativos tienen una equidistribución de los mismos, por lo que tienen una misma población de procedencia, estimándose un p-valor bajo H_0 para el valor observado de Friedman (chi-cuadrado) o probabilidades del azar para este resultado, igual o mayor que las correspondientes al nivel de significación alfa preseleccionado ($p \geq \alpha$).

ñ) Más de dos muestras relacionadas, una variable cualitativa nominal dicotómica, si se desea determinar si los datos de esta variable pueden provenir de la misma población, empleamos el método no paramétrico prueba no paramétrica de Cochran y un posible planteamiento de hipótesis estadísticas para el contraste sería:

- H_0: las poblaciones de las que provienen las más de dos muestras están equidistribuídas en los valores de la variable cualitativa nominal dicotómica.
- H_1: las poblaciones de las que provienen las más de dos muestras no están equidistribuídas en los valores de la variable cualitativa nominal dicotómica.

Pasos a seguir:

- Se construye una tabla para incluir los datos necesarios de la variable en estudio (Tabla 4.30).

Tabla 4.30. Representación tabulada de los elementos teóricos necesarios para realizar el contraste de hipótesis en la prueba de Cochran

Número de observaciones (n) según la categoría para el total de las muestras relacionadas	Muestra 1	Muestra 2	Muestra ...	Muestra M	Total de respuestas positivas
1	$X_{1,1}$	$X_{1,2}$	....	$X_{1,m}$	S_1
2	$X_{2,1}$	$X_{2,2}$	.....	$X_{2,m}$	S_2
....	....	.....	....	....	....
N =	$X_{N,1}$	$X_{N,2}$	.....	$X_{N,m}$	S_N
Total	T_1	T_2	....	T_m	T
Proporción	P_1	P_2	....	P_m	P

- El estadístico para el contraste (de prueba) sería:

$$\text{Q de Cochran (observado-calculado)} = \frac{N^2(m-1)}{m} \cdot \frac{\sum_{j=1}^{m}(p_j - \bar{p})^2}{N\bar{p}(1-\bar{p}) - \sum_{n=1}^{N}(P_n - \bar{p})^2}$$

o también:

$$\text{Q de Cochran (observado-calculado)} = (m-1)\cdot\frac{m\sum_{j=1}^{m}T_j^2 - T^2}{mT - \sum_{n=1}^{N}S_n^2}$$, sigue una distribución teórica de probabilidades $\chi^2_{m-1;\,1-\alpha}$ con m – 1 grados de libertad para el valor de alfa preseleccionado.

donde:

m: total de muestras empledas;

N: total de observaciones realizadas para todas las muestras;

X: representa en cada celda de la tabla, una de las dos respuestas posibles que dan los individuos en relación a una variable cualitativa dicotómica, donde corresponde el valor 0 al fracaso y el valor 1 al éxito;

S_1: representa el número total de sujetos con respuesta positiva (éxito) para la variable dicotómica en la primera observación de las m muestras;

S_n: representa el número total de sujetos con respuesta positiva (éxito) para la variable dicotómica en la observación n de las m muestras;

T_1: representa el número total de sujetos con respuesta positiva (éxito) para la variable dicotómica de la primera muestra;

T: representa el número total de sujetos con respuesta positiva (éxito) de la variable dicotómica en todas las muestras;

$P_1 = T_1/N$: proporción de sujetos con respuesta positiva en la muestra 1 para el total de observaciones de la muestra 1:

$$P_j = \frac{T_j}{N};$$

$$P_m = \frac{S_m}{m}$$

$$p = \frac{1}{m}\sum_{j=1}^{m}p_j = \frac{1}{m}\sum_{n=1}^{N}P_n = \frac{T}{Nm}$$

– Valor p (p-valor), bajo H_0:

p-valor = Pr[χ^2_{m-1} (teórico-esperado) ≥ Q de Cochran (observado-calculado)]

En algunos textos como el de Polo Vega (2011) se utiliza una metodología simplificada al referirse a la prueba de Cochran aunque siguiendo la propuesta realizada por Cochran (1950) y el estadístico de contraste o de prueba empleado como Q de Cochran se expresa como sigue:

$$Q = \frac{J(J-1)\sum T_{+j}^2 - (J-1)T^2}{J\cdot T - \sum T_{i+}^2}$$, sigue una distribución teórica de probabilidades $\chi^2_{j-1;\,1-\alpha}$ con j – 1 grados de libertad para el valor de alfa preseleccionado.

donde:

T_{+j}: número de éxitos en cada muestra relacionada (fila) de todos los individuos;
T_{i+}: número de éxitos de cada individuo (columna) en el total de muestras relacionadas;
T: número total de éxitos;
J: número de muestras relacionadas.
Nota: El número de éxitos se refiere a la respuesta positiva o esperada de las posibles de la variable nominal en el experimento.
Una forma práctica y sencilla de organizar los datos de las muestras relacionadas para este contraste de hipótesis siguiendo los elementos necesarios de la expresión matemática anterior del estadístico de Cochran (Tabla 4.31).

Tabla 4.31. Representación tabulada de los elementos teóricos necesarios para realizar el contraste de hipótesis en la prueba de Cochran

Muestras relacionadas	Orden de individuos					Número de éxitos en cada muestra relacionada (T_{+j})
	1	2	3	4	n	
1						T_{+1}
2						T_{+2}
3						T_{+3}
...	...	...	...	...	...	...
Total de muestras relacionadas (J)	Número de éxitos de cada individuo en el total de muestras relacionadas (T_{i+}).					Número total de éxitos en las J muestras relacionadas (T)

Regla de decisión

- Si Q de Cochran (observado-calculado) > $\chi^2_{k-1,1-\alpha}$ o $\chi^2_{j-1,1-\alpha}$ (teórico-esperado), entonces Q de Cochran observado se ubica en el área de rechazo para H_0, por lo que existe evidencia suficiente para rechazar H_0 y aceptar la alternativa, es decir, las más de dos muestras relacionadas al no tener equidistribución de los valores de la variable no se corresponden con una misma población de procedencia, estimándose un p-valor bajo H_0 para el valor observado de la Q de Cochran o las probabilidades del azar para este resultado menores que las correspondientes al nivel de significación alfa preseleccionado ($p < \alpha$).
- Si Q de Cochran (observado-calculado) ≤ $\chi^2_{k-1,1-\alpha}$ o $\chi^2_{j-1,1-\alpha}$ (teórico-esperado), entonces Q de Cochran observado no se ubica en el área de rechazo para H_0, por lo que no existe evidencia suficiente para rechazar H_0 y aceptar la alternativa, es decir, en las más de dos muestras relacionadas no se puede plantear que no tienen equidistribución de los valores de la variable por lo que se corresponden con una misma población de procedencia, estimándose un p-valor bajo H_0 para el valor observado de la Q de Cochran o las probabilidades del azar para este resultado igual o mayor que las correspondientes al nivel de significación alfa preseleccionado ($p \geq \alpha$).

p) Más de dos muestras no relacionadas, una variable cuantitativa aleatoria que no sigue ni se aproxima a la distribución normal estándar (tiene distribución libre), si se desea determinar si los datos de esta variable cuantitativa pueden provenir de la misma población se emplea el método no paramétrico prueba no paramétrica de Kruskal-Wallis y un posible planteamiento de hipótesis estadísticas para el contraste sería:

- H_0: las poblaciones de las que provienen las más de dos muestras están equidistribuidas en los valores de la variable cuantitativa o que no existen diferencias entre las medianas de los grupos formados.
- H_1: las poblaciones de las que provienen las más de dos muestras no están equidistribuidas en los valores de la variable cuantitativa o que existen al menos dos medianas diferentes de los grupos formados.

Pasos a seguir

- Ordenar en forma ascendente los valores de la variable cuantitativa aleatoria correspondiente a todas las observaciones en los sujetos de todas las muestras.
- Asignar a cada valor de las observaciones ordenadas un valor de rango (R_i) correspondiente, comenzando por el 1.
- Distribuir los rangos correspondientes en cada muestra i y calcular el valor de su sumatoria (ΣR_i).
- Calcular el estadístico de prueba o contraste (H) según la expresión:

$$H\,(\text{observado} - \text{calculado}) = \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(N+1)$$

, que sigue una distribución teórica $\chi^2_{k-1,\,1-\alpha}$ con k – 1 grados de libertad para un valor de alfa según el nivel de confianza preseleccionado para muestras grandes (N ≥ 15) y cuando el número de grupos (k > 5).

donde:
k: número total de muestras (siempre k > 2);
ni: es el tamaño de cada muestra i;
N: total de observaciones (sujetos) (n1 + n2 + … + nk);
Ri: representa los rangos para los valores de la variable en la muestra i.
Otra forma de expresar el estadístico de contraste lo constituye:

$$H \text{ de Kruskall } (\text{observado} - \text{calculado}) = (N-1)\frac{\sum_{i=1}^{g} n_i(\overline{r}_i - \overline{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{r}_{ij} - \overline{r})^2}$$

, que sigue una distribución teórica $\chi^2_{n-1,\,1-\alpha}$ con n – 1 grados de libertad para un valor de alfa prefijado para muestras grandes (N ≥ 15) y cuando el número de grupos (k > 5).
donde:
n_i: número total de muestras independientes para las observaciones i (siendo n > 2);

g: representa el número total de observaciones en cada muestra;
j: representa el número de muestras independientes;
i: representa el valor de la variable cuantitativa en cada observación para la muestra independiente;
N: representa el total de observaciones (valores de la variable cuantitativa en sujetos) de todas las muestras independientes;
r_{ij}: rango u orden correspondiente al valor de la variable cuantitativa de cada observación (en el sujeto) i para cada muestra independiente j;
$\bar{r}$: es el rango medio correspondiente para todos los valores de la variable cuantitativa en el total de las muestras independientes;
$\bar{r}_i$: rango medio correspondiente para todos los valores de la variable cuantitativa de la muestra independiente i.

Regla de decisión

Se compara el valor observado del estadístico H con su valor crítico (esperado-teórico) según el nivel de significación alfa preseleccionado por el investigador (véase el anexo 1) en el caso de muestras pequeñas (N <15 o k ≤ 5), donde se rechaza H_0 si H (observado) > H (esperado-tabulado) según el valor de alfa seleccionado. En caso contrario, para muestras grandes (N ≥ 15) y cuando el número de grupos k > 5, se aplica el criterio o regla de decisión en este contraste:

- Si H o K (observado-calculado) > $\chi^2_{k-1,1-\alpha}$ o $\chi^2_{n-1,1-\alpha}$ (teórico-esperado), esto se corresponde con la ubicación de chi-cuadrado observado (estima a H o K) en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, entonces hay evidencias suficientes para rechazar H_0, es decir, en las más de dos poblaciones estimadas de las más de dos muestras o grupos independientes formados, los valores de la variable cuantitativa aleatoria no están equidistribuídos, hay al menos una muestra con mediana diferente a las otras de forma estadísticamente significativa al estimarse un p-valor donde se cumple H_0 para el chi-cuadrado observado o una probabilidad relativa al azar para este resultado menor que el nivel de significación alfa preseleccionado para el modelo teórico de referencia ($p < \alpha$).
- Si H o K (observado- calculado) ≤ $\chi^2_{k-1,1-\alpha}$ o $\chi^2_{n-1,1-\alpha}$ (teórico-esperado), esto no se corresponde con la ubicación de chi-cuadrado observado (estima a H o K) en el intervalo de los valores del área crítica o de rechazo de H_0 según el modelo teórico de referencia, lo que aumenta la probabilidad de cometer el error tipo 1, entonces no hay evidencias suficientes para rechazar H_0, es decir, en las más de dos poblaciones estimadas de las más de dos muestras o grupos formados, los valores de la variable cuantitativa aleatoria están equidistribuídos, no hay al menos una muestra con mediana diferente a las otras de forma estadísticamente significativa al estimarse un p-valor donde se cumpliría H_0 para el chi-cuadrado observado o una probabilidad relativa al

azar para este resultado igual o mayor que el nivel de significación alfa para el modelo teórico de referencia ($p \geq \alpha$).

Pruebas para el contraste de normalidad en la distribución de los datos de una muestra: son pruebas que calculan la probabilidad de encontrar la distribución normal de los datos de una variable cuantitativa de una muestra (o una más alejada de la normalidad), si en la población de la que procede la muestra esa variable sigue una distribución normal perfecta. Los requisitos para aceptar que una variable sigue una distribución normal:

- Prácticamente el 100 % de los valores de la variable deben estar entre la media y tres desviaciones estándar por encima y por debajo: $\bar{x} \pm 3$ desviaciones estándar incluyen el valor máximo y el mínimo de los datos de la variable.
- La distribución de los valores debe ser simétrica: la simetría en valor absoluto debe ser menor que dos veces su error estándar [|Valor calculado de la asimetría| < 2 (error estándar de la asimetría)].
- La distribución debe ser mesocúrtica: la curtosis en valor absoluto, deberá ser menor que dos veces su error estándar [|Valor calculado de la Curtosis | < 2 (error estándar de la Curtosis)].

Ejemplos de pruebas empleadas para evaluar la ''normalidad'' en la distribución de los datos de las variables aleatorias:

- Prueba de D'Agostino (D'Agostino, Belanger y D'Agostino Jr., 1990): tiene como probable planteamiento para el contraste de hipótesis:
  - H_0: Los valores de la variable cuantitativa siguen una distribución Normal en las muestras para cada categoría del factor estudiado (variable cualitativa multinominal).
  - H_1: Los valores de la variable cuantitativa no siguen una distribución Normal en las muestras para cada categoría del factor estudiado (variable cualitativa multinominal).

Como condición se requiere que el tamaño de cada muestra ($n \geq 10$). Los pasos a seguir son:

- Ordenar los valores de las observaciones de menor a mayor en cada muestra de la categoría de la variable cualitativa multinominal (factor) y establecer así mismo un rango para cada uno de estos valores según las observaciones realizadas.
- Calcular en cada muestra (para cada categoría) la media, desviación típica (estándar), un estadístico T y el estadístico de contraste D (observado), donde:

$$D\ (\text{observado-calculado}) = \frac{T}{n^2 S}$$

donde:

$$T = \sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right) x_i = \sum_{i=1}^{n} i x_i - \frac{n(n+1)}{2} \bar{x}$$

i: valor del rango correspondiente a cada valor de x de la variable en cada observación n;

$\bar{x}$: media de los valores de x de la variable para las observaciones n;

S: desviación típica o estándar para los valores de x de la variable en las n observaciones.

- Una vez calculado el estadístico D (observado-calculado) en cada muestra se compara con el valor del D (teórico-esperado) respectivo para la misma muestra, que se calcula mediante la tabla de valores críticos de D'Agostino para este estadístico, partiendo del número de observaciones muestral n y el nivel de significación alfa seleccionado por el investigador (véase el anexo 1).

Regla de decisión

Si D (observado-calculado) > $D_{n,\alpha}$ (teórico-esperado), entonces el valor observado de D queda ubicado en el área de rechazo de H^0, por lo que existen evidencias para no aceptar la normalidad en la distribución de los valores observados de la variable cuantitativa en la muestra conformada para cada categoría del factor o variable cualitativa multinominal en estudio según el nivel de confianza seleccionado.

– Prueba de Shapiro, Francia: tiene como probable planteamiento para el contraste de hipótesis:
  - H_0: los valores de la variable cuantitativa siguen una distribución normal en las muestras para cada categoría del factor estudiado (variable cualitativa multinominal).
  - H_1: los valores de la variable cuantitativa no siguen una distribución normal en las muestras para cada categoría del factor estudiado (variable cualitativa multinominal).

    Como condición se requiere que el tamaño de cada muestra (n) ≥ 10. Los pasos a seguir son:
  - Ordenar los valores de las observaciones de menor a mayor en cada muestra de la categoría de la variable cualitativa multinominal (factor) y establecer así mismo un rango (orden ascendente desde 1 en adelante) para cada una. El estadístico de contraste a emplear es:

$$Z\ (\text{observado-calculado}) = \frac{\ln(1-W') - \hat{\mu}}{\hat{\sigma}}$$

donde:

$$W' = \frac{\left(\sum_{i=1}^{n}(m_i - \bar{m})(x_i - \bar{x})\right)^2}{\left(\sum_{i=1}^{n}(m_i - \bar{m})^2\right)\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)}$$: cuadrado del coeficiente de correlación de Pearson entre los valores de la muestra xi y los valores mi;

n: tamaño de la muestra;

x_1,..., x_n: valores de la muestra;

$\bar{x}$: media de la muestra;

$m_i = \Phi^{-1}\left(\frac{r(i) - 0{,}375}{n + 0{,}25}\right)$, con i = 1,..., n;

r(i): rango que ocupa la observación xi en la muestra ordenada, en caso de empate el rango de las observaciones repetidas es el promedio de sus rangos;

$\bar{m}$: a media de los valores m_i;

$\hat{\mu} = -1{,}2725 + 1{,}0521(v - u)$;

$\hat{\sigma} = 1{,}0308 - 0{,}26758\left(v + \frac{2}{u}\right)$;

u = ln (n);

v = ln (u);

Φ: función de distribución normal estándar.

– Valor de p (p-valor) bajo H_0:
  p-valor = Pr[Z (observado-calculado) ≤ Z (teórico-esperado)]

Regla de decisión

– Para Z (observado-calculado) < 0:
  • Si Z (observado-calculado) < Z (teórico-esperado), entonces se ubica el valor de Z observado en el área crítica o de rechazo en la distribución de valores Z teóricos, por lo que hay evidencias para no aceptar la normalidad en la distribución de los valores cuantitativos observados de la variable estudiada en relación a la muestra conformada para la categoría del factor (variable cualitativa multinominal o politómica en estudio) al estimarse una probabilidad (p-valor) bajo H_0 o una probabilidad relacionada con el azar para el Z observado, menor que el nivel de significación alfa seleccionado.
  • Si Z (observado-calculado) ≥ Z (teórico-esperado), entonces no se ubica el valor de Z observado en el área crítica o de rechazo en la distribución de valores Z teóricos, por lo que hay evidencias para aceptar la normalidad en la distribución de los valores cuantitativos observados de la variable estudiada en relación a la muestra conformada para la categoría del factor (variable cualitativa multinominal o politómica en estudio) al estimarse una probabilidad (p-valor) bajo H_0 o una probabilidad relacionada con el azar para el Z observado igual o superior al nivel de significación alfa seleccionado.
– Para Z (observado-calculado) ≥ 0:
  • Si Z (observado-calculado) > Z (teórico-esperado), entonces se ubica el valor de Z observado en el área crítica o de rechazo en la distribución de valores Z teóricos, por lo que hay evidencias para no aceptar la normalidad en la distribución de los valores cuantitativos observados de la variable

estudiada en relación a la muestra conformada para la categoría del factor (variable cualitativa multinominal o politómica en estudio) al estimarse una probabilidad (p-valor) bajo H_0 o una probabilidad relacionada con el azar para el Z observado menor que el nivel de significación alfa seleccionado.

- Si Z (observado-calculado) ≤ Z (teórico-esperado), entonces no se ubica el valor de Z observado en el área crítica o de rechazo en la distribución de valores Z teóricos, por lo que hay evidencias para aceptar la normalidad en la distribución de los valores cuantitativos observados de la variable estudiada en relación a la muestra conformada para la categoría del factor (variable cualitativa multinominal o politómica en estudio) al estimarse una probabilidad (p-valor) bajo H_0 o una probabilidad relacionada con el azar para el Z observado igual o superior al nivel de significación alfa seleccionado.

– Prueba de Kolmogorov-Smirnov (Thode Jr., 2002):

- Pertenece a las pruebas para evaluar la bondad del ajuste entre una función de distribución empíríca (muestral) y una función de distribución de referencia (hipotética) que sigue la normalidad (bajo la hipótesis nula).
- Evalúa la significación en la distancia máxima o diferencia máxima de las probabilidades para la distribución de los valores de la variable aleatoria cuantitativa según el número de observaciones, entre la función de distribución empírica (muestral) y la hipotética de referencia bajo H_0.

  Uno de los planteamientos de hipótesis a plantear para comprobar la normalidad en la distribución de los datos de una variable aleatoria cuantitativa en una muestra pudiera ser:

- H_0: la función de distribución acumulativa empírica o muestral de los valores de la variable aleatoria cuantitativa es similar a una función de distribución acumulativa hipotética o de referencia normal.
- H_1: la función de distribución acumulativa empírica o muestral de los valores de la variable aleatoria cuantitativa no es similar a una función de distribución acumulativa hipotética o de referencia normal.

  Elementos a tener en cuenta en el contraste de hipótesis:

- No se necesita que los datos de la variable aleatoria cuantitativa se encuentren agrupados y se puede aplicar a muestras de tamaño pequeño.
- Se basa en una comparación entre las funciones de distribución acumulativa de una variable cuantitativa que se observan en la muestra ordenada y la distribución propuesta bajo la hipótesis nula.
- Si la comparación entre ambas distribuciones de los datos de la variable cuantitativa revela una diferencia suficientemente grande entre las funciones de distribución muestral y la propuesta, entonces la hipótesis nula (H_0) probablemente sea rechazable.

- Se denotan por $X_{(1)}$, $X_{(2)}$, …, $X_{(n)}$ a las observaciones ordenadas de una muestra aleatoria de tamaño n y se define la función de distribución acumulativa muestral como:

$$S_n(x) = \begin{cases} 0 & x < x_{(1)} \\ k/n & x_{(k)} \le x < x_{(k+1)} \\ 1 & x \ge x_n \end{cases}$$

donde:

$S_n(x)$: es la proporción del número de valores de la variable cuantitativa en la muestra que son iguales o menores a x y equivale a $F_n(x)$.

$F_0(x)$: expresa la proporción del número de valores de la variable cuantitativa empleados de referencia bajo H0.

El estadístico de contraste de Kolmogorov-Smirnov (D_n) sería:

$D_n = \max_x |S_n(x) - F_0(x)|$ o

D_n (observado-calculado) de Kolmogorov-Smirnov = máx (x) $|F_n(x) - F_0(x)|$

Autores como ThodeJr. (2002) lo expresan como:

$$D^+ = \max_{i=1,\ldots,n}\left[\frac{i}{n} - p(i)\right]$$

$$D^- = \max_{i=1,\ldots,n}\left[p(i) + \frac{(i-1)}{n}\right]$$

para referirse a la máxima diferencia entre las probabilidades acumuladas de los valores de la variable aleatoria cuantitativa de la distribución empírica o muestral (i/n) y la probabilidad acumulada de la función de distribución hipótetica o de referencia bajo H_0 (p (i)), para el número de observaciones (n) realizadas, siendo D+ (para los valores máximos positivos) y D– (para los valores máximos negativos).

donde:

D_n: expresa la máxima diferencia que equivale a decir la máxima desviación entre las probabilidades acumuladas de ambas funciones (muestral e hipotética de referencia) evaluadas para los datos de la variable cuantitativa;

n: tamaño de la muestra empleada.

Regla de decisión en el contraste de hipótesis

Si D_n (observado-calculado) ≤ ($D_{n,\alpha}$ (teórico-esperado) de Kolmogorov-Smirnov (véase el anexo 1) no se puede rechazar H0, por lo que la variable sigue una distribución con ajuste a la normalidad, en caso contrario si se rechaza H_0 y se acepta la hipótesis alternativa.

Nota: derivado de esta prueba y como fruto de los trabajos de Lilliefors (1967) y Stephens (1974) se ha empleado también el estadístico de prueba:

$D^* = \left(\sqrt{n} - 0{,}01 + \frac{0{,}85}{\sqrt{n}} \right)$; que incluye el D (de Kolmogorov-Smirnov) en su fórmula, donde se utiliza el valor D* (teórico-esperado) de referencia igual a 0,895 para un nivel alfa de 0,05 para cualquier tamaño de muestra a emplear, lo que se toma en cuenta en su comparación con el valor D* (observado-calculado) en el contraste de hipótesis.

También fue desarrollado el estadístico de prueba V de Kuiper según la expresión:

$$V^* = \left(\sqrt{n} + 0{,}05 + \frac{0{,}82}{\sqrt{n}} \right) V$$

donde:

$V = D^+ + D^-$, empleando también el estadístico D (Kolmogorov-Smirnov) en su expresión matemática.

– Prueba de Anderson-Darling (Thode Jr., 2002): pertenece a las pruebas para evaluar la bondad del ajuste de la función de distribución acumulativa empírica (muestral) de los datos de una variable cuantitativa respecto a una función de distribución acumulativa de referencia (hipotética) bajo la hipótesis nula de normalidad en su distribución. Evalúa la significación de la diferencia de las probabilidades acumuladas en forma cuadrática para una función de distribución empírica o muestral en relación con las probabilidades acumuladas de una función de distribución hipotética o de referencia con normalidad bajo H_0, pero empleando una función de ponderación en este caso para un número de observaciones determinado.

  Uno de los planteamientos de hipótesis a plantear para comprobar la normalidad en la distribución de los datos de una variable cuantitativa en una muestra, a partir de la bondad en el ajuste de las probabilidades acumuladas de estos datos para la función de distribución muestral respecto a las probabilidades acumuladas esperadas de una función de distribución de referencia, pueden ser:

  - H_0: los valores de la variable cuantitativa en la función de distribución muestral siguen una distribución similar a la de una función de distribución hipotética normal.
  - H_1: los valores de la variable cuantitativa en la función de distribución muestral no siguen una distribución similar a la de una función de distribución hipotética normal.

    Se emplea una función ponderada o de ponderación en la comparación de las probabilidades acumuladas de las distribuciones empírica y la hipotética o de referencia en el número de observaciones predeterminado por medio de la evaluación de la bondad en el ajuste como expresión de la normalidad en la distribución de los valores de la variable cuantitativa estudiada, siendo expresada esta función ponderada como:

$\psi(p) = [p(1-p)]^{-1}$, lo que fue empleado en la formulación matemática de estadístico de prueba A2 (Anderson-Darling, 1954) como sigue:

$$A^2 = -n - n^{-1}\sum_{i=1}^{n}[2i-1]\left[\log(p_{(i)}) + \log(1-p_{(n-i+1)})\right]$$

Regla de decisión en el contraste de hipótesis

- Si A^2 (observado-calculado) de Anderson-Darling ≤ A^2 (teórico-esperado) (véase el anexo 1) no se puede rechazar H0, por lo que la variable sigue una distribución ajustada a la normalidad, en caso contrario si se rechaza H_0 y se acepta la hipótesis alternativa.

  Nota: derivado del procedimiento de Anderson-Darlingy como fruto de los trabajos de Stephens (1986) propuso esta modificación:

  $A^{2*} = \left(1{,}0 + \frac{0{,}75}{n} + \frac{2{,}25}{n^2}\right)A^2$, lo que permite utilizar diferentes tamaños de muestra y poder comparar los valores observados-calculados de A^{2*} con los respectivos valores tabulados de referencia para el nivel alfa previsto como en el caso anterior para el contraste de las hipótesis.

  También se ha desarrollado con el empleo de una función de ponderación a partir de los trabajos de Anderson-Darling el estadístico de prueba W^2 de Cramer-von Mises:

  $W^2 = \frac{1}{12n} + \sum\left(p_{(i)} - \frac{2i-1}{2n}\right)^2$, con una respectiva modificación a partir de diferentes tamaños de muestra cuando se emplean los valores críticos tabulados:

  $W^{2*} = \left(1{,}0 + \frac{0{,}5}{n}\right)W^2$

  Los distintos valores para los puntos porcentuales críticos o que demarcan la significación de los distintos estadísticos descritos hasta el momento relacionados con los trabajos de Anderson-Darling se presentan de forma resumida (Tabla 4.32).

Tabla 4.32. Puntos porcentuales superiores (correspondiendo a los valores críticos o esperados-teóricos) para evaluar la normalidad, por el método de contraste de hipótesis empleando la función de distribución empírica modificada (con los respectivos estadísticos de prueba modificados en cada una)

Estadístico de prueba empleado	90%	95%	97,5%	99%
D^*	0,819	0,895	0,955	1,035
V^*	1,386	1,489	1,585	1,693
A^{2*}	0,656	0,787	0,918	1,092
W^{2*}	0,104	0,126	0,148	0,178

Ejemplos de pruebas de asímetría y Curtosis para la forma de distribución de los datos de una variable aleatoria (D' Agostino, 1990):

- Contraste de asimetría:
  - Estadístico de prueba o contraste:

$$Z_1 = \frac{1}{\sqrt{\ln w}} \ln\left\{\frac{Y}{\alpha} + \left[\left(\frac{Y}{\alpha}\right)^2 + 1\right]^{\frac{1}{2}}\right\}$$

donde:

$$W^2 = -1 + \left[2\left(\beta_2(g_1) - 1\right)\right]^{\frac{1}{2}}$$

$$Y = g_1\left\{\frac{(n+1)(n+3)}{6(n-2)}\right\}^{\frac{1}{2}}$$

$$\beta_2(g_1) = \frac{2(n^2 + 27n - 10)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$$

$g_1 = \dfrac{m_3}{m_2^{3/2}}$: coeficiente de asimetría;

$m_k = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$: momento central de orden k, k = 2, 3;

$X_1,\ldots, X_n$: valores de la muestra;

n: tamaño de la muestra;

$$\alpha = \left(\frac{2}{W^2 - 1}\right)^{1/2}$$

Φ: función de distribución normal estándar.

  - Valor de p (p-valor) donde se cumple H_0:

$$\text{p-valor} = \begin{cases} 2\Phi(Z_1) & \text{si } Z_1 < 0 \\ 2\left[1 - \Phi(Z_1)\right] & \text{si } Z_1 \geq 0 \end{cases}$$

- Contraste de Curtosis:
  - Estadístico de prueba o contraste:

$$Z_2 = \frac{1}{\sqrt{\frac{2}{9A}}}\left\{\left(1 - \frac{2}{9A}\right) - \left[\frac{1 - \frac{2}{A}}{1 + X\sqrt{\frac{2}{A-4}}}\right]^{1/3}\right\}$$

donde:

$$A = 6 + \frac{8}{\sqrt{\beta_1(b_2)}}\left[\frac{2}{\sqrt{\beta_1(b_2)}} + \left(1 + \frac{4}{\beta_1(b_2)}\right)\right]^{1/2}$$

$$\sqrt{\beta_1(b_2)} = \frac{6(n^2-5n+2)}{(n+7)(n+9)}\left[\frac{6(n+3)(n+5)}{n(n-2)(n-3)}\right]^{1/2}$$

$$b_2 = \frac{m_4}{m_2^2}$$

$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$: momento central de orden k, k = 2, 4;

$X_1,\ldots, X_n$: valores de la muestra;

n: tamaño de la muestra;

$$X = \frac{b_2 - E(b_2)}{\sqrt{Var(b_2)}} \ ;$$

$$E(b_2) = \frac{3(n-1)}{n+1} \ ;$$

$$Var(b_2) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)} \ ;$$

Φ: función de distribución normal estándar.

- Valor de p (p-valor); donde se cumple H_0:

$$\text{p-valor} = \begin{cases} 2\Phi(Z_2) & \text{si } Z_2 \leq 0 \\ 2\left[1-\Phi(Z_2)\right] & \text{si } Z_2 > 0 \end{cases}$$

– Contraste conjunto de asimetría y Curtosis (Royston, 1991):

- Estadístico de prueba o contraste:
  $K^2 = -2\ln(P)$, que sigue una distribución chi-cuadrado (X^2_2) con dos grados de libertad para el nivel de significación alfa correspondiente; donde:
  $P = 1 - \Phi(Z)$;

$$Z = \begin{cases} Z_c & \text{si } Z_c < -1 \\ a_1 + b_1 Z_c & \text{si } Z_c < Z_t \\ a_2 + b_2 Z_c & \text{si } Z_2 > 0 \end{cases}$$

$$Z_c = \Phi^{-1}\left(\exp\left[-\frac{1}{2}\left(Z_1^2 + Z_2^2\right)\right]\right);$$

$Z_t = 0{,}55n^{0{,}2} - 0{,}21$;

n: tamaño de la muestra;

$a_1 = [-5 + 3{,}46\ln(n)]\exp[-1{,}37\ln(n)]$;

$b_1 = 1 + [0{,}854 + 0{,}148\ln(n)]\exp[-0{,}55\ln(n)]$;

$$a_2 = a_1 - \left(\frac{2{,}13}{1-2{,}37\ln(n)}\right)Z_t \ ;$$

$$b_2 = \frac{2{,}13}{1 - 2{,}37\ln(n)} + b_1;$$

Φ: función de distribución normal estándar.

- Valor de p (p-valor), donde se cumple H_0:

$$\text{p-valor} = \Pr(\chi_2^2 > K^2)$$

Resumen de algunas pruebas de significación empleadas en el enfoque de riesgo

A partir del empleo de las tablas de contingencia para el análisis descriptivo inicial e inferencia posteriormente, a continuación se presentan las distintas pruebas de significación o para el contraste de hipótesis estadísticas relacionadas con las mismas en el enfoque de riesgo de los estudios observaciones epidemiológicos.

Para tablas 2x2 simples

Prueba de significación para determinar la asociación de dos variables (exposición y enfermedad)

En estudio cohorte con datos por tasas de incidencia (Colectivo de autores, 2006):

– Si V ≥ 5 se utiliza la aproximación normal para su estimación como:

- Estadístico para contrastar:

H_0: $I_0 = I_1$ *vs.* H_1: $I_0 \neq I_1$

$$z = \begin{cases} \dfrac{a - E - \frac{1}{2}}{\sqrt{V}} & \text{si } a > E \\ \dfrac{a - E + \frac{1}{2}}{\sqrt{V}} & \text{si } a \leq E \end{cases}$$

- Valor p:

$$\text{p-valor} = \begin{cases} 2\Phi(z) & \text{si } z < 0 \\ 2\left[1 - \Phi(z)\right] & \text{si } z \geq 0 \end{cases}$$

– Si V < 5 hay que utilizar un método exacto:

- Valor p exacto para contrastar:

H_0: $I_0 = I_1$ *vs.* H_1: $I_0 \neq I_1$

$$\text{p-valor} = \begin{cases} 2\sum_{k=0}^{a}\binom{n}{k}p^k(1-p)^{n-k} & \text{si } a < np \\ 2\sum_{k=a}^{n}\binom{n}{k}p^k(1-p)^{n-k} & \text{si } a \geq np \end{cases}$$

donde:

$E = \frac{nT_1}{T}$;

$V = \frac{nT_1T_2}{T^2}$;

$p = \frac{T_1}{T}$;

z_α: cuantil de la distribución normal que deja una cola α a la izquierda;
Φ: función de distribución de probabilidades de la normal estándar.

Prueba de significación para determinar la asociación o independencia de dos variables (exposición y enfermedad)

En estudio cohorte por incidencia acumulada, estudio tranversal y de caso- control no apareado (Schlesselman, 1982):

- Estadísticos para contrastar:

  H_0: $p_0 = p_1$ *vs.* H_1: $p_0 \neq p_1$:

  - Estadístico $\chi 2$ sin corregir de Pearson (χ^2 Pearson):

    $\chi^2_{\text{Pearson}} = \frac{n(ad-bc)^2}{m_1m_0n_1n_0}$, sigue una distribución (χ^2_1) con 1 grado de libertad

  - Estadístico $\chi 2$ con la corrección de Yates (X^2 Yates):

    $\chi^2_{\text{Yates}} = \frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{m_1m_0n_1n_0}$, sigue una distribución (χ^2_1) con 1 grado de libertad.

- Valor p (p-valor) bajo H_0 para cada uno de estos estadísticos:

  $$\text{p-valor} = \Pr\left(\chi_1^2 \geq \chi^2\right)$$

  donde:
  n: número total de sujetos;
  χ^2_1: variable chi-cuadrado con un grado de libertad.

Prueba de la probabilidad exacta de Fisher

Se emplea cuando las frecuencias espearadas para las diferentes celdas, al menos una de ellas, es menor de 5.

Puede ser empleado en la comparación de dos proporciones binomiales procedentes de dos muestras independientes en tablas de 2x2 con valores de frecuencias esperadas menores de 5 en alguna celda.

Se puede emplear tanto para pruebas de homogeneidad en dos muestras como de Independencia o asociación de dos variables.

Permite determinar la probabilidad (p-valor) exacta de obtener valores de la celda a, fuera de las a + b observaciones para un número N (a + b + c + d = total de observaciones) de las cuales a + c se consideran con éxito dentro del evento estudiado, al comparar dos proporciones binomiales en tablas de contingencia 2x2 simples, esto se puede realizar empleando la probabilidades de la distribución hipergeométrica y estimar dicho valor p (Rosner, 2000):

- Valor p unilateral:
  - p-valor para la cola derecha se estima como probabilidades (+) a partir de la condición:

$$\text{p-valor} = \begin{cases} P(+) \text{ si } \dfrac{ad}{(b+1)(c+1)} > \dfrac{bc}{(a+1)(d+1)} \\ P(-) \text{ en caso contrario} \end{cases}$$

$$P(+) = \sum_{P(a+k) \le P(a)} P(a+k)$$

cuando se quiere contrastar $p_1 > p_2$.

  - p-valor para la cola izquierda, se estima como probabilidades (–) a partir de la condición:

$$\text{p-valor} = \begin{cases} P(-) \text{ si } \dfrac{ad}{(b+1)(c+1)} < \dfrac{bc}{(a+1)(d+1)} \\ P(-) \text{ en caso contrario} \end{cases}$$

$$P(-) = \sum_{P(a-k) \le P(a)} P(a-k)$$

cuando se quiere contrastar $p_1 < p_2$.

- Valor p bilateral:

p-valor = P (+) + P (–) – P (a)

donde:

$P(+) = \sum_{P(a+k) \le P(a)} P(a+k)$: probabilidades de las tablas estructuradas diferentes a la tabla observada inicial empleando valores mayores o iguales a los de la celda a (de la tabla inicial);

$P(-) = \sum_{P(a-k) \leq P(a)} P(a-k)$: probabilidades de las tablas estructuradas diferentes a la tabla observada inicial empleando valores menores o iguales a los de la celda a (de la tabla inicial);

P (a + k) y P (a – k): se calculan a partir de la función de probabilidad de una distribución hipergeométrica de parámetros N = a + b + c + d, R = a + b y n = a + c.

La explicación del método es como sigue:

- Se tiene una tabla de contingencia 2x2 simple (Tabla 4.33).

Tabla 4.33. De contingencia 2x2 simples que relaciona el hábito de fumar y la presencia de enfermedad cardiovascular, sin valores asignados en las celdas

Hábito de fumar	Presencia de enfermedad cardiovascular	No enfermedad cardiovascular	Totales
Fumador	a	b	a + b
No fumador	c	d	c + d
Totales	**a + c**	**b + d**	**N = a + b + c + d**

La probabilidad de obtener los valores para una tabla 2x2 simple observada se pueden calcular con la expresión matemática denominada probabilidad exacta (Pr) para los valores observados de las celdas a,b,c,d en una tabla 2x2 simple:

$Pr(a, b, c, d) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$, donde el signo ! significa factorial del valor o expresión que le antecede.

- Se asignan valores a las celdas de la tabla anterior (Tabla 4.34).

Tabla 4.34. De contingencia 2x2 simple que relaciona el hábito de fumar y la presencia de enfermedad cardiovascular, con valores asignados en las celdas

Hábito de fumar	Presencia de enfermedad cardiovascular	No enfermedad cardiovascular	Totales marginales
Fumador	a = 2	b = 5	a + b = 7
No fumador	c = 3	d = 1	c + d = 4
Totales	**a + c = 5**	**b + d = 6**	**N = a + b + c + d = 11**

La probabilidad exacta a obtener para los valores de las celdas empleados para esta tabla 2x2 simple es:

$$Pr(2, 5, 3, 1) = \frac{7!4!5!6!}{11!2!5!3!1!} = \frac{5040(24)(120)(720)}{39\ 916\ 800(2)(120)(6)} = \frac{1{,}0450944 \cdot 10^{10}}{5{,}7480192 \cdot 10^{10}} = 0{,}182$$

Los pasos que se pueden seguir para el cálculo de la probabilidad exacta de Fisher (que incluye la sumatoria de las probabilidades para los distintos valores en las celdas para las tablas formadas, manteniendo sus valores totales margi-

nales) en pruebas de hipótesis de Independencia entre dos variables cualitativas dicotómicas sería:

- Teniendo en cuenta que su estrategia básica en el contraste de hipótesis (H_0: $p_1 = p_2$ *vs.* H_1: $p_1 \neq p_2$), siendo p_1 los individuos con efecto presente y factor de riesgo presente entre el total de individuos de la muestra y p_2 los individuos con efecto presente y factor de riesgo ausente entre el total de individuos de la muestra, es enumerar todas las posibles tablas a formar manteniendo constantes los totales marginales en ellas tal y como se encuentran en la tabla observada con los valores iniciales y calcular la probabilidad exacta para cada una de las tablas formadas según la expresión matemática anterior (basada en las probabilidades de la distribución hipergeométrica), cuyo valor esperado E (X) y su varianza [Var (X)] para una variable aleatoria X (para la celda a o $C_{1,1}$), o sea, fila 1, columna 1, se expresa como:

$$E(X) = \frac{M_1 N_1}{N}$$

$$Var(X) = \frac{M_1 M_2 N_1 N_2}{N^2 (N-1)}$$

  - Se reestructuran las filas y columnas de la tabla observada con los valores iniciales de modo que el total marginal inferior por las filas quede en la primera fila y el total marginal por columna inferior en la primera columna.
  - Comenzar asignando valor de 0 a la primara celda ($C_{1,1}$) o celda a, tomando en lo siguiente cada celda restante los valores correspondientes que permitan sin variar los valores totales marginales por filas y columnas que presentó la tabla inicial observada y calcular la probabilidad para estos nuevos valores a partir de la distribución de probabilidades hipergeométrica.
  - En lo siguiente construir nuevas tablas incrementando tabla a tabla los valores de la celda a, obteniéndose tablas hasta que alguno de los valores de las celdas restantes adquiera el valor 0 y entonces calcular igualmente la probabilidad para los valores de cada nueva tabla estructurada.
  - Si se toma como ejemplo la tabla observada inicial con los valores de a, b, c, d con sus respectivos valores totales marginales a + b, c + d, a + c, b + d y el número total de observaciones (N = a + b + c + d) y se denomina tabla (a) como la observada (inicial) y la tabla (k) la última posible obtenida, cuando alguna otra celda toma un valor de 0 debido al aumento de la celda a y que se debe mantener los totales marginales, se tiene que:
    - Para contrastar la hipótesis bilateral a partir de la estimación de la probabilidad exacta de Fisher se tiene:
      H_0: $p_1 = p_2$ *vs.* H_1: $p_1 \neq p_2$
      p-valor Fisher (dos colas) = 2 · mín [Pr (0) + Pr (1) +… + Pr (a) + Pr (a) + Pr (a + 1) + …Pr (k)

donde:

mín: menor valor de las probabilidades;

p-valor: duplicado de la menor área bajo la curva, bien de la cola derecha o de la izquierda.

Otra forma que puede ser aplicada en su estimación es:

$$\text{p-valor Fisher (dos colas)} = \sum_{\{i:\ Pr(i) \leq Pr(a)\}} Pr(i)$$

donde:

Pr (i): probabilidades obtenidas para los valores de cada tabla y se sumarían todas las Pr (i) menores o iguales a la Pr (a) de la tabla observada con los valores iniciales en cada celda.

- Para contrastar la hipótesis unilateral:

  H_0: $p_1 = p_2$ *vs.* $H_{1:}$ $p_1 < p_2$

  p-valor Fisher = Pr (0) + Pr (1) + … + Pr (a), que corresponde al área bajo la curva para la cola izquierda.
- Para contrastar la hipótesis unilateral:

  H_0: $p_1 = p_2$ *vs.* H_1: $p_1 > p_2$

  p-valor Fisher = Pr (a) + Pr (a + 1) + … Pr (k), que corresponde al área bajo la curva para la cola derecha.

  Nota: en este caso el valor de p (p-valor) expresa la probabilidad de obtener una tabla con valores tan extremos o más que los obtenidos con la tabla observada inicial.

Regla de interpretación

– Si p-valor (observado-calculado) exacto de Fisher < nivel de significación (alfa) de acuerdo con el tipo de contraste de hipótesis a realizar (bilateral o unilateral), entonces las probabilidades estimadas exactas de poder cumplirse H_0 con los datos observados o más extremos que los mismos son mímimas por lo que existen evidencias suficientes para el rechazo de H_0, por ende se concluye que existen diferencias significativas en las proporciones formadas a partir de las dos variables cualitativas dicotómicas por lo que tienen asociación estadísticamente significativa ambas variables para el nivel de significación alfa preseleccionado.

– Si p-valor (observado-calculado) exacto de Fisher ≥ nivel de significación (alfa) de acuerdo con el tipo de contraste de hipótesis a realizar (bilateral o unilateral); las probabilidades estimadas exactas de poder cumplirse H_0 con los datos observados o más extremos que los mismos no son mímimas por lo que no existen evidencias suficientes para el rechazo de H_0; por endese concluye que no existen diferencias significativas en las proporciones formadas a partir de las 2 variables cualitativas dicotómicas por lo que tienen independencia ambas variables para el nivel de significación alfa preseleccionado.

Prueba de significación para determinar la asociación de dos variables (factor de riesgo y enfermedad)

Para en estudio de caso-control pareado (1 caso-1 control) empleando la prueba de McNemar (Colectivo de autores, 2006):

- Si mp (1 – p) ≥ 5: el estadístico para contrastar:

  H_0: OR = 1 *vs.* H_1: OR ≠1:

  $$\chi^2_{\text{Mc Nemar}}\ (\text{observado-calculado}) = \chi^2_{\text{McN}} \frac{\left(|b-c|-1\right)^2}{b+c}, \text{sigue una distribución teórica}$$

  χ^2_1 con un grado de libertad.

  - Valor p (p-valor), bajo H_0:

    $$\text{p-valor} = \Pr\left(\chi^2_1 \geq \chi^2\right) .$$

    donde:

    χ^1_2: variable chi-cuadrado con 1 grado de libertad;

    m = b + c: número de pares discordantes.

- Si mp (1– p) < 5), emplear el prueba exacta para p-valor: el valor p exacto para contrastar:

  H_0: OR = 1 *vs.* H_1: OR ≠1:

  $$\text{p-valor} = \begin{cases} \min\left\{2\sum_{k=0}^{b}\binom{m}{k}\left(\frac{1}{2}\right)^m, 1\right\} & \text{si } p \leq \frac{1}{2} \\ \\ \min\left\{2\sum_{k=b}^{m}\binom{m}{k}\left(\frac{1}{2}\right)^m, 1\right\} & \text{si } p > \frac{1}{2} \end{cases} .$$

  Nota: el término min expresa el mínimo valor de la expresión matemática que le continúa.

Prueba para evaluar la significación en la asociación de dos variables (exposición *versus* enfermedad)

En estudio de caso-control pareado de k controles por caso (1< K< 5) o prueba de asociación de Mantel-Haenszel cuando se tiene más de un control por caso (Colectivo de autores, 2006). Si se tienen n (k + 1) tuplas de 1 caso y k controles y se representan tabularmente (Tabla 4.35):

- Estadístico para contrastar:

  H_0: OR = 1 *vs.* H_1: OR ≠1:

  $$\chi^2_{\text{(Mantel-Haenszel)}}(\text{observado-calculado}) = \chi^2_{\text{M-H}} = \frac{\left(|T_1 - T_2| - \frac{1}{2}\right)^2}{T_3}, \text{sigue una distribución}$$

  χ^2_1 con un grado de libertad.

  donde:

k: número de controles por caso;

a_i: número de casos expuestos según el número de control (i);

c_i: número de casos no expuestos según el número de control (i);

$$T_1 = \sum_{i=0}^{k} a_i \ ;$$

$$T_2 = \frac{\sum_{i=0}^{k} i(a_{i-1} + c_i)}{k+1} \ ;$$

$$T_3 = \frac{\sum_{i=0}^{k} i(k+1-i)(a_{i-1} + c_i)}{(k+1)^2}.$$

– Valor p (p-valor), bajo H_0:

$$\text{p-valor} = \Pr\left(\chi_1^2 \geq \chi_{M-H}^2\right) \ .$$

Tabla 4.35. De contingencia para estudio de casos y controles, con k controles expuestos

		Número de controles expuestos			
		0	**1**	**...**	**k**
Casos	Expuestos	a_0	a_1	...	a_k
	No expuestos	c_0	c_1	...	c_k

Para tablas que relacionan a dos muestras pareadas (casos y controles)

Con más de un control por caso para evaluar la significación en la homogeneidad de las proporciones de casos y controles respecto a la categoría presente (de alguna variable dicotómica o factor de riesgo como la exposición) se puede emplear la prueba de Mantel-Haenszel con este objetivo (Tabla 4.35) (Fleiss, 1981):

– El estadístico de contraste (de prueba):

$$\chi^2_{\text{(Mantel-Haenzel)}}(\text{observado-calculado}) \doteq {}^{2}_{\text{M-H}} = \left(\frac{p_2 - p_1}{\text{s.e.}(p_2 - p_1)}\right)^2 = \frac{\left[(m-1)B - mA\right]^2}{mB - \sum_{i=1}^{N} n_i^2},$$

sigue una distribución χ^2_1 con un grado de libertad.

donde:

$A = \sum_{i=1}^{N} X_i'$: número total de sujetos controles que tienen el factor presente o categoría positiva de la variable;

$B = \sum_{i=1}^{N} n_i'$: número total de sujetos (casos y controles) que tiene el factor presente o categoría positiva de la variable;

B – A: número de casos que tiene el factor presente;

$p_1 = \frac{A}{N(m-1)}$: proporción en la cual el factor está presente entre los controles;

$p_2 = \frac{B-A}{N}$: proporción en la cual el factor está presente entre los casos;

m: número de constituyentes de la combinación caso-controles;

M – 1: número de controles por caso a emplear;

N: número total de combinaciones de caso-controles (el número total de observaciones).

- Valor p (p-valor), bajo H_0:

  p-valor = Pr [χ^2_1 (eórico-esperado) ≥ $\chi^2_{M\text{-}H}$ (observado-calculado)].

Tabla 4.36. Representación tabular de los datos al emplear la prueba de Mantel-Haenszel al relacionar dos muestras pareadas, con más de un control por caso

Número de observaciones para combinaciones de caso-controles	Casos con factor 1, sin factor 0	Número de Controles con factor presente (X_i) (desde 0 hasta m – 1)	Total de casos y controles teniendo el factor presente (n_i) (desde 0 hasta m)	$(n_i)^2$
1				
2				
	...	...	...	...
N				
Total por columnas	**∑B – A**	**∑A**	**∑ B**	**∑ $(n_i)^2$**

Para tablas que relacionan a dos muestras pareadas (casos y controles)

Con los datos de una variable cualitativa nominal politómica con ordinalidad para cada muestra, por ejemplo, para tres categorías, evaluando la significación en la homogeneidad de las muestras según la categoría en la variable cualitativa politómica, se emplea la prueba de Stuart-Maxwell por chi-cuadrado (Tabla 4.37) (Fleiss, 1981):

- Estadístico para el contraste (de prueba):

  $\chi^2_{(\text{Stuart-Maxwell})}(\text{observado-calculado}) = \frac{\bar{n}_{23}d_1^2 + \bar{n}_{13}d_2^2 + \bar{n}_{12}d_3^2}{2(\bar{n}_{12}\bar{n}_{13} + \bar{n}_{12}\bar{n}_{23} + \bar{n}_{13}\bar{n}_{23})}$ sigue una distribución χ^2_2 con dos grados de libertad

  donde:

  $\bar{n}_{ij} = \frac{n_{ij} + n_{ii}}{2}$: mitad de la suma de los valores para la misma categoría en ambas muestras representados por las filas i y las columnas j;

dn: diferencia entre el total de cada fila según el orden numérico y el respectivo total de columna del mismo orden numérico.

– Valor p (p-valor), bajo H_0:
p-valor = Pr[χ^2_2 (teórico-esperado) ≥ $\chi^2_{\text{(Stuart-Maxwell)}}$ (observado-calculado)]

Tabla 4.37. Representación tabular de los datos al emplear la prueba de Stuart-Maxwell, al relacionar dos muestras pareadas de caso y control con una variable cualitativa nominal politómica

Categoría de la variable politómica para casos	Categoría de la variable politómica para controles			
	Categoría 1	Categoría 2	Categoría 3	Total
Categoría 1	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$	
Categoría 2	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$	
Categoría 3	$n_{3,1}$	$n_{3,2}$	$n_{3,3}$	
Total				

Para tablas que relacionan más de dos muestras pareadas con los datos de una variable cualitativa nominal dicotómica

Evaluando la significación de la homogeneidad de las muestras según las categorías en la variable cualitativa dicotómica se emplea la prueba de Cochran por chi-cuadrado (Tabla 4.37) (Fleiss, 1981):

– Estadístico para el contraste (de prueba):

$$\text{Q de Cochran (observado-calculado)} = \frac{N^2(m-1)}{m} \cdot \frac{\sum_{j=1}^{m}(p_j - \bar{p})^2}{N\bar{p}(1-\bar{p}) - \sum_{n=1}^{N}(p_n - \bar{p})^2}$$

o también:

$$\text{Q de Cochran} = (m-1) \cdot \frac{m\sum_{j=1}^{m} T_j^2 - T^2}{mT - \sum_{n=1}^{N} S_n^2}$$, sigue una distribución χ^2_{m-1} con m − 1 grados de libertad

donde:

m: total de muestras empleadas;

N: total de observaciones realizadas para todas las muestras;

x: cada X se corresponde a un valor 0 si la respuesta de la variable dicotómica empleada es negativa (fracaso) y valor 1 si la respuesta es positiva (éxito);

S_1: representa el número total de sujetos con respuesta positiva (éxito) para la variable dicotómica en la primera observación de las m muestras;

S_n: representa el número total de sujetos con respuesta positiva (éxito) para la variable dicotómica en la observación n de las m muestras;

T_1: representa el número total de sujetos con respuesta positiva (éxito) para la variable dicotómica de la primera muestra;
T: representa el número total de sujetos con respuesta positiva (éxito) de la variable dicotómica en todas las muestras;

$P_1 = \frac{T_1}{N}$: proporción de sujetos con respuesta positiva en la muestra 1 para el total de observaciones de la muestra 1;

$P_j = \frac{T_j}{N}$;

$P_m = \frac{S_m}{m}$;

$\bar{p} = \frac{1}{m}\sum_{j=1}^{m} p_j = \frac{1}{N}\sum_{n=1}^{N} p_n = \frac{T}{Nm}$.

– Valor p (p-valor), bajo H_0:
p-valor = Pr[χ_{2m-1} (teórico-esperado) ≥ Q de Cochran (observado-calculado)]

Tabla 4.38. Representación tabular de los datos al emplear la prueba de Cochran, al relacionar más de dos muestras pareadas con una variable cualitativa nominal dicotómica

Número de observaciones (n) según la categoría para el total de las muestras relacionadas	**Muestra 1**	**Muestra 2**	**Muestra …**	**Muestra m**	**Total de respuestas positivas**
1	$X_{1,1}$	$X_{1,2}$	….	$X_{1,m}$	S_1
2	$X_{2,1}$	$X_{2,2}$	…..	$X_{2,m}$	S_2
….	….	…..	….	….	….
N =	$X_{N,1}$	$X_{N,2}$	…..	$X_{N,m}$	S_N
Total	$\mathbf{T_1}$	$\mathbf{T_2}$	….	$\mathbf{T_m}$	**T**
Proporción	P_1	P_2	….	P_m	P

En algunos textos como el de Polo Vega (2011) se utiliza una metodología simplificada al referirse a la prueba de Cochran, aunque siguiendo la propuesta realizada por Cochran (1950) y el estadístico de contraste o de prueba empleado como Q de Cochran se expresa como:

$$Q = \frac{J\left[(J-1)\sum T_{+j}^2 - (J-1)T^2\right]}{J\cdot T - \sum T_{i+}^2}$$

donde:

T_{+j}: número de éxitos en cada muestra relacionada (fila) de todos los individuos;

T_{i+}: número de éxitos de cada individuo (columna) en el total de muestras relacionadas;

T: número total de éxitos;

J: número de muestras relacionadas.

Nota: el número de éxitos se refiere a la respuesta positiva o esperada de las posibles de la variable nominal en el experimento.

Existe otra forma de representación tabular de los datos al emplear la prueba de Cochran (Tabla 4.39).

Tabla 4.39. Representación tabular al relacionar más de dos muestras pareadas con una variable cualitativa nominal dicotómica

Muestras relacionadas	**Orden de individuos**					**Número de éxitos en cada muestra relacionada (T_{+j})**
	1	**2**	**3**	**4**	**n**	
1						T_{+1}
2						T_{+2}
3						T_{+3}
...	...	...	...	...	...	...
Total de muestras relacionadas (J)	**Número de éxitos de cada individuo en el total de muestras relacionadas (T_{i+})**					**Número total de éxitos en las J muestras relacionadas (T)**

Para tablas 2x2 estratificadas

Prueba para evaluar la significación de la homogeneidad en el comportamiento de las medidas de asociación del efecto E

En los diferentes estratos o pruebas de homogeneidad de Wald (Colectivo de autores, 2006):

- El estadístico χ_2 para contrastar la homogeneidad de la medida del efecto (E), es decir, la medida de asociación específica entre los diferentes estratos:
  - Opción 1:

    $\chi^2_{(Wald)} = \chi^2_w = \sum_i \frac{(Ln(E_i) - Ln(E))^2}{Var[Ln(E_i)]}$, se utiliza cuando E representa la razón de tasas de incidencia, riesgo relativo o la *odds ratio.*
  - Opción 2:

    $\chi^2_{(Wald)} = \chi^2_w = \sum_i \frac{(E_i - E)^2}{Var(E_i)}$, se utiliza cuando E representa la diferencia de tasas de incidencia o la diferencia de riesgos (y sigue una distribución χ^2_{K-1} con k – 1 grados de ibertad

donde:

K: número de estratos;

E_i: medida del efecto en el estrato i;

E: es la estimación combinada de M-H o la estimación ponderada de la medida del efecto y se estima como:

- Para la opción 1:

$$E = \exp\left\{\frac{\sum_i W_i \text{Ln}(E_i)}{\sum_i W_i}\right\}$$

$$\text{Var}\left[\text{Ln}(E)\right] = \frac{1}{\sum_i W_i}$$

$$W_i = \frac{1}{\text{Var}\left[\text{Ln}(E_i)\right]}$$

- Para la opción 2:

$$E = \frac{\sum_i W_i E_i}{\sum_i W_i}$$

$$\text{Var}(E) = \frac{1}{\sum_i W_i}$$

$$W_i = \frac{1}{\text{Var}(E_i)}$$

χ^2_{k-1}: variable χ^2 con k – 1 grados de libertad.

– Valor p (p-valor), bajo H_0:

$$\text{p-valor} = \Pr\left(\chi^2_{k-1} \geq \chi^2_w\right)$$

Prueba de asociación o independencia entre exposición y enfermedad

Entre estratos o pruebas de asociación de Mantel-Haenszel (Colectivo de autores, 2006), estadístico para contrastar la hipótesis nula de no asociación entre la exposición y la enfermedad según estratos:

– Estadístico de prueba o contraste:

$\chi^2_{\text{(Mantel-Haenszel)}}(\text{observado-calculado}) = \chi^2_{\text{M-H}} = \frac{\left(\left|a - E(a)\right|\right)^2}{V(a)}$, sigue una distribución χ^2_1 con 1 grado de libertad.

– Valor p (p-valor), bajo H_0:

$$\text{p-valor} = \Pr\left(\chi^2_1 \geq \chi^2_{\text{M-H}}\right)$$

donde:

$a = \sum_{i=1}^{k} a_i$: número total de casos expuestos;

χ^2_1: variable χ^2 con 1 grado de libertad.

E (a): enfermos o casos esperados para a;

V (a): variación de a;

Estas dos variables dependen del tipo de estudio y se describen en cada caso a continuación:

- Para estudios transversales de caso-control no apareado y de cohorte por incidencia acumulada:

$$E(a) = \sum_{i=1}^{k} \frac{n_i^1 (a_i + c_i)}{n_i};$$

$$V(a) = \sum_{i=1}^{k} \frac{n_i^1 n_i^0 (a_i + c_i)(b_i + d_i)}{(n_i)^2 (n_i - 1)}$$

donde:

$n_i^1 = a_i + b_i$: número de sujetos expuestos del estrato i;

$n_i^0 = c_i + d_i$: número de sujetos no expuestos del estrato i;

$n_i = n_i^1 + n_i^0$: total de sujetos del estrato i.

- Para estudios de cohorte con datos de tasa de incidencia:

$$E(a) = \sum_{i=1}^{k} \frac{n_i T_i^1}{T_i}$$

$$V(a) = \sum_{i=1}^{k} \frac{n_i T_i^1 T_i^0}{(T_i)^2}$$

Prueba para evaluar la significación de la homogeneidad de las *odds ratio*

Entre los diferentes estratos o pruebas de homogeneidad de Breslow-Day en estudios de casos y controles no pareados (Colectivo de autores, 2006):

– Estadístico de prueba o contraste:

$$\chi^2 \text{ (observado-calculado)} = \sum_i \frac{(a_i - A_i)^2}{V_i} - \frac{\left(\sum_i a_i - \sum_i A_i\right)^2}{\sum_i V_i}$$, sigue una distribución χ^2_{k-1} con k–1 = grados de libertad.

donde:

k: número de estratos;

a_i: número de casos expuestos en el estrato i;

A_i: raíz positiva de la ecuación cuadrática:

$$OR_{M-H} = \frac{A_i(m_i^0 - n_i^1 + A_i)}{(m_i^1 - A_i)(n_i^1 - A_i)}$$

A_i es uno de estos valores:

$$A_i = \frac{-B \pm \sqrt{B^2 + 4(1-OR_{M-H})C}}{2(1-OR_{M-H})}$$

$$B_i = m_i^0 - n_i^1 + OR_{M-H}(m_i^1 + n_i^1)$$

$$C_i = m_i^1 n_i^1 OR_{M-H}$$

$n_i^1 = a_i + b_i$: número de sujetos expuestos del estrato i;
$m_i^1 = a_i + c_i$: número de casos del estrato i;
$m_i^0 = b_i + d_i$: número de controles del estrato i;
OR_{M-H}: es el estimador combinado de Mantel-Haenszel del OR;

$$V_i = \frac{1}{A_i} + \frac{1}{m_i^1 - A_i} + \frac{1}{n_i^1 - A_i} + \frac{1}{m_i^0 - n_i^1 + A_i} \ ;$$

χ^2_{k-1}: variable χ^2 con k – 1 grados de libertad.

– Valor p (p-valor), bajo H_0:

$$\text{p-valor} = Pr(\chi^2_{k-1} \geq \chi^2)$$

Para tablas 2xn simples

Prueba para evaluar la significación de la asociación entre dos variables (exposición y enfermedad)

A lo largo de los diferentes niveles de exposición en estudios de Cohorte con datos de tasas de incidencia (Colectivo de autores, 2006):

– Estadístico de prueba o contraste:

$\chi^2(\text{observado-calculado}) = \sum_{i=1}^{N} \frac{(a_i - E_i)^2}{E_i}$, sigue una distribución χ^2_{N-1} con N – 1 grados de libertad

donde:

N: número de niveles de exposición;

$$E_i = \frac{aT_i}{T};$$

$$a = \sum_{i=1}^{N} a_i \ ;$$

$$T = \sum_{i=1}^{N} T_i \ ;$$

χ^2_{N-1}: variable χ^2 con N – 1 grados de libertad.

– Valor p (p-valor), bajo H_0:

$$\text{p-valor} = \Pr\left(\chi^2_{N-1} \geq \chi^2\right)$$

Prueba para evaluar la significación de la tendencia lineal en el crecimiento de la tasa de incidencia

Entre los diferentes niveles de exposición en estudio de cohorte (datos por tasas de incidencia) o prueba de tendencia lineal (Colectivo de autores, 2006):

– Estadístico de prueba o contraste:

$\chi^2(\text{observado-calculado}) = \frac{U^2}{V}$, sigue una distribución teórica χ^2 con un grado de libertad.

donde:

$$U = V_1 - \frac{a}{T}V_2 ;$$

$$V = \frac{a}{T^2}\left\{TV_3 - (V_2)^2\right\} ;$$

$$V_1 = \sum_{i=1}^{N} a_i x_i ;$$

$$V_2 = \sum_{i=1}^{N} T_i x_i ;$$

$$V_3 = \sum_{i=1}^{N} T_i x_i^2 ;$$

x_i: puntuación asignada a la i-ésima categoría de exposición;

χ^2_1: variable χ^2 con 1 grado de libertad.

– Valor p (p-valor), bajo H_0:

$$\text{p-valor} = \Pr\left(\chi^2_1 \geq \chi^2\right)$$

Prueba para evaluar la significación de la homogeneidad de las *odds ratio*

A lo largo de los diferentes niveles de exposición en estudios transversales de caso-control no apareado o de cohortes con datos de incidencia acumulada (Colectivo de autores, 2006):

– Estadístico de prueba o contraste:

χ^2 (observado-calculado) $= \frac{n(n-1)}{m_1 m_0}\sum_{i=1}^{N}\frac{(a_i - E_i)^2}{n_i}$, sigue una distribución teórica χ^2_{N-1} con N – 1 grados de libertad

donde:

$E_i = \frac{m_1 n_i}{n}$;

χ^2_{N-1}: variable χ^2 con N – 1 grados de libertad.

– Valor p (p-valor), bajo H_0:

$$\text{p-valor} = \Pr\left(\chi^2_{N-1} \geq \chi^2\right)$$

Prueba para evaluar la significación para tendencia lineal en el comportamiento de las *odds ratio*

Entre los diferentes niveles de exposición en estudios transversales, de caso-control no apareado o de cohorte (datos de incidencia acumulada) o llamado pruebaa de tendencia lineal (Colectivo de autores, 2006):

– Estadístico de prueba o contraste:

$\chi^2(\text{observado-calculado}) = \frac{U^2}{V}$, sigue una distribución teórica chi-cuadrado (χ^2_1) con un grado de libertad

donde:

$$U = T_1 - \frac{m_1}{n} T_2;$$

$$V = \frac{m_1 m_0}{n^2 (n-1)} \left\{ n T_3 - (T_2)^2 \right\};$$

$$T_1 = \sum_{i=1}^{N} a_i x_i;$$

$$T_2 = \sum_{i=1}^{N} n_i x_i;$$

$$T_3 = \sum_{i=1}^{N} n_i x_i^2;$$

x_i: puntuación asignada a la i-ésima categoría de exposición;

χ^2_1: variable χ^2 con 1 grado de libertad.

– Valor p (p-valor), bajo H_0:

$$\text{p-valor} = \Pr\left(\chi^2_1 \geq \chi^2\right)$$

Para tablas 2xn estratificadas

Prueba para evaluar la significación de la tendencia lineal en el comportamiento de las *odds ratio*

Entre los diferentes niveles de exposición para cada estrato en el estudio de caso-control no apareado o llamado prueba de tendencia lineal en los estratos (Colectivo de autores, 2006):

– Estadístico de prueba o contraste:

$\chi^2(\text{observado-calculado}) = \frac{U^2}{V}$, sigue una distribución teórica χ^2_1 con 1 grado de libertad

donde:

$U = \sum_{j=1}^{k} U_j$;

$V = \sum_{j=1}^{k} V_j$;

U_j y $V_{j:}$ valores correspondientes al estrato j, calculados como se describe en la prueba de tendencia para tablas sin estratificar;

χ^2_1: variable χ^2 con 1 grado de libertad.

– Valor p (p-valor), bajo H_0:

$$\text{p-valor} = \Pr\left(\chi^2_1 \geq \chi^2\right)$$

Para tablas MxN (M, N ≥ 2)

Prueba de significación para determinar la asociación o independencia de dos variables así como pruebas de homogeneidad

Para dos o más poblaciones estimadas en relación a una variable cualitativa nominal (dicotómica o politómica en estudio cohorte por Incidencia acumulada, estudio tranversal y de caso-control no apareado o llamado prueba de χ^2 de Pearson (Rosner, 2000; Colectivo de autores, 2006):

– Estadístico de prueba o contraste:

$\chi^2(\text{observado-calculado}) = \sum_i \sum_j \frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}}$, sigue una distribución teórica χ^2 con $(M - 1) \cdot (N - 1)$ grados de libertad

donde:

M: número de filas (i);

N: número de columnas (j).

– Valor p (p-valor), bajo H_0:

$$\text{p-valor} = \Pr\left(\chi^2_g \geq \chi^2\right)$$

Esta prueba solo es aplicable si se satisfacen dos condiciones:

– No hay más de un 20 % de celdas con frecuencia esperada menor que 5.

– Ninguna celda tiene frecuencia esperada menor que 1.

En el caso de tablas MxN de 2x2 se debe utilizar las fórmulas de tablas 2x2 simples para calcular el estadístico χ^2 con y sin corrección por continuidad y para la prueba exacta de Fisher.

Usos de los estadísticos de prueba en contrastes de hipótesis mediante pruebas paramétricas

Estadístico de prueba Z ~ N (0, 1) (distribución normal estándar)

- Se emplea para evaluar la significación estadística del contraste de hipótesis para la diferencia de dos medias, una muestral que estima una poblacional versus la media de una población de referencia de la que se conoce su desviación estándar poblacional empleando una muestra grande (n ≥ 30).
- Se emplea para evaluar la significación estadística en el contraste de hipótesis para la diferencia de dos medias, correspondientes a dos poblacionales independientes estimadas a partir de muestras no apareadas grandes (n ≥ 30) con una varianza poblacional estimada conocida y sin diferencias significativas de las mismas para las dos poblaciones estimadas.
- Se emplea para evaluar la significación estadística en el contraste de hipótesis para la diferencia de dos proporciones, una muestral que estima la poblacional versus una poblacional de referencia a partir de una muestra grande (npq > 5).
- Se emplea para evaluar la significación estadística en el contraste de hipótesis para la diferencia de dos proporciones, correspondientes a dos poblaciones estimadas a partir de muestras no apareadas grandes (npq > 5).
- Se emplea para evaluar la significación estadística en el contraste de hipótesis para la comparación de dos tasas de incidencia poblacionales estimadas de dos muestras grandes dada la posibilidad de aproximación de la normal a las distribuciones de frecuencias de variables cualitativas expresadas en forma de evento por persona-tiempo de seguimiento.
- Se emplea para evaluar la significación estadística en el contraste de hipótesis para la concordancia entre dos pruebas diagnósticas o dos variables cualitativas dicotómicas a partir de los valores del índice kappa.

Estadístico de prueba t-student (t)

- Se emplea para evaluar la significación estadística en el contraste de hipótesis para la diferencia de dos medias, una muestral que estima una poblacional versus una

media poblacional de referencia a partir de una muestra pequeña (n < 30) y si no se conoce la desviación estándar de la población de referencia.
- Se emplea para evaluar la significación estadística en el contraste de hipótesis para la diferencia de dos medias, correspondientes a dos poblacionales independientes estimadas a partir de muestras pequeñas (n < 30) con o sin desviaciones estándar iguales.
- Se emplea para evaluar la significación estadística en el contraste de hipótesis para la diferencia de dos medias relacionadas o apareadas, es decir, las correspondientes a una misma población con empleo de datos obtenidos en dos momentos diferentes, estimadas a partir de muestras preferentemente grandes.
- Se emplea para evaluar la significación estadística en el contraste de hipótesis para la correlación lineal simple entre dos variables cuantitativas que tienen distribución con normalidad o la aproximación normal.
- Se emplea para evaluar la significación estadística en el contraste de hipótesis para la correlación por rangos ordenados entre dos variables cuantitativas de libre distribución.
- Se emplea para evaluar la significación estadística en el contraste de hipótesis para el ajuste de linealidad del coeficiente de regresión lineal.

Estadístico de prueba chi-cuadrado (χ^2)

- Se emplea para evaluar la significación estadística en el contraste de hipótesis para la diferencia entre una varianza poblacional estimada de una muestral versus una varianza poblacional de referencia o hipotética.
- Se emplea para evaluar la significación estadística en el contraste de hipótesis para la independencia entre dos variables cualitativas, con dos o más nominaciones o categorías en cada variable, en una muestra, utilizando tablas de contingencia o cruzadas para su análisis.
- Se emplea para evaluar la significación estadística en el contraste de hipótesis para la homogeneidad entre dos o más muestras de una variable cualitativa (dicotómica o politómica), utilizando tablas de contingencia o cruzadas para su análisis.
- Se emplea para evaluar la significación estadística en el contraste de hipótesis para la independencia entre dos variables cualitativas con la presencia de estratificación empleando el procedimiento de Mantell-Haenszel y la representación de los datos mediante las tablas de contingencia o cruzadas.
- Se emplea para evaluar la significación en el contraste de hipótesis para la homogeneidad de los diferentes estratos en el análisis de asociación versus independencia de dos variables cualitativas dicotómicas en el enfoque de riesgo para los estudios epidemiológicos observacionales.
- Se emplea para evaluar la significación en el contraste de hipótesis para la tendencia lineal en los diferentes niveles de exposición al relacionar dos variables cualitati-

vas: una dicotómica (dependiente) y otra ordinal (independiente, expresada por por niveles de exposición) en el enfoque de riesgo para los estudios epidemiológicos observacionales.
- Se emplea para evaluar la significación estadística en el contraste de hipótesis para el cociente compuesto por el valor del coeficiente de regresión (β) y su error estándar (prueba de Wald) para cada variable independiente explicativa en el análisis de regresión logística y en la regresión de Cox.
- Se emplea para evaluar la significación estadística en el contraste de hipótesis para la bondad de ajuste del modelo en el análisis de la regresión logística mediante su expresión matemática por el método de Hosmer-Lemechov.

Estadístico de prueba F de Snedecor-Fisher

- Se emplea para evaluar la significación estadística en el contraste de hipótesis para el ajuste de linealidad del coeficiente de regresión lineal simple en muestras grandes ($n \geq 30$).
- Se emplea para evaluar la significación estadística en el contraste de hipótesis para la diferencia entre dos varianzas poblacionales estimadas de dos muestrales independientes.
- Se emplea para evaluar la significación estadística en el contraste de hipótesis para la diferencia entre más de dos varianzas poblacionales estimadas de dos muestrales independientes por la prueba de Levene.
- Se emplea para evaluar la significación estadística en el contraste de hipótesis para la diferencia de más de dos medias poblaciones estimadas con mediciones tomadas de forma repetidas o no para la variable cuantitativa dependiente a partir de más de dos categorías de la variable independiente, mediante el análisis de varianzas (ANOVA de una vía o un factor).
- Se emplea para evaluar la significación estadística en el contraste de hipótesis para las cargas discriminantes relacionadas con los coeficientes discriminantes en el modelo de la función discriminante mediante el método multivariado o de análisis discriminante.
- Se emplea para evaluar la significación estadística en el contraste de hipótesis en relación a que la cantidad de variacion explicada por el modelo de regresión múltiple es más que la variacion explicada por la media, es decir, que el coeficiente de determinación R^2 es mayor que cero como parte del análisis del ajuste del modelo.

Para recordar

En la tabla 4.38 se muestra un resumen que relaciona a los estadígrafos o estadísticos de prueba más frecuentemente empleados en los contrastes de hipótesis de la inferencia bioestadística con el número de muestras y el de los grados de libertad a emplear.

Tabla 4.40. Resumen de algunos tipos de estadísticos de prueba o contraste que requieren tomar en cuenta los grados de libertad según el tipo de contraste de hipótesis a realizar

Tipo de estadístico de prueba o contraste	Número de muestras	Grados de libertad para determinar el valor crítico téórico esperado
t-student (diferencia de medias)	Una.	GI = n – 1 (n es el tamaño de la muestra).
t-student (diferencia de medias)	Dos no pareadas, con σ desconocidas, pero supuestamente iguales.	GI = $n_1 + n_2 - 2$ (n_1 y n_2 son los tamaños de las muestras 1 y 2)
t-student (diferencia de medias)	Dos no pareadas, con σ desconocidas, pero no supuestamente iguales	El menor valor de las restas: $n_1 - 1$ o $n_2 - 1$
t-student (diferencia de medias)	Dos pareadas	GI = n – 1 (n es el número de pares de valores de la variable)
t-student (correlación lineal simple)	Una	GI = n – 2 (n es el número de pares de valores de las dos variables)
t-student (correlación por rangos)	Una	GI = n – 2 (n es el número de pares de valores de las dos variables)
t-student (regresión lineal simple)	Una	GI = n – 2 (n es el número de pares de valores de las dos variables)
Chi-cuadrado (χ^2) (diferencia de varianza)	Una	GI = n – 1 (n es el es el tamaño de la muestra)
Chi-cuadrado de Pearson (independencia *versus* asociación de dos variables dicotómicas)	Una	Por tablas 2x2 simple G l = (f – 1) (c – 1) = 1
Chi-cuadrado de Pearson (independencia *versus* asociación de dos variables cualitativas con una politómica)	Una	Por tablas 2x > 2 simple GI = (f – 1) (c – 1)
Chi-cuadrado de Pearson (homogeneidad muestral) con una variable nominal	Dos	Por tablas 2x > 2 simple GI = (f – 1) (c – 1) o GI = k – 1 (k es el número de grupos de datos o muestras)

Tabla 4.40. Continuación...

Tipo de estadístico de prueba o contraste	Número de muestras	Grados de libertad para determinar el valor crítico téórico esperado
Chi-cuadrado de Mantell-Haenzsel (independencia *versus* asociación de dos variables dicotómicas con estratificación por variable dicotómica)	Una	GI = k – 1 (k es el número de estratos)
Chi-cuadrado de McNemar (independencia *versus* asociación de dos variables cualitativas dicotómicas)	Una	Por tablas 2x2 simple GI = (f – 1) (c – 1) = 1
Chi-cuadrado de McNemar (homogeneidad muestral)	Dos	Por tablas 2x2 simple GI = (f – 1) (c – 1) = 1
Chi-cuadrado (asociación *versus* independencia de dos variables entre los N niveles)	Una	Por tablas 2xN simple, GI = N (n es el número de niveles de la variable independiente)
Chi-cuadrado (tendencia lineal entre los N niveles)	Una	GI = 1
chi-cuadrado de Wald (diferencia del valor observado y el esperado para el coeficiente de regresión para cada variable predictora en los estudios multivariados)	Una	GI = k – 1 (k es el número de categorías de la variable predictora)
Chi-cuadrado de bondad de ajuste del modelo de regresión logística (Hosmer Lemechow)	Una	GI = grupos formados – 2
F-Snedecor (diferencia de dos varianzas)	Una	GI numerador = $n_1 - 1$ GI denominador = $n_2 - 1$
F-Snedecor (diferencia de más de dos medias igual a ANOVA de un factor, una vía o una entrada)	Una	GI numerador (entre grupos) = t – 1 o k – 1 (t o k so los números de grupos de datos a emplear) GI denominador (intragrupos) = N – t o N – k (N es la sumatoria de los tamaños de todas las muestras n o grupos de datos)
F de Fisher-Snedecor (diferencia de más de dos varianzas igual a ANOVA por prueba de Levene).¿	Una	GI numerador (entre grupos) = k – 1 (k es el número de grupos a comparar) GI denominador (dentro de grupos) = $\sum_{j=1}^{k}(n_j - 1)^2$

Tipo de estadístico de prueba o contraste	Número de muestras	Grados de libertad para determinar el valor crítico téórico esperado
F de Fisher-Snedecor (ajuste del modelo o ecuación de regresión lineal simple)	Una	GI numerador (variación explicada por la regresión) = 1 GI numerador (no explicada o residual) = n – 2 (n es el número de pares x, y de observaciones)

Opciones de técnicas bioestadísticas a realizar en el contraste de hipótesis estadísticas

Por el enfoque o paradigma frecuentista

Para una variable en una muestra (univariado)

- Descripción y representación gráfica o tabulada de los estimadores muestrales para cada tipo de variable.
- Estimación del parámetro poblacional a partir del estimador muestral respectivo por intervalos de confianza para variables cuantitativas, cualitativas y las expresadas como evento por personas-tiempo (tasa de incidencia).
- Comparación entre un parámetro poblacional estimado a partir del muestral (media, proporción o desviación estándar) versus un respectivo parámetro poblacional tomado de referencia para un nivel de confianza seleccionado.

Para dos variables en una muestra (bivariada)

- Descripción y representación gráfica o tabulada de los estimadores muestrales respectivos en las dos variables.
- Estimación por intervalos de confianza de:
  - La diferencia de dos medias poblacionales estimadas a patir de los datos de la variable cuantitativa continua que se corresponden a los dos grupos formados a partir de la variable cualitativa dicotómica.
  - La diferencia o razón de dos varianzas poblacionales estimadas con empleo de una variable independiente: cualitativa dicotómica y una dependiente: cuantitativa continua.
  - La diferencia de dos proporciones poblacionales estimadas a partir de las muestrales, que expresan un evento de la variable dependiente para dos categorías de

la independiente o para una categoría de la independiente y dos tipos de eventos en la variable dependiente, empleando dos variables cualitativas dicotómicas.
    - El coeficiente de regresión lineal simple para la pendiente de la recta estimada a partir de la muestral o del coeficiente de regresión lineal simple poblacional estimado a partir del muestral, al analizar la asociación de dos variables cuantitativas continuas.
    - El coeficiente de correlación lineal simple estimado a partir del muestral, al analizar la asociación de dos variables cuantitativas continuas para un determinado nivel de confianza.
- El contraste de hipótesis para evaluar la significación de:
    - La diferencia de dos medias poblacionales estimadas a partir de los datos de la variable cuantitativa continua que se corresponden a los dos grupos formados a partir de la variable cualitativa dicotómica para un determinado nivel de confianza.
    - La diferencia de más de dos medias poblacionales estimadas a partir de los datos de la variable cuantitativa continua que se corresponden a los más de dos grupos formados a partir de la variable cualitativa politómica u ordinal (se comporta como una nominal en este caso) para un determinado nivel de confianza.
    - La diferencia o razón de dos varianzas poblacionales estimadas con empleo de una variable independiente: cualitativa dicotómica y una dependiente: cuantitativa continua, para un determinado nivel de confianza.
    - La diferencia de dos proporciones poblacionales estimadas a partir de las muestrales, que expresan un evento de la variable dependiente para dos categorías de la independiente o para una categoría de la independiente y dos tipos de eventos en la variable dependiente, empleando dos variables cualitativas dicotómicas para un determinado nivel de confianza.
    - El coeficiente de regresión lineal simple para la pendiente de la recta estimada a partir de la muestral o del coeficiente de regresión lineal simple poblacional estimado a partir del muestral, al analizar la asociación de dos variables cuantitativas continuas para un determinado nivel de confianza.
    - El coeficiente de correlación lineal simple estimado a partir del muestral, al analizar la asociación de dos variables cuantitativas continuas para un determinado nivel de confianza.
    - La evaluación de la significación en la asociación *versus* independencia entre dos variables cualitativas dicotómicas o con alguna variable politómica para estudios no estratificados y estratificados con un determinado nivel de confianza.
    - La evaluación de la significación en la tendencia lineal entre los diferentes niveles de exposición en la variable independiente para el evento en la variable dependiente dicotómica, según el nivel de confianza seleccionado.

– La relación de asociación entre dos variables cualitativas nominales (una independiente y otra dependiente) por medio de:
  • Razón de riesgos o riesgo relativo (crudo, estratificado, ponderado y combinado) a partir de dos variables cualitativas dicotómicas.
  • Diferencia de riesgos o riesgo atribuible absoluto (crudo, estratificado, ponderado, combinado) a partir de dos variables cualitativas dicotómicas.
  • Riesgo atribuible proporcional (en expuestos y poblacional) a partir de dos variables cualitativas dicotómicas.
  • Razón de momios, chances, oportunidades u *odds ratio* (crudo, estratificado, ponderado y combinado) a partir de dos variables cualitativas dicotómicas.

Para más de dos variables en una muestra (multivariada)

– Descripción y representación gráfica o tabulada de los estimadores muestrales en más de dos variables.
– Determinar las relaciones de: dependencia e interdependencias entre las más de dos variables por medio de:
  • Realizar la reducción de variables en factores o componentes a partir de las interdependencias entre las mismas por el análisis de componentes factoriales o componentes principales.
  • Realizar el procesamiento estratificado de las matrices de datos por técnica del análisis estratificado.
  • Formación de grupos o racimos de variables a partir de las interdependencias de las mismas por técnica del Análisis de conglomerados (clúster análisis).
  • Realizar el análisis de patrones de interactividad de variables independientes y dependientes por técnica del *path* de análisis o comportamiento de patrones.
  • Formación de grupos excluyentes en la variable dependiente cualitativa (a partir de la estimación de una función discriminante) según el comportamiento de las variables independientes cuantitativas por medio de la técnica del análisis discriminante.
  • Correlación lineal múltiple para determinar el grado, intensidad o magnitud de la asociación de todas las variables independientes cuantitativas en conjunto con la dependiente cuantitativa.
  • Regresión lineal múltiple para modelar la probabilidad de predecir el valor de la variable dependiente cuantitativa a partir de los valores de las independientes cuantitativas al unísono.
  • Regresión logística para modelar la predicción en cuanto a la ocurrencia de un evento de la variable dependiente cualitativa a partir de una razón de chances para la variable dependiente en relación a las categorías de las variables independientes cualitativas o cuantitativas.

- Regresión de Cox para modelar la probabilidad de predecir la ocurrencia de un evento expresado para un lapso de tiempo determinado, a partir de una razón de tasas instantáneas entre las categorías de las variables predictivas independientes (factores de riesgo) para el evento en la variable dependiente.
- Análisis de supervivencia o sobrevida, a partir de la influencia de un grupo de variables independientes por medio del empleo de tablas de vida y el rango logarítmico (*log-rank*).
- Anális de la varianza (ANOVA) para determinar la diferencia entre las medias de los grupos formados por la variable independiente cualitativa politómica empleando los datos de la variable dependiente cuantitativa, en caso de incluir la participación de covariables cuantitativas entre las independientes, realizar el análisis de covarianzas (ANCOVA).
- Análisis múltiple de varianzas (MANOVA) para determinar las diferencias entre los grupos formados por las variables independientes cualitativas categóricas empleando los datos de las más de dos variables dependientes cuantitativas combinadas linealmente, en caso de incluir la participación de covariables cuantitativas entre las independientes realizar el análisis múltiple de covarianzas (MANCOVA).

Por el enfoque o paradigma bayesiano

Se realiza la descripción y representación gráfica tabulada de los estimadores muestrales en una muestra, la estimación por intervalos de probabilidad de las probabilidades a *posteriori* del valor del parámetro estimado o para la diferencia o razón de los mismos (véase el capítulo 5).

Métodos bivariados de análisis de regresión lineal simple y de la correlación lineal simple

Los conceptos de regresión y correlación fueron expuestos por primera vez por el científico inglés Sir Francis Galton en sus informes de investigación sobre la herencia, primero en chicharos y posteriormente sobre la estatura humana entre 1888 y 1889, quien afirmó que la estatura de un individuo adulto, sin importar la estatura de sus padres, tiende a revertirse hacia la estatura promedio de la población. Con el paso del tiempo su utilización de la palabra reversión fue sustituída por la palabra regresión y este parece ser el origen del término.

El análisis de regresión se emplea cuando el investigador sabe o sospecha que existe una relación entre las variables, porque hay alguna teoría o investigación previas que la

han descubierto. Tal es el caso que el investigador desea comprobar experimentalmente esta relación y el objeto de análisis de regresión es encontrar la ecuación que relacione las variables y que mejor se ajuste a los datos experimentales. En general permite encontrar la forma más probable de la relación entre las variables y su objetivo final es predecir o estimar el valor de una variable, que se corresponde a un valor dado de una o más variables.

El análisis de correlación se emplea cuando el investigador sospecha que ambas magnitudes (variables) están relacionadas, pero no tiene idea de una ecuación que las combine. En general permite medir la intensidad de la relación que se establece entre las variables y su objetivo final es determinar el grado y la manera en que dos o más variables cambian simultáneamente.

Cuando ambas técnicas estadísticas se combinan el análisis de la regresión permite encontrar la ecuación matemática que mejor describe la relación entre las variables en estudio, mientras que el de correlación permite determinar "el grado de ajuste" de la ecuación matemática encontrada a los datos experimentales, lográndose los resultados más óptimos para variables cuantitativas continuas.

Una buena correlación entre dos variables cuantitativas no necesariamente significa una relación causa-efecto entre las mismas, este último aspecto depende más bien del procedimiento del método científico y es competencia del investigador su decisión.

Para realizar estas técnicas se precisa de emplear dos tipos de variables:

- Independiente, explicativa, predictora: es la que se ubica en el eje X del diagrama de dispersión, es la que está bajo el control del investigador, por lo que sus valores pueden ser seleccionados por este durante el experimento.
- Dependiente, criterio, predicha, explicada: es la que se ubica en el eje Y del diagrama de dispersión, es la que escapa del control del investigador y sus valores pueden depender de muchas causas, como las características individuales de los sujetos experimentales, el desarrollo del proceso experimental, el azar entre otros.

En la figura 4.11 se muestran algunas "nubes" de puntos (pares x, y) experimentales y los modelos de regresión.

Como se puede observar el modelo de regresión puede ser lineal y no lineal, de acuerdo con el tipo de función matemática que se utilice para describir la relación entre las variables. En caso del modelo de regresión en que se considere solo una variable independiente, ejerciendo influencia sobre los valores de la variable dependiente, el mismo se denomina *simple*, mientras que si dos o más variables independientes son empleadas en el mismo, el análisis de regresión es múltiple (Fig. 4.12).

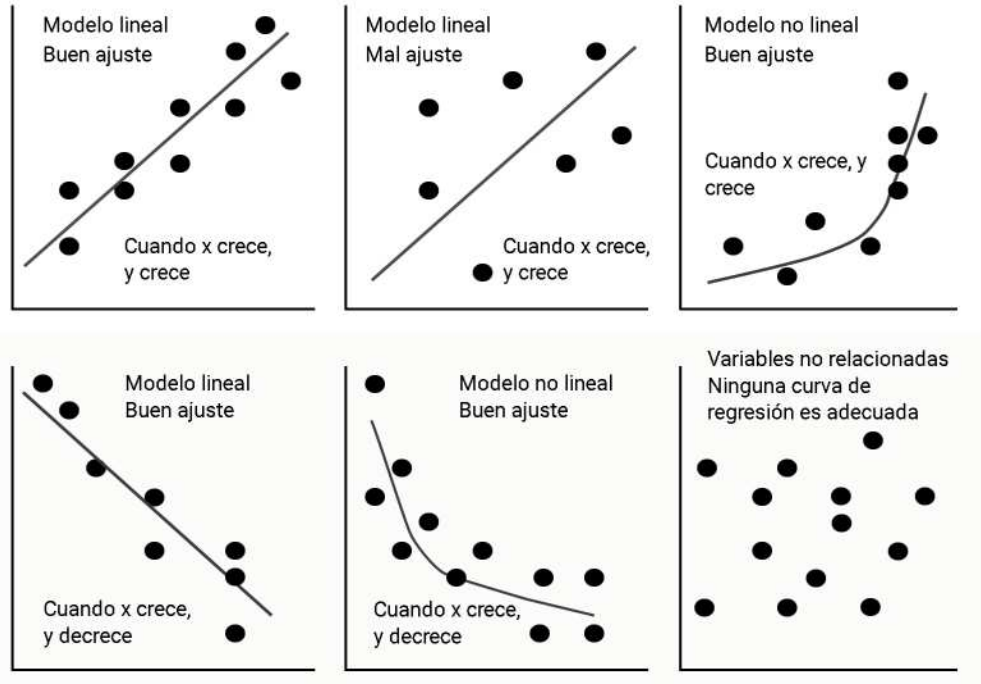


Fig. 4.11. Modelos de regresión más empleados.

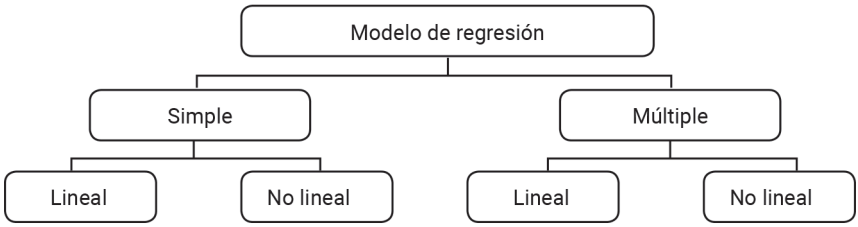


Fig. 4.12. Clasificación de los modelos de regresión.

El modelo de regresión lineal simple es el más sencillo y uno de los de mayor utilidad y empleo práctico. Los supuestos que requiere para su empleo son:

- Los valores de la variable X son fijados o seleccionados previamente por el investigador.
- La variable X se mide sin error (aunque se sabe que ningún procedimiento de medición es perfecto, se considera insignificante el error en la medición de X).
- Para cada valor de X existe una subpoblación o conjunto de valores de Y (estas subpoblaciones de Y siguen una distribución normal con iguales varianzas).
- Todas las medias de subpoblaciones de valores de Y se encuentran sobre la misma línea recta. A esto se le conoce como suposición de linealidad.
- Los valores de Y son independientes entre sí.

Todas estas suposiciones se resumen mediante la ecuación:

$Y = \alpha + \beta X + e$

donde:

α y β: coeficientes de regresión de la población, identificados geométricamente como el intercepto y la pendiente de la recta respectivamente en la que se suponen están las medias de las subpoblaciones de Y;

e: error total e indica cuanto se desvía el valor de Y observado de la media de la subpoblación de la cual se extrajo.

Estos son los pasos a seguir para realizar un análisis de regresión lineal simple y los aspectos a tener en cuenta:

- Se lleva a cabo cuando se sospeche o se espere la existencia de una relación lineal entre dos variables de interés, cuya evidencia inicial la puede aportar la realización de un diagrama de dispersión (para ambas variables: X, Y).
- En caso de existir evidencias de una relación lineal entre ambas variables cuantitativas a partir de la evidencia aportada por el gráfico de dispersión antes mencionado, se procede a calcular la ecuación de regresión lineal simple para los pares de valores de las variables (X, Y) de la muestra, la que representa el gráfico de una recta siguiendo la expresión matemática: y = a + bx, en la cual para obtener los valores de ay b, que equivalen a decir en la notación que se emplea a continuación a b_0 y b_1, respectivamente, empleando el procedimiento denominado método de mínimos cuadrados. Si se expresa esta recta obtenida por este método como: $\hat{y} = b_0 + b_1 x$, que constituye un estimado de la recta de regresión lineal poblacional $y = \beta_0 + \beta_1 x$. Los valores de b_0 y b_1 se calculan por medio de las expresiones:

$$b_0 = \overline{y} - b_1 \overline{x} \quad \text{o}$$

$$b_0 = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

$$b_1 = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} \quad \text{o}$$

$$b_1 = \frac{SP_{xy}}{SC_x} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

utilizando los valores calculados de la media de los valores de ambas variables muestrales X y Y

donde:

SP: expresa la suma de productos, en este caso para las diferencias de los valores observados de X y Y respecto a los de sus medias respectivas;

SC: expresa la suma de cuadrados, en este caso para las diferencias de los valores observados de X respecto al de su media;
n: número de observaciones recogidas en la muestra.
Cuando se conoce el valor del coeficiente de correlación lineal simple de Pearson muestral r que estima el poblacional, dada su semejanza en la formulación matemática respecto a la pendiente o valor b de la recta de regresión estimada se pueden relacionar mediante la expresión en forma de ecuación:

$$b_1 = r\frac{Sy}{Sx} \ ;$$

así como de forma inversa:

$$r = b_1\frac{Sx}{Sy} \ ;$$

donde:

$Sx = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n-1}}$: expresa la desviación estándar de los valores observados de la variable predictora X;

$Sy = \sqrt{\frac{\sum(y_i - \overline{y})^2}{n-1}}$: expresa la desviación estándar de los valores observados de la variable de respuesta Y;
n: número total de observaciones.

– Calcular cada valor predicho $\hat{y}i$ de la recta de regresión muestral (recta de mínimos cuadrados), por medio de la sustitución en la misma de cada valor observado de la variable independiente X, según el número de observaciones realizadas (n).
Al obtener cada valor predicho de la variable Y por la ecuación de la recta de regresión o modelo estimado con base en algún valor dado respectivo de la variable X, se debe tomar en cuenta que:
  - Si no existe una correlación lineal entre los valores de X observada y Y predicha, el mejor valor predicho de la variable Y es $\overline{y}$ y no se debe utilizar la ecuación de regresión para hacer predicciones, en cambio si existe una correlación lineal entre los mismos, el mejor valor predicho de la variable Y se calcula sustituyendo el valor de X en la ecuación de regresión $\hat{y}$.
  - Cuando se utilice la ecuación de regresión para hacer predicciones, se debe permanecer en el ámbito de los datos muestrales disponibles, no emplear datos antiguos dado que esta ecuación de regresión en este caso no necesariamente es válida siempre y no se deben hacer predicciones acerca de una población distinta de la población de donde se obtuvieron los datos muestrales.
– El siguiente aspecto a tomar en cuenta es confeccionar una tabla que relacione estos valores calculados (Tabla 4.41).

Tabla 4.41. Valores calculados relacionados con la recta de regresión lineal simple para las variables X y Y de la muestra de estudio

Número de observaciones por pares de (X, Y)	Valor de X	Valor de Y	Valor de la media de X ($\bar{x}$)	Valor de la media de Y ($\bar{y}$)	Valor de (b_0)	Valor de (b_1)	Valor de ($\hat{y}$)	Valor del error no explicado por la regresión o residual ($y - \hat{y}$)
1								
2								
...								
n								

– Como el objetivo fundamental del método de mínimos cuadrados consiste en minimizar la suma de los cuadrados de las diferencias entre cada valor experimental (muestral) observado y_i y el valor teórico predicho por la ecuación de la recta de regresión lineal simple correspondiente ($\hat{y}_i$), denominada suma de cuadrados del error no explicado (SCNE), entonces se deriva la ecuación de esta última respecto a cada uno de los coeficientes de regresión de la muestra: a o b_0 y b o b_1 (según se elija la nomenclatura) y se igualan a cero las ecuaciones resultantes (condición de mínimo de la primera derivada). Las ecuaciones resultantes por un procedimiento matemático, que no es objetivo de este libro, se llega a las ecuaciones resultantes que permiten calcular los valores de a y b correspondientes a la recta de "mejor ajuste" para la cual es mínima la sumatoria de cuadrados del error no explicado por la regresión (SCNE) o residual, esta última definida como:

$$SC_{NE} = \sum e_i^2 NE = \sum (y_i - \hat{y}_i)^2$$

La expresión matemática para el cálculo de la pendiente (b) de la recta de mejor ajuste de regresión lineal simple sería:

$$b = \frac{\sum_{i=1}^{n} x_i \cdot y_i - \frac{\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

y para el intercepto (a) de la recta de regresión lineal simple sería:

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}$$

La esencia de este método es presentada de forma gráfica (Fig. 4.13.

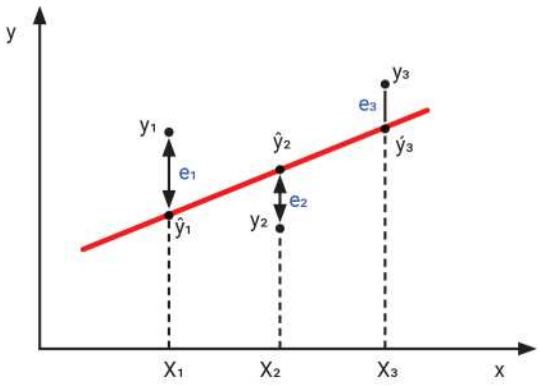


Fig. 4.13. Error no explicado por el modelo de regresión lineal simple o de la ecuación de regresión estimada.

Como puede observarse gráficamente para cada punto experimental de ambas variables (X, Y), puede definirse un "error no explicado" por el modelo teórico o ecuación de la recta denominado e_i, correspondiente a la diferencia $(y_i - \overline{y}_i)$. La sumatoria de los cuadrados de estas diferencias constituye la suma de los cuadrados predicha o del error no explicado por la regresión que constituye una medida de la falta de ajuste del modelo a los datos experimentales, por lo que mientras menor sea sumatoria de cuadrados del error no explicado por la regresión, mejor es el ajuste de la ecuación lineal a los datos experimentales, que es en definitiva lo que se consigue con el método de mínimos cuadrados.

De modo general si una gráfica con los valores de los residuos o residuales $(y_i - \overline{y}_i)$, para cada valor de x_i) no revela ningún patrón, la ecuación de regresión es una buena representación de la asociación entre las dos variables y una recta satisface la propiedad de mínimos cuadrados si la suma de los cuadrados de los residuales (del error no explicado por la regresión) es la menor suma posible.

– Una vez obtenida en el paso anterior la ecuación de regresión es necesario evaluarla para determinar si en realidad describe adecuadamente la relación entre las dos variables en estudio y si puede ser utilizada con fines de predicción y estimación:
  La medida más conocida de la eficiencia de una recta de regresión es el coeficiente de correlación lineal simple de Pearson al cuadrado o coeficiente de determinación (r^2) cuya expresión matemática sería:

$$r^2 = \frac{\sum_{i=1}^{n}(\overline{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{SC_E}{SC_T}$$

  donde:

$SC_E = \sum_{i=1}^{n} e_E^2 = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$: suma de cuadrados explicada por la regresión;

$SC_{NE} = \sum_{i=1}^{n} e_{NE}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$: suma de cuadrados no explicada por la regresión;

$SC_T = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \overline{Y}_i)^2$: suma de cuadrados total, o sea, la explicada y la no explicada por la regresión en conjunto ($SC_T = SC_{NE} + SC_E$);

$e_i = Y_i - \overline{Y}$: error total, es la distancia vertical entre cada punto experimental Y_i y la recta $\overline{Y}$) en síntesis es la suma del error explicado y el no explicado por la regresión para una Y_i en particular ($e_i = e_E + e_{NE}$);

$e_{NE} = Y_i - \hat{Y}_i$: error no explicado por la regresión o el error por falta de adaptación del modelo (es la distancia vertical entre el punto observado Y_i y la recta de regresión $\hat{Y}_i$, siendo representativa de la "porción" de la desviación total que no se ha explicado o tomada en cuenta por la introducción de la recta de regresión;

$e_E = \hat{Y}_i - \overline{Y}$: error explicado por la regresión o el error de adaptación del modelo es la distancia vertical entre la recta de regresión $\hat{Y}_i$ y la recta teórica $\overline{Y}$ y representa en cuanto disminuye la desviación total a medida que la recta de regresión se ajusta a los puntos experimentales.

Se presentan los diferentes tipos de errores relacionados o no con la regresión (Fig. 4.14).

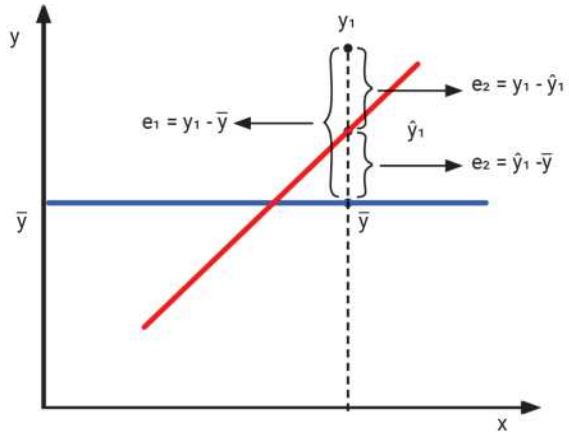


Fig. 4.14. Desviación del punto experimental Y_i respecto a la recta de regresión $\hat{Y}$ y la recta para la media de las Y_i denominada $\overline{Y}$.

De modo que es razonable pensar que si una ecuación de regresión describe correctamente la relación entre dos variables, la suma de cuadrados explicada o de adaptación del modelo debe constituir una gran proporción de la suma total de cuadrados, donde esta proporción es el coeficiente de determinación r^2 de la muestra. Como las sumas

de cuadrado son magnitudes positivas y tenemos que $SC_E \leq SC_T$, siempre se cumple que $0 \leq r^2 \leq 1$ y los valores de r^2 se interpretan como:

- Si $r^2 = 0$, dado que $SC_E = 0$, la regresión lineal no explica la dependencia entre la variable independiente X y la variable dependiente Y.
- Si $r^2 = 1$, dado que $SC_E = SC_T$, el modelo de regresión lineal se ajusta totalmente a los datos. La ecuación refleja perfectamente la relación entre las variables X y Y.
- Si $0 < r^2 < 1$, dado que $SC_E < SC_T$, la regresión explica parte del error, quedan errores no explicados o residuos, lo que es lo más frecuentemente observado en los estudios con esta técnica.

  Mientras más cercano está de 1, significa que hay mejor ajuste entre el modelo lineal y los datos experimentales, esto implica que la recta de regresión describe más eficientemente la relación entre las dos variables en estudio.
- Comprobar la significación de la regresión lineal simple a partir de su modelo o ecuación de regresión: aunque el coeficiente de determinación es el parámetro más utilizado para evaluar la recta de regresión se requieren otros procedimientos o pruebas de significación que permitan satisfacer las necesidades de confirmación del ajuste de la ecuación de regresión a los datos observados o experimentales en el marco de un análisis de regresión lineal simple. A continuación se exponen las más frecuentemente empleadas como parte de este análisis:
  - Prueba de ajuste de normalidad de aproximación a la distribución normal estándar de los datos experimentales de las dos variables cuantitativas, una de las más empleada es la prueba de Kolmogorov-Smirnov en la que se espera el no rechazo de la hipótesis nula (H_0) que expresa la presencia de una distribución con un patrón normal para las variables.
  - Prueba de las rachas para comprobar la alietoriedad de los residuos del error no explicado por la regresión, en la que se espera un valor ni muy bajo ni muy alto del número de rachas, lo que no permite el rechazo de H_0, que expresa la presencia de alietoriedad para los errores o residuos.
  - Prueba de Durbin-Watson para verificar la independencia de los errores o residuos, es decir, que los valores de los diferentes residuos o errores para los valores Y_i no guardan relación entre ellos, en la que se espera la presencia de un valor entre 1,5 y 2,5 de su estadístico correspondiente, para no rechazar la H_0 de independencia de los residuos entre sí.
  - Determinación del error estándar del estimado (S_{YX}) es una medida de las diferencias (o distancias) entre los valores muestrales observados de la variable Y y los valores predichos $\hat{Y}_i$, que se obtienen por medio de la ecuación de regresión, que se expresa como:

    $S_{YX} = \sqrt{\frac{SC_{ne}}{n-2}}$ con n – 2 grados de libertad, con propiedades similares a la varianza, que miden la dispersión de los puntos alrededor de la recta de regresión, donde

se emplea como criterio para su interpretación que mientras menor sea su valor, mejor es el ajuste de la recta a los datos experimentales o de la muestra. También puede ser expresada como:

$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}} \quad \text{o}$$

$$s_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}}$$

- Prueba del ANOVA simple de Fisher es la más importante prueba de significación respecto a las pruebas antes mencionadas. Esta prueba se basa en el cálculo del cuadrado medio de cada uno de los errores definidos antes. El cuadrado medio de cada tipo de error se obtiene como el cociente de la suma de cuadrados del error y los grados de libertad correspondientes como sigue:

$CM_{NE} = S^2_{X,Y} = \frac{SC_{NE}}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$, con grados de libertad $(\theta) = n - 2$

$CM_E = S^2_{\hat{Y}} = \frac{SC_E}{\theta_e} = \sum (\hat{y}_i - \overline{y}_i)^2$, con grados de libertad $(\theta_e) = 1$.

Dado que S_{YX} tiene propiedades de la desviación estándar, los cuadrado medio tienen propiedades de varianzas, de esto se deriva que se puede aplicar la prueba de razón de varianzas (Fisher), como método de comprobación de la diferencia entre los cuadrado medio explicado y no explicado por la regresión. A modo de resumen se presentan los elementos estadísticos involucrados en el ANOVA simple para el análisis de regresión lineal simple (Tabla 4.42).

Tabla 4.42. ANOVA simple empleado en la regresión lineal simple

Fuente de variación	SC	θ	Cm = SC/θ	F
Error explicado (e_E)	$\sum (\hat{y}_i - \overline{y})^2$	1	$\sum (\hat{y}_i - \overline{y})^2$	CM_e
Error no explicado (e_{NE})	$\sum_{i=1}^{n} (y_i - \hat{y})^2$	n – 2	$\frac{\sum (y_i - \hat{y}_i)^2}{n-2}$	CM_{ne}

Para el contraste de hipótesis de diferencias entre ambos cuadrados medios de os tipos de errores explicados y no eplicados por la regresión se emplea el esstadístico de contraste o prueba F de Snedecor-Fisher como:

$F\ (\text{observado-calculado}) = \frac{CM_e}{CM_{ne}}$, sigue una distribución $F_{1,\, n-2,\, 1-\alpha}$.

donde:

CM_e: cuadrado medio explicado por la regresión;

CM_{ne}: cuadrado medio no explicado por la regresión;
θ: grados de libertad en el numerador = 1 y en el denominador = n – 2.
El valor de p (p-valor), bajo H_0:
p-valor = $F_{1, n-2, 1-\alpha}$ (teórico-esperado) ≥ F (observado-calculado).
La regla de decisión respecto al contraste de hipótesis tendría como criterio:

– Si F (observado-calculado) > $F_{1, n-2, 1-\alpha}$ (teórico-esperado) para la tabla de la distribución de probabilidades de F (Snedecor), entonces se ubica el valor observado de F en la región de rechazo de H_0, por lo que existen evidencias para rechazar H_0 y plantear que el cuadrado medio explicado por la regresión (CM_e) es significativamente mayor que el cuadrado medio no explicado por la regresión (CM_{ne}), por lo que el ajuste del modelo de regresión lineal simple a los datos muestrales (experimentales) es suficientemente bueno.

Como último paso para culminar el análisis de regresión lineal simple debemos realizar inferencias acerca de los valores de α (intercepto) y b (pendiente) de la recta de regresión lineal simple muestral que estimarían los coeficientes en la recta de regresión lineal simple poblacional α y β respectivamente. En general las inferencia s respecto al intercepto son de menor valor, por lo que nos centraremos en el valor de la pendiente y analizar la significación estadística del mismo.
El contraste de hipótesis se plantea generalmente como:
H_0: b = 0 *vs.* H_1: b ǂ 0.
El estadístico de contraste o prueba sería:

$$\text{t-student (observado-calculado)} = \frac{b - \beta}{Sb}, \text{ sigue una distribución t-student } _{n-2;\alpha}$$

donde:
b: pendiente de la recta de regresión lineal simple muestral o experimental;
β: valor hipotético o teórico poblacional contra el cual se quiere comparar b, en este caso se toma como 0 esta pendiente;
Sb: error estándar de b.
El valor de p (p-valor), bajo H_0:
p-valor = t-student $_{n-2, \alpha}$ teórico-esperado ≥ t-student (observado-calculado).
La regla de decisión respecto al contraste de hipótesis tendría como criterio:

– Si t-student (observado-calculado) > t-student $_{n-2, \alpha}$ (teórico-esperado) para la tabla de la distribución de probabilidades de t-student, entonces se ubica el valor observado de t-student en la región de rechazo de H_0, esto supone una variación significativa de los valores de la variable Y respecto a los valores de la variable X, por lo que los valores de la variable X tiene mucha utilidad para predecir o estimar los valores de la variable Y, todo esto favorece el criterio de un buen ajuste del modelo de regresión lineal simple a los datos muestrales (experimentales).

La estimación por intervalo de confianza para la pendiente puede ser consultada en el capítulo 3 referido para los intervalos de confianza en las distintas pruebas de hipótesis.

Para el caso del intercepto (a) se procede de manera similar empleando el estadístico de prueba o contraste:

$$\text{t-student(observado-calculado)} = \frac{a - \delta}{Sa}, \text{ sigue una distribución t-student}_{n-2;\,a}$$

donde:

δ: valor hipotético o teórico poblacional contra el cual se quiere comparar a;

Sa: error estándar de a.

El valor de p bajo H_0 y la regla de decisión respecto al contraste de hipótesis es similar al caso anterior de la pendiente.

El intervalo de predicción puede ser estimado en el análisis de regresión lineal simple:

- Para el valor individual de una variable Y:

$$y - E < y < \hat{y} + E$$

donde:

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \dot{x})^2}{n\left(\sum x^2\right) - \left(\sum x\right)^2}} \;;$$

$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}$$

x_0: representa el valor dado de X;

$t_{\alpha/2}$: tiene n – 2 grados de libertad.

El intervalo de confianza para el intercepto y la pendiente (β_0 y β_1) respectivamente, de la recta de regresión poblacional estimada ($y = \beta_0 + \beta_1 x$):

- $b_0 - E < \beta_0 < b_0 + E$

donde:

$$E = t_{\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}}$$

- $b_1 - E < \beta_1 < b_1 + E$

donde:

$$E = t_{\alpha/2} \cdot \frac{s_e}{\sqrt{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}}$$

Correlación entre variables

El modelo de correlación se emplea para caracterizar la relación entre dos variables cuando en un experimento, ambas variables se consideran aleatorias y se obtienen al azar observaciones de las mismas, sin preseleccionar valores de ninguna de estas. Este modelo pone en igualdad de condiciones a ambas variables y generalmente no las distingue refiriéndolas como dependiente y la otra como independiente dado que en este procedimiento puede analizarse la relación tanto de la variable X *versus* Y, como viceversa.

Si se persigue el objetivo fundamental del análisis de correlación se debe obtener una medida de la intensidad, fuerza o grado de relación entre las dos variables. Entre estas medidas se tienen el coeficiente de correlación lineal simple de Pearson, el coeficiente de correlación por rangos de Spearman y el coeficiente de correlación de Kendall, por lo que los dos primeros son los más empleados.

Coeficiente de correlación lineal simple de Pearson

- Se emplea con variables cuantitativas o numéricas que siguen una distribución normal.
- Su expresión matemática es la raíz cuadrada del coeficiente de determinación (r^2) como:

  $r = \sqrt{r^2}$;

  donde:

  $$r^2 = \frac{\sum_{i=1}^{n}(\overline{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{SC_E}{SC_T}$$

- Si se emplea el gráfico de dispersión para dos variables cuantitativas se pueden apreciar distintos valores del r a partir de la "intensidad" de la relación entre las variables X y Y (Fig. 4.15).

Los posibles valores y signo que puede tomar (r) y la forma de interpretar sus resultados serían como:

- El valor de r = ±1 significa que las observaciones de ambas variables están fuertemente correlacionadas, mientras que cuando r está próximo a 0 prácticamente no existe relación lineal entre ambas variables.
- El signo de r es el signo de la pendiente b y el mismo indica si la relación entre las dos variables es directa o inversa.
- Carece de unidades de medidas.

– Es invariable para transformaciones lineales (cambio de origen y escala) de las variables.
– Sus valores oscilan entre –1 y 1.

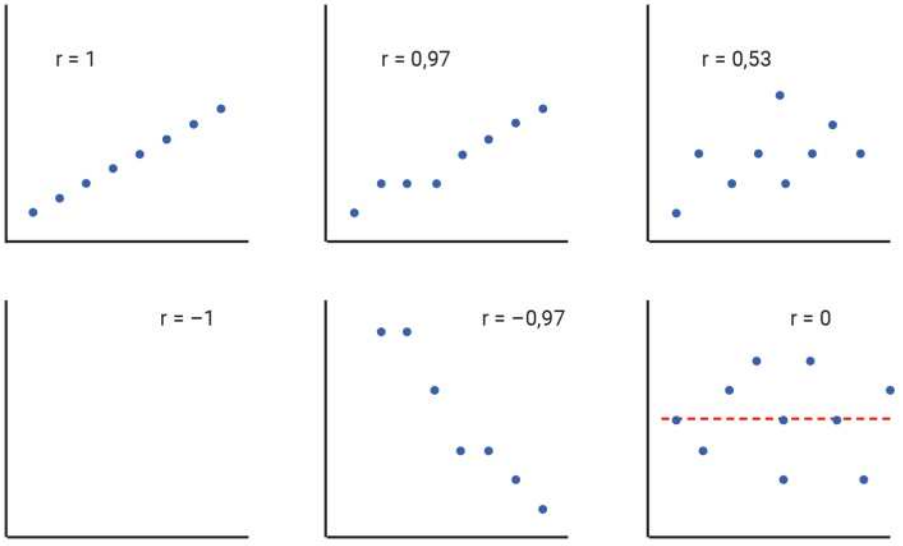


Fig. 4.15. Ejemplos de gráficos de dispersión con diferentes valores del coeficiente de correlación lineal simple de Pearson (r).

Para evaluar la significación del r se emplea un contraste de hipótesis empleando el estadístico de contraste t-student, lo que fue abordado anteriormente en este capítulo, así como la consulta para la estimación por intervalo de confianza para este coeficiente en el capítulo 3.

El coeficiente de correlación lineal simple muestral r se utiliza para calcular la proporción de la variación total en y que está explicada por la relación lineal entre los datos de la variable x y los de y.

Apoyados en toda la información recogida anteriormente en el análisis de regresión lineal simple y de la recta de regresión estimada, el coeficiente de determinación (r^2) es la cantidad de variación en y que está explicada por la recta de regresión, es decir, es la proporción de la variación en y que está explicada por la relación lineal de x, y, lo que se expresa como:

$$r^2 = \frac{\text{variación explicada}}{\text{variación total}}$$

donde:

$SC_E = \sum_{i=1}^{n} (\hat{y} - \overline{y})^2$: variación explicada;

$SC_T = \sum (y - \overline{y})^2$: variación total.

Coeficiente de correlación por rangos o jerarquías de Spearman

Como características fundamentales se puede resumir que:

- Para su utilización se requieren dos conjuntos de jerarquías o rangos que pueden asignarse a los valores de las variables X y Y, que representan a las variables independiente y dependiente respectivamente.
- Los supuestos en que se basa esta prueba son: que los valores de X y Y se hayan medido por pares y que no se requiere comprobar la normalidad en la distribución de los valores de ambas variables, por lo que la prueba de significación a emplear para el mismo deberá ser no paramétrica.
- Los pasos a seguir para su cálculo son:
  - Clasificar por rangos o jerarquías los valores de las variables X y Y desde 1 hasta n (número total de observaciones).
  - Calcular d_i para cada pareja de observaciones, restando el rango de Y_i del de X_i.
  - Efectuar el cálculo de r_S propiamente mediante la siguiente expresión:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

- Para evaluar la significación del r_S, dependiendo el número de observaciones se puede emplear un contraste de hipótesis empleando el estadístico de contraste t-student, lo que fue abordado anteriormente en este capítulo.

Para recordar

- El análisis de correlación y regresión es una herramienta importante del arsenal con que cuenta el investigador para describir y descubrir las relaciones de interdependencias entre las variables que constituyan objeto de estudio en cualquier investigación cuantitativa.
- Una vez establecida la función que describe la relación entre las variables y estando satisfechas las pruebas de significación necesarias, la ecuación de regresión puede ser utilizada para predecir o estimar el comportamiento futuro de las variables dependientes analizadas entre otras aplicaciones.
- Al igual que las diferentes opciones de técnicas de procesamiento estadístico de la información, el éxito de la aplicación del análisis de regresión y correlación, depende sobre todo del investigador, para lo que deben tenerse estas precauciones:
  - Antes de reunir los datos, debe revisar cuidadosamente las suposiciones que fundamentan los análisis de regresión y correlación para poder seleccionar el tipo de análisis más adecuado.
  - Los valores de cada variable deben estar pareados y medidos sobre los mismos individuos o individuos relacionados entre sí.

- Los valores elevados detectados de coeficientes que indican una relación "importante" entre variables no necesariamente entrañan una relación causa-efecto en lo cual toma parte esencial el criterio del investigador y el tipo de diseño investigativo.
- La ecuación de regresión de la muestra no debe emplearse para predecir o estimar fuera del intervalo representado en la muestra de estudio, debido que se puede caer en riesgos de interpolación. Por eso la relación entre dos variables puede ser lineal dentro del intervalo estudiado y así mismo, de otro tipo fuera de este.

Premisas y aspectos necesarios para realizar un análisis de varianza (ANOVA) de una vía o de un factor

Uno de los métodos paramétricos más utilizados para el análisis de los datos derivados de experimentos donde se comparan más de dos grupos o muestras es el análisis de varianzas (ANOVA), desarrollado fundamentalmente por Ronald Fisher en el siglo XX. Según Tukey constituye un instrumento fundamental para adentrarnos en la naturaleza de la variación de los acontecimientos, permite discernir mejor las causas de los fenómenos y los efectos de los factores involucrados al tiempo que puede representar una herramienta básica en el diseño de experimentos. Constituye una técnica bioestadística en la que la varianza total de un conjunto de datos se divide en dos o más componentes y cada uno de ellos se asocia con una fuente específica de variación, así como durante el análisis es posible encontrar la magnitud con la que contribuye cada una de esas fuentes a la variación total.

El ANOVA engloba un grupo de modelos que, independientemente de su grado de complejidad, siguen la misma distribución teórica esperada y se basan en los mismos supuestos. La distribución teórica en que se asienta el ANOVA es la función F de Fisher-Snedecor, donde las pruebas de significación a emplear se basan en comparar el valor muestral empírico observado F (observado-calculado), que es esencialmente una razón de varianzas, contra un valor crítico-esperado F (tabulado) según los grados de libertad respectivos en el numerador y denominador y el nivel de confianza seleccionado por el investigador. Los posibles valores de la distribución de F que pueden corresponderse con áreas o regiones de rechazo de la hipótesis nula (H_0) o no, referida a la igualdad de todas las medias de los grupos comparados (Fig. 4.16).

Como la distribución F es asimétrica, las pruebas de hipótesis basadas en ella son de una sola cola, lado o extremo.

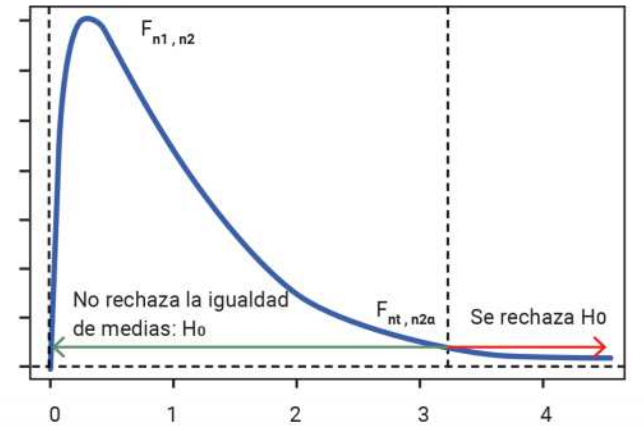


Fig. 4.16. Regiones de aceptación y rechazo de H_0 de la distribución F.

El modelo más elemental del ANOVA es el de un factor o una vía y se define como aquel en que la variable analizada (Y) se considera dependiente de un solo factor, de tal manera que las causas de su variabilidad son englobadas en una componente aleatoria que se denomina *error experimental* Y = Factor ± error. El término factor se acostumbra a sustituir por tratamiento, que también puede referirse genéricamente a elementos a comparar como medicamentos, dosis, métodos de análisis, formulaciones, niveles, horarios, procesos investigativos, entre otros.

Otra forma de refererirnos al método de análisis por el ANOVA consiste en introducir una cantidad μ que sea el valor esperado o promedio para una persona cualquiera de la población (sin tener en cuenta los diferentes tratamientos o niveles recibidos por cada una) y considerar los efectos t_i introducidos por los tratamientos de modo que el modelo ANOVA de un factor puede escribirse como:

$Y_{ij} = \mu + t_i + e_{ij}$,

donde:

Y_{ij}: respuesta al tratamiento de cada individuo j (variable dependiente);

μ: media poblacional o gran media, promedio de todas las posibles respuestas de los individuos;

t_i: efecto del tratamiento i sobre la respuesta Y del individuo j (variable independiente);

e_{ij}: cantidad en que difiere cada respuesta individual de la media poblacional o error no explicado asociado a cada variación (sigue una distribución normal que se expresa como $e \sim N(0, \sigma^2)$.

El estadístico de prueba o contraste a calcular es la F de Fisher-Snedecor (F_{EXP}) que se expresa como:

$F_{EXP} = \frac{S_E^2}{S_D^2}$, que sigue una distribución F con grados de libertad en el numerador t – 1 y en el denominador= N – t, para (1 – α) % como nivel de confianza

donde:

S_E^2: intervarianza, expresa la variabilidad entre los diferentes tratamientos, niveles, muestras o grupos de datos a comparar;

S_D^2: intravarianza, expresa la variabilidad dentro de cada tratamiento, nivel, muestra o grupo de datos a comparar;

t: número de tratamientos, niveles, muestras o grupos de datos a comparar;

N: número total de mediciones de los grupos a comparar u observaciones realizadas para todos los grupos de datos.

Para el cálculo de estos términos se desglosa la variabilidad total de la muestra en términos independientes, partiendo de la proposición:

$SC_T = SC_D + SC_E$

donde:

$SC_T = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2$: suma de cuadrados total;

$SC_D = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$: suma de cuadrados dentro de cada nivel;

$SC_E = \sum_{i=1}^{t} n_i(\overline{y}_i - \overline{y})^2$: suma de cuadrados entre los niveles;

n_i: cantidad de mediciones en cada nivel;

y_{ij}: cada uno de los valores medidos o respuestas individuales para la variable dependiente;

$\overline{y}_i$: media de los valores medidos o respuestas individuales para la variable dependiente en cada nivel, grupo de datos o muestra;

$\overline{y}_i$: media de todos los valores medidos o respuestas individuales para la variable dependiente en el total de niveles, grupos de datos o muestras.

Una vez calculadas las magnitudes anteriores se pueden obtener las cuasivarianzas total, la intervarianza y la intravarianza respectivamente según las expresiones:

$$S_T^2 = \frac{SC_T}{N-1};$$

$$S_E^2 = \frac{SC_E}{t-1};$$

$$S_D^2 = \frac{SC_D}{N-t}.$$

Se emplean los valores de la intervarianza y la intravarianza para a partir de su razón obtener el valor de F (observado-calculado o experimental) (F_{EXP}), que se compara con el F (tabulado) según los grados de libertad t – 1 en el numerador y N – t en el denominador para un nivel de confianza (1 – α)% seleccionado por el investigador (véase el anexo 1).

Los supuestos necesarios para llevar a cabo el ANOVA de una vía o un factor en la comparación de más de dos grupos de datos a partir de sus medias son:

1. El primer aspecto es que se haya realizado una asignación aleatoria de los individuos de la población para conformar los más de dos grupos o muestras según las categorías del factor estudiado (variable cualitativa multinominal independiente) que representaría a más de dos muestras que estimarían más de dos poblaciones. Las observaciones de cada muestra son independientes y también las de las muestras entre sí. La alietoriedad es razonable también si la técnica de muestreo aplicada es adecuada y probabilística. Ejemplos de pruebas aplicables en este sentido:
   a) Prueba no paramétrica de las rachas o prueba de independencia de las observaciones de una muestra:
      - Tiene como probable planteamiento en el contraste de hipótesis:
        H_0: existe alietoriedad en la extracción de la muestra de estudio *versus* H_1: no existe alietoriedad en la extracción de la muestra de estudio.
      - Descripción del procedimiento: si se tiene una muestra de tamaño (n), en este caso representada para una de las más de dos categorías de la variable independiente polinominal o factor a emplear con los respectivos valores de los datos de una variable cuantitativa se realiza el ordenamiento de estos valores y se divide la muestra en dos grupos de datos en base a dos categorías, con un número de observaciones en cada una n_1 y n_2 que tienen un tamaño similar ($n_1 = n_2 \pm 1$). Este procedimiento se realiza para cada una de las muestras que corresponde a cada categoría de la variable independiente.

        Se denomina rachas a la sucesión de valores de la misma categoría. En función de las cantidades n_1 y n_2 se espera que el número de rachas no sea muy pequeño ni muy grande para poder afirmar un orden aleatorio para los valores de la variable cuantitativa en la muestra de estudio.

        Una forma práctica de dividir la muestra (correspondiente con cada categoría de la variable independiente) es:

        - Si las observaciones pueden ser consideradas como cantidades numéricas que se dividen en dos categorías del mismo tamaño y se considera la mediana de las observaciones como el valor que divide la muestra en: observación inferior a la mediana (n_1), observación superior a la mediana (n_2), de esta forma se tienen dos muestras hipóteticas con sus respectivos valores ordenados numéricamente y se le asigna una, por ejemplo, una letra (E: valores mayores o iguales a la mediana y D: valores menores a la mediana).

          Se define la variable aleatoria R o G como el número de rachas (R observado). Su distribución está tabulada para los casos $n_1 + n_2 \leq 20$ (consultar tabla de referencia en el Anexo 1) para R teórico-esperado en sus valores críticos inferior y superior). La regla de decisión en el contraste de hipótesis sería:

– Si R (observado) mayor es superior que el valor crítico superior tabulado (teórico-esperado) para el número de rachas o si R (observado) menor es inferior al valor crítico inferior tabulado, entonces se rechaza H_0, por lo que no existe alietoriedad en el orden de los valores de la variable cuantitativa, al realizar la extracción de la muestra.

La distribución acumulativa de R se encuentra tabulada en forma extensa, pero si tanto n_1 como n_2 son mayores que 10 ($n_1 + n_2 > 20$), la distribución de R se encuentra, en forma adecuada, aproximada por una distribución normal con media [E (R)] y varianza [Var (E)] expresadas respectivamente como:

$$E(R) = \frac{2n_1n_2}{n_1 + n_2} + 1$$

$$Var(R) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)} + 1$$

Se estima:

$$Z\ (\text{observado-calculado}) = \frac{R\ (\text{observado}) - R\ (\text{esperado})}{\sqrt{Var(R)}}$$

Regla de decisión

– Para el caso que $n_1 + n_2 \leq 20$ y si:

R (observado-calculado) ≤ $R_{n1,\ n2,\ \alpha/2}$ (teórico-esperado) en su valor crítico inferior o R (observado-calculado) ≥ $R_{n1,\ n2,\ \alpha/2}$ (teórico-esperado) en su valor crítico superior, entonces quiere decir que hay menos rachas o muchas más respectivamente que las esperadas y con esto no pueden entremezclarse los valores muestrales lo que va en contra de la distribución aleatoria en los datos y por eso se rechaza H_0, por lo que existen evidencias estadísticamente significativas para rechazar la aleatoriedad del orden de los datos de la variable respecto a la extracción de la muestra, lo que impide el empleo del ANOVA de una vía en ese caso.

– Para el caso que $n_1 + n_2 > 20$ y si:

|Z (observado calculado)| ≤ Z (teórico-esperado) según alfa seleccionado, entonces se ubica fuera del área crítica de rechazo el valor de Z observado, lo que corresponde a que el número de rachas observadas y esperadas similares y con esto pueden entremezclarse los valores muestrales lo que va a favor de la distribución aleatoria en los datos y por esto no se puede rechazar H_0, se asume la existencia de aleatoriedad en la extracción de los datos de la muestra, lo que permite el empleo del ANOVA de una vía en este caso.

2. El segundo aspecto es que se demuestren que los valores de la variable cuantitativa se distribuyen siguiendo una distribución normal en cada una de las más de dos muestras o de los más de dos grupos a comparar, definidos a partir de las categorías de la variable cualitativa multinominal (factor). Ejemplo de pruebas aplicables en este sentido:

a) Prueba de D' Agostino: tiene como probable planteamiento para el contraste de hipótesis:
   - H_0: los valores de la variable cuantitativa siguen una distribución normal en las muestras para cada categoría del factor estudiado (variable cualitativa multinominal).
   - H_1: los valores de la variable cuantitativa no siguen una distribución normal en las muestras para cada categoría del factor estudiado (variable cualitativa multinominal).

   Como condición se requiere que el tamaño de cada muestra n ≥ 10. Pasos a seguir:
   - Ordenar los valores de las observaciones de menor a mayor en cada muestra correspondiente con cada categoría de la variable cualitativa multinominal (factor) y establecer un rango para cada uno de dichos valores según el número de las observaciones realizadas.
   - Calcular en cada muestra (para cada categoría) la media, desviación típica (estándar), un estadístico T y el estadístico de contraste D (observado-calculado), donde:

$$D\ (\text{observado-calculado}) = \frac{T}{n^2 S}$$

donde:

$$T = \sum_{i=1}^{n}\left(i + \frac{n+1}{2}\right)x_i = \sum_{i=1}^{n} i x_i - \frac{n(n+1)}{2}\bar{x}$$

i: valor del rango correspondiente a cada valor **i** de la variable x en cada observación n;

$\bar{x}$: media de los valores i de la variable x para las observaciones n.

S: desviaciín típica o estándar para los valores i de la variable x en las n observaciones.

Una vez calculado el estadístico D (observado-calculado) de las muestras se compara con el valor D (teórico-esperado) tabulado para las respectivas muestras y se determinaeste último mediante la tabla de valores críticos de D'Agostino (véase el anexo 1), partiendo del número de observaciones muestral n y el nivel de confianza (1 – a) % seleccionado por el investigador.

Regla de decisión

- Si D (observado-calculado) > $D_{n,\alpha}$ (teórico-esperado), entonces el valor observado de D quedaubicado en el área de rechazo de H_0, por lo que existen evidencias para no aceptar la normalidad en la distribución de los valores observados de la variable cuantitativa en la muestra conformada para cada categoría del factor o variable cualitativa multinominal en estudio, según el nivel de confianza seleccionado, este resultado impide aplicar ANOVA de una vía.

– Si D (observado-calculado) ≤ $D_{n,\alpha}$ (teórico-esperado) entonces no se puede rechazar H_0.

b) Prueba de Shapiro-Francia: tiene como probable planteamiento para el contraste de hipótesis:

– H_0: los valores de la variable cuantitativa siguen una distribución normal en las muestras para cada categoría del factor estudiado (variable cualitativa multinominal).

– H_1: los valores de la variable cuantitativa no siguen una distribución normal en las muestras para cada categoría del factor estudiado (variable cualitativa multinominal).

Como condición se requiere que el tamaño de cada muestra (n) ≥ 10. Pasos a seguir:

– Ordenar los valores de las observaciones de menor a mayor en cada muestra de la categoría de la variable cualitativa multinominal (factor) y establecer así mismo un rango (orden ascendente desde 1 en adelante) para cada una:

• El estadístico de contraste a emplear es:

$$Z\ (\text{observado-calculado}) = \frac{\ln(1-W') - \hat{\mu}}{\hat{\sigma}}$$

donde:

$W' = \dfrac{\left(\sum_{i=1}^{n}(m_i - \bar{m})(x_i - \bar{x})\right)^2}{\left(\sum_{i=1}^{n}(m_i - \bar{m})^2\right)\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)}$: cuadrado del coeficiente de correlación de Pearson entre los valores de la muestra x_i y los valores m_i;

n: tamaño de la muestra.

$x_1,\ldots, x_n$: valores de la muestra;

$\bar{x}$: media de la muestra;

$$m_i = \Phi^{-1}\left(\frac{r(i) - 0{,}375}{n + 0{,}25}\right)\text{, con } i = 1,\ldots, n;$$

r (i): rango que ocupa la observación x_i en la muestra ordenada. En caso de empate el rango de las observaciones repetidas es el promedio de sus rangos;

$\bar{m}$: media de los valores m_i;

$$\hat{\mu} = -1{,}2725 + 1{,}0521(v - u)$$

$$\hat{\sigma} = 1{,}0308 - 0{,}26758(v + 2/u)$$

u = ln (n);

v = ln (u);

Φ: función de distribución normal estándar.

• Valor de p (p-valor) bajo H_0:

p-valor = Pr[Z (observado-calculado) ≤ Z (teórico-esperado)]

Regla de decisión

- Para Z (observado-calculado) < 0:
  - Si Z (observado-calculado) < Z (teórico-esperado), entonces se ubica el valor de Z observado en el área crítica o de rechazo en la distribución de valores Z teóricos, por lo que hay evidencias para no aceptar la normalidad en la distribución de los valores cuantitativos observados de la variable estudiada en relación a la muestra conformada para la categoría del factor (variable cualitativa multinominal o politómica en estudio) al estimarse una probabilidad (p-valor) bajo H_0 o una probabilidad relacionada con el azar para el Z observado menor que el nivel de significación alfa seleccionado, lo que impide aplicar el ANOVA de una vía.
  - Si Z (observado-calculado) ≥ Z (teórico-esperado), entonces no se ubica el valor de Z observado en el área crítica o de rechazo en la distribución de valores Z teóricos, por lo que hay evidencias para aceptar la normalidad en la distribución de los valores cuantitativos observados de la variable estudiada en relación a la muestra conformada para la categoría del factor (variable cualitativa multinominal o politómica en estudio) al estimarse una probabilidad (p-valor) bajo H_0 o una probabilidad relacionada con el azar para el Z observado igual o superior al nivel de significación alfa seleccionado, lo que no impide aplicar el ANOVA de una vía.
- Para Z (observado-calculado) ≥ 0:
  - Si Z (observado-calculado) > Z (teórico-esperado), entonces se ubica el valor de Z observado en el área crítica o de rechazo en la distribución de valores Z teóricos, por lo que hay evidencias para no aceptar la normalidad en la distribución de los valores cuantitativos observados de la variable estudiada en relación a la muestra conformada para la categoría del factor (variable cualitativa multinominal o politómica en estudio) al estimarse una probabilidad (p-valor) bajo H_0 o una probabilidad relacionada con el azar para el Z observado menor que el nivel de significación alfa seleccionado, lo que impide aplicar el ANOVA de una vía.
  - Si Z (observado-calculado) ≤ Z (teórico-esperado), entonces no se ubica el valor de Z observado en el área crítica o de rechazo en la distribución de valores Z teóricos, por lo que hay evidencias para aceptar la normalidad en la distribución de los valores cuantitativos observados de la variable estudiada en relación a la muestra conformada para la categoría del factor (variable cualitativa multinominal o politómica en estudio) al estimarse una probabilidad (p-valor) bajo H_0 o una probabilidad relacionada con el azar para el Z observado, igual o superior al nivel de significación alfa seleccionado; lo cual no impediría aplicar el ANOVA de una vía.

3. El tercer aspecto es demostrar que las varianzas para la variable cuantitativa de todos los grupos o muestras a comparar son homogéneas (homocedasticidad), es decir, no existen diferencias significativas entre las varianzas de las respectivas muestras o de los más de dos grupos de datos a comparar: Algunos de los ejemplos de pruebas para el contraste de hipótesis en este sentido son:
   a) Prueba de Barlett: para la mayoría de los autores es la prueba más utilizada para contrastar si más de dos muestras aleatorias son homocedásticas (tienen igual varianzas) y brinda la posibilidad de comparar grupos de datos o muestras con diferentes tamaños. Tiene como probable planteamiento para el contraste de hipótesis:
      - $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$
      - $H_1 : \sigma_1^2 \neq \sigma_j^2$ para algún i, j

      El estadístico de prueba o de contraste es χ^2, que adopta la expresión:

      $\chi^2_{\text{Barlett}}(\text{observado-calculado}) = \frac{1}{k}\left[(N-t)\ln\hat{S}_D^2 - \sum_{i=1}^{t}\ln\hat{S}_i^2\right]$, que sigue una distribución teórica $\chi^2_{t-1,\,1-\alpha}$ con t – 1 grados de libertad para el valor de alfa según el nivel de confianza preseleccionado

      donde:

      N: número de observaciones (sujetos) en toda la muestra;

      t: número total de muestras o grupos según las categorías del factor (variable cualitativa multinominal);

      $\hat{S}$: cuasivarianza de cada muestra.

      También se ha utilizado por diferentes autores la siguiente expresión para el estadístico de Barlett:

      $$\chi^2_{\text{Barlett}} = \frac{\left[\ln\frac{\sum\sigma^2(n-1)}{\sum(n-1)}\sum(n-1)\right] - \sum\ln\sigma^2(n-1)}{1+\frac{K+1}{3(K-1)(N-K)}}$$

      donde:

      α^2: varianza de cada muestra que estima una poblacional (se asume como igual a S^2);

      n: tamaño de cada muestra;

      K: número de muestras comparadas;

      N: (Σn_i): tamaño total de cada muestra.

Regla de decisión

- Si χ^2_{Barlett} (observado-calculado) > $\chi^2_{t-1,\,1-\alpha}$ (teórico-esperado), esto ubica el valor de χ^2 observado en el área crítica o de rechazo en la distribución de valores F teóricos, por ende hay evidencia para rechazar H_0 y se puede aceptar H_1,

por lo que existen evidencias para estimar que al menos las varianzas de dos de las muestras son significativamente diferentes entre sí, al estimarse un valor de probabiliadades (p-valor) bajo H_0 para el valor χ^2 de Barlett observado o que se relacionan con el azar, inferiores al valor de probabilidad del modelo teórico de referencia según el valor alfa prefijado ($p < \alpha$), lo que impide aplicar el ANOVA de una vía.

- Si χ^2_{Barlett} (observado-calculado) $\leq \chi^2_{t-1,\,1-\alpha}$ (teórico-esperado), esto no ubica el valor de χ^2 de Barlett observado en el área crítica o de rechazo en la distribución de valores F teóricos, por ende no se puede rechazar H_0, por lo que no existen evidencias para estimar que al menos las varianzas de dos de las muestras son significativamente diferentes entre sí, al estimarse un valor de probabiliadades (p-valor) bajo H_0 para el valor χ^2 de Barlett observado o que se relacionan con el azar iguales o superiores al valor de probabilidad del modelo teórico de referencia según el valor alfa prefijado ($p \geq \alpha$), lo que no impide aplicar el ANOVA de una vía.

b) Prueba de Cochran: aunque implica cálculos más sencillos, está algo restringida para el caso de tamaños de muestras o de los grupos de datos sean iguales y que se cumpla la alietoriedad y normalidad de las observaciones. Es principalmente útil cuando se desea detectar si una varianza es mucho mayor que las otras. Tiene como probable planteamiento para el contraste de hipótesis:

- $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$
- $H_1 : \sigma_1^2 \neq \sigma_j^2$ para algún i, j

El estadístico de prueba o de contraste es R de Cochran, que adopta la expresión:

$$R_{\text{Cochran}}(\text{observado} - \text{calculado}) = \frac{\text{máx}\left\{\hat{S}_i^2\right\}_{i=1}^{t}}{\sum_{i=1}^{t}\hat{S}_i^2}$$

donde:

$\hat{S}_i^2$: cuasivarianza muestral para el nivel i o grupo de datos i y se calcula por la expresión:

$$\hat{S}_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - x_{i\cdot})^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}x_{ij}^2 - \frac{n_i}{n_i - 1}\overline{x}_{i\cdot}^2$$

N = nt: número total de observaciones;
n: tamaño de cada muestra, nivel o grupo de datos (el mismo);
t: número total de muestras, niveles o grupos de datos.
Una vez calculado el valor del estadístico R de Cochran se compara con el valor teórico-esperado de las tablas de referencia (véase el anexo 1) según los grados de libertad n – 1, el tamaño de muestra n y el nivel de significación alfa seleccionado.

Regla de decisión

- Si R de Cochran (observado-calculado) > R de Cochran $_{n-1,t,1-\alpha;}$ (teórico-esperado) esto ubica el valor de R de Cochran observado en el área crítica o de rechazo en la distribución de valores F teóricos, por ende hay evidencia para rechazar H_0 y aceptar H_1, por lo que existen evidencias para estimar que al menos las varianzas de dos de las muestras son significativamente diferentes entre sí, al estimarse un valor de probabiliadades (p-valor) bajo H_0 para el valor R de Cochran observado o que se relacionan con el azar inferiores al valor de probabilidad del modelo teórico de referencia según el valor alfa prefijado ($p < \alpha$), lo que impide aplicar el ANOVA de una vía.
- Si R de Cochran (observado-calculado) ≤ R de Cochran $_{n-1,t,1-\alpha;}$ (teórico-esperado), esto no ubica el valor de R de Cochran observado en el área crítica o de rechazo en la distribución de valores F teóricos, por ende, hay evidencia para no rechazar H_0, por lo cual no existen evidencias para estimar que al menos las varianzas de dos de las muestras son significativamente diferentes entre sí, al estimarse un valor de probabiliadades (p-valor) bajo H_0 para el valor R de Cochran observado o que se relacionan con el azar iguales o superiores al valor de probabilidad del modelo teórico de referencia según el valor alfa prefijado ($p \geq a$), lo que no impide aplicar el ANOVA de una vía.

c) Prueba de Levene: tiene como probable planteamiento para el contraste de hipótesis:

- $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$
- $H_1 : \sigma_1^2 \neq \sigma_j^2$ para algún i, j

Los pasos a seguir en la prueba de Levene son:

- Calcular la diferencia (en valor absoluto) entre cada valor y la media de su grupo de datos:

$$D_{ij} = \left| x_{ij} - \overline{x}_j \right|$$

donde:

x_{ij}: valor de la variable X correspondiente al sujeto i, perteneciente al grupo j;

$\overline{x}_j$: media del grupo j.

- Calcular la media de las diferencias de cada grupo:

$$\overline{D}_j = \frac{\sum D_{ij}}{n_j}$$

donde:

$\sum D_{ij}$: suma de las diferencias;

n_j: tamaño de cada grupo.

- Calcular la media total de las diferencias:

$$\bar{D}_t = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} D_{ij}}{N}$$

- Calcular la suma de cuadrados intragrupos o dentro de grupos:

$$SC_{intra} = \sum_{i=1}^{n}\sum_{j=1}^{k}(D_{ij} - \bar{D}_j)^2$$

- Calcular la suma de cuadrados intergrupos:

$$SC_{inter} = \sum_{j=1}^{k} n_j(\bar{D}_j - \bar{D}_t)^2$$

- Calcular los grados de libertad:
  GL intergrupos = k – 1; siendo k el número de grupos de datos.

$$GL\ intragrupos = \sum_{j=1}^{k}(n_j - 1)^2$$

- Calcular la media cuadrática intergrupos:
  $MC_{inter} = SC_{inter}/GL_{inter}$
- Calcular la media cuadrática intragrupos:
  $MC_{intra} = SC_{intra}/GL_{intra}$
- Calcular el estadístico de prueba o contraste:
  $F = MC_{inter}/MC_{intra}$
- Calcular el valor de F (teórico-esperado crítico) por la tabla para la distribución de F de Snedecor según grados de libertad en el numerador (k – 1) y en el denominador $\sum_{j=1}^{k}(n_j - 1)^2$ para un nivel de significación alfa seleccionado por el investigador.
- Comparar el F (observado- calculado) *versus* F (tabulado-crítico).

Regla de decisión

- Si F (observado-calculado) > F (teórico-esperado) según los grados de libertad correspondientes y para (1 – α) % como nivel de confianza, esto ubica el valor de F observado en el área crítica o de rechazo en la distribución de valores F teóricos, por ende, hay evidencia para rechazar H_0 y se puede aceptar H_1, por lo que existen evidencias para estimar que al menos las varianzas de dos de las muestras son significativamente diferentes entre sí, al estimarse un valor de probabilidades (p-valor) bajo H_0 para el valor de F observado o que se relacionan con el azar inferiores al valor de probabilidad del modelo teórico de referencia según el valor alfa prefijado ($p < \alpha$), lo que impide aplicar el ANOVA de una vía.
- Si F (observado-calculado) ≤ F (teórico- esperado) según los grados de libertad correspondientes y para (1 – α) % como nivel de confianza, esto no ubica

el valor de F observado, por ende, no se puede rechazar H_0, por lo que no existen evidencias para estimar que al menos las varianzas de dos de las muestras son significativamente diferentes entre sí al estimarse un valor de probabilidades (p-valor) bajo H_0 para el valor de F observado o que se relacionan con el azar iguales o superiores al valor de probabilidad del modelo teórico de referencia según el valor alfa prefijado ($p \geq \alpha$), lo que no impide aplicar el ANOVA de una vía.

Uno de los propósitos fundamentales del ANOVA es conocer si al menos una de las muestras difiere significativamente del resto. Cuando esta información no es suficiente para los objetivos del estudio se dispone de las llamadas pruebas *pos hoc* o pruebas para comparaciones múltiples de medias, que permiten establecer entre que grupos específicamente se producen las diferencias más importantes.

Existen diferentes tipos de clasificaciones de las pruebas para comparaciones múltiples de medias, lo que ofrece una variedad de dimensiones para caracterizarlas: según el número de comparaciones, según el tipo de contraste, según la manera de comparación, según el proceso de cálculo, según el tipo de tasa o error, según el estadístico o la distribución teórica empleada en sus cálculos, al tiempo que pueden aparecer combinadas estas clasificaciones de acuerdo con el tipo de prueba a emplear.

A continuación se presentan algunas pruebas empleadas para la comparación múltiple de medias a partir del criterio de clasificación según el tipo de estadístico de contraste empleado y la distribución teórica utilizada para su cálculo (Tabla 4.43).

Tabla 4.43. Algunas pruebas para la comparación múltiple de medias

Tipo de distribución teórica que emplea	Tipo de pruebas para comparaciones múltiples de medias
T de Student	– Dunn Bonferroni – Dunnett – Holm-Shaffer
Rango estudentizado	– Tukey – Newman-Keuls – Duncan – Ryan – Peritz
F de Fisher-Snedecor	– Scheffé – F de Newman-Keuls – F de Ryan

La prueba de Dunn Benferroni consiste en formar todos los pares posibles de medias de los tratamientos e ir comprobando, para cada par formado, la hipótesis nula que

supone que entre las medias comparadas no existen diferencias. Se utiliza cuando se quieren comparar varios grupos respecto a un grupo control.

La prueba de Tukey utiliza un valor denominado *diferencia verdaderamente significativa* contra el que se comparan todas las diferencias entre las medias. Este valor se obtiene mediante la expresión:

$$HSD = q\sqrt{\frac{CM_{error}}{r}}$$

donde:

r: número de "réplicas" (cantidad de mediciones efectuadas en cada grupo de datos o tratamientos);

CM_{error} /r: cuadrado medio del error para el número de réplicas (cuasivarianza intragrupos = S^2_D);

q: estadístico de amplitud de student (se encuentra tabulado).

Para el caso en que cualquier diferencia del par de medias, cuyo valor absoluto exceda el de diferencia verdaderamente significativa, entonces se declara como significativa dicha diferencia de medias.

La prueba de Duncan es conocida como "nueva prueba de rangos múltiples", utiliza un valor denominado "mínima diferencia significativa", contra el que se comparan todas las diferencias entre las medias.

Los pasos a seguir para realizar la prueba de Duncan son:

1. Se calculan las medias para los valores de las respuestas de la variable dependiente de cada tratamiento, nivel categoría o grupo de datos seleccionados y se ordenan numéricamente en forma descendente (Tabla 4.44).

Tabla 4.44. Representación de tratamientos y media de las respuestas para cada uno

Tratamientoscategorías o grupos de datos (T)	T_4	T_1	T_3	T_2	T_n
Media de las respuestas en cada Tratamiento o grupo de datos ($\overline{Y}_{ij}$)					

Nota: en este caso para T_4 se obtiene la mayor de las medias en las respuestas de la variable dependiente ($\overline{Y}_{ij}$).

2. Se construye una tabla colocando en la primera columna desde la primera media (el mayor valor de las medias empledas) de las respuestas hasta la penúltima y en la primera fila se coloca desde la segunda media de las respuestas hasta la última (paso 1). A continuación se calculan las diferencias entre las medias de manera ordenada a lo largo de cada fila, lo que constituye los pasos 2, 3, 4, … n (Tabla 4.45.

Tabla 4.45. Representación de las medias de las respuestas de cada grupo o tratamiento y diferencias entre las mismas según los pasos.

Media de las respuestas en cada grupo de datos ($\bar{Y}_{ij}$)	T_1	T_3	T_2	Pasos
T_4	T_4-T_1	T_4-T_3	T_4-T_2	2
T_1	T_1-T_1	T_1-T_3	T_1-T_2	3
T_3	T_3-T_1	T_3-T_3	T_3-T_2	4

Nota: para este ejemplo o caso particular se asume que se emplearon cuatro tratamiento o grupos de datos y que para el tratamiento o grupo de datos (T_4) se obtiene la mayor de las medias calculadas para las respuestas en la variable dependiente, entre todos los grupo de datos o tratamientos.

3. Se calculan los valores de rango significativo mínimo para cada paso según la fórmula (Tabla 4.44):

   $$LSR = q\sqrt{\frac{CM_{error}}{r}}$$ o

   $$R_p = qS_{\bar{x}}$$

   donde:

   $$S_{\bar{x}} = \sqrt{\frac{CM_{error}}{r}}$$

   CM_{error}/r: cuadrado medio del error para el número de réplicas (cuasivarianza intragrupos S^2_D);
   q: rango studentizado que se extrae de las tablas para valores críticos de Duncan según los grados de libertad θ = tr – t, el nivel de significación alfa seleccionado y el paso correspondiente de la tabla conformada para este análisis (véase el anexo 1);
   r: número de "réplicas" (cantidad de mediciones efectuadas en cada grupo de datos o tratamientos).

Tabla 4.46. Representación de las medias de las respuestas de los tratamientos, sus diferencias según pasos y el valor del rango significativo mínimo

Media de las respuestas en cada grupo de datos ($\bar{Y}_{ij}$)	T_1	T_3	T_2	Pasos	Rango significativo mínimo para cada paso ($R_p = qS_{\bar{x}}$)
T_4	T_4-T_1	T_4-T_3	T_4-T_2	2	
T_1	T_1-T_1	T_1-T_3	T_1-T_2	3	
T_3	T_3-T_1	T_3-T_3	T_3-T_2	4	

4. Se comparan los valores calculados de las diferencias entre cada par de medias pertenecientes a cada tratamiento o grupo de datos contra el valor de rango signifi-

cativo mínimo para cada paso de la tabla 4.42. Como conclusión se puede decir que si la diferencia entre un par de medias determinado es mayor o igual que el valor de rango significativo mínimo correspondiente a su paso, entonces la diferencia entre este par es significativa.

Nociones sobre el empleo del ANOVA en diseños de estudio experimentales

El ANOVA unifactorial puede clasificarse como simple, doble o triple en correspondencia con la presencia de tratamientos sin acompañamiento o falta de bloques o grupos homogéneos en que puede ser subdividida la muestra de estudio (ANOVA simple o unilateral) o contrariamente con la presencia de bloques más tratamientos (ANOVA doble o bilateral) o junto a los mismos (tratamientos y bloques) la presencia del factor fila (ANOVA triple o trilateral).

El ANOVA tiene gran aplicación en diseños experimentales simples en que se emplea una sola variable independiente o factor, es decir, en los que solo se considera la influencia de una sola variable independiente o tratamiento sobre las respuestas de un individuo (variable dependiente). Esto significa que sobre cada unidad experimental (individuo) solo se aplica un tratamiento de manera que su respuesta individual se vea modificada por una sola causa. En cambio, también se pueden realizar diseños experimentales en que se utilicen más de una variable independiente relacionada con una dependiente y se pueden analizar las interacciones entre las primeras llamados *diseños experimentales complejos* en los que generalmente se requieren emplear métodos multivariados para el análisis estadístico.

Los diseños experimentales simples más estudiados en ciencias biomédicas son el diseño completamente al azar, diseño en bloques al azar y el diseño en cuadrado latino.

Diseño completamente al azar

Es el diseño donde la regla de asignación de las unidades experimentales a las condiciones de estudio es completamente al azar.

Las ventajas fundamentales del diseño completamente al azar son:

- El análisis estadístico es el más sencillo (ANOVA simple).
- Permiten una gran flexibilidad en cuanto al número de tratamientos y réplicas.
- -Si durante el experimento se pierden datos esto no trae mayores consecuencias.

La principal limitación o desventaja consiste en que la población investigada debe ser homogénea y de no cumplirse este requisito, puede cometerse un error tipo II (aceptar una hipótesis nula falsa, que debió ser rechazada).

Debido a que en experimentos donde se compara una gran cantidad de tratamientos es usual que no se disponga del número suficiente de unidades experimentales homo-

géneas para efectuar las réplicas necesarias y, por tanto, no es aconsejable utilizar el diseño completamente al azar, se han desarrollado otros diseños, que sin dejar de ser simples, permiten controlar el efecto de la no homogeneidad de la población.

El modelo teórico en que se basa este diseño es:

$Y_{ij} = \mu + t_i + e_{ij}$

donde:

Y_{ij}: respuesta de cada individuo j al tratamiento i;

μ: media general o gran media, promedio de todas las posibles respuestas de los individuos;

t_i: efecto del tratamiento i sobre la respuesta del individuo j;

e_{ij}: cantidad en que difiere una respuesta individual de la media poblacional (se conoce como error no explicado asociado a cada medición).

Cuando se conduce un experimento por un diseño completamente al azar se asume que la respuesta de cada individuo ante un tratamiento determinado depende solamente de la acción de este tratamiento y del error que ocurra durante la medición de esta respuesta.

El tamaño de muestra recomendado para este diseño es:

$n = tr$

donde:

t: número de tratamientos a comparar;

r: número de réplicas o número de individuos que reciben cada tratamiento.

Se apoya en la realización de la prueba de Fisher con el estadístico de prueba F-Snedecor-Fisher, mediante la razón de varianzas:

$$F = \frac{S^2_{TRATAMIENTOS}}{S^2_{ERROR}} = \frac{CM_{TRATAMIENTOS}}{CM_{ERROR}}$$

donde:

S^2_E: en el planteamiento general del ANOVA equivale a $S^2_{TRATAMIENTOS}$ en los diseños experimentales, representando la Intervarianza, que expresa la variabilidad entre los diferentes tratamientos o niveles, es decir, la explicada por los tratamientos:

$$S^2_E = \frac{SC_E}{t-1} \text{ y } SC_E = \sum_{i=1}^{t} n_i(\overline{y}_i - \overline{y})^2$$

S^2_D : en el planteamiento general del ANOVA equivale a S^2_{ERROR} en los diseños experimentales, representando la intravarianza, que expresa la variabilidad dentro de cada tratamiento o nivel,; es decir, la no explicada por los tratamientos:

$$S^2_D = \frac{SC_D}{t-1} \text{ y } SC_D = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$$

$S_T^2 = \frac{SC_T}{N-1} = S_D^2 = \frac{SC_D}{t-1} + S_E^2 = \frac{SC_E}{t-1}$: representa la variabilidad total de la muestra estudiada:

$$SC_T = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 = SC_{TOTAL} = SC_{TRATAMIENTOS} + SC_{ERROR}$$

n_i: cantidad de mediciones en cada nivel;

y_{ij}: cada uno de los valores medidos o respuestas individuales;

$\overline{y}_i$: media muestral de las respuestas de cada nivel;

$\overline{y}$. media muestral de todas las observaciones o gran media.

En la tabla 4.47 se representa una manera de ubicar los datos necesarios para la realización del ANOVA simple del diseño completamente al azar.

Tabla 4.47. Organización tabular de los datos en el ANOVA simple del diseño completamente al azar

Tratamientos (T_i)	Respuesta de los individuos (Y_{ij})	Σ (Y_i)	n_i
A			
B			
C			
D			
E			
		Total (Y) = Σ (Y_i)	N = Σ n_i

En la tabla 4.48 se muestran los elementos a tomar en cuenta para el cálculo de F en el ANOVA simple del diseño completamente al azar.

Tabla 4.48. Resumen de los aspectos para el cálculo de F en el ANOVA simple en el diseño

Fuente de variación	SC	θ	CM = SC/θ	F
Tratamientos	$SC_{tratamientos} = \sum_{i=1}^{t}\frac{T_i^2}{n_i} - FC$	t – 1	$\frac{SC_{tratamientos}}{t-1}$	$\frac{CM_{tratamientos}}{CM_{error}}$
Error	$SC_{total} - SC_{tratamientos}$	n – t	$\frac{SC_{error}}{n-1}$	
Total	$SC_{total} = \sum y_{ij}^2 - FC$	n – 1	-	

donde:

θ: grados de libertad;

T_i: sumatoria de todas las respuestas de los individuos *j* que recibieron el tratamiento i;

$FC = \frac{(\sum y_i)^2}{n}$: factor de corrección.

Regla de decisión

– Si $F_{\text{tratamientos}}$ (observado-calculado) > F crítico $_{t-1,\ N-t;\ \alpha}$, entonces se rechaza H_0 respecto a la igualdad de las medias en las respuestas de los tratamientos.

Diseño en bloques al azar

Es aquel diseño donde las unidades experimentales a las que se aplica los tratamientos son subdivididas en grupos homogéneos a los que se llama *bloques*, de acuerdo con alguna causa de variabilidad interna de la población estudiada.

Al clasificar los individuos (unidades de observación) se restringe la aleatorización, pero dentro de cada bloque los individuos son homogéneos y los tratamientos se asignan al azar. De ahí la denominación de este diseño en bloques al azar.

Para que se cumplan los requisitos anteriores, en cada bloque se debe realizar una réplica completa del experimento, por lo que se necesita un individuo por cada tratamiento, de manera que cada uno de los tratamientos aparezca en cada bloque y cada bloque reciba todos los tratamientos.

El tamaño de muestra experimental recomendado para este diseño es:

$n = tr$

donde:

t: número de tratamientos a comparar;

r: número de réplicas o número de bloques en que se agrupan los individuos que conforman la muestra.

El objetivo de utilizar un diseño en bloques al azar es aislar y eliminar el término del error debido a la causa de variabilidad interna entre las unidades experimentales, también llamado *efecto de los bloques*, lo que aumenta la potencia de la prueba (probabilidad de rechazar una hipótesis nula falsa).

En este caso el modelo lineal incluye un nuevo término β_j relacionado precisamente con el efecto de los bloques:

$$y_{ij} = \mu + t_i + \beta_j + e_{ij}$$

La ecuación planteada indica que los efectos del tratamiento y del bloque son independientes entre sí, entre ellos no existe interacción. En otras palabras, una combinación particular tratamiento-bloque no produce un efecto mayor o menor que la suma de sus efectos individuales.

La técnica estadística que se utiliza para analizar los datos de un diseño en bloques al azar se llama *analisis de varianza bilateral* o ANOVA doble, porque cada observación se clasifica en base a dos criterios: el tratamiento que recibe y el bloque a que pertenece. Además, se calculan dos razones de varianzas de Fisher (Tabla 4.49).

Tabla 4.49. Organización tabular de los datos en el ANOVA doble del diseño en bloques al azar

Tratamientos	Bloques I II III			ΣTratamientos
A	$Y_{A,I}$	$Y_{A,II}$	$Y_{A,III}$	
B	$Y_{B,I}$	$Y_{B,II}$	$Y_{B,III}$	
C	$Y_{C,I}$	$Y_{C,II}$	$Y_{C,III}$	
D	$Y_{D,I}$	$Y_{D,II}$	$Y_{D,III}$	
ΣBloques				G = Total de ΣBloques = Total de ΣTratamientos

En la tabla 4.50 se muestran los elementos a tomar en cuenta para el cálculo de F en el ANOVA doble del diseño en bloques al azar.

Tabla 4.50. Resumen de los aspectos para el cálculo de F en el ANOVA en el diseño en bloques al azar

Fuente de variación	SC	θ	CM = SC/θ	F
Tratamientos	$\frac{\sum(\sum y_{ij})^2_{TRAT}}{r} - FC$	t – 1	$\frac{SC_{TRAT}}{t-1}$	$F_{TRAT} = \frac{CM_{TRAT}}{CM_{ERROR}}$
Bloques	$\frac{\sum(\sum y_{ij})^2_{BLOQUES}}{t} - FC$	r – t	$\frac{SC_{BLOQUES}}{n-1}$	$F_{BLOQUES} = \frac{CM_{BLOQUES}}{CM_{ERROR}}$
Error	$SC_{TOT} - SC_{TRAT} - SC_{BLOQUES}$	(t – 1) (r – 1)	$\frac{SC_{ERROR}}{(t-1)(r-1)}$	-
Total	$\sum y_{ij}^2 - FC$	tr – 1	-	

Regla de decisión

- Si $F_{Tratamientos}$ (observado-calculado) > $F_{Tratamientos\ t-1;\ (t-1)\ (r-1);\ \alpha}$ (crítico), entonces se rechaza H_0 respecto a la igualdad de las medias en las respuestas de los tratamientos.
- Se compara $F_{Bloques}$ (observado-calculado) > $F_{Bloques\ t-1;\ (t-1)\ (r-1);\ \alpha}$ (crítico) entonces se rechaza H_0 respecto a la igualdad del comportamiento de los bloques empleados para los distintos tratamientos.

En el diseño en bloques al azar los datos perdidos deben ser estimados aproximadamente, debido a que una condición indispensable para su aplicación es que dentro de cada bloque haya una réplica completa del experimento (que constituye una desventaja

respecto al diseño completamente al azar). Para estimar un dato perdido en un diseño en bloques al azar se utiliza la fórmula de Yates:

$$y_{ij} = \frac{rB + tT - G}{(r-1)(t-1)}$$

donde:

G: suma general de todos los resultados o respuestas (gran suma);

B: suma de los resultados del bloque donde se perdió el dato;

T: suma de los resultados del tratamiento donde se perdió el dato.

Mediante esta fórmula pueden estimarse 1 o 2 datos. Si la cantidad de datos perdidos es superior debe repetirse el experimento.

Existe una variante del diseño en bloques al azar denominado *diseño de mediciones repetidas* en el cual para una misma variable se hacen mediciones sobre cada individuo en dos o más ocasiones, para este caso los individuos constituyen los bloques (Tabla 4.51).

Tabla 4.51. Organización tabular de los datos en el ANOVA doble del diseño de mediciones repetidas

Orden de individuos en el estudio	Orden de los días de mediciones a los individuos de la muestra			
	Día A	Día B	Día C	Día n
1	Y_{1A}	Y_{1B}	Y_{1C}	Y_{1n}
2	Y_{2A}	Y_{2B}	Y_{2C}	Y_{2n}
3	Y_{3A}	Y_{3B}	Y_{3C}	Y_{3n}
4	Y_{4A}	Y_{4B}	Y_{4C}	Y_{4n}
5	Y_{5A}	Y_{5B}	Y_{5C}	Y_{5n}
6	Y_{6A}	Y_{6B}	Y_{6C}	Y_{6n}
n	Y_{nA}	Y_{nB}	Y_{nC}	Y_{nn}

donde:

1, 2, 3..., n: representan los bloques o factor de bloqueo en el estudio;

A, B, C.., n: representa los valores de la variable independiente;

Y: representa los valores de la variable dependiente o de respuesta.

Las ventajas de este diseño son:

- Controlar las variaciones extrañas entre los individuos, ya que en este diseño cada individuo sirve como su propio control.
- Se necesitan menos individuos que en otros diseños en los que se utilizan diferentes individuos para cada medición. Esta característica es de gran utilidad cuando resulta difícil o caro encontrar voluntarios o animales para una investigación.

El diseño de mediciones repetidas presenta las siguientes desventajas:

- Efecto acarreado: si se evalúan dos o más tratamientos, la respuesta ante cada nuevo tratamiento puede estar afectada por un efecto residual del tratamiento anterior (se minimiza dejando pasar más tiempo entre las mediciones).

– Efecto de posición: diferentes secuencias de los mismos tratamientos pueden provocar diferentes respuestas de los individuos (se minimiza aleatorizando las secuencias).

Diseño en cuadrado latino

Es aquel diseño donde están presentes dos o más fuentes de variación lo suficientemente importantes como para ser designadas factores de bloque (variables aparte de la independiente). Se emplea un diseño con el mismo número de filas y columnas para los factores de bloque cruzados, donde existen unidades experimentales en todas las celdas (intercepciones de filas y columnas).

El modelo matemático correspondiente a este diseño es:

$y_{ij} = \mu + t_i + \beta_j + \delta_k + e_{ij}$

donde:

$\beta_j + \delta_k$: representan el efecto de las filas y las columnas respectivamente;

El procesamiento estadístico se realiza mediante el análisis ANOVA trilateral o ANOVA triple, dado que cada observación se clasifica en base a tres criterios: el tratamiento, el factor bloque y el factor fila.

La principal ventaja de este diseño consiste en que la igualdad del número de tratamientos, filas y columnas simplifica las fórmulas para la realización de los cálculos. Sin embargo, esta característica también puede considerarse una desventaja ya que limita la flexibilidad del experimento (Tabla 4.52).

Tabla 4.52. Organización tabular de los datos en el ANOVA triple del diseño en cuadrado latino a modo de ejemplo

Factor de bloque para las filas (FBF)	Factor de bloque para columnas (FBC)			
	FBC 1	FBC 2	FBC 3	FBC 4
FBF 1	A/ Y_i	B/Y_i	C/Y_i	D/Y_i
FBF 2	C/Y_i	D/Y_i	A/Y_i	B/Y_i
FBF 3	B/Y_i	A/Y_i	C/Y_i	D/Y_i
FBF 4	D/Y_i	B/Y_i	A/Y_i	C/Y_i

Donde:

A,B,C,D: representan los tipos de tratamientos recibidos por la muestra de estudio (variable independiente cualitativa) de acuerdo a los FBF y FBC respectivo para cada celda de intersección fila x columna;

Y_i: representan la respuesta al tratamiento específico recibido por los individuos (i) de la muestra de estudio (variable dependiente cuantitativa) de acuerdo a los FBF y FBC respectivo para cada celda de intersección fila x columna.

El tamaño de muestra necesario para este diseño es t^2, lo que implica que el número de individuos necesarios para el experimento se incremente geométricamente al aumentar el número de tratamientos a comparar y hace muy difícil su aplicación cuando el número de tratamientos a comparar es elevado, mientras que si el cuadrado latino es muy pequeño, pueden cometerse inexactitudes en la estimación del error experimental. Por esto el diseño en cuadrado latino se utiliza generalmente en el rango de 5x5 a 8x8. No se recomienda utilizar diseño en cuadrado latino menores de 3x3 o mayores que 12x12.

El diseño en cuadrado latino puede considerarse una extensión del diseño en bloques al azar, para dos factores de bloque, pero solo debe utilizarse cuando el investigador está seguro de que no existen interacciones entre los factores de bloque utilizados.

El diseño en cuadrado latino tiene algunas desventajas que no deben dejar de considerarse al decidir cómo conducir un experimento donde exista más de un factor de bloque. Por esto los investigadores suelen considerar algunos de los dos factores de bloqueo como un segundo factor o variable independiente y desarrollan el experimento mediante un diseño complejo que comprende las interacciones entre los factores en estudio.

Una vez abordados los aspectos conceptuales esenciales relacionados con la aplicación del ANOVA en algunos de los más frecuentes tipos de diseños experimentales resulta muy facilitador para su mejor comprensión su presentación resumida en forma tabular (Tabla 4.53).

Tabla 4.53. Resumen de aspectos relacionados con el ANOVA en los diseños experimentales.

Tipo de ANOVA en diseños experimentales	Subtipos de ANOVA a emplear	Componentes incluídos para cada tipo de diseño	Tipo de análisis en el ANOVA a realizar
Simple	Diseño completamente al azar	Tratamientos	Simple
	Diseño en bloques al azar	Tratamientos más un factor bloques	Doble
	Diseño de mediciones repetidas	Tratamientos más un factor bloques (bloques constituídos a partir de los días ordenados de tratamientos).	Doble
	diseño en cuadrado latino	Tratamientos más dos factores bloques de filas y columnas	Triple
Complejo	Emplea dos o más variables independientes o factores, es decir, además de la variable tratamientos se utiliza al menos otra variable independiente y puede incorporar factores de bloques donde se puede aalizar interacciones entre variables independientes.		

Métodos de análisis multivariados

Para el desarrollo y generalización de investigaciones sustantivas que analizan grandes bases de datos, tan importante como el desarrollo de la informática, es el conocimiento y aplicación de un grupo de técnicas estadísticas que se agrupan bajo el nombre de "métodos de análisis multivariados".

En su sentido más general se refiere a todas las técnicas estadísticas que analizan simultáneamente medidas múltiples de cada individuo u objeto experimental. De acuerdo con este criterio, cualquier análisis simultáneo de más de dos variables puede ser considerado como un análisis multivariado.

Los métodos de análisis multivariados son un conjunto de métodos estadísticos cuya finalidad es analizar simultáneamente conjuntos de datos multivariantes donde existen varias variables medidas para cada caso y constituyen una herramienta que ofrecen una visión conjunta e integrada de la realidad al emplear una o más de una variable dependiente y más de una independiente en la generalidad de ellos.

Los objetivos generales de los métodos de análisis multivariados para el análisis estadístico son:

- Obtener información del estudio de conjuntos de datos multivariantes que el análisis estadístico unidimensional y bidimensional no es capaz de brindar.
- Ayudar al investigador a tomar decisiones óptimas en el contexto en que se encuentre teniendo en cuenta la información disponible en el conjunto de datos multivariante analizado.

Las ventajas de los métodos multivariados se pueden sintetizar:

- Permiten investigar la relación entre una variable dependiente y una independiente mientras se "controla" o más bien se ajusta esta relación según el efecto de otras variables independientes.
- Permiten realizar pruebas de significación estadística de diversas variables manteniendo al mismo tiempo la probabilidad (alfa) escogida de cometer un error de tipo 1.
- Se puede utilizar para comparar por separado la capacidad de dos o más variables independientes para estimar los valores de la variable dependiente.
- En la mayoría de los métodos de análisis multivariados de dependencias se emplea un contraste de hipótesis general para el modelo completo, expresado a partir de la relación significativa o no, entre las variables dependientes con las independientes, así como el contraste de hipótesis especificas dentro del modelo como por ejemplo para los coeficientes de regresión estimados en los distintos métodos de regresión.

Los supuestos básicos a tener en cuenta en la ejecución de los diferentes métodos multivariados son:

– La normalidad en la distribución de las variables incluidas en el modelo: evaluada por prueba de chi-cuadrado con p grados de libertad; donde p es el número de variables del estudio.
– La homocedasticidad: evaluada principalmente por la prueba M de Box.
– La linealidad en las relaciones que se establecen entre las variables del estudio.
– La independencia en los valores que toman las variables, es decir, los valores que toman las variables en un caso no están influidos por los valores que toman en otro caso.

Los tipos de técnicas a emplear en los métodos multivariados se pueden dividir en:
– Métodos de dependencia.
– Métodos de interdependencia.

Métodos de dependencia

Suponen que las variables analizadas están divididas en dos grupos: las variables dependientes y las variables independientes.

El objetivo consiste en determinar si el conjunto de variables independientes afecta al conjunto de variables dependientes y de qué forma. Como ejemplo se tiene:
– Con variables cuantitativas o métricas en la dependiente: regresión múltiple, correlación múltiple, ANOVA, ANCOVA, MANOVA, MANCOVA, correlación canónica y análisis de sobrevida.
– Con variables cualitativas o no métricas en la dependiente: análisis discriminante, regresión logística, análisis estratificado, análisis conjunto (todos sin dependencia del tiempo) y regresión de Cox (con dependencia del tiempo).

Las variables cualitativas ordinales deben ser transformadas a cualitativas nominales y estas pueden o no ser transformadas a *dummy* para su empleo en los métodos multivariados de dependencias.

Cuando los valores de las variables cuantitativas siguen una distribución que es significativamente diferente con la aproximación a la normal se les puede transformar mediante el ordenamiento numérico de sus datos y el otorgamiento de rangos a esos valores ordenados, convirtiéndose así en una nueva variable cuantitativa "ordenada u ordinal por rangos" que se emplea en métodos no paramétricos.

A continuación se presenta de forma resumida las distintas combinaciones de los tipos de variables: dependientes e independientes empleadas en los distintos métodos multivariados de dependencias según el objetivo que persiga el investigador (Tabla 4.54).

Tabla 4.54. Combinaciones de variables dependientes e independientes en los distintos tipos de métodos multivariados de dependencias

Tipo de método multivariado de dependencias	Tipo de variables dependiente	Tipo de variables independientes
Regresión lineal múltiple	Cuantitativa continua	Cuantitativas continuas o cualitativas transformadas a indicadoras, ficticias o *dummy*
Correlación lineal múltiple	Cuantitativa continua	Cuantitativas continuas o cualitativas transformadas a indicadoras, ficticias o *dummy*
ANOVA (análisis de varianza de más de una vía o de una vía y factorial con datos apareados-repetidos o no)	Cuantitativa continua	Cualitativas nominales
ANCOVA (análisis de la covarianza)	Cuantitativa continua	Cualitativas nominales combinadas con covariable cuantitativa continua para "control o ajuste"
Regresión logística	Cualitativa nominal	Cuantitativas o cualitativas transformadas en variables indicadoras, ficticias o *dummy*
Análisis discriminante	Cualitativa nominal	Cuantitativas o cualitativas transformadas a *dummy*
Análisis estratificado	Cualitativa nominal (expresada en proporciones no dependiente del tiempo)	Cualitativas nominales (expresada en proporciones no dependiente del tiempo)
Regresión de Cox	Cuantitativa (tiempo hasta la presencia de un efecto o evento según cierta condición, expresada por una razón tasas instántaneas o *hazard ratio*)	Cuantitativas o cualitativas nominales (factores de riesgo para un tiempo determinado que se convierten en variables *dummy*)
Regresión multivariada de Poisson	Cuantitativa (número de sujetos con evento y cierta condición, expresada por medio de una razón de densidad o tasas de incidencia para sujetos con un evento y cierta condición según el total de sujetos con la condición por tiempo de seguimiento)	Cantitativas o cualitativas (transformadas a *dummy*)
Análisis conjunto	Cualitativaso cuantitativas	Cualitativas

Tipo de método multivariado de dependencias	Tipo de variables dependiente	Tipo de variables independientes
Análisis de sobrevida	Cuantitativa continua (tiempo sin la ocurrencia de un evento o efecto)	Cualitativas nominales (factor de riesgo presente o no expresado para un periodo de tiempo de seguimiento)
MANOVA (análisis múltiple de varianzas)	Cuantitativas continuas	Cualitativas nominales
MANCOVA (análisis múltiple de covarianzas)	Cuantitativas continuas	Cualitativas nominales y covariable cuantitativa
Correlación canónica	Cuantitativas continuas	Cuantitativas continuas
Análisis de ecuaciones estructurales	Cuantitativa	Cuantitativas o cualitativas

Métodos de interdependencia

No distinguen entre variables dependientes e independientes. El objetivo consiste en identificar qué variables están relacionadas, cómo lo están y por qué; es decir, son técnicas de clasificación. Como ejemplo se tienen:

- Con variables cuantitativas o métricas: componentes factoriales y principales, clúster (conglomerados) y escalas multidimensionales.
- Con variables cualitativas o no métricas: análisis por correspondencias y modelos log-lineales.

A continuación se presenta de forma resumida los tipos de variables y el objetivo que se persigue en cada tipo de método multivariado de interdependencias (Tabla 4.55).

Tabla 4.55. Tipos de métodos multivariados de interdependencias

Tipo de método multivariado de interdependencias	Tipo de variable a emplear	Objetivo del método
Componentes principales.	Cuantitativas (métricas)	Relación entre variables. Forma grupos de variables observables
Análisis factorial.	Cuantitativas (métricas).	Relación entre variables. Forma grupos de variables no observables
Análisis de correspondencias	Cualitativas (no métricas)	Relación entre variables. Forma grupos de categorías de variables y se aplica a tablas de contingencia multidimensionales presentando las filas y columnas simultáneamente

Tipo de método multivariado de interdependencias	Tipo de variable a emplear	Objetivo del método
Análisis clúster o por conglomerados	Cuantitativas o cualitativas	Relación entre casos. Forma grupos de sujetos muy similares entre sí y muy diferentes del resto en función de alguna cualidad
Escalamiento multidimensional	Cuantitativas o cuantitativas	Relación entre objetos. Transforma juicios de semejanzas o preferencias en distancias representadas en un espacio o mapa multidimensional y forma grupos de objetos muy cercanos entre sí, si son similares en sus dimensiones y muy alejados si ocurre lo contrario

Todas las técnicas multivariantes, excepto el análisis clúster y el análisis multidimensional, se basan en la inferencia estadistica de los valores de una población o la relación entre variables de una muestra escogida aleatoriamente de esa población.

A pesar de que estos métodos son mas exigentes en cuanto a su comprensión matemática permiten un mejor entendimiento del fenómeno objeto de estudio, obteniéndose una información sintetizada sobre las relaciones entre las variables estudiadas que los métodos estadísticos univariantes y bivariantes son incapaces de conseguir y ayudan al investigador a tomar decisiones óptimas en el contexto en que se encuentra, teniendo en cuenta la información disponible en el conjunto de datos multivariantes analizados.

Las etapas para la ejecución de los métodos multivariados se pueden resumir en:

1. Objetivos del análisis: se define el problema especificando los objetivos y las técnicas multivariantes que se van a emplear.
2. Diseño del análisis: se determina el tamaño muestral, las ecuaciones a emplear para el mismo y las técnicas de estimación a emplear, con todo lo que se procede a observar los datos.
3. Hipótesis del análisis: se evalúan las hipótesis subyacentes a la técnica multivariada seleccionada. Estas hipótesis pueden ser de normalidad, linealidad, independencia, homocedasticidad, entre otras, así como decidir que metodología se va a seguir con los datos perdidos.
4. Realización del análisis: se estima el modelo y se evalúa el ajuste de los datos.
5. Interpretación de los resultados: los resultados obtenidos pueden llevar a nuevas especificaciones de las variables o del modelo con lo cual se puede volver de nuevo a los pasos 3 y 4.
6. Validación del análisis: consiste en establecer la validez de los resultados obtenidos con la muestras, analizando si se pueden generalizar a la población de la que procede. Como procedimiento se puede dividir la muestra en varias partes en las cuales

el modelo se vuelve a estimar y se comparan los resultados. Otras técnicas que se pueden emplear en esta etapa son las técnicas de remuestreo (*bootstrap* y *jacknife*).

Método multivariado de regresión logística

La regresión logística es la variante de la regresión que corresponde al caso en que se valora la contribución de diferentes factores (variables independientes) en la ocurrencia de un evento simple (en la variable dependiente).

En el marco epidemiológico con ella se pueden "controlar" varias variables potencialmente confusoras (de cualquier naturaleza) a la vez.

El objetivo primordial de la regresión logística es el de modelar cómo influye en la probabilidad de aparición de un suceso, habitualmente dicotómico, la presencia o no de diversos factores y el valor o nivel de los mismos. También puede ser usada para estimar la probabilidad de aparición de cada una de las posibilidades de un suceso con más de dos categorías (politómico).

Como recomendaciones generales para el empleo de la regresión logística se tiene que:

- Las variables explicativas o independientes deben tener una relación monótona con la probabilidad del evento que se estudia. Vale decir cuando el valor de una variable independiente crece la probabilidad del desenlace aumenta o disminuye, es decir, no ha de pasar de una tendencia a la opuesta en algún punto del recorrido de la variable independiente.
- Las variables independientes involucradas en el modelo no deben estar muy correlacionadas entre sí. Si la correlación entre dos variables es alta, entonces los resultados de la regresión logística son poco confiables. Concretamente, los errores estándares se incrementan indebidamente y puede ocurrir, incluso, que el proceso iterativo para la estimación no converja.
- Debe recordarse que el conjunto de variables *dummy* o indicadoras independientes constituye un todo indisoluble con el cual se suple a una variable nominal. Cualquier decisión que se adopte o valoración que se haga concierne al conjunto íntegro, por ejemplo, si una de las variables *dummy* es significativa, entonces toda la variable nominal correspondiente lo es.
- Es muy importante distinguir entre un contexto explicativo y uno predictivo. Debe tenerse en cuenta, en este caso, que una variable puede tener valor predictivo aunque no sea parte del mecanismo causal que provoca el fenómeno en estudio.
- En lo posible ha de procurarse que haya en la base de datos al menos 10 sujetos con cada una de las respuestas posibles para la variable independiente, es decir, se hace necesario disponer de al menos 10 (k + 1) casos para estimar un modelo con k variables independientes.
- La definición de mejor modelo de la regresión logística depende del tipo y el objetivo del estudio. En un modelo con finalidad predictiva se considera como mejor modelo

aquel que provoca predicciones más fiables, mientras que en un modelo que pretende estimar la relación entre dos variables (corrigiendo el efecto de otras) se considera mejor aquel con el que se consigue una estimación más precisa y significativa del coeficiente de la variable de interés.

- El procedimiento para escoger el modelo (ecuación de regresión logística) consta de tres mecanismos básicos. Primero empezar con una sola variable independiente e ir añadiendo nuevas variables según un criterio prefijado (procedimiento hacia adelante) o bien empezar con el modelo máximo e ir eliminando de él variables según un criterio prefijado (procedimiento hacia atrás). El tercer método, denominado en la literatura *stepwise*, combina los dos anteriores y en cada paso se puede tanto añadir una variable como eliminar otra que ya estaba en la ecuación.
- El criterio para decidir en cada paso si se escoge un nuevo modelo frente al actual, viene dado por el logaritmo del cociente de verosimilitudes de los modelos. Si al añadir una nueva variable al modelo no mejora la verosimilitud de forma apreciable, en sentido estadístico, esta variable no se incluye en la ecuación.
- La función de verosimilitud de un modelo de regresión logística es una medida de cuán compatible es este con los datos realmente observados.
- La regresión logística se emplea habitualmente en uno de los diseños epidemiológicos más utilizados: los estudios de casos y controles para calcular *odds ratio* ajustadas por factores de confusión. Estas *odds ratio* representan una aproximación razonable al riesgo relativo, ya que los análisis de casos y controles suelen elegirse como el diseño más apto para estudiar enfermedades raras y cuando en la variable dependiente el evento no es frecuente (menos del 10 % presente en la muestra de estudio) puede interpretarse la *odds ratio* como una razón de probabilidades.
- Existe un tipo de estudio de casos y controles (apareados) en los que no se debe aplicar la regresión logística convencional, sino que se ha de utilizar un tipo especial de regresión logística: la regresión logística condicional.
- Algunas formulaciones necesarias para recordar y tomar en cuenta en las bases de la regresión logística son:
  - Si p = odds/1 + odds, entonces:
    Odds $_{(casos)}$ = p/(1 − p) = [(Casos con factor o condición)/(Total con factor)] / {1 − [(Casos con factor o condición)/(Total con factor)]} o también:
    Odds $_{(para\ enfermar\ o\ de\ casos)}$ = Ventaja de enfermar = [Enfermos con cierta condición/Total con condición o factor]/[Enfermos sin cierta condición/Total sin condición o factor] y además:
    $\ln(Odds) = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$ y $\ln (Odds) = \ln (p/1 - p) = logit\ (p)$
  - Si: OR = (Odds para el evento o casos) / (Odds para no evento o no casos); entonces:
    ln OR = ln (Odds para el evento o casos)/(Odds para no evento o no casos) y si:ln OR = β, esto equivale decir que OR = Exp β, lo que es de utilidad al entrar en los

aspectos propiamente de los parámetros y estadísticos fundamentales a emplear en el análisis de los resultados de la regresión logística.

- Para la ejecución del análisis de la regresión logística siempre se debe señalar la categoría de referencia en cada variable independiente predictora contra la cual se comparan las otras categorías de dicha variable y la categoría principal o evento en la variable dependiente al estimar la razón de *odds* en la misma para el modelo estimado.
- La *odds* referente al evento en la variable dependiente expresa las veces que es mayor la presencia del evento predicho para una categoría determinada respecto a la categoría de referencia en la variable independiente específica motivo de análisis en el modelo de regresión logística.
- Los propósitos posibles de este método multivariado son:
  - Contribución a la descripción de cierto proceso.
  - Búsqueda de explicaciones causales.
  - Construcción de un modelo para la predicción, que exprese la probabilidad de ocurrencia del evento de que se trate en función de un conjunto de variables independientes.
- La forma de llevar a cabo la aplicación de la regresión logística en la construcción de un modelo para la predicción es como sigue:

  Si x_1, x_2, ...x_k son k variables independientes de naturaleza cuantitativa o cualitativa e Y es una variable cualitativa con respuesta dicotómica 0 a 1, entonces el modelo múltiple de regresión logística está dado por:

$$\Pr\left\{Y=\frac{1}{x_1,\ldots,x_k}\right\}=\frac{\exp\left[-\left(\alpha+\beta_1x_1+\ldots+\beta_kx_k\right)\right]}{1+\exp\left[-\left(\alpha+\beta_1x_1+\ldots+\beta_kx_k\right)\right]}$$

  La variable Y se codifica como 1 si se produce cierto desenlace o evento y como 0 en caso opuesto, de modo que la regresión logística expresa una probabilidad del evento Pr (Y = 1) en función de ciertas variables relevantes X_1, ..., X_k a los efectos del problema que se haya planteado.

  Si se ajusta un modelo que incluya una variable nominal X con C clases o categorías, esta debe ser sustituída por C-1 variables *dummy* (dado que la categoría que no se introduce es la que se reserva como estrato o categoría de referencia) y a cada una corresponde su respectivo coeficiente.

  La manera de definir estas variables:

  - Si el sujeto pertenece a la primera categoría, entonces, las C-1 variables *dummy* valen 0.
  - Si el sujeto está en la segunda categoría, entonces la primera variable *dummy* vale 1 y las demás valen 0. Así sucesivamente hasta llegar a la última categoría.

En la tabla 4.56. se muestra, a partir de una variable nominal X con sus respectivas categorías, la generación de las variables *dummy* Z que son empleadas directamente en la regresión logística.

Tabla 4.56. Representación de las variables *dummy* Z_{n-1} en relación a una variable nominal X con las categorías C_n

Variable nominal X	Dummy		
	Z_1	Z_2	Z_3
C1	0	0	0
C2	1	0	0
C3	0	1	0
C4	0	0	1

Para incluir en el modelo la interacción de dos variables X y Z basta incluir el producto de las dos, XxZ.

El resultado fundamental del método consiste en hallar los coeficientes β_0, β_1,..., β_r, que mejor se ajustan a la representación funcional:

$$P(Y=1)=\frac{1}{1+\exp(-\beta_0-\beta_1 x_1-...\beta_r x_r}$$

donde:

exp(.): representa la función exponencial.

El modelo se estima por el método de máxima verosimilitud utilizando el algoritmo de Newton Raphson según Jones (1975) y como resultado se obtienen para valorar:

– Los coeficientes de regresión estimados de cada variable explicativa o independiente (β_k) con su error estándar:

$EE(\hat{\beta}_k)=\sqrt{\hat{V}(\hat{\beta}_k)}$ para $\hat{V}(\hat{\beta}_k)$, k = 0, 1, 2, ..., r en relación al evento de interés en la variable dependiente.

Cada coeficiente de regresión beta (β_k) expresa el logaritmo neperiano de la *odds ratio* (lnOR) de que ocurra un fenómeno o evento, por unidad de cambio de la variable independiente, es decir, expresa el aporte del cambio de una determinada categoría con respecto a la categoría de referencia en la variable independiente (predictora) específica, para predecir el valor del logaritmo neperiano del *odds ratio* respectivo, una vez ajustado por las otras variables predictoras el análisis de esta variable en el modelo o ecuación logística. El ajustar significa equiparar o asumir semejantes a los grupos formados a partir de las categorías que se comparan en la variable predictora analizada, en cuanto a los valores o categorías de las otras variables independientes o factores por las que se ajusta y que están incluidas en el modelo. Su intervalo de confianza (IC) se estima como:

$IC(\beta_k) = Exp(\beta_k \pm Z \cdot EE\beta_k) = OR_k \cdot Exp(\pm Z \cdot EE\beta_k)$

La significación estadística de cada coeficiente de regresión β_k (para cada variable independiente en relación con el evento de la dependiente) en la regresión logística se determina por el contraste de hipótesis como:

H_0: $\beta_k = 0$ *vs.* H_1: $\beta_k \neq 0$, k = 0,1,..., r:

- Estadístico de contraste o de prueba:
  χ^2_{Wald} (observado-calculado) = $[(\beta_k)/(EE\beta_k)]^2 \sim \chi^2_{K-1,\,1-a}$, siendo el número de grupos (gr) o categorías (k) del factor pronóstico o variable independiente predictora menos 1 los grados de libertad para un nivel de confianza (1 – a) 100 %; o sea, GL = K – 1 para (1 – a) · 100 %.
- Valor p (bajo H_0 o donde se cumple H_0):
  p valor = Pr [$\chi^2_{gr-1,\,1-a}$ (teórico-esperado) ≥ χ^2_{Wald} (observado-calculado)].
  También puede ser evaluada la significación del coeficiente de regresión (β_k) por la aproximación de la normal estándar (Z) ~N (0, 1) para muestras grandes en la prueba de Waldsegún Silva (1995) como:
- Contraste de hipótesis:
  H_0: $\beta_k = 0$ *vs.* H_1: $\beta_k \neq 0$, k = 0,1,..., r
- Estadístico de contraste (de prueba):
  Z (observado-calculado) = $\beta_k/EE\beta_k \sim Z \sim N(0, 1)$.
- Valor p (bajo H_0 odonde se cumple H_0):
  p valor = Pr [$Z_{1-a/2}$ (teórico –esperado) ≥ |Z (observado-calculado)|].
  Para variables independientes (predictoras) con más de dos categorías, el valor p (p-valor) global, indica si la presencia del evento de interés en la variable dependiente mejora significativamente al añadir esta variable independiente al conjunto de variables independientes predictoras o no, según el valor de alfa prefijado; es decir, si: p-valor global ≥ alfa no deberá incluirse en el modelo dicha variable predictora; mientras que el valor p (especifico) para cada categoría respecto a la de referencia en cada variable independiente (predictora) indica si tiene dicha categoría o no significación estadística en la predicción del evento en la variable dependiente.

– *Odds ratio* estimado según cada variable predictora, explicativa o independiente (con su intervalo de confianza) en relación al evento en la variable dependiente, a partir de los coeficientes de regresión B_k o beta (β_k) estimados (producto de la comparación de alguna categoría *versus* la de referencia en cada variable predictora si es cualitativa, previa conversión a variables *dummy*) como sigue:

- *Odds ratio* (OR_k) de la variable predictora k (k = 1, 2, ..., r)
  (OR_k) = exp (β_k); donde exp = 2,7182.
- Intervalo de confianza (IC) [(1 – a) 100 %] para (OR_k):
  IC[(1 – α) 100 %] Inferior (OR_k) = Exp ($\beta_k - Z_{1-\alpha/2}\, EEB_k$).
  IC[(1 – α) 100%] Superior (OR_k) = Exp ($\beta_k + Z_{1-\alpha/2}\, EE\, B_k$).
  donde:
  $Z_{1-a/2}$: percentil de la distribución normal estándar N (0,1), que deja a la izquierda una cola 1 – a/2;
  (1 – α) 100 %: l nivel de confianza.

Interpretación de los estadígrafos resultantes a mostrar en una regresión logística

En los paquetes estadísticos se pueden encontrar los siguientes aspectos:

– Coeficiente de regresión (β): expresa el peso que tiene la variable predictora independiente (K) en el resultado de la OR global para el modelo de regresión específico, es decir, por cada unidad que varíela variable independiente predictora en la relación entre dos de sus dos categorías, con una como referencia, el valor del coeficiente de regresión es aportado al resultado de la OR_k. La manera matemática de estimar el valor del coeficiente de regresión b en el modelo de la regresión logística es:

$$\ln OR_k = \ln\left(\frac{\text{odds}_{\text{sujetos con evento o no/total de sujetos con de categoría a evaluar en la variable predictora}}}{\text{odds}_{\text{sujetos con evento o no/total de sujetos con de categoría de referencia en la variable predictora}}}\right) = b$$

si se desarrolla esta ecuación para los odds se observa que:

$$\ln OR_k = \ln\left(\frac{P_{\text{sujetos con evento y categoría a evaluar/total con categoría a evaluar}} / 1 - P_{\text{sujetos con evento y categoría a evaluar/total con categoría a evaluar}}}{P_{\text{sujetos con evento y categoría de referencia/total con categoría de referencia}} / 1 - P_{\text{sujetos con evento y categoría de referencia/total con categoría de referencia}}}\right)$$

De forma similar puede ser:

$$\ln OR_k = \ln\left(\frac{P_{\text{sujetos con evento y categoría a evaluar/total con categoría a evaluar}} / 1 - P_{\text{sujetos sin evento y categoría a evaluar/total con categoría a evaluar}}}{P_{\text{sujetos con evento y categoría de referencia/total con categoría de referencia}} / P_{\text{sujetos sin evento y categoría de referencia/total con categoría de referencia}}}\right)$$

Esto parte de la relación entre la probabilidad P y la *odds* o razón de ventaja para cierto evento según una categoría o clase específica, lo que se expresa mediante la fórmula: P = odds/1 + odds y odds = P/1 – P.

La interpretación más sencilla de la regresión logística es que cada coeficiente de regresión b (al estimar uno poblacional, se emplea la letra B o β), expresa el valor del logaritmo neperiano de la *odds ratio*, es decir, expresa lo que varía el logaritmo neperiano de las veces más frecuente de que ocurra un evento o fenómeno para una categoría respecto a otra de la variable independiente o predictora, estando ajustadas el resto de las variables independientes incluidas en el modelo o ecuación de regresión logística, es decir, a igualdad de los valores de las observaciones en el resto de las variables independientes a partir de las categorías comparadas en la variable independiente específica en el análisis.

– Error estándar (del valor del coeficiente de regresión β): expresa la variabilidad de β, es la raíz cuadrada de la varianza del coeficiente de regresión β.

– Prueba de Wald: se realiza mediante la fórmula matemática $[(B_k)/(EEB_k)]^2$ y sigue una distribución de probabilidad por χ^2, con grados de libertad igual al número total de categorías o grupos para la variable independiente o factor menos 1, para un nivel de confianza $(1 - \alpha)\%$. Esta prueba permite comparar los valores observados y esperados del coeficiente de regresión en relación a la variable predictora específica y poder evaluar su significación respecto al valor ² = 0, bajo la hipótesis de nulidad.

– Valor p (significación bajo H_0 para el estadístico de Wald): expresa la probabilidad de cumplirse la H_0 en relación al valor observado-calculado del estadístico de Wald y se expresa como:

  p-valor = $Pr[\chi^2_{grupos - 1,1 - \alpha}$ (teórico-esperado) ≥ χ^2 (observado-calculado)] bajo H_0

  o mediante la aproximación de la distribución normal estándar Z ~ N (0, 1) se expresa como:

  p-valor = Pr[Z (teórico-esperado) ≥ Z (observado-calculado) bajo H_0 .

– Valor del *odds ratio* (OR_k) global EXP(β_k): expresa globalmente las veces que es diferente una oportunidad o chance para el evento en la variable dependiente, estimada a partir del cociente de las *odds* calculadas a su vez a partir del cociente de las probabilidades del evento o no, en dos categorías (la de referencia y la propia a evaluar) para el conjunto de variables predictoras.
  - Si $OR_k > 1$ se interpreta que la oportunidad $Odds_1$ para el evento según la categoría de la variable independiente empleada corresponde al numerador de la razón OR_k, es mayor con respecto a la $Odds_2$ para el evento según la categoría de referencia de la variable independiente corresponde al denominador de la razón OR_k, por lo que entonces la variable independiente a partir de las dos categorías empleadas se comporta como factor de riesgo en el modelo para el evento.
  - Si $OR_k < 1$ se interpreta que la oportunidad $Odds_1$ para el evento según la categoría de la variable independiente empleada corresponde al numerador de la razón OR_k, es menor respecto a la $Odds_2$ para el evento según la categoría de referencia de la variable independiente corresponde al denominador de la razón OR_k, por lo que entonces la variable independiente a partir de las dos categorías empleadas se comporta como factor de protección en el modelo para el evento.
  - Si $OR_k = 1$ (razón de oportunidades es nula), es decir, no se presentan diferencias entre las ventajas para el evento o no a partir de la relación entre las dos categorías de la variable predictora involucrada en el análisis de regresión logística.

– Intervalo de confianza (IC) (1 – α) 100 %, para la OR_k estimado en relación con cada posible factor o variable predictora ($_K$):
  - IC[(1 – a)100 %] Inferior (OR_k) = Exp ($\beta_k - Z_{1 - \alpha/2} EE_k$).
  - IC[(1 – a)100 %] Superior (OR_k) = Exp ($\beta_k + Z_{1 - \alpha/} EE_k$).

  Interpretación del resultado del IC[(1 – a)100 %] (OR_k):
  - El valor nulo de una OR_k es 1 cuando el intervalo de confianza al 95 % (más empleado en investigaciones) incluya la hjpóresis nula ($OR_K = 1$), se sabe que las diferencias en la incidencia del evento o suceso en la variable dependiente (a partir del cociente de ambos *chances* o *hazards* en la variable independiente) es sugerente que no deben tener significación las diferencias en las *odds* (probabilidades no menor al 5 % que se cumpla la H_0) para los dos grupos o dos categorías a comparar, es decir, la categoría específica *versus* la de referencia en la variable independiente y, por tanto, no es muy probable que esta variable sea un buen predictor del suceso o evento en cuestión en la variable dependiente.

Los elementos anteriores involucrados en la estimación del modelo de la regresión logística (ecuación de regresión logística) se pueden presentar tabularmente en los paquetes estadísticos (Tabla 4.57).

Tabla 4.57. Elementos esenciales para el análisis de la regresión logística

Variables independientes (k) (solo están para las categorías no de referencia)	Coeficiente de regresión estimado para cada variable o factor (B_k)	Error estándar de B (EEB_k)	Estadístico de la prueba de Wald $[(B_k)/(EEB_k)]^2$	Grados de libertad (número de categorías no referencia de cada variable)	p-valor (Sig.)	exp^B (HR_k)	IC (HR_k)
1							
2							
n							
Constante (β_0) en la ecuación logística.							

Bases para el ajuste del modelo

- Verosimilitud del modelo (V): expresa el grado en que el modelo arroja resultados coherentes con los datos usados para su construcción, es el producto de todas las probabilidades (predichas por el modelo) de que los n sujetos de la muestra empleada para su construcción tengan la condición que realmente tienen. Se formula como:
  $V = [p_1 p_2 \ldots p_d][(1 - p_{d+1})(1 - p_{d+2}) \ldots (1 - p_n)]$
  donde:
  - Los primeros d factores corresponden a sujetos con la condición y los restantes n – d a los que no la tienen.
  - Su valor siempre es mayor de cero.
  - Si su valor se aproxima a 1 el modelo es exitoso, eficiente.
  - Si su valor se aproxima a 0 el modelo es deficiente.
- Lejanía del modelo (L): expresa el grado en que el modelo arroja resultados no coherentes (cuan alejados están) con los datos usados para su construcción. Se divide en:
  - (L): es la que corresponde propiamente al modelo que se ha ajustado para las variables independientes incluidas.
  - (L0): es la que corresponde al "modelo nulo" en que no se ha incorporado ninguna variable independiente.
    La lejanía del modelo nulo es más grande que la de cualquier modelo ampliado. Se formula la lejanía (L) como:

$L = -2\ln V$

donde:

- ln es el logaritmo natural de la verosimilitud.
- Como $V < 1$ su logaritmo natural siempre es negativo, de modo que la lejanía (L) siempre es un número positivo (debido a la multiplicación de valores con signos negativos).

  Cuanto mayor sea la verosimilitud (V), o sea, más próxima a 1, entonces más próximo a cero es el valor de su logaritmo natural, pero con signo negativo y por ello su multiplicación por –2 en la fórmula anterior da un resultado cada vez mas próximo a 0, pero positivo en la lejanía (L). De esto se deriva que el grado de ajuste de un modelo es mejor cuanto más próxima a 1 es la verosimilitud y, en consecuencia, cuanto más se aproxima a cero la Lejanía.

– Cociente o razón de verosimilitudes (CV): expresa la diferencia entre ambas lejanías y mide "el aporte" que hacen las variables incorporadas al modelo, es útil para determinar si hay una diferencia significativa entre incluir en el modelo todas las variables y no incluir ninguna, o sea, valorar si las variables X_1, X_2 … X_r tomadas en conjunto, contribuyen efectivamente a "explicar" las modificaciones que se producen en P (Y = 1). Se formula como:

$$CV = L_0 - L = -2\ln V_0 + 2\ln V = -2\ln(V_0 - V) = -2\ln\left(\frac{V_0}{V}\right)$$

Relevancia del cociente o razón de verosimilitudes (Colectivos de autores, 2014)

Cociente o razón de verosimilitudes (Silva, 1995)

– Planteamiento del contraste de Hipótesis:
H_0: $\beta_1 = \beta_2 = \ldots = \beta_r = 0$ *vs*. H_1: $\beta_1 = \beta_2 = \ldots = \beta_r \neq 0$.

– Estadístico de contraste (de prueba):
$R = 2(\text{Ln}V_F - \text{Ln}V_I)$ es un estadístico que se distribuye como una variable χ^2 con r grados de libertad

donde:

r: número de variables presentes en el modelo ampliado;

Ln: logaritmo natural o neperiano;

V_I o V_o: verosimilitud inicial (sin las variables incluidas en el modelo) y se formula:

$$V_I = \exp\left[n_0 \ln\left(\frac{n_0}{n}\right) + n_0 \ln\left(\frac{n_1}{n}\right)\right]$$

n_0: número de observaciones con Y = 0;

n_1: número de observaciones con Y = 1;

$n = n_0 + n_1$: número total de observaciones;

V_F o V: verosimilitud final (con las variables incluidas en el modelo) y se formula:

$$VF = \prod_{i=1}^{n} \hat{p}_i^{y_i} (1-\hat{p}_i)^{1-y_i}$$

y_i: valor de la variable Y en la i-ésima observación, i = 1,..., n;

$\hat{p}_i$: probabilidad predicha por el modelo final para la i-ésima observación, i = 1,..., n;

K: número de variables explicativas en el modelo y se corresponde con el número de grados de libertad;

χ^2_k: variable χ^2con k grados de libertad.

– Valor p (p-valor), bajo H_0 (valores de probabilidad donde se cumpliría la hipótesis nula):

$$\text{p-valor} = \Pr\left\{\chi_k^2 > R\right\}$$

El empleo del cociente o razón de verosimilitudes permite determinar si hay una diferencia significativa entre incluir en el modelo todas las variables y no incluir ninguna; es decir, sirve para valorar si las variables X_1, X_2 ... X_r tomadas en conjunto, contribuyen efectivamente a "explicar" las modificaciones que se producen en P (Y = 1) y así mismo, permite valorar el aporte atribuible a cierto conjunto de variables adicionadas a las de un primer ajuste. En efecto, si se ajusta un modelo que produce cierto valor CV1 (cociente de verosimilitudes 1) y se ajusta otro al que se agregaron hvariables, el cual produce un cociente CV2, entonces CV2-CV1 permite evaluar si la adición de las variables hace un aporte significativo.

Bases para la calidad del ajuste del modelo (Colectivo de autores, 2014)

Cuando se quiere obtener un modelo de regresión, de cualquier tipo, una precaución importante a los efectos de sacar conclusiones es la de corroborar que este modelo se ajusta efectivamente a los datos usados, lo que es aplicable a la regresión logística. A continuación se muestran algunos coeficientes que facilitan el análisis de la calidad del ajuste del modelo estimado para este tipo de regresión.

Coeficiente de determinación $(R)^2$ (Mittlböck y Schemper,1996)

– Es un indicador empleado para expresar el grado de explicación de la variabilidad en la variable de respuesta (dependiente) conseguido con el modelo, por parte de las variables explicativas (independientes).
– Se halla necesariamente entre 0 y 1.
– Alcanza el valor 1 cuando el modelo atribuye probabilidad 1 a los sujetos de la muestra que efectivamente tienen el evento.
– Alcanza el valor 0 cuando el modelo atribuye probabilidad 0 a los sujetos de la muestra que no tienen el evento.
– Se aproxima a 0 en la medida que las probabilidades atribuidas por el modelo disten más, respectivamente, de 1 y 0. Se formula como:

$$R^2 = \frac{\left(\sum_{i=1}^{n}(y_j - \overline{p})(p_j - \overline{p})\right)^2}{\sum_{i=1}^{n}(y_j - \overline{p})^2 \sum_{i=1}^{n}(p_j - \overline{p})^2}$$

donde:

n: dos veces el número de configuraciones distintas de las variables explicativas;

y_j: valor de la respuesta (0 o 1) en el grupo j;

p_j: probabilidad predicha en el grupo j;

p: es la proporción de sujetos con respuesta Y = 1 en la muestra.

Coeficiente de Cox y Snell (Cox y Snell, 1989)

Puede ser considerado, en cierto sentido, variaciones del coeficiente de determinación y se formula:

$$R^2_{CS} = 1 - \left(\frac{\ln V_I}{\ln V_F}\right)^{\frac{2}{n}}$$

Coeficiente de Nagelkerke (Nagelkerke, 1991)

Puede ser considerado, en cierto sentido, variaciones del coeficiente de determinación y se formula:

$$R^2_N = \frac{R^2_{CS}}{1 - (\ln V_I)^{\frac{2}{n}}}$$

donde:

$VI = \exp\left[n_0 \ln\left(\frac{n_0}{n}\right) + n_0 \ln\left(\frac{n_1}{n}\right)\right]$: verosimilitud inicial;

$VF = \prod_{i=1}^{n} \hat{p}_i^{y_i}(1-\hat{p}_i)^{1-y_i}$: verosimilitud final;

n_0: número de observaciones con Y = 0;

n_1: número de observaciones con Y = 1;

$n = n_0 + n_1$: número total de observaciones;

r: número de variables explicativas;

y_i: valor de la variable Y en la i-ésima observación, i = 1,...,n.

$\hat{p}_i$: probabilidad predicha por el modelo final para la i-ésima observación, i = 1,..., n;

$\overline{p}$: proporción de observaciones con Y = 1.

Prueba de bondad de ajuste de Hosmer y Lemeshow (Lemeshow y Hosmer, 1982; Colectivo de autores, 2014)

Una vez obtenida la ecuación logística, la que constituye el modelo predictor estimado por la regresión logística con la incorporación de los coeficientes de regresión de las variables independientes, se le debe dar solo un carácter preliminar, ya que antes de considerarla

como definitiva debe evaluarse cómo se ajusta este modelo estimado a los datos de la muestra de estudio. Ajustarse bien supone que globalmente las predicciones realizadas a partir del modelo estimado o ecuación de regresión logística no están lejos de la realidad observada a partir de los datos muestrales. Existen diferentes técnicas para verificar la bondad de ajuste, pero para modelos que incluyan al menos una variable cuantitativa predictora independiente la prueba más aceptada es la prueba de Hosmer-Lemeshow.

Esta prueba consiste en observar detalladamente cómo se separan los valores observados (número de sujetos con el suceso o evento en la variable de respuesta Y = 1 y número de sujetos sin el suceso Y = 0) respecto de los valores esperados o estimados en cada caso según las predicciones del modelo de regresión logística ajustado.

El procedimiento previo para realizar la prueba de Hosmer-Lemeshow consiste inicialmente en determinar las probabilidades predichas por el modelo estimado de la regresión logística para cada sujeto de la muestra (sustituir los datos obtenidos de cada sujeto para cada variable predictora empleada en la ecuación de regresión logística) y con los resultados obtenidos de probabilidad en los sujetos de tener el evento en la variable de respuesta o dependiente, definir los deciles de riesgo detener el evento, conformándose los respectivos intervalos de este modo:

- Se ordenan las n probabilidades p_j.
- El i-ésimo decil (i = 1, 2, ..., 9) es la probabilidad que ocupa la posición d(i), siendo d(i) el primer índice que cumple:

  $$\sum_{j=1}^{d(i)} m_j \geq \frac{m \times i}{10}$$

- Los 10 grupos están definidos por los intervalos:

  $$\left[p_1, p_{d(1)}\right], \left(p_{d(1)}, p_{d(2)}\right], \ldots, \left(p_{d(9)}, p_n\right]$$

  A veces se obtienen menos de 10 grupos porque hay deciles que coinciden, denotados por:

  p_j: las probabilidades predichas por el modelo de presentar el evento en la variable dependiente o de respuesta;

  n: el número de valores distintos de p_j;

  m_j: el número total de observaciones con probabilidad predicha p;

  m: el número total de observaciones;

  y_j: el número de observaciones con probabilidad predicha p_j y con valor 1.

  Se comparan el número de sujetos según sus valores observados y esperados en la variable de respuesta o dependiente en 10 categorías correspondientes aproximadamente a los deciles de riesgo o probabilidad predicha de presentar el evento en la variable de respuesta o dependiente. Se crean tabularmente 20 casillas en total (10 categorías y dos casillas en cada categoría, una para quienes no tienen el evento, Y = 0 y otra para quienes lo tienen, Y = 1).

El valor esperado para quienes tienen el evento en la variable dependiente Y = 1 se obtiene sumando los valores pronosticados o predichos obtenidos a partir de la aplicación de la ecuación de regresión logística para los sujetos observados incluídos en cada categoría o intervalo determinado según los deciles de riesgo previamente establecidos. Los valores esperados para quienes no tienen el evento se obtienen restando el valor esperado de quienes tienen el evento del total de sujetos observados de la categoría o decil correspondiente.

El estadístico de contraste o de prueba propuesto por los autores para efectuar la prueba de hipótesis o el contraste de hipótesis respecto a la comparación de los valores observados y esperados por el modelo estimado en la muestra de estudio es el chi-cuadrado de Hosmer-Lemeshow:

$$\chi^2_{\text{Hosmer-Lemeshow}}(\text{observado-calculado}) = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i} + \sum_{i=1}^{g} \frac{(O_i^* - E_i^*)^2}{E_i^*}$$, que sigue una distribución χ^2 con grados de libertad igual al número de grupos o deciles – 2 (10 grupos en este caso y 8 grados de libertad) para el valor de alfa prefijado

donde:

$$O_i = \sum_{d(i-1)<j\leq d(i)} y_j \ ;$$

$$O_i^* = \sum_{d(i-1)<j\leq d(i)} m_j - y_j \ ;$$

$$E_i = \sum_{d(i-1)<j\leq d(i)} m_j p_j \ ;$$

$$E_i^* = \sum_{d(i-1)<j\leq d(i)} m_j (1-p_j)$$

O_i: valor observado de sujetos con el evento ausente Y = 0, para la categoría i o el decil i;

O_i^* : valor observado de sujetos con el evento presente (Y = 1), para la categoría ió el decil i;

E_i: valor esperado o suma de probabilidad estimada por el modelo de los sujetos con el evento ausente (Y = 0), para la categoría ió el decil i;

E_i^* : valor esperado o suma de probabilidad estimada por el modelo de los sujetos con el evento presente (Y = 1), para la categoría i o el decil i;

d: representa cada decil de riesgo seleccionado, según la probabilidad estimada por el modelo para los sujetos de la muestra de estudio;

g: representa el valor de la última categoría i o decil i.

La representación gráfica de los elementos mencionados a utilizar en la determinación de la bondad en el ajuste del modelo de predicción estimado respecto a los datos muestrales, mediante la prueba de Hosmer-Lemechow, se muestra a continuación (Tabla 4.58).

Tabla 4.58. Representación resumida de los elementos a utilizar en la prueba de Hosmer-Lemeshow para evaluar el ajuste del modelo estimado en la regresión logística

		Evento en la variable dependiente: ausente (Y = 0)		Evento en la variable dependiente: presente (Y = 1)		Total
		Observado	Esperado	Observado	Esperado	
		$O_i = \sum_{d(i-1)<j\leq d(i)} y_j$	$E_i = \sum_{d(i-1)<j\leq d(i)} m_j p_j$	$O_i^* = \sum_{d(i-1)<j\leq d(i)} m_j - y_j$	$E_i^* = \sum_{d(i-1)<j\leq d(i)} m_j(1-p_j)$	
Paso 1	1					
	2					
	3					
	4					
	5					
	6					
	7					
	8					
	9					
	10					
Total						

– Valor de p (p-valor) bajo H_0:
p-valor = Pr [($\chi^2_{g-2,\ 1-\alpha}$teórico-esperado ≥ χ^2 de Hosmer Lemechow (observado-calculado)].

Regla de decisión

Dado que en este caso la hipótesis nula (H_0) plantea que no hay diferencias entre las distribuciones de los valores observados muestrales y los pronosticados o predichos por el modelo estimado en relación a la variable de respuesta o predictora para el resultado del evento (Y = 1), entonces:

– Si χ^2 de Hosmer-Lemechow (observado-calculado) ≤ $\chi^2_{g-2,\ 1-\alpha}$ (teórico-esperado), entonces no se ubica en el área de rechazo al valor estimado de χ^2 de Hosmer-Lemechow observado-calculado, por lo que no hay razones suficientes para rechazar H_0, es decir, que el valor de p (para cumplirse H_0) a partir del χ^2 observado-calculado es mayor o igual que el valor de alfa prefijado y con esto aumenta la probabilidad de cometer el error tipo 1, lo que sugiere que el ajuste del modelo a partir de los datos observados (reales) es adecuado para el nivel de confianza preseleccionado por el investigador.
Más allá de la significación estadística de la chi-cuadrado, resulta muy importante valorar una a una las categorías o deciles de riesgo y ver si hay disparidad entre

observados y esperados. Esa observación indica las regiones de la predicción del modelo que se ajusta peor a los datos. Cuando no se ajuste bien el modelo a los datos, deben especificarse de otra manera las variables independientes, recategorizarlas, plantearse introducir interacciones (términos de producto) o términos cuadráticos y repetir de nuevo esta prueba de Hosmer-Lemeshow.

Curva ROC

Al igual que la prueba de Hosmer-Lemechow, la construcción y análisis de curvas ROC (Receiver Operating Characteristic), que fue un método desarrollado inicialmente para evaluar la efectividad en la recepción de señales de radares), permiten evaluar la bondad de ajuste del modelo de regresión logística estimado, empleando la probabilidad pronosticada en relación con los datos observados (reales) en la muestra de estudio.

Solo si se pretende utilizar el modelo logístico con carácter predictivo, debe construirse una curva ROC y hallar el área bajo la curva con su correspondiente límite de confianza. Esta área proporciona, además, la capacidad de discriminación del modelo logístico, cuanto mayor sea, más eficiente es el modelo.

El poder discriminatorio es la capacidad de clasificar correctamente a los sujetos según su evento, equivale a la capacidad de conseguir que los sanos sean clasificados como sanos y los enfermos sean clasificados como enfermos empleando el modelo de regresión logística.

La curva ROC se obtiene representando, en un cuadrado de lado 1, los valores de especificidad (porcentaje de sujetos observados con respuesta en la variable dependiente para el evento Y = 0 y que son clasificados incorrectamente con la presencia de esta condición del evento Y = 1 por el modelo estimado) se ubica en el eje de las abscisas o eje x, frente a sensibilidad (porcentaje de sujetos observados que cumplen con la condición Y = 1, que son clasificados correctamente con la presencia de esta condición del evento Y = 1 por el modelo estimado) lo cual queda ubicado en el eje de las ordenadas o eje y, para cada punto de corte considerado en todo su trayecto.

Si se fija un "punto de corte", un valor cualquiera entre 0 y 1, se puede clasificar las n probabilidades predichas o pronosticadas (resultados obtenidos al aplicar la ecuación de regresión logística como modelo estimado a los valores de las variables para cada sujeto) en una tabla de contingencia 2x2, por una parte se tienen los valore resultante que están por debajo o por arriba de este punto de corte (prueba positiva o negativa respectivamente predicha por el modelo estimados) y por otra las que corresponden a sujetos que presentan el evento (respuesta Y = 1) y las que corresponden a quienes no lo presentan (respuesta Y = 0) de forma real entre los sujetos observados (Tabla 4.59).

Tabla 4.59. Relación entre lo observado en la muestra de estudio y lo pronosticado por el modelo de la regresión logística para la variable dependiente

Observado en la muestra	Pronosticado por el modelo (por la ecuación de regresión logística)		Porcentaje correcto
	Evento de la variable dependiente pronosticado		
	Ausente	Presente	
Evento de la variable dependiente: ausente	A (verdaderos negativos)	B (falsos positivos)	A/(A + B) · 100 %
Evento de la variable dependiente: presente	C (falsos negativos)	D (verdaderos positivos)	D/(D + C) · 100 %
Porciento global			[A/(A + B)] + [D/(D + C)]/2 · 100 %

Donde:

- Sensibilidad = Verdaderos positivos/Total de observados con el evento = D/(D + C)% (para el caso de la curva ROC interesa este valor para estimar el valor de la abscisa o eje x en cada punto de corte).
- Especificidad = Verdaderos negativos/Total de observados sin el evento = A/(A + B)%
- Complementario de la Especificidad = 1 – Especificidad = 1 – [A/(A + B)]% (para el caso de la curva ROC a partir de esto interesa el valor de [B (A + B)]% para estimar el valor de la ordenada o eje y en cada punto de corte).

Desde esta perspectiva puede considerarse el modelo de regresión logística como un medio para definir una prueba diagnóstica cuantitativa, en este caso para definir el poder discriminatorio del modelo estimado o ecuación de regresión logística.

En la curva ROC se presentan ambas características (sensibilidad y especificidad) del modelo logístico cuando el punto de corte se va variando.

A medida que se va cambiando el punto de corte (el cual toma valores desde 1 hasta 0 partiendo del origen de coordenadas) se obtienen diferentes valores de sensibilidad y de (1 – especificidad) que son las coordenadas para representar cada punto o par x, y gráficamente. La unión de estos puntos conforma la curva ROC.

El procedimiento para obtener estos pares de sensibilidad y 1 – especificidad parte de comparar los valores estimados por el modelo de regresión logística para cada individuo (al sustituir los valores de sus datos para cada variable independiente en la ecuación previamente estimada de regresión logística) *versus* el valor del punto de corte específico seleccionado en un rango que va desde 1 a 0. Los valores predichos por el modelo para cada sujeto, que sean superiores al valor de cada punto de corte, se asumen como suje-

tos con respuesta predicha por el modelo estimado de tipo afirmativa para el evento en la variable de respuesta o dependiente (Y = 1). De este número de sujetos se determina cuantos fueron observados con respuesta afirmativa para el evento en la variable dependiente (Y = 1), cuyo dato es empleado junto al total de sujetos observados con Y = 1 en la muestra para realizar el cálculo de la sensibilidad, lo que es a su vez empleado como la coordenada y u ordenada en el gráfico; mientras que los sujetos restantes que fueron observados con respuesta no afirmativa para el evento en la variable dependiente (Y = 0), cuyo dato es empleado junto al total de sujetos observados con Y = 0 en la muestra para realizar el cálculo del complementario de la especificidad (1 – especificidad), lo que es a su vez empleado como la coordenada x o abscisa, obteniéndose de esta forma en conjunto las coordenadas x, y de cada par o punto ordenado (1 –especificidad, sensibilidad) para cada punto de corte específico, los que al unirse forman el gráfico de la curva ROC de cada modelo estimado específico de la regresión logística.

Cada curva ROC empieza en el punto con coordenada (0, 0), que corresponde al punto de corte con valor 1 y termina en (1, 1) que se obtiene al considerar el 0 como el último punto de corte a emplear. Si el modelo tiene capacidad predictiva nula la curva coincide con la diagonal principal del cuadrado y el área bajo la curva toma su valor mínimo de 0,5. Por el contrario, un modelo perfecto tiene una curva ROC con área 1.

Formulario a emplear en la estimación de la curva ROC (Silva, 1997):

- Área bajo la curva ROC (Δ):

$$\Delta = \frac{1}{ua}\sum_{i=1}^{n}\left(u_i A_i + \frac{u_i a_i}{2}\right)$$

- Error estándar del área bajo la curva ROC [EE(Δ)]:

$$EE(\Delta) = \sqrt{\frac{1}{ua}\left[\Delta(1-\Delta) + (a-1)(U-\Delta^2) + (n-1)(v-\Delta^2)\right]}$$

- Intervalo de confianza inferior o límite inferior para el área bajo la curva ROC con nivel de confianza (1 – α)%:

$$\text{IC Inferior} = \Delta - Z_{1-\alpha/2}EE(\Delta)$$

- Intervalo de confianza superior o límite superior para el área bajo la curva ROC con nivel de confianza (1 – α)%:

$$\text{IC Superior} = \Delta + Z_{1-\alpha/2}EE(\Delta)$$

donde:

m: número de categorías en que se dividen las n observaciones a partir de las probabilidades predichas por el modelo;

a_i: número de observaciones con Y = 1 en la i-ésima categoría, i = 1,..., m;

u_i: número de observaciones con Y = 0 en la i-ésima categoría, i = 1,..., m;

$a = \sum_{i=1}^{m} a_i$ es el número total de observaciones con Y = 1;

$u = \sum_{i=1|}^{m} u_i$: número total de observaciones con Y = 0;

$A_i = a - \sum_{j=1}^{i} a_j$, i = 1,..., m

$U_i = \sum_{j=1}^{i-1} u_j$, j = 2,..., m y $U_1 = 0$

$U = \frac{1}{ua^2} \sum_{i=1}^{m} u_i \left(A_i^2 + A_i a_i \frac{a_i^2}{3} \right)$ y

$V = \frac{1}{au^2} \sum_{i=1}^{m} a_i \left(U_i^2 + U_i u_i \frac{u_i^2}{3} \right)$

$z_{1-\alpha/2}$: percentil de la distribución normal estándar N (0, 1), que deja a la izquierda una cola de probabilidad $1 - \alpha/2$.

$1 - \alpha$: nivel de confianza.

Valores predictivos:

- Valor predictivo positivo (VPP):

$$VPP = \frac{S\hat{p}_i}{S\hat{p}_i + (1-E)(1-\hat{p}_i)}$$

- Valor predictivo negativo (VPN):

$$VPN = \frac{(1-S)\hat{p}_i}{(1-S)\hat{p}_i + E(1-\hat{p}_i)}$$

– Intervalo de confianza *bootstrp* para el valor predictivo positivo con nivel de confianza $(1 - \alpha)$% (Efron-Tibshirani, 1993):

$$\left(\frac{S\hat{p}^*_{i,\frac{\alpha}{2}}}{S\hat{p}^*_{i,\frac{\alpha}{2}} + (1-E)(1-\hat{p}^*_{i,\frac{\alpha}{2}})}, \frac{S\hat{p}^*_{i,1-\frac{\alpha}{2}}}{S\hat{p}^*_{i,1-\frac{\alpha}{2}} + (1-E)(1-\hat{p}^*_{i,1-\frac{\alpha}{2}})} \right)$$

– Intervalo de confianza *bootstrp* para el valor predictivo negativo con nivel de confianza $(1 - \alpha)$% (Efron-Tibshirani, 1993):

$$\left(\frac{(1-S)\hat{p}^*_{i,\frac{\alpha}{2}}}{(1-S)\hat{p}^*_{i,\frac{\alpha}{2}} + E(1-\hat{p}^*_{i,\frac{\alpha}{2}})}, \frac{(1-S)\hat{p}^*_{i,\frac{\alpha}{2}}}{(1-S)\hat{p}^*_{i,\frac{\alpha}{2}} + E(1-\hat{p}^*_{i,\frac{\alpha}{2}})} \right)$$

donde:

S: sensibilidad;

E: especificidad;

$\hat{p}_i$: probabilidad predicha por el modelo para la i-ésima observación, i = 1,..., n;

$\hat{p}^*_{i,\frac{\alpha}{2}}$: percentil de orden $\alpha/2$ de las β = 1,000 estimaciones bootstrap; $(\hat{p}_i^{*(b)}, b = 1,...,B)$ de la probabilidad predicha por el modelo, i = 1,..., n;

$\hat{p}^{*}_{i,1-\frac{\alpha}{2}}$: es el percentil de orden 1– $\alpha/2$ de las β = 1000 estimaciones bootstrap

$(\hat{p}_i^{*(b)}, b = 1,...,B)$ de la probabilidad predicha por el modelo, i = 1,..., n;

1 – α: nivel de confianza.

Validación del modelo de la ecuación de regresión logística estimada

Para que la aplicación de un modelo estimado por regresión logística sea fructuoso en otro contexto, el modelo debe ser validado con datos procedentes de ese otro contexto. El acto de corroborar que tiene este mérito, o sea, que ocurran las predicciones que se supone que hace, se conoce como "validación" del modelo. Para esto se procede en esencia de este modo:

- Se construye el modelo usando una muestra 1 de tamaño n_1.
- Se busca una muestra 2, independiente de la primera, de tamaño n_2, de la que se tiene toda la información, tanto los datos "de entrada" X_1, X_2 … X_r, como el valor de Y para cada uno de sus elementos.
- Se aplica el modelo usando una muestra 1 de tamaño n_1 a cada vector X_1, X_2 … X_r de la muestra 2 y se obtienen n_2 valores de P_i.
- Se valora el grado en que los n_2 valores de P_i obtenidos "se parecen" a los respectivos valores de Y.

De forma general, en cualquier caso, para realizar la validación ha de proveerse una nueva base de datos y aplicar el modelo que se acaba de construir a cada uno de los sujetos de esta base. Con los verdaderos desenlaces acaecidos y con las estimaciones resultantes de la aplicación del modelo se aplica la prueba de bondad de ajuste de Hosmer-Lemeshow y luego se estima el número esperado de casos con la condición mediante la suma de las probabilidades obtenidas. La comparación de los valores esperados bajo el modelo que se valida y los resultados objetivamente producidos previamente en el modelo inicial, tanto en un caso como en el otro, permite conformar un juicio sobre la validez del modelo.

La aplicación práctica del modelo estimado y validado de la regresión logística en un sujeto o individuo se puede poner de manifiesto si se desea conocer la probabilidad del evento en la variable dependiente seleccionada en este modelo a partir del comportamiento de las respectivas variables independientes incluidas. Para esto se realiza la sustitución de los valores observados del sujeto para las respectivas variables independientes incluidas en este modelo se multiplican estos valores por los correspondientes coeficientes de regresión de estas variables y los valores obtenidos se suman para obtener un resultado que se corresponde con el valor del logaritmo neperiano de la OR o razón de odds para la ocurrencia o no del evento en la variable dependiente. A continuación se calcula el antilogaritmo para este resultado mediante la expresión:

$\exp^{\beta_0 + \beta_1 X_1 + \ldots + \beta_r X_r}$

Este valor *odds ratio* u OR obtenido expresa las veces más frecuente que este sujeto presenta el evento en la variable dependiente para el modelo de regresión logística empleado, a partir de las respectivas variables independientes utilizadas. En caso de que se desee conocer la probabilidad o riesgo absoluto que el modelo de regresión logística predice para un sujeto de tener una respuesta en la variable dependiente para el evento de tipo presente, se puede efectuar el cálculo del logaritmo neperiano del *odds ratio* según las variables escogidas del modelo y los valores para estas variables de ese sujeto. El resultado obtenido de la *odds ratio* es sustituido en la fórmula que vincula esta con la probabilidad P y así obtener el valor de probabildad que predice el modelo de regresión logística según: P = odds/1 + odds, cuyo valor P calculado es expresado en porciento al multiplicarlo por 100.

Método multivariado de análisis de sobrevida o supervivencia

Inicialmente se debe hacer referencia concepto de análisis de sobrevida, que consiste en el conjunto de métodos destinados a estudiar la ocurrencia de distintos eventos, entre ellos la muerte, en grupos de individuos seguidos a lo largo del tiempo, es decir, de forma longitudinal.

Concepto de curvas de sobrevida

Las curvas de sobrevida son representaciones gráficas que dan la proporción de individuos, de un conjunto inicial, que se mantienen con vida en diferentes momentos a lo largo de un tiempo de observación.

Son muy utilizadas para tener una imagen visual de la forma en que la mortalidad afecta a distintos grupos de individuos, por ejemplo, en el seguimiento de distintos tratamientos, en la búsqueda de diferencias que puedan tener importancia en la futura selección de las mejores opciones terapéuticas, en la ocurrencia de recaídas, las intervenciones o reintervenciones quirúrgicas, los ingresos hospitalarios, entre otros, aunque también puede estar influenciada por circunstancias médicamente favorables como la remisión de determinados procesos o síntomas. Cualquiera de estas situaciones se denomina evento y las técnicas estudian la incidencia de eventos y el tiempo transcurrido hasta el evento. El tiempo que transcurre hasta la presentación de los eventos se suele designar como tiempo de sobrevida, tiempo de seguimiento o tiempo al evento, aunque son utilizados también los términos *failure time*, para el caso de ensayos terapéuticos se emplea *trial time*, comenzando este tiempo a contarse de cero en el momento de la inclusión del individuo en el seguimiento.

Es frecuente observar la proporción de pacientes libre de eventos en diferentes momentos del seguimiento lo que permite calcular la proporción de individuos libre de

eventos o sobreviviente, por ejemplo, a los seis meses, al año, entre otros, mientras en otras oportunidades se informa la proporción de individuos dentro de un grupo que han sufrido un evento durante un tiempo de observación o seguimiento determinado, lo que se conoce como la incidencia del evento durante este intervalo.

Para estimar la probabilidad de supervivencia individual acumulada a lo largo del tiempo suele utilizarse el método de Kaplan-Meier, que no asume una distribución particular de los datos (no paramétrico).

Construcción de curvas de sobrevida o de supervivencia

El método de Kaplan-Meier es el seleccionado frecuentemente en la construcción de curvas de sobrevida o de supervivencia, que no asume una distribución particular de los datos (es un método no paramétrico), proporciona herramientas de análisis efectivo y procede de esta forma: teniendo n_j individuos en seguimiento, si en un día cualquiera se producen d_j eventos (uno o más eventos), la probabilidad de evento estimada para el tiempo t_j es:

$q_j = d_j/n_j$

La probabilidad de sobrevida (p_j) en un tiempo dado ($_j$) se estima como la proporción entre el número de individuos que se hallan con vida en el momento del control, es decir, sin el evento (d_j) que equivale a ($n_j - d_j$) y el total de individuos ingresados al estudio (n_j).

$p_j = 1 - q_j = 1 - (d_j/n_j) = (n_j - d_j)/n_j$

El complemento de la probabilidad de sobrevida (p_j), es la probabilidad de óbito o evento (q_j) igual a ($1 - p_j$):

$q_j = d_j/n_j$

La sobrevida a lo largo de varios periodos (l_t) se estima multiplicando las probabilidades para cada uno de ellos (p_j), que se tratan como probabilidades independientes:

$l_t = \Pi p_j = \Pi[(n_j - d_j)/n_j]$, expresando el símbolo Π la multiplicación de un grupo de probabilidades de sobrevida de los periodos anteriores hasta el tiempo ($_j$).

En los días en los que no hubo eventos $d_j = 0$ y por lo tanto $p_j = (n_j - 0)/n_j = 1$ y como multiplicar por 1 no altera el producto, entonces los días en que no hubo eventos no entran en los cálculos.

El estimador (l_j) se conoce como producto límite, por ser un producto donde a los intervalos tienden a un límite mínimo.

Las curvas de vida por el Método de Kaplan-Meier requieren previamente para su conformación de las llamadas *tablas de vida*, que recogen toda la información necesaria como sigue:

– Ordenar los datos de menos a más según el tiempo de supervivencia (de observación) empleando el código: 1 = muerte, 0 = sobrevive (Tabla 4.60).

Tabla 4.60. Modelo para la recogida de datos iniciales en el estudio de sobrevida

Número de orden del paciente	Muerte	Años

– Construir una tabla de sobrevida propiamente (Tabla 4.61) con:
  - La primera columna (t_j) corresponde a los tiempos de observación (por ejemplo, medidos en años). Se inicia un nuevo tiempo solo cuando alguien fallece.
  - La segunda columna (n_j) corresponde al número de individuos que están a riesgo defallecer en ese periodo de tiempo. Son los que entran vivos en ese periodo. Se incluye el individuo o individuos que van a morir al inicio de ese periodo.
  - La tercera columna (d_j) se corresponde a los que mueren en ese periodo de tiempo.
  - La cuarta columna (s_j) son los supervivientes ($s_j = n_j - d_j$).

Tabla 4.61. Representación gráfica para la recolección de los datos necesarios en el estudio de sobrevida

Tiempos (t_j)	Entran (n_j)	Mueren (d_j)	Sobreviven (s_j)

– Calcular para cada tiempo (expresado como el intervalo entre el tiempo inicial y el que le sigue) el cociente de los sujetos sobrevivientes (s_j) entre los que tienen riesgo de fallecer, (n_j) que representa la probabilidad de sobrevida para el periodo (p_j) (Tabla 4.62).

Tabla 4.62. Representación gráfica para la recolección de los datos necesarios en la determinación de la probabilidad de sobrevida

Tiempos en intervalos (t_j)	A riesgo o que entran (n_j)	Mueren (d_j)	Sobreviven (s_j)	Cociente $(s_j)/(n_j) = (p_j)$.

– Multiplicar en cada tiempo los cocientes $[(s_j)/(n_j)]$ por los de los tiempos anteriores. La supervivencia acumulada para cada tiempo es precisamente este producto:
  $S_j = \Pi[(s_j)/(n_j)]$
  donde:
  Π: indica la multiplicación por el cociente $[(s_j)/(n_j)\ ] = (p_j)$ del intervalo anterior de tiempo (Tabla 4.63).

Tabla 4.63. Representación gráfica para la recolección de los datos necesarios para la estimación de la supervivencia acumulada

Tiempos en intervalos (t_j)	A riesgo o que entran (n_j)	Mueren (d_j)	Sobreviven (s_j)	Cociente $[(s_j)/(n_j)]$.	Supervivencia acumulada (Kaplan-Meier) $S_j = \Pi[(s_j)/(n_j)]$

– Representar gráficamente los datos anteriores empleando el estimador de Kaplan-Meier (supervivencia acumulada de los sujetos para cada intervalo de tiempo) en un sistema de ejes con el tiempo en las abscisas (x) y la proporción de sobrevivientes en las ordenadas (y), se inician en tiempo cero y con el 100 % de los individuos con vida, presentando una forma escalonada (con tantos escalones como días o intervalos con eventos se presenten), descendente hacia la derecha y abajo, a medida que la proporción de sobrevivientes disminuye con el tiempo (Fig. 4.17). Se puede estimar la supervivencia acumulada mediana, que es aquel tiempo en el que aún están vivos el 50 % de los sujetos incluidos en el estudio, en el caso que al concluir el tiempo de seguimiento hayan sobrevivido menos del 50 % de los que tuvieron riesgo de fallecer.

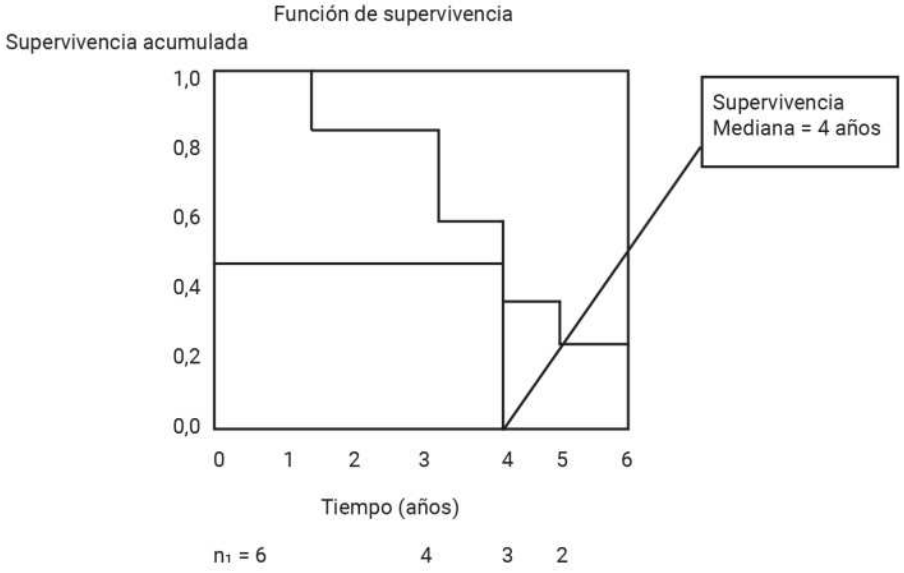


Fig. 4.17. Representación de una función de supervivencia, empleada en el análisis de sobrevida o de curvas de vida.

A partir de la supervivencia acumulada (estimador) en cada muestra o grupo y su error estándar para un tiempo determinado se puede estimar también la supervivencia poblacional por intervalos de confianza como:

$IC_{(1-\alpha)\%}$ de $S_{pob} = S_t^{EXP \pm Z\alpha/2\, EE_t}$

donde:

EE: error estándar transformado de la sobrevida acumulada para un intervalo de tiempo en la muestra o del grupo respectivo y se estima como:

$$EE_t = \sqrt{\frac{1}{\left[\ln(S_r)\right]^2} \cdot \sum \frac{n_i - s_i}{n_i s_i}}$$

S_{pob}: supervivencia o sobrevida poblacional;

S_t: supervivencia acumulada;

EXP: valor base del logaritmo neperiano;

$Z_{\alpha/2}$: percentil de la distribución normal para el respectivo error alfa ($\alpha/2$).

Estas curvas de sobrevida son la expresión gráfica de toda una elaboración teórica que permite su cálculo numérico, incluyendo los métodos para la construcción de modelos matemáticos que las representan, las técnicas para la comparación de varias curvas entre sí y también para la evaluación de covariables que puedan tener influencia sobre la producción de los eventos.

En el caso de que los individuos que por algún motivo no completan el tiempo de seguimiento establecido (observaciones censuradas o *censored observations*), cuyas dos causas fundamentales son las pérdidas de contacto o retiradas durante el seguimiento (*lost to folow up* o *withdrawals*) y el fin del protocolo antes de que algún o algunos individuos hayan cumplido su seguimiento, hay técnicas de análisis para tratar con estos casos sin descartarlos del estudio, que se basan en que el individuo en cuestión estaba vivo, es decir, es sobreviviente (información aprovechable) por lo menos hasta el momento del último contacto. Sin embargo; a pesar de que estas técnicas son efectivas, la recomendación es hacer todo lo posible por no perder ningún caso.

Cada dato censurado influye disminuyendo el denominador de los cocientes $(s_j)/(n_j)$ siguientes. Aunque un dato censurado no provoque un salto sí hace que tenga mayor magnitud el siguiente salto o escalón.

Si en el intervalo de tiempo en que hay eventos, también se presentan pérdidas de seguimiento, estas se descuentan de n_j después de haber calculado p_j, como si se hubieran producido luego de los eventos y con esta convención el procedimiento evita correcciones más o menos complicadas en el cálculo de la sobrevida.

Comparación de curvas de sobrevida empleando el método no paramétrico del *log-rank* o prueba generalizada de Savage

Este método permite comparar la incidencia del evento en los intervalos de tiempo para las respectivas curvas de sobrevida, entre dos o más grupos de datos o de muestras estudiadas.

La expectativa lógica bajo la hipótesis nula es que no existe diferencia en la incidencia de eventos entre los grupos o muestras a comparar, o sea, que la incidencia de eventos sea proporcional al número de individuos de cada grupo partiendo de la comparación entre los valores observados y esperados de cada grupo.

Para evaluar la significación estadística en la diferencia de los comportamientos de la sobrevida entre diferentes grupos o muestras de estudio se toma como planteamiento de hipótesis:

- $H_{0:}$ los grupos comparados presentan igual supervivencia globalmente, o sea, no hay diferencias persistentes *versus* H_1: los grupos comparados no presentan igual supervivencia globalmente.
- El estadístico de prueba a emplear sería:
  χ^2 (experimental-observado) = $[(O_A - E_A)^2/E_A] + [(O_B - E_B)^2/E_B] + \ldots + [(O_n - E_n)^2/ E_n]$, que sigue una distribución de probabilidades χ^2 con n – 1 grados de libertad para el nivel de confianza $(1 - \alpha)$% seleccionado por el investigador
  donde:
  A, B ...: representan los grupos a comparar;
  n: total de grupos a comparar;
  O: significa el valor observado de eventos igual a muertes;
  E: significa el valor esperado de eventos igual a muertes.
- Valor de p o p-valor (bajo H_0, donde se cumpliría H_0):
  p-valor = $Pr[\chi^2_{n-1,\,1-\alpha}$ (teórico-esperado) ≥ χ^2 (experimental – observado)].

  Regla de decisión
  - Si $\chi^2_{n-1,\,1-\alpha}$ (teórico-esperado) < χ^2 (experimental-observado), se ubica el χ^2 observado en el área crítica o de rechazo de H_0, por lo que existen evidencias para poder rechazar H_0 y plantear una heterogeneidad estadísticamente significativa en los comportamientos de las probabilidades de sobrevida entre los diferentes grupos estudiados representados de forma graficada por las curvas de sobrevida siguiendo el método de Kaplan y Meier por medio del método no paramétrico del *log-rank* para el nivel de confianza seleccionado.
  - Si $\chi^2_{n-1,\,1-\alpha}$ (teórico-esperado) ≥ χ^2 (experimental-observado), no ubicaríamos el chi-cuadrado observado en el área crítica o de rechazo de H_0, por lo que no existen evidencias suficientes para rechazar H_0, de esta forma no se puede plantear una heterogeneidad estadísticamente significativa en los comportamientos de las probabilidades de sobrevida entre los diferentes grupos estudiados representados de forma graficada por las curvas de sobrevida siguiendo el método de Kaplan y Meier por medio del método no paramétrico del *log-rank* para el nivel de confianza seleccionado.
- Para el caso de comparación de solo dos grupos (dos curvas de sobrevida) se puede emplear la razón:
  HR = [(OA/EA)]/(OB/EB), conocida como *hazard ratio* o razón de riesgos y expresa una razón de riesgo del evento en el grupo A (valor observado y esperado) con respecto al grupo B (valor observado y esperado).

La significación estadística de h (razón de riesgos para el evento en cada grupo) frente a la hipótesis nula de riesgos iguales para ambos grupos (h = 1) puede ser evaluada empleando el estadístico chi-cuadrado sobre las mismas bases de decisión que en el caso del método no paramétrico del *log-rank* tal y como se expresa en algunos paquetes estadísticos.

Además de la prueba del *log-rank*, existen otros métodos de comparación de curvas de supervivencia como:

- Prueba generalizada de Wilcoxon (también llamado de Breslow).
- Prueba de Tarone-Ware.
- Prueba de Pet-Peto-Prentice.
- Prueba de Fleming-Harrington.

Todos estas pruebas se basan en comparar las muertes observadas en cada grupo con las esperadas, sí la mortalídad fuese igual en todos los grupos (para H_0), pero difieren entre sí, porque en cada uno se hace una ponderación distinta y se le da distinto peso a las diferencias según ocurran más precoz o más tardíamente a lo largo del seguimiento, aunque no constituye objeto de exposición en este texto.

Comparación de curvas de sobrevida (método paramétrico)

El método de los riesgos proporcionales de Cox es el más frecuentemente utilizado donde el parámetro (de asociación) estimado es la razón de riesgos o razón de tasas de incidencia instantánea o *hazard ratio* (HR), que está determinada por:

HR (*hazard ratio*)= [Riesgo o tasa instantánea (en función del tiempo) para el evento, hazard en grupo 1 o para categoría 1]/[Riesgo o tasa instantánea (en función del tiempo para el evento, hazard en grupo 2 o para categoría 2 (de referencia)], siendo:

Riesgo o tasa instantánea (Hazard) = Número de sujetos con el evento o suceso para un tiempo determinado en un grupo o categoría/Número de sujetos que están a riesgo para este tiempo en el mismo grupo o categoría.

Se postula una distribución de probabilidades para el tiempo de sobrevida y se estima sus parámetros a partir de los datos de las muestras o grupos de individuos, es decir, se elige una fórmula o modelo matemático que se espera que describa la forma de transcurrir el tiempo de sobrevida y a partir de los datos reales de los grupos se estiman los parámetros correspondientes a la distribución previamente elegida.

El modelo como función del tiempo, describe la Sobrevida observada en los distintos grupos, en forma análoga a lo que hace una ecuación de regresión cuando estima la variable independiente.

Los modelos de la sobrevida permiten una serie de inferencias y comparaciones, así como la introducción de métodos multivariados para controlar los efectos de diversas covariables, siendo el tiempo la variable en función de la cual se expresan las características de la sobrevida.

La estimación de los parámetros de estos modelos se realiza en general por el método de máxima verosimilitud (*maximum likelihood*).

Si una función F(t) da cuenta de la proporción de eventos observados en un tiempo, la proporción de sobrevivientes hasta cada instante t, está dada por la función:

$S(t) = 1 - F(t)$

Si S(t) se denomina función de sobrevida (*survivor function*) y su representación gráfica es similar a las curvas de sobrevida, en este caso la misma deriva de un modelo matemático ajustado a los datos.

Hay diversos modelos de sobrevida y una forma de describirla es a través de la llamada función de riesgo λ(t) (*hazard function*), que relaciona el riesgo de muerte en un instante f(t), con la proporción de sobrevivientes hasta ese mismo momento S(t):

$\lambda(t) = f(t)/s(t)$

Existen varios modelos para λ(t),el más sencillo es el correspondiente a λ = constante y es donde la función de sobrevida toma la forma:

$S(t) = e^{-\lambda t}$

El empleo del método de los riesgos proporcionales (*proportional hazards*) determina que la función λ(t) queda expresada como λ(t, x), donde x representa las distintas variables x_1, x_2, … x_n, siendo los datos analizados en relación al tiempo y a las n variables. La expresión general de la función de riesgo puede escribirse:

$\lambda(t, x) = \lambda_0 e^{\beta x}$, donde λ_0 es el componente del riesgo dependiente del tiempo y βx representa la función de regresión $\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$, donde las x_i son los factores o variables de interés pronóstico. Cada variable x_i posee su coeficiente de regresión, que representa la parte que tiene en el riesgo total del modelo. Dado que los riesgos implicados por cada conjunto de valores de las x_i no es afectado por el tiempo, cuyo efecto está descripto por λ_0, estos modelos se conocen como de riesgos proporcionales.

El método en que el riesgo para cada conjunto de variables se considera independiente del tiempo y relativo con respecto al valor que tomen las variables se conoce como método de los riesgos proporcionales de Cox.

Como el modelo de Cox no toma en cuenta el tiempo, no permite hacer predicciones con respecto al tiempo de presentación de los eventos, sino estimar el riesgo que implica cada cambio en las covariables con respecto al conjunto de los datos, que serían estimaciones respecto al riesgo relativo.

Cada coeficiente β_i mide el cambio en el riesgo relativo por cada incremento en una unidad de la variable independiente correspondiente.

En el modelo de Cox de riesgos proporcionales para el caso de una variable continua, cuando hay diferentes predictores independientes constituye un efectivo modelo de regresión múltiple para la estimación de los riesgos relativos, siempre bajo el presupuesto de que el efecto del tiempo es el mismo para todos los individuos considerados.

Método multivariado de regresión de Cox

Está indicado el uso del método multivariado de la regresión de Cox cuando la variable dependiente está relacionada con la supervivencia de un grupo de sujetos o en general, con el tiempo que trascurre hasta que se produce en ellos un suceso o evento. Entre los aspectos que la caracterizan están:

- El evento de interés puede ser la muerte, el fallo de una prótesis, la incidencia de una enfermedad o la ocurrencia de una complicación en quién ya tiene una patología de base. Lo importante es que se trate de un suceso o evento que solo ocurra una vez como máximo.
- Para sucesos o complicaciones que se repiten más de una vez en algunos pacientes durante el seguimiento, es dudosamente válida la regresión de Cox y probablemente hay que recurrir a la regresión de Poisson.
- Se usa para valorar simultáneamente el efecto independiente de una serie de variables explicativas o factores pronósticos sobre la supervivencia, es decir, sobre la tasa de mortalidad o sobre la tasa de ocurrencia de otro fenómeno que vaya ocurriendo tras un periodo de tiempo variable en cada sujeto.
- Es la extensión multivariable del análisis de supervivencia para evaluar de manera general variables dependientes del tipo "tiempo hasta un suceso o evento" y emplea modelos de regresión próximos al modelo de regresión logística.
- Es una técnica que permite identificar y evaluar la relación entre un conjunto de variables explicativas y la tasa de ocurrencia del suceso de interés.
- Permite predecir las probabilidades de supervivencia o, en general, de permanencia libre del evento para un determinado sujeto a partir del patrón de valores que presenten sus variables pronosticas.
- Asume que el suceso o evento de interés debe ser irreversible, debe poder ocurrir una sola vez y la censura (casos que se pierden o retiran del segimiento) no debe ser informativa.
- El *hazard ratio* (HR) o razón de tasas para el evento según el tiempo transcurrido para el mismo entre grupos es el parámetro de asociación empleado en la comparación de dos grupos, dos muestras o dos categorías para un factor pronóstico o variable independiente empleada, más que un riesgo relativo, lo que realmente estima es una razón de densidades de incidencia o razón de tasas, su utilidad es expresar la rapidez relativa con la cual un paciente pasa de un estado en la variable dependiente a otro. Esto se expresa con el término *hazard*, que es la tasa instantánea de ocurrencia de la enfermedad o el suceso de que se trate.
- Asume que la razón de tasas instantáneas es constante a lo largo del tiempo, los *hazards* son proporcionales, de ahí su nombre de *proportional hazards model*.

El parámetro de asociación razón de hazard, *hazard ratio* (HR) o razón de tasas instantáneas, está determinado por:

- Tasa instantánea del evento o *hazard* en grupo 1 o para la categoría 1 de la variable independiente o factor/Tasa instantánea o *hazard* en grupo 2 o para la categoría 2 (de referencia) de esta variable independiente o factor:

Tasa instantánea (Hazard) = Número de sujetos con el evento o suceso para un tiempo determinado en un grupo formado según la categoría de la variable independiente/ Número de sujetos que están a riesgo para este tiempo en el mismo grupo formado.

- El modelo estimado de la regresión de Cox promedia de manera ponderada las *hazard ratio* de los diversos momentos en los que se produce alguna muerte. Viene a ser como hacer muchas regresiones logísticas, una para cada momento en que se observa alguna muerte.
- No se trata solo de saber el efecto sobre la supervivencia a un tiempo determinado, por ejemplo, supervivencia a los cinco años, sino de valorar cuál es el efecto sobre la función de supervivencia (a lo largo de todo el periodo, sea cual sea, el punto temporal que se elija).
- Permite afirmar que una supervivencia más ventajosa puede ser atribuida a un determinado tratamiento, porque comprueba que a igualdad de diferentes variables de caracterización general de las muestras o grupos, los pacientes que fueron tratados con cierto método sobrevivieron más o no tuvieron el evento seleccionado para la variable dependiente en cualquier punto dentro del seguimiento llevado a cabo en los mismos.
- Asume que hay cierta constancia o consistencia en la *hazard ratio* a lo largo del tiempo, es decir, si a los tres meses el *hazard* de un grupo es el doble que el del otro, en los otros momentos del tiempo también habrá una *hazard ratio* aproximadamente igual a 2. A este supuesto se le llama proporcionalidad de los *hazards* y de aquí viene el otro nombre que tiene la regresión de Cox (*proportional hazards model*).
- La ecuación de la regresión de Cox es el logaritmo del *hazard* o de la tasa instantánea del evento para un tiempo [$\ln(\lambda_t)$] es la variable dependiente y se expresa como: $\ln(\lambda_t) = a + b_1 x_1 + b_2 x_2 + ... + b_k x_k$; donde el hazard = λ_t, no es un dato ni cuantitativo ni cualitativo, sino que es una función variable con el tiempo para cierto evento o suceso, que va cambiando de un tiempo a otro.
- La tasa se diferencia del riesgo en que tiene en cuenta el tiempo, por ejemplo, fallecimientos por unidad de tiempo de seguimiento, mientras que el riesgo es una proporción y solo tiene en cuenta el número de sujetos inicialmente a riesgo de fallecer. La tasa instantánea o *hazard* de fallecer, por ejemplo, en el instante t se obtiene según el modelo de Cox al tomar los respectivos antilogaritmos, empleando el valor e de la base del logaritmo neperiano o natural y exponenciando el miembro derecho de la ecuación anterior como sigue:

  $(\lambda_t) = e^{a + b_1 x_1 + b_2 x_2 + ... + b_p x_p} = e^a x\ e^{b_1 x_1 + b_2 x_2 + ... + b_k x_k,}$

  Si $e^a = \lambda_{0t}$, entonces:

  $(\lambda_t) = \lambda_{0t} \cdot e^{b_1 x_1 + b_2 x_2 + ... + b_k x_k}$

Cuando las variables independientes (factores predictores) $x_1 = 0, x_2 = 0 \ldots x_k = 0$, entonces $(\lambda_t) = \lambda_{0t}$, la cual es la tasa instantánea basal (*baseline hazard*) que expresa una función que va cambiando con el tiempo y corresponde a los sujetos para los cuales el resto de la parte de derecha de la ecuación $(b_1x_1 + \ldots + b_kx_k)$ vale 0.

La regresión de Coxal no compara riesgos propiamente dichos (proporciones), sino tasas instantáneas, permite estimar la rapidez con la cual se pasa de un estado a otro para el evento o suceso ajustando por las demás variables del modelo. La HR expresa cuántas veces es más rápida la ocurrencia del evento o suceso por ejemplo la muerte u otro fenómeno en un grupo que en el empleado como referencia; es decir, para una categoría dada con aspecto a otra de referencia para cierto evento o suceso. Viene a ser como un cociente entre dos velocidades.

Una *hazard ratio* de *2* significa que se multiplica por *2* la velocidad con que ocurre una enfermedad o el suceso en los sujetos que están expuestos al factor de riesgo. Una *hazard* de 1 significa que el efecto del factor es nulo. Un valor de 0,5 significa que esa exposición en vez de aumentar el riesgo reduce a la mitad la velocidad de ocurrencia del suceso.

Interpretación de los estadígrafos resultantes a mostrar en una regresión de Cox

En los paquetes estadísticos se pueden encontrar estos aspectos:

– Coeficiente de regresión β_K o B_k: expresa el peso que tiene la variable predictiva independiente $(_K)$ en el resultado de la *hazard ratio*, es decir, por cada unidad que varíe la relación entre los dos grupos o categorías en la variable predictora independiente para un tiempo determinado, el valor del coeficiente de regresión es aportado al resultado de la *hazard ratio*, teniendo ajustadas las demás variables independientes.

La forma matemática de relacionarse el coeficiente de regresión de Cox y el *hazard ratio* se presenta:

$\ln HR = b;$

$e^b = HR;$

$$\ln HR = \ln \frac{\text{Hazard grupo 1}}{\text{Hazard grupo 2}} = \ln \frac{\text{Hazard categoría no referencia del factor}}{\text{Hazard categoría referencia del factor}} = b \ ;$$

$$\ln HR = \ln \frac{\text{Taza instantánea (categoría no referencia de la variable independiente}}{\text{Taza instantánea (categoría referencia de la variable independiente}} = b \ ;$$

$$\ln HR = \ln \frac{\text{Número sujetos con evento categoría no referencia/Total sujetos con evento categoría no referencia para t}}{\text{Número sujetos con evento categoría referencia/Total sujetos con evento categoría no referencia para t}} = b$$

El *hazard ratio* es un cociente entre el riesgo o tasa instantánea de ocurrir el evento por ejemplo fallecer en expuestos al factor y el riesgo o tasa instantánea de fallecer en quienes no están expuestos al factor. Este concepto se parece mucho al riesgo relativo o cociente entre dos riesgos, pero cuando se habla de riesgos se suele hacer referencia simplemente a proporciones. En cambio, la regresión de Cox lo que estima

no son proporciones, sino tasas expresadas mediante la densidad de incidencia. Por eso el *hazard ratio* más que un riesgo relativo lo que realmente estima es una razón de densidades de incidencia o razón de tasas.

- Error estándar [$EE(B_K)$] del valor del coeficiente de regresión (B_K): expresa la variabilidad de β en la relación de dos categorías de la variable independiente k, teniendo una categoría como referencia.
- Prueba de Wald: se realiza al dividir el valor del coeficiente de regresión (B_k) de cada variable independiente predictora entre su error estándar [$EE(B_k)$] y elevar al cuadrado la cantidad resultante:
  Wald (χ^2) = $[(B_K)/(EEB_K)]^2 \sim \chi^2_{gr-1,\,1-\alpha}$ (gr = número de grupos o categorías del factor pronóstico menos 1 = (grados de libertad) para un nivel de confianza $(1-\alpha)$% prefijado. Expresa la diferencia entre el valor observado y esperado para el coeficiente de regresión β_K estimado a partir de dos categorías del factor predictor k tomando una de estas como referencia, en el contraste de hipótesis H_0: $\beta_K = 0$ *versus* $H_1 = \beta_K \neq 0$). Es empleado para valorar la significación estadística individual de cada variable o factor predictor k en el tiempo en relación a la variable de respuesta empleada en el modelo estimado en el tiempo.
- Valor p o p-valor (expresa el valor de probabilidad de cumplirse H_0,para el valor observado del estadístico Wald (χ^2) y se expresa como:
  p-valor = Pr[(χ^2_{gr-1} (teórico-esperado) ≥ χ^2 (observado-calculado)] bajo H_0
- Valor de *hazard ratio* (HR_k) global ponderada = $EXP(\beta_K)$: expresa globalmente las veces que es diferente (superior por ejemplo) la tasa instantánea de ocurrencia de un suceso o evento para una categoría de una variable predictora en un tiempo determinado respecto a la tasa para ESE suceso en otra categoría (la de referencia) en la misma variable predictora para el mismo tiempo:
  - Si $HR_k > 1$, significa que para la razón de tasas instantáneas del evento específico (*hazard*) en la variable dependiente, la categoría del numerador del factor predictor o variable independiente tiene un efecto perjudicial para la aparición del evento específico en la variable de respuesta o dependiente y esto puede predecir una menor sobrevida si, por ejemplo, el evento que se trata es la muerte en el tiempo analizado.
  - Si $HR_k < 1$, significa que para la razón de tasas instantáneas del evento específico (*hazard*) en la variable dependiente, la categoría del numerador del factor predictor o variable independiente tiene un efecto predictor no perjudicial para la aparición del evento específico en la dependiente y esto predice una mayor sobrevida si, por ejemplo, el evento que se trata es la muerte en el tiempo analizado.
  - Si $HR_k = 1$, significa no diferencias entre las *hazards* para ambas categorías comparadas de la variable independiente predictora para la aparición del evento en la variable dependiente en el tiempo analizado.
- Intervalo de confianza $(1-\alpha)$ 100 %, para la *hazard ratio* estimado en relación a cada posible factor o variable predictora:

- IC $(1 - \alpha)$% inferior $(HR_k) = \exp(\beta_{k,\,t} - ZEEB_{k,\,t})$.
- IC $(1 - \alpha)$ superior $(HR_k) = \exp(\beta_{k,\,t} + ZEEB_{k,\,t})$.
  El resultado para el *hazard ratio* $HR_{k,}$ a partir del IC [(1 - α)%] se interpreta de esta forma: el valor nulo de una HR_k es 1. Cuando el intervalo de confianza por ejemplo al 95 % (más empleado) incluya la hipótesis nula (HR = 1), se sabe que las diferencias en la incidencia del evento o suceso, por medio de la razón entre ambas tasas de incidencia instántaneas o *hazard*, no son significativamente diferentes (al 5 %) para los dos grupos formados o dos categorías a comparar, es decir, la categoría específica *versus* la de referencia en la variable independiente predictora y, por tanto, no es muy probable que esta variable sea un buen predictor del suceso o evento en cuestión en el periodo de tiempo observado de seguimiento (Tabla 4.64).

Tabla 4.64. Forma tabulada y resumida en los paquetes estadísticos

Variables independientes (k) (solo están para las categorías no de referencia)	Coeficiente de regresión estimado para cada variable o factor (B_k)	Error estándar de B (EEB_k)	Estadístico de la Prueba de Wald: $[(B_k)/(EEB_k)]^2$	Grados de libertad (número de categorías no referencia de cada variable)	p-valor (Sig)	exp^B (HR_k)	IC (HR_k)

Aplicación práctica al estimar un modelo de predicción por medio de la regresión de Cox

Una vez estimado un modelo para la razón de tasas de incidencias instantáneas para un tiempo determinado (*hazards ratio*), a partir de la aplicación de la ecuación para la regresión de Cox, si se quiere evidenciar su aplicación práctica en un determinado sujeto para conocer la predicción para el evento en ese modelo se procede de la esta manera: se utilizan los valores de los coeficientes de regresión de cada factor o variable predictora y de la tasa instantánea basal (*baseline hazard*) para el modelo estimado y se introducen los valores del sujeto para cada una de las variables predictoras y con ello se efectúa la operación sumatoria matemática para estimar cual sería el logaritmo de la "razón de tasas instántaneas global ponderada o razón de hazard (promedio ponderada), que al aplicarle su antilogaritmo se obtiene su valor estimado para el sujeto. Este resultado de la razón de hazard expresa globalmente cuantas veces la tasa instantánea para el evento o suceso es superior en el grupo o para la categoría seleccionada respecto a la tasa del grupo o para la categoría de referencia de los factores o variables independientes empleadas.

Para recordar

- La prueba del *log rank* puede tener limitaciones en la comparación de curvas de sobrevida al mostrar resultados que no lleguen a ser válidos porque estén afectados por factores de confusión no controlados y solo valoran una única variable (la que constituye los grupos) como posible factor pronóstico, pero puede haber muchas variables que interese valorar a la vez.
- La aplicación de una técnica de análisis multivariante como la regresión de Cox significa que se tienen en cuenta simultáneamente muchas variables independientes en el análisis de los datos y la principal ventaja que ofrece es que permiten controlar de modo eficiente muchos factores de confusión al mismo tiempo, cuando se trata de estudiar asociaciones potencialmente causales entre una determinada exposición y un efecto o desenlace.
- La regresión de Cox requiere para su ejecución que el tamaño muestral sea igual o mayor que: el número de variables independientes (explicativas) más 1 por 10.
- Aunque un modo de aproximarse al cálculo de *hazard ratio* es calcular razones de densidades de incidencia, la diferencia fundamental entre una razón de tasas de incidencia instantánea y una razón de densidades de incidencias se debe a que el modelo de Cox (que emplea *hazard ratio*) da más importancia (pondera o pesa más) a las comparaciones entre los *hazards* para los eventos que ocurren al principio, mientras la razón de densidades de incidencia no tiene en cuenta cuándo ocurre cada evento, simplemente los suma.

La regresión de Cox y la logística guardan estas semejanzas:

- Se mueven siempre en escala multiplicativa (análoga al riesgo relativo) y no en escala aditiva.
- Usan métodos de máxima verosimilitud para calcular los parámetros.
- Usan la prueba de Wald para valorar la significación estadística individual de cada término (variable independiente) introducido en el modelo. Tanto en la regresión logística corno en regresión de Cox la prueba de Wald se presenta como una elevación al cuadrado del cociente entre el coeficiente de regresión y su error estándar $(b/EE)^2$ y sigue una χ^2 con tantos grados de libertad como categorías tenga esa variable.
- Tienen como mejor prueba de hipótesis la prueba de razón de verosimilitudes (–2ln[L] o –2 log de la verosimilitud), que sigue una χ^2, con tantos grados de libertad como variables de más haya en el modelo más amplio, en comparación con el más reducido.
- Producen un listado con exactamente la misma apariencia en la mayoría de paquetes estadísticos.
- El parámetro para medir la asociación entre exposición y efecto se halla elevando al número Exp = 2,7182, el coeficiente de regresión expresándose como [Exp(B)].
- Sirven para controlar la confusión usando los mismos procedimientos (se controla un factor de confusión añadiéndolo al modelo).

- Valoran la interacción con el mismo procedimiento (términos multiplicativos añadidos al modelo).
- Pueden utilizar los mismos algoritmos automáticos (*stepwise*) de selección de variables, aunque estos procedimientos están en general desaconsejados, salvo para seleccionar más rápidamente si hay alguna interacción significativa.
- Exigen que se compruebe si la forma de la relación entre una variable independiente cuantitativa y el efecto sigue un patrón monotónico. Esto debe chequearse construyendo cuartiles antes de introducir como cuantitativa la variable independiente. Si no hay patrón monotónico se debe introducir categorizada esta variable.
- Requieren aproximadamente que hayan ocurrido unos 10 eventos por cada variable independiente que se desee introducir en el modelo.

La regresión de Cox y la logística tienen estas diferencias:

- La regresión de Cox tiene en cuenta el tiempo que tarda en producirse un evento.
- La regresión logística estima *odds ratios* que exageran siempre el riesgo relativo.
- En la regresión de Cox no hay ordenada en el origen.
- La regresión de Cox asume la proporcionalidad de los hazards.
- La regresión de Cox exige analizar visualmente las curvas de supervivencia para valorar si realmente los hazards son proporcionales (se van separando las curvas con el tiempo).
- La regresión de Cox permite un ajuste alternativo más fino de algunas variables que esté muy claro que actúan como factor de confusión.
- Existe la posibilidad de hacer un modelo de Cox con covariables dependientes del tiempo. Esto permite, por una parte, comprobar el supuesto de si los hazards son proporcionales y, por otra parte, adaptarse en cierto modo a una falta de proporcionalidad de los hazards, mediante un término de producto entre el tiempo y la variable cuyo *hazard ratio* cambia con el tiempo.
- La regresión de Cox se puede usar en estudios de casos y controles emparejados para sustituir a la regresión logística condicional. Se hacen estratos por la pareja (o relación caso-control), se crea un valor de seudotiempo que sea superior en los controles que en los casos y se usa como evento la categoría correspondiente a los casos. Esto permite preservar el emparejamiento (por introducirlo en los estratos) y es equivalente a la regresión logística condicional.

Método multivariado regresión lineal múltiple o de regresión múltiple

El análisis de regresión lineal múltiple es una técnica estadística general utilizada para analizar las relaciones entre una única variable dependiente cuantitativa (criterio) y varias variables independientes (predictoras), es decir, se emplea cuando se desea estudiar cómo influyen varios factores (o variables independientes) en una variable de respuesta

(la variable dependiente) con el objetivo de predicción de un resultado o explicación. Entre los aspectos que la caracterizan están:

- El modelo de regresión múltiple es una generalización a varias variables (más de dos) de un modelo de regresión líneal simple y su objetivo fundamental es encontrar la ecuación de una recta que permita predecir o estimar los valores de una variable dependiente cuantitativa que correspondan a los valores conocidos de varias variables cuantitativas.
- Utiliza las variables independientes cuyos valores son conocidos para predecir la única variable criterio seleccionada por el investigador.
- Cada variable predictor es ponderada (sopesada) de forma que las ponderaciones indican su contribución relativa a la predicción conjunta. AI calcular las ponderaciones, el procedimiento del análisis de regresión asegura la máxima predicción a partir del conjunto de variables independientes. Estas ponderaciones facilitan también la interpretación de la influencia de cada variable en la realización de la predicción.

Un modelo de regresión lineal múltiple se construye con uno de estos tres objetivos:

- Control de la confusión: que mida la relación entre una variable (exposición) y su respuesta en presencia de otras variables que puedan influir. Se busca estimar con el menor sesgo posible los coeficientes de regresión relativos a la variable exposición. La alternativa a la causalidad es casi siempre la confusión por la existencia de una causa común de la exposición y del efecto. Se debe ajustar por esas variables.
- Predicción: que permita predecir la respuesta de nuevos sujetos con el menor error posible. El criterio de selección es maximizar el coeficiente de determinación ajustado (R^2) corregido o ajustado.
- Descripción: que describa lo mejor posible la muestra estudiada. En este caso se trata de eliminar las variables redundantes. En esta situación se exploran modelos que maximicen el coeficiente de determinación R^2, pero eliminando las variables que solo producen mínimos incrementos en R^2.

Los supuestos de la regresión múltiple incluyen:

- Las X_i (variables independientes o predictoras) son no aleatorias (fijas).
- Para cada conjunto de X_i, existe una subpoblación j de valores de Y (variable dependiente), donde la variable Y es una variable aleatoria y sus valores siguen una distribución normal.
- Las varianzas de las subpoblaciones j de Y son todas iguales (homocedásticas).
- Los valores de las subpoblaciones j de Y son independientes.

El conjunto de variables independientes ponderadas es conocido también como valor teórico de la regresión, una combinación lineal de las variables independientes con sus respectivos coeficientes que predice mejor la variable dependiente (criterio) y la ecuación

de regresión, es el ejemplo de valor teórico mas ampliamente reconocido entre todas las técnicas multivariantes.

La ecuación para la regresión lineal multivariada o múltiple sería:

$y_j = \beta_0 + \beta_i x_{1j} + \beta_i x_{2j} + ... + \beta_k x_{kj} + e_j$

donde:

y_j: valor representativo de una de las subpoblaciones de los valores de Y;

β_i: coeficientes de regresión parcial poblacional estimados respecto a cada variable variable independiente para predecir la variable dependiente Y. Este coeficiente para cada una de las k variables independientes β_k se interpreta como el cambio cuantificable en la variable dependiente y, estimada mediante su valor medio, por cada unidad de cambio en la variable predictora respectiva (x_1, x_2 ...x_k), independientemente al valor de las otras variables independientes cuantitativas o en caso de ser cualitativa alguna de las variables independientes incluidas en el modelo de regresión múltiple, se asume este resultado de la variable independiente cuantitativa seleccionada para el análisis, de forma ajustada (para la igualdad de sus valores en los niveles o categorías de las otras variables independientes incluidas en esta ecuación de regresión múltiple, para evitar que pudiesen comportarse como factores de confusión. El tipo de signo delante del valor de β_i – o + se interpreta como que disminuye o aumenta respectivamente, el valor medio estimado de la variable dependiente Y al aumentar en una unidad el valor de la X_i empleada (en caso de una variable cuantitativa predictora) o al cambiar de la categoría seleccionada respecto a la de referencia en la X_i utilizada (si se tratara de una variable cualitativa predictora);

X_{1j}, X_{2j} ..... X_{kj}: corresponde con los valores numéricos concretos de las k variables independientes predictoras para las subpoblaciones j de valores de Y;

e_j: variable aleatoria de media 0, cuya varianza es la que es común a las subpoblaciones de valores de Y;

β_0: constante en la ecuación, representa el intercepto; es decir, representa el punto donde el plano de regresión corta el eje Y;

k: orden de las n variables independientes.

Las estimaciones de los parámetros $\beta_0, \beta_1, \beta_2, ..., \beta_k$ se obtienen por extensión del método de los mínimos cuadrados a más de una variable independiente. Esto significa que se minimiza la suma de las desviaciones elevadas al cuadrado de los valores observados de Y respecto a la superficie de regresión resultante.

Bajo ciertas circunstancias es posible incluir datos no métricos (cualitativos) para las variables independientes (transformando los datos ordinales o los nominales en variables ficticias, indicadoras o *dummy*), es decir, al aplicar el análisis de regresión múltiple: los datos deben ser métricos o apropiadamente transformados y, antes de estimar la ecua-

ción de regresión, el investigador debe decidir que variable va a ser dependiente y cuáles de las restantes variables serán independientes, o sea, mientras la variable dependiente (Y) es una variable cuantitativa y debe estar medida en un nivel por intervalos o de razón, las variables independientes (X) pueden ser tanto cuantitativas como cualitativas (es posible introducir variables categóricas con dos o más de dos categorías) transformadas a variables "dummies" indicadoras o ficticias.

Si la variable cualitativa categórica independiente está formada por k categorías se pueden analizar introduciendo en el modelo simultáneamente k – 1 variables ficticias. Estas variables "dummies" son dicotómicas y toman los códigos 0 o 1. El valor 0 se asigna a la categoría que se toma como referencia y hay una variable *dummy* por cada una de las otras categorías, que solo valen 1 cuando el sujeto pertenece a la respectiva categoría. Generalmente se le da el valor 0 a la categoría en que se espera un nivel menor, basal o la referencia natural. Cuando las variables categóricas tienen más de dos categorías, una de las categorías (que no se introduce en el modelo) es la que queda como estrato de referencia y frente a ella se comparan todas las demás.

El procedimiento utilizado para calcular una regresión múltiple es estimar la superficie que mejor se ajusta a la nube de puntos observados. El método se denomina *ajuste por mínimos cuadrados*. Es un método que minimiza las distancias elevadas al cuadrado, desde cada punto observado hasta el plano (residuales al cuadrado).

La aproximación tradicional al problema de construcción de un modelo implica buscar el modelo más parsimonioso (con el menor número de variables) que explique bien los datos y se recomienda que no se introduzca más de una variable independiente por cada 20 sujetos que haya en la base de datos.

De forma general como requisitos para la ejecución de un análisis de regresión lineal múltiple se debe tener en cuenta:

- Aspectos de la selección y operacionalización de las variables independientes y la dependiente, así como preveer transformaciones necesarias, por ejemplo, la creación de variables ficticias, indicadoras o *dummy* en el caso de variables cualitativas no métricas a incluir en el estudio o transformaciones logarítmicas en caso de no seguir una distribución normal los valores de sus datos.
- Aspectos referentes al diseño investigativo como la obtención de una muestra con un tamaño adecuado, que permita contar con una potencia estadística y nivel de significación deseados (basada en la probabilidad de detectar como estadísticamente significativo un nivel especifico del coeficiente de determinación múltiple (R^2) o de un coeficiente de regresión múltiple beta (β) en las k variables independientes para un nivel de significación especificado y a su vez facilitar una generalización en los resultados en función del ratio (razón) del número de observaciones en relación al número de las variables independientes. Una norma general es que

este ratio o razón nunca debe caer por debajo de cinco, lo que significa que no debe existir menos de cinco observaciones para cada variable independiente presente en el valor teórico. La potencia de la regresión múltiple se refiere a la probabilidad de detectar como estadísticamente significativo un nivel especifico de R^2 o un coeficiente de regresión beta para un nivel de significación prefijado y un tamaño de muestra específico.
- Aspectos referentes a los supuestos necesarios a cumplir en los métodos multivariados, que garanticen los mejores y verdaderamente representativos resultados de la muestra como normalidad en la distribución de los valores de las variables, homocedasticidad o varianza constante, linealidad en las relaciones entre las variables e independencia de los datos de las variables.

La estimación y evaluación del modelo o ecuación de regresión múltiple específico constituye la pieza fundamental en el análisis multivariado de regresión múltiple para el investigador y se sugiere realizar tomando en cuenta estos aspectos a partir de un análisis teórico de las posibles relaciones de asociación entre las variables involucradas en el estudio a realizar, la operacionalización de estas, la obtención de una muestra con un tamaño adecuado y el cumplimiento de los supuestos necesarios, seguidamente se comienza la estimación del modelo de regresión múltiple, donde los pasos a seguir son varios.

Primer paso del modelo de regresión múltiple

Se debe escoger un método para determinar las variables que participarán en la estimación del modelo de regresión múltiple, bien sea por el método de estimación secuencial (progresiva-regresiva o por etapas) o por el método de combinación. Realizar un análisis de correlación lineal bivariado entre cada variable independiente y la dependiente (seleccionar las que más correlación ofrezcan entre sí), así como un análisis de correlación entre dos a dos variables independientes para determinar si existe correlación entre dos o más variables que evidencia presencia de colinealidad o multicolinealidad, lo que debe tenerse en cuenta para evitar la inclusión de esas variables independientes en el modelo de regresión múltiple al conspirar con la capacidad predictiva de cada variable independiente en el modelo o ecuación de regresión múltiple estimada, todo esto se realiza empleando una matriz de correlación entre las distintas variables independientes con la dependiente y de las independientes entre sí (Tabla 4.65). Se obtiene así un primer mapa de ruta que orienta sobre qué resultados se encontrarán. Cuando dos variables independientes tengan entre sí un coeficiente de correlación lineal simple de gran magnitud ($r > 0,90$ en valor absoluto) hay que decidir cuál de las dos es la mejor candidata o bien usar la media de ambas, pero no introducir nunca las dos, pues se producen casi siempre problemas de colinealidad.

Tabla 4.65. Representación gráfica de la matriz para la correlación entre variables.

Tipo de variable	Variable dependiente (Y) con el valor del coeficiente de correlación lineal)	Variable independiente (X_1)	Variable independiente (X_2)	Variable independiente (X_n)
Y (dependiente)	-	-	-	-
X_1 (independiente)	-	-	-	-
X_2 (independiente)	-	-	-	-
X_n (independiente)	-	-	-	-

Para recordar

- La colinealidad es la asociación, medida como correlación, entre dos variables independientes y multicolinealidad se refiere a la correlacion entre tres o más variables independientes (evidenciada cuando se hace la regresión de una variable respecto de las otras).
- El impacto de la multicolinealidad consiste en reducir el poder predictivo de cualquier variable independiente individual en la medida en que está asociado con las otras variables independientes.
- Conforme se incrementa la colinealidad, la varianza única explicada por cada variable independiente se reduce y el porcentaje de predicción compartida aumenta. Dado que solo se puede calcular una vez que se ha realizado la predicción compartida, la predicción global aumenta mucho más lentamente conforme se añaden variables independientes con un nivel de multicolinealidad alta.
- El medio más simple y obvio de identificar la colinealidad es un examen de la matriz de correlación de las variables independientes. La presencia de una elevada correlación (generalmente de 0,90 en adelante) es la primera indicación de una elevada colinealidad. La ausencia de elevados valores de correlación no asegura una falta de colinealidad. La colinealidad puede deberse a los efectos combinados de dos o más variables independientes.

Dos de las medidas más comunes para evaluar la colinealidad de parejas o de múltiples variables son el valor de tolerancia y su inverso: el factor de inflación de la varianza. Estas medidas nos dan el grado en el que cada variable independiente se explica por otras variables independientes y aparece en la mayoría de los paquetes estadísticos donde puede ser consultada.

Algunas sugerencias para tratar la multicolinealidad son:

- Omitir una o varias variables independientes correlacionadas e identificar otras variables independientes para ayudar con la predicción. Sin embargo, el investigador debe tener cuidado cuando sigue esta opción para evitar la creación de un error de especificación cuando se eliminan una o más variables independientes.

– Utilizar el modelo con las variables correlacionados solo para predecir, es decir, no intentando interpretar los coeficientes de regresión.
– Utilizar las correlaciones simples entre cada variable independiente y cada variable dependiente para entender la relación entre la variable independiente-dependiente.
– Utilizar un método más sofisticado de análisis, como una regresión bayesiana (o un caso especial de regresión de tipo cresta) o una regresión de componentes principales para obtener un modelo que refleje más claramente los efectos simples de las variables independientes.

Las recomendaciones para el orden de introducción de las variables independientes con el objetivo de obtener la mayor varianza aportada por las mismas a la variable dependiente son:
– Ingresar las variables independientes de acuerdo con la fuerza de su correlación con la variable dependiente, de la más alta a la más baja.
– Seguir el orden en que se han introducido en estudios previos.
– Proceder de acuerdo a lo acumulado por la teoría.
– Seguir un orden cronológico en que fueron tomadas, si su medición fue por etapas, entonces tomar de la primera a la última para el caso.
– Como criterio de referencia o guía para la selección de las variables a tomar en cuenta para conformar el modelo, el investigador emplea R (coeficiente de correlación lineal simple estimado), R^2 (coeficiente de correlación lineal al cuadrado o de determinación) y el error estándar de la estimación, al analizar los resultados de la correlación entre las variables antes mencionadas y siguiendo muy estrechamente los aspectos de colinealidad antes expresados.

Segundo paso del modelo de regresión múltiple

Se debe estimar inicialmente la probable ecuación de regresión múltiple dando la oportunidad de incluirse en el valor teórico a las variables independientes explicativas anteriormente mejor correlacionadas con la dependiente (criterio) y evaluar la significación estadística del modelo conjunto (valor teórico) en la predicción de la variable dependiente o de respuesta, a partir del valor del coeficiente de determinación R^2 o del coeficiente de determinación ajustado $R^2_{ajustado}$ estimados para la ecuación de regresión múltiple, el error estándar de la estimación y de los coeficientes de regresión beta (β_i) de cada variable independiente X_i, así como los estadísticos de prueba respectivos para evaluar su significación estadística respectiva t-student y ratio o razón F.

Para la estimación de la ecuación o modelo de regresión lineal múltiple se sigue, por ejemplo, el método por pasos (*stepwise*) y se estima el coeficiente de regresión estimado

(βi) para cada variable independiente X_i de las k variables independientes que fueron seleccionadas a partir de sus mayores valores del coeficiente de correlación linealsimple estimad o R y del de determinación R^2, según el análisis de correlación lineal simplepreviamente realizado y el error estándar de este coeficiente ($EE\beta_i$) respectivo, así como los valores del estadístico t-student o su raíz cuadrada (razón F) (según la versión del estadístico de Wald en la regresión múltiple), que permiten evaluar la significación de cada coeficiente de regresión β_i en relación con su error estándar ($EE\beta_i$) para las variables independientes k antes seleccionadas, las que son candidatas a su inclusión en el modelo y se emplean para el contraste de hipótesis de la predicción de la variable dependiente cuantitativa continua, a partir de cada variable independiente cuantitativa o cualitativa transformada a *dummy* con una tasa de error tipo 1 igual a alfa preseleccionada por el investigador como:

H_0: $\beta_k = 0$ *vs*. H_1: $\beta_k \neq 0$.

Se emplea como estadístico de prueba o contraste F de Snedecor parcial o una t-student:

– Para el estadístico de contraste F de Snedecor parcial (el estadístíco de Wald) $[\beta_i/EE_i]^2$, sigue una distribución de probabilidades $F_{(n, m; 1-\alpha)}$ para un nivel de significación alfa de cometer el error tipo 1, según el nivel de confianza $(1 - \alpha)$% predeterminado;

  donde:

  n (grados de libertad en el numerador): número de variables independientes en el modelo, lo que es igual a k o sea, n = k;

  m (grados de libertad en el denominador): número de observaciones (N) en el modelo, menos el número de variables independientes ken el mismo menos 1, o sea, m = N – k – 1;

  $1 - \alpha$: representa el nivel de confianza seleccionado.

  El valor de p o p (valor) = Pr [(F-Snedecor $_{(n, m; 1-\alpha)}$ (teórica-esperada) ≥ F-Snedecor (observada-calculada)], bajo H_0.

– En el caso de la t-student como equivale a la raíz cuadrada de F-Snedecor en este caso t-student (como el estadístico de Wald) $[\beta_k/EE_k]$, sigue una distribución de probabilidades t-student, con (N – k –1) grados de libertad, siendo:

  N: número de observaciones muestrales para el modelo;

  K: número de variables independientes en el modelo.

  Valor de p o p (valor) = Pr[(t-student $(_{N-k-1; 1-\alpha)}$ (teórica-esperada) ≥ t-student (observada-calculada)], bajo H_0.

Regla de decisión

En cuanto a la significación de la predicción de la variable dependiente Y a partir de cada variable independiente X_i por medio del correspondiente coeficiente de regre-

sión lineal múltiple β_i poblacional estimado del muestral b_i se tiene que si el valor de F-Snedecor o la t-student (observado-calculado) > valor de la F-Snedecor o t-student (teórico-esperado) a partir de la tabla de distribución de probabilidades correspondiente, según los grados de libertad y el nivel de significación alfa seleccionado por el investigador, el valor observado de estos estadísticos de prueba queda ubicado en el área de rechazo de H_0, estimándose un p-valor en que se cumple H_0 menor al nivel de significación alfa, por lo que hay evidencias para plantear que existe una relación de predicción significativa de la variable dependiente Y a partir de la independiente X_i específica en cuestión, estando ajustadas en un mismo nivel de valores de la variable X_i para el resto de las categorías de las variables independientes si fuesen cualitativas en el nivel de confianza preseleccionado.

El coeficiente de regresión múltiple poblacional estimado (β_K) a partir del muestral b_k, expresa cuanto cambia numéricamente (aumentando o disminuyendo, según el signo + o –, respectivamente) el valor medio en la variable dependiente cuantitativa continua o de respuesta Y, por cada unidad que cambia el valor de la variable independiente cuantitativa predictora X_k, independientemente o sobrepuesto a los cambios medios de la variable dependiente o de respuesta Y asociados con los valores predichos por todas las demás variables independientes cuantitativas en el modelo estimado de regresión múltiple. En el caso de emplearse en el modelo estimado de regresión múltiple anterior (ecuación de regresión múltiple) algunas variables independientes predictoras cualitativas, estas deben ser transformadas a variables indicadoras, ficticias o *dummy*, donde cada valor de β_i expresa lo que varía como valor medio Y_i, a partir del cambio o paso de una categoría analizada respecto a la tomada de referencia en la variable independiente X_i. Si se selecciona alguna variable predictora independiente *dummy* en el modelo estimado de regresión múltiple comentado anteriormente, aparte de la variable independiente analizada X_i para estimar el β_i parcial se debe aclarar que es más apropiado expresar que el valor del β_i parcial estimado en ese caso está ajustado a la condición de iguales valores en las observaciones de la variable independiente analizada X_i, para los distintos niveles las categorías o en definitiva grupos formados de la otra variable independiente cualitativa transformada antes mencionada, es decir, el ajuste en la predicción realizada a partir del resultado del β_i parcial, se efectúa equiparando o asemejando los valores en la variable independiente predictora específica a analizar X_i, para los grupos formados a partir de las categorías que se comparan en las variables independientes predictoras específicas a analizar diferente a X_i.

La hipótesis nula para cada una de las pruebas de hipótesis de cada coeficiente β_i de regresión múltiple es que la nueva variable independiente que se introduce a partir de la primera no mejora la predicción (no disminuye hasta hacer significativo el p-valor estimado respectivo de la variable introducida en el modelo). La prueba de hipótesis

que emplea su valor y su error estándar mediante un cociente va comparando en cada caso la capacidad predictiva de dos modelos: uno solo con las primeras dos variables independientes y otro con tres, que incluya la variable de interés.

La contribución de cualquier variable independiente particular solo es la contribución que se superpone a las contribuciones de todas las otras variables independientes.

Los coeficientes de regresión estimados en la regresión múltiple (β_K) son el reflejo de la relación entre la subpoblación para la variable dependiente Y y cada una de las k variables independientes, lo que se controla o se ajusta para la relación entre la variable predictora seleccionada y la dependiente según los efectos de las otras variables predictoras, por eso la regresión múltiple se puede utilizar para evitar el posible efecto de confusión sobre la predicción de la variable cuantitativa continua Y, a partir de una variable explicativa independiente X_i, mediante el ajuste que se realiza para las otras variables explicativas cualitativas incluidas en el modelo estimado de regresión múltiple.

Cuantas más variables independientes se examinen en este procedimiento, mayor es la probabilidad de que al menos un coeficiente de regresión parezca significativo aunque no exista una relación entre esas variables en la población de la que se ha extraído la muestra, por tanto, asociaciones estadísticamente significativas, entre la variable dependiente y las independientes que no se esperaba tuvieran importancia antes de analizar los datos, deben interpretarse con mucho cuidado.

Si cada una de nuestras variables predictoras o independientes son estandarizadas antes de estimar la ecuación de regresión, se encuentran diferentes coeficientes de regresión. Los coeficientes resultantes de los datos estandarizados se denominan *coeficientes beta* o *estandarizados*. Su valor reside en que eliminan el problema de tratar con diferentes unidades de medida y reflejan el impacto relativo sobre la variable de respuesta o criterio de un cambio en una desviación estándar de cada variable. Ahora que se tiene una unidad común de medida, se pueden determinar que variable es la más influyente.

Se deben tener en cuenta precauciones cuando se utilizan los coeficientes beta estandarizados. En primer lugar deben utilizarse como guía de la importancia relativa de las variables individuales independientes únicamente cuando la colinealidad es mínima. En segundo lugar, los valores beta estandarizados pueden interpretarse solo en el contexto de las otras variables de la ecuación. En resumen, utilizar los coeficientes beta solo como guía de la importancia relativa de las variables predictor incluidas en tu ecuación y solo en el rango de valores para el que realmente existe una muestra de datos.

La adición de nuevas variables independientes al modelo de regresión múltiple puede mejorar la predicción de la variable dependiente siempre y cuando se logre mejorar el valor del coeficiente de determinación R^2 preferentemente el ajustado o corregido, sea

significativo el valor del coeficiente de regresión para esa variable independiente β_k y esta no se corresponda con fenómenos de colinealidad o multicolinealidad con otras variables independientes.

Para cada modelo probable de la regresión múltiple a probar, que incluye las variables independientes X_k y la dependiente Y, los elementos relacionados con los coeficientes β_k de regresión múltiple se pueden representan de forma tabulada en los paquetes estadísticos (Tabla 4.66).

Tabla 4.66. Representación resumida de los elementos relacionados con el coeficiente de regresión múltiple estimado β para cada modelo de regresión múltiple

Número de modelo de regresión múltiple	**Coeficientes no estandarizados**		**Valor de t-student**	**p-valor (significación)**
	β	**Error típico de β**		
1 Valor de la constante				
Variable independiente (X_1)				
Variable independiente (X_2				
Variable independiente (X_k)				
2 Valor de la constante				
Variable independiente (X_1)				
Variable independiente (X_2)				
...				
Variable independiente *(X_k)*				
n Valor de la constante n				
Variable independiente (X_1)				
...				
Variable independiente (X_k)				

Al compararse dos modelos a estimar, uno sin ajustar (estimación cruda) y el otro ajustado (estimación ajustada) por un posible factor de confusión tras introducir la variable potencialmente confusora en el modelo, si hay un cambio suficientemente grande en la magnitud del coeficiente de regresión β para la variable independiente entre los dos modelos (al pasar del crudo al ajustado), entonces el factor por el que se ajustó es un factor de confusión y el modelo válido es el ajustado. Este criterio supone una mejoría respecto al planteamiento anterior de guiarse por la mera significación estadística (valor p). Ahora ya no debe enfocarse solo en los valores p, sino las magnitudes del efecto, antes y después de ajustar. Es un procedimiento práctico que se ha recomendado y usado muy frecuentemente (de hecho es lo que más se utiliza), pero que tampoco es ideal si se aplica de manera "automática" o dogmática y sin tener en cuenta otros criterios. Debe

pensarse siempre en términos de relaciones causales y ajustar solo por las variables que supongan realmente una posible causa común y no estén contando la misma historia que otra variable que ya está en el modelo. Lo ideal es conocer muy bien el estado de la ciencia sobre las posibles relaciones causales entre las variables que se manejan y recurrir a gráficos causales.

El modelo estimado final de la regresión múltiple es el que se conoce como modelo de efectos principales. Sobre él deben valorarse las interacciones una a una. Para esto se crean términos de producto y se van introduciendo, añadiéndolos uno a uno modelo principal. Cuando se comprueba que su introducción en el modelo no provoca cambios significativos en el coeficiente de regresión se elimina y se prueba el siguiente.

Validez de la ecuación de regresión múltiple o modelo teórico

Al igual que se hace con la ecuación de la regresión lineal simple, antes de utilizar en la vida práctica la ecuación de regresión múltiple para predicciones en sujetos es conveniente determinar su validez. Para esto se debe centrar en tres aspectos: el coeficiente de determinación o de correlación lineal múltiple al cuadrado, el análisis de varianza (ANOVA) y los coeficientes de regresión múltiple parciales (β_i).

Primeramente la exactitud en la predicción del valor de la variable dependiente, de respuesta o criterio a partir de las variables independientes incluidas (con sus respectivos coeficientes de regresión beta) en el modelo de regresión (ecuación de regresión estimada) se evalúa estimando el coeficiente de determinación (R^2) según la fórmula:

$$\text{Coeficiente de determinación } (R^2) = \frac{\text{Suma de los cuadrados de la regresión}}{\text{Suma total de los cuadrados}}$$ llevando a la expresión matemática se tiene que:

$$\text{Coeficiente de determinación } (R^2) = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2} = \frac{SC_E}{SC_T}$$

donde:

SC_E: suma de los errores al cuadrado explicado por regresión ($\sum(\hat{y}_i - \overline{y})^2$);

SC_T: suma de los errores al cuadrado total ($\sum(y_i - \overline{y})^2$);

y_i: valor observado i de la variable dependiente Y.

$\hat{y}_i$: valor medio predicho o estimado i por la ecuación de regresión múltiple para la variable dependiente Y.

$\overline{y}$: valor de la media observada para la variable dependiente Y en toda la muestra.

Este resultado es una medida de la bondad de ajuste de la superficie de regresión al indicar que proporción del total de la variación en los valores de Y observados,

es explicada por la regresión de la variable dependiente Y a partir de las variables predictoras X_1, X_2,..., X_k.

De forma práctica los datos se pueden recoger y quedar contenidos para su utilización en los respectivos cálculos (Tabla 4.67).

Tabla 4.67. Representación gráfica de la información (datos necesarios) para efectuar la evaluación en la significación estadística en la predicción de la variable dependiente del modelo de regresión múltiple

Número de observación u orden de individuos en la muestra	Valor observado de la variable dependiente	Valor observado de la variable independiente (1)	Valor observado de la variable independiente (2)	Valor del error de predicción (valor observado de la dependiente menos el valor predicho por la ecuación de regresión de la dependiente)	Valor del error de predicción al cuadrado
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
Totales:-	**-**	**-**	**-**	**-**	**-**

También puede emplearse en el sentido de la significación del modelo conjunto a partir del valor teórico para predecir la variable criterio el coeficiente de determinación (R^2) ajustado, que permite aclarar, según su aumento o no, si una variable independiente introducida mejora la predicción o no, respectivamente, de la variable Y de una forma más real y relacionada que el coeficiente de determinación (R^2) no ajustado, al tomar en cuenta aspectos en su ecuación que no incluye el coeficiente de determinación (R^2):

$$R^2_{\text{ajustado}} = R^2 - \frac{p(1-R^2)}{N-1-p}$$

donde:

p: número de variables predictoras independientes;

N: tamaño muestral, que se corresponde al número de observaciones para la variable dependiente Y, a partir de las p variables independientes.

Para cada modelo probable de la regresión múltiple, que incluye las variables independientes p y la dependiente Y, los elementos antes mencionados se representan gráficamente en los paquetes estadísticos (Tabla 4.68).

Tabla 4.68. Representación resumida de los tipos de coeficientes relacionados con la correlación utilizados en el análisis del modelo de regresión múltiple

Número de modelo estimado de regresión múltiple	Coeficiente de correlación lineal múltiple (R)	Coeficiente de determinación (R^2)	Coeficiente de determinación ajustado o corregido ($R^2_{ajustado}$)	Error típico de estimación
1	-	-	-	-
2	-	-	-	-
n	-	-	-	-

La significación estadística en la exactitud del ajuste del modelo puede ser estimada a partir del contraste de hipótesis basado en que la cantidad de variación explicada por el modelo de regresión múltiple es mayor que la variación explicada por la media, es decir, que R^2 es mayor que cero, en el que se utiliza el *ratio* o razón F como estadístico de contraste o de prueba. La fórmula matemática del estadístico F se define como:

$$F\ (\text{observado-calculado}) = \frac{\left[\dfrac{\text{Suma de los errores al cuadrado (regresión)}}{\text{Grados de libertad (regresión)}}\right]}{\left[\dfrac{\text{Suma de los errores al cuadrado (total)}}{\text{Grados de libertad (residuos)}}\right]}$$

Desarrollando esta expresión se puede obtener que:

$$F = \frac{\dfrac{\sum(\hat{y}_i - \overline{y})^2}{\#\beta - \#\beta_0 - 1 = \beta}}{\dfrac{\sum(y_i - \overline{y})^2}{N - (\#\beta + \#\beta_0)}} = \frac{CM_{\text{Regresión o explicado}}}{CM_{\text{Total}}}$$

donde:

Grados de libertad (regresión): número de los coeficientes de regresión lineal múltiple estimados β_i, incluido la constante $\beta_0 - 1$;

Grados de libertad (residuos): tamaño muestral menos el número de los coeficientes de regresión lineal múltiple estimados β_i, incluido la constante β_0;

SC_E: suma de los errores al cuadrado explicado por la regresión ($\sum(\hat{y}_i - \overline{y})^2$);

SC_{Total}: suma de los errores al cuadrado total ($\sum(y_i - \overline{y})^2$);

CM_E: cuadrado medio con error explicado;

CM_{total}: cuadrado medio con error explicado más cuadrado medio con error no explicado o residual;

y_i: valor observado i de la variable dependiente Y;

$\hat{y}_i$: valor medio predicho o estimado i por la ecuación de regresión múltiple o el modelo estimado para la variable dependiente Y;

$\overline{y}$: valor de la media observada para la variable dependiente Y en toda la muestra;

F: estadístico de contraste *ratio* o razón F de Fisher, que sigue una distribución de probabilidad $F_{k, n-k-1; 1-\alpha}$;
N: tamaño muestral.
Valor de p (p-valor) = Pr[F (observado-calculado) ≤ F (teórico-esperado)], según grados de libertad y valor de alfa prefijado, donde no es significativamente diferente de 0 el coeficiente de determinación R^2, bajo H_0.

Otra manera de realizar el contraste de hipótesis general de la predicción de la variable dependiente cuantitativa (criterio) a partir de la influencia del conjunto sumatorio de las variables independientes cuantitativas sería:

- H_0: no se puede utilizar el conjunto de variables independientes para estimar los valores de la variable dependiente.
- H_1: se puede utilizar el conjunto de variables independientes para estimar los valores de la variable dependiente.

El empleo del ANOVA permite, en este caso, al comprobar la significación del modelo de regresión lineal múltiple dar respuesta a este contraste (Tabla 4.69), donde los grados de libertad para la prueba de Fisher en este método dependen de la cantidad de variables independientes (k) incluidas en el análisis.

Tabla 4.69. Resumen de los elementos constituyentes esenciales del ANOVA empleado para el análisis de la regresión lineal múltiple

Fuente de variación	SC	Grados de libertad	CM	F
Debido a la regresión (error explicado)	$\sum(\hat{y}_i - \overline{y})^2$	k	$\frac{SC_E}{k}$	$\frac{CM_E}{CM_{NE}}$
Respecto a la regresión (error no explicado)	$\sum(y_i - \hat{y}_i)^2$	n − k − 1	$\frac{SC_{NE}}{(n-k-1)}$	
Total	$SC_E + SC_{NE}$	n − 1		

Donde:
SC_E: suma de cuadrados con error explicado por la regresión;
SC_{NE}: suma de cuadrados con error no explicado por la regresión;
SC_{Total}: suma de cuadrados total ($SC_E + SC_{NE}$);
CM_E: cuadrado medio con error explicado (según la regresión);
CM_{NE}: cuadrado medio con error no explicado (respecto a la regresión o residual);
y_i: valor observado i de la variable dependiente Y;
$\hat{y}_i$: valor medio predicho i por la ecuación de regresión múltiple o el modelo estimado para la variable dependiente Y;
$\overline{y}$: valor de la media observada para la variable dependiente Y en toda la muestra;

k: número total de variables independientes predictoras;

n: número total de observaciones;

F: corresponde al estadístico de contraste *ratio* o razón F de Fisher, que expresa cuantas veces es mayor la varianza explicada por la regresión respecto a la no explicada o residual, el mismo sigue una distribución de probabilidad $F_{k, n-k-1; 1-\alpha}$.

Valor de p (p-valor) = Pr[F (observado-calculado) ≤ F (teórico-esperado)], según los grados de libertad y el valor de alfa prefijado.

Como regla de decisión en cuanto a la significación del modelo de regresión múltiple se tiene que:

- Si el valor de F (observado-calculado) > valor de F (teórico-esperado) para la tabla de distribución de probabilidades correspondiente, según los grados de libertad en el numerador y denominador respectivos y el nivel de significación alfa seleccionado por el investigador, entonces se rechaza la H_0 y se acepta la significación estadística del modelo de regresión lineal múltiple, lo que equivale a decir que en la población de la que se extrajo la muestra la variable dependiente está relacionada de forma lineal con las variables independientes como un solo grupo, es decir, conocidos los valores de las variables independientes para un sujeto se predice mejor el valor medio de la variable dependiente con la ecuación de regresión múltiple que con la media calculada de Y para toda la muestra.
- Si la prueba no resulta significativa (p-valor ≥ α), no existe relación lineal entre la variable dependiente y las variables independientes como un solo grupo.

Respecto a los coeficientes de regresión β_k estimados parciales se realizan pruebas de hipótesis para evaluar la "fuerza predictiva" de la relación lineal entre la variable dependiente Y y las variables independientes X_i de manera individual; se pretende probar la hipótesis nula (H_0): $\beta_k = 0$ *versus* la hipótesis alternativa (H_1): $\beta_k \neq 0$, que fue expuesto al abordar la estimación de la ecuación de regresión múltiple.

La estimación de cada β_i parcial, según cada variable independiente X_i para la predicción de Y, se realiza siguiendo el método de los mínimos cuadrados, de una forma matemática muy similar a como fue descrita al ser abordada la regresión lineal simple; pero adaptada al análisis de la regresión lineal múltiple. La formulación empleada para este objetivo sería:

$$b_i = \frac{SP_{xy}}{SC_x} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$$

donde:

b_i: coeficiente de regresión muestral que estima el poblacional β_i;

SP: suma de productos, para la diferencia de los valores de X_i y Y_i respecto a sus medias respectivas;

SC: suma de cuadrados para la diferencia de los valores de X_i respecto a su media.

Tercer paso del modelo de regresión múltiple

Se debe determinar si cualquiera de las observaciones es capaz de ejercer una indebida influencia sobre los resultados a partir de la evaluación de la ecuación de predicción o de regresión múltiple con los coeficientes de regresión. Las observaciones que caen fuera de las pautas generales del conjunto de datos o que ejercen una fuerte influencia en los resultados de la regresión se clasifican en tres casos:

- Atípicos: son observaciones que tienen grandes valores residuales y pueden identificarse solo respecto a un modelo de regresión específico. Los atípicos han sido tradicionalmente la única forma de observación influyente considerada en los modelos de regresión, y se han desarrollado métodos de regresión específicos (regresión robusta) para tratar específicamente con el impacto de atípicos sobre los resultados de la regresión.
- Puntos de apalancamiento: son observaciones diferentes del resto de las observaciones de los valores de las variables independientes. Su impacto es particularmente destacado en los coeficientes estimados de una o más variables predictoras o independientes.
- Observaciones influyentes: en sentido amplio, incluyen todas las observaciones que tienen un efecto desproporcionado sobre los resultados de la regresión, no solo incluyen potencialmente atípicos y puntos de apalancamiento, sino que pueden incluir otras observaciones. Se basan en alguna de las cuatro condiciones:
  - Un error en la entrada de observaciones o datos.
  - Una observación válida aunque excepcional que es explicable por una situación extraordinaria.
  - Una observación excepcional sin una explicación plausible.
  - Una observación ordinaria en sus características individuales, pero excepcional en su combinación de características.

Pueden recomendarse varios cursos de acción para tratar con las observaciones influyentes de los diferentes tipos. Para un error en la observación se puede corregir el dato o eliminar la observación.

Con observaciones válidas, pero excepcionales (segunda condición), está autorizada la eliminación del caso a menos que las variables que reflejan la situación extraordinaria se incluyan en la ecuación de regresión. La observación inexplicable (tercera condición) presenta en un problema especial, dado que no existe razón para eliminar el caso, aunque su inclusión no puede justificarse.

Finalmente, la observación que es ordinaria en una variable, aunque excepcional en su combinación de características (cuarta condición) indica modificaciones para la base conceptual del modelo de regresión múltiple y debería retenerse.

Evaluación del valor teórico con respecto a los supuestos del análisis de regresión lineal múltiple

Una vez estimado el valor teórico correspondiente al modelo teórico de la regresión lineal múltiple, mediante la ecuación de regresión, que incluye la combinación lineal de todas las variables independientes con su respectivo coeficiente de regresión (β_i) y la constante (β_0), debe ser evaluado para comprobar el cumplimiento de los supuestos del análisis de regresión.

La cuestión básica es si en el proceso de cálculo de los coeficientes de regresión y predicción de la variable dependiente o criterio se cumplen los supuestos del análisis de la regresión como son:

- La linealidad del fenómeno medido.
- La varianza constante del término de error.
- La independencia de los términos de error.
- La normalidad de la distribución del término de error.

Para la comprobación de estos supuestos en el valor teórico del modelo se hace necesario emplear los métodos basados en el análisis de los residuos por gráficos o por pruebas estadísticas de significación como la t-student, donde la medida principal del error de predicción del valor teórico es el residuo, que expresa la diferencia entre los valores observados y las predicciones de la variable dependiente, criterio empleando la ecuación de regresión múltiple o modelo estimado. El no cumplimento de supuestos puede llevar a la transformación logarítmica de las variables para permitir su ejecución. Aunque el análisis de regresión se ha mostrado bastante robusto, incluso cuando se incumple el supuesto de normalidad, los investigadores estimarían el análisis de regresión tanto con las variables originales como con las transformadas para evaluar las consecuencias de la no normalidad de las variables independientes sobre la interpretación de los resultados.

Cuarto paso del modelo de regresión lineal múltiple

El último paso en el análisis multivariante de la regresión lineal múltiple es la Validación de los resultados obtenidos partiendo de un análisis de división de la muestra y el empleo del estadístico PRESS.

Después de identificar el mejor modelo de regresión, el paso final consiste en asegurarse de que represente a la población general (generalización) y que sea apropiada para situaciones en las cuales será utilizada (transferibilidad). La mejor guía es ver en qué medida se ajusta a un modelo teórico o a un conjunto de resultados validados previamente sobre el mismo asunto.

La aproximación más apropiada para la validación empírica de una ecuación de regresión múltiple obtenida es contrastar su modelo de regresión mediante la extracción de una nueva muestra de la población general, lo que asegura la representatividad de la población.

El modelo original (ecuación de regresión inicial estimada) puede predecir nuevos valores de Y a partir de los nuevos datos de las observaciones para las X_i con la nueva muestra, además de su respectivo ajuste predictivo; pero también se puede estimar un modelo separado con la nueva muestra para a continuación compararla con la ecuación original sobre las características de las variables incluidas: signo, valor e importancia relativa (por los coeficientes de regresión beta) de las variables independientes y su precisión predictiva o ajuste. En ambos casos se determina la validez del modelo original comparándolo con los modelos de regresión estimados con la nueva muestra.

En el caso de no poder obtener otra muestra de la misma población se puede dividir la muestra en dos partes: una submuestra de estimación para crear el modelo de regresión o ecuación de regresión y una submuestra de validación/duración utilizada para "contrastar" la ecuación mediante los procedimientos tanto aleatorios como sistemáticos para dividir el conjunto de datos.

Una aproximación alternativa a la obtención de muestras adicionales para validar el modelo es emplear la muestra original de forma especializada mediante el cálculo del estadístico PRESS, una medida similar al R^2 (coeficiente de determinación) utilizada para evaluar la precisión predictiva del modelo de regresión estimado, en el caso de querer comparar entre modelos con diferentes números de predictores o coeficientes de regresión múltiple se utiliza el R^2 ajustado, que es útil en la comparación de modelos entre diferentes conjuntos de datos, en la medida en que se compensan los diferentes tamaños muestrales.

Predicción con el modelo de regresión lineal múltiple

Las predicciones del modelo siempre pueden realizarse aplicando el modelo estimado para un nuevo conjunto de valores de las variables independientes y calculando los valores de la variables dependientes o criterio, no obstante, se deben tomar en cuenta estas recomendaciones:

- Se deben calcular siempre los intervalos de confianza de las predicciones, además de los puntos estimados para observar el rango esperado de los valores de la variable dependiente.
- Se debe asegurar de que las condiciones y relaciones medidas en el momento en que la muestra original fue tomada no han cambiado.
- No se debe utilizar el modelo para estimar más allá del rango de las variables independientes que se encuentran en la muestra.

Ventajas y desventajas de la regresión lineal múltiple

Aunque la estimación del modelo teórico o ecuación para la regresión lineal simple una vez validada puede tener una aplicación bioestadística importante en la comprensión y predicción de ciertos fenómenos en el área investigativa en general y traer ciertas ventajas, no obstante, pueden observarse algunas desventajas a tener en cuenta:

- Permite el análisis de varios factores o variables independientes que influyen en un desenlace cuantitativo.
- En estudios no experimentales se puede controlar el sesgo producido por características que hacen a los grupos heterogéneos (confusión) y, por tanto, no comparables entre sí.
- La introducción de más variables reduce el porcentaje no explicado por la regresión e incrementa lo que es capaz de explicar.
- Permite describir como un factor modifica el efecto de otro (interacción).
- Es difícil escoger el mejor modelo ya que a veces hay varios candidatos buenos.
- Algunas veces es difícil interpretar el significado del modelo en términos de la vida real.
- Es más difícil visualizar el aspecto del modelo ya que solo es posible representar gráficamente tres dimensiones.
- Los cálculos son prácticamente imposibles sin un ordenador con un paquete estadístico.

Respecto al probable papel de confusor de una variable independiente en la relación entre otra variable independiente para cierto efecto (variable dependiente):

- Solo hay confusión en la relación entre dos variables cuando ambas compartan una causa común.
- No hay ningún valor p ni ningún contraste de hipótesis que indique que una variable actúa como factor de confusión.
- La confusión no depende de la significación estadística, pero si está muy condicionada por el tamaño de muestra. Si hay un cambio suficientemente grande (del 10 % o mayor) en la magnitud del coeficiente para la variable independiente entre dos modelos, o sea, sin ajustar y con ajuste por el posible confusor (al pasar del modelo crudo al ajustado), entonces el factor por el que se ajustó es un factor de confusión.
- La confusión no se evalúa por valores p, sino por la magnitud del efecto y el conocimiento biológico.
- Lo más importante que hay que hacer para valorar la confusión es basarse en las relaciones causales entre variables que se esperan a partir del conocimiento previo o estado del arte para la relación entre variables de dependencia.

Respecto al probable papel de interacción de una variable independiente con otra para cierto efecto (variable dependiente):

- Se habla de interacción biológica entre dos o más factores causales cuando existe una interdependencia entre sus mecanismos de acción para curar, prevenir o controlar una enfermedad de manera que la incidencia de enfermedad atribuible a este conjunto de factores causales es diferente a la incidencia esperable de sus efectos individuales.
- Se habla de interacción en escala aditiva cuando el efecto conjunto de dos variables es significativamente superior (sinergia) a la suma de sus efectos individuales; también cabe la posibilidad de una interacción negativa (antagonismo). Hay antagonismo cuando el efecto conjunto es significativamente inferior a la suma de los efectos individuales. En regresión múltiple la escala es aditiva.
- Se habla de interacción en escala multiplicativa si el efecto conjunto de dos variables es significativamente distinto a la multiplicación de cada uno de sus efectos individuales.
- La interacción se plantea como un efecto significativamente distinto cuando hay acción conjunta del que sería esperable al combinar las acciones por separado, es decir, para la interacción sí hay una prueba estadística y lo que realmente importa es su valor p.

En la conformación del modelo de regresión múltiple se debe tener en cuenta el conocimiento biológico del problema, los hallazgos previos en otras investigaciones y la experiencia manejando grandes bases de datos, además de estos aspectos:

- Gráficos de dispersión.
- Matriz de correlaciones.
- Regresiones univariantes de cada variable independiente con la dependiente.
- Uso de regresiones no paramétricas.
- Selección de variables candidatas para el modelo multivariante.
- Valorar la contribución de cada variable incluida en el modelo por medio del coeficiente de determinación (R^2) ajustado.
- Valorar la colinealidad entre las variables independientes por medio del análisis de la tolerancia, que es la proporción de la varianza en una variable independiente no explicada por las otras variables independientes de la ecuación, donde una variable con una tolerancia muy baja contribuye con poca información a un modelo (es colineal) y se estima por:

Tolerancia (de k) = $1 - R_k^2$ (para esta variable independiente cuando es pronosticada por las otras variables independientes ya incluidas en el análisis).

- Inclusión de términos de interacción entre variables: se crean términos de producto y se van introduciendo, añadiéndolos uno a otros al modelo principal evaluándolos a partir de su p-valor.
- Comprobar los residuales del modelo: deben seguir una distribución normal para poder considerar el modelo como válido.

Relación entre la regresión lineal múltiple y la correlación lineal múltiple

Si todas las variables independientes continuas de un grupo de observación son el resultado de un muestreo aleatorio de alguna población de interés se puede estimar el grado de asociación o la fuerza de la asociación entre la variable dependiente y todas las variables independientes, mediante el método de la correlación lineal múltiple.

En el método de la correlación lineal múltiple la variable dependiente debe estar medida en un nivel por intervalos o de razón, mientras las independientes en cualquier nivel de medición, en caso de ser categóricas se transforman a variables *dummy*.

El resultado del análisis de la correlación lineal múltiple se puede expresar mediante el coeficiente de correlación múltiple o coeficiente de determinación (R^2), que expresa la fuerza de asociación entre cada variable independiente X del conjunto de las independientes X_i con la dependiente Y y la significación del conjunto se evalúa mediante los contrastes de hipótesis anteriormente descritos al abordar la significación del modelo de regresión estimado.

A medida que se añaden variables independientes al valor teórico, los cálculos deben considerar también las intercorrelaciones entre las variables independientes X_i, aparte de las correlaciones entre estas y la dependiente Y.

Si las variables independientes están correlacionadas, entonces "comparten" algo de su poder predictivo. Dado que se usa solo la predicción del conjunto del valor teórico (variables independientes seleccionadas por el investigador para la ecuación o modelo a estimar), la varianza compartida de la variable dependiente Y con cada una de las independiente X_i no debe ser "contabilizada dos veces" mediante el uso de las correlaciones directas entre cada variable independiente X y la dependiente Y. Por lo que se calculan las correlaciones adicionales para representar estos efectos compartidos: la primera es el coeficiente de correlación parcial, que es la correlación de una variable independiente (X) y una variable dependiente (Y) cuando se han suprimido los efectos de las otras variables independientes tanto de X_i e Y; una segunda forma de correlación es la correlación semiparcial, que refleja la correlación entre una variable dependiente e independiente mientras se controlan los efectos predictivos de todas las otras variables independientes X_i sobre X. Las dos formas de correlación difieren en que la correlación parcial elimina los efectos de las otras variables independientes sobre la X_i seleccionada e Y, mientras que la correlación semiparcial elimina solo los efectos sobre la independiente X seleccionada. La correlación parcial representa el aumento en el efecto predictivo de una variable independiente X_i para la dependiente Y, respecto al efecto colectivo del resto de las variables independientes y se utiliza para identificar variables independientes que tienen el mayor aumento de poder predictivo a partir de los valores del coeficiente de determinación R^2 al examinar la correlación y regresión lineal simple, dado un conjunto de variables independientes ya incluidos en

el valor teórico, lo que se empleado tanto antes de conformar la probable ecuación de regresión o el modelo estimado a probar, como al determinar cuáles variables deben permanecer en el modelo junto a la valoración de los coeficientes de regresión y las respectivas pruebas de significación cuando se utiliza el método de introducción de variables paso a paso o *stepwise*, por ejemplo.

La correlación semiparcial representa la relación única prevista por una variable independiente una vez que se ha obtenido la predicción compartida con el resto de las variables independientes, por tanto, la correlación semiparcial se utiliza para el "reparto" de la varianza entre las variables independientes. Elevando al cuadrado la correlación semiparcial se obtiene la varianza única de Y explicada o prevista por la variable independiente.

A continuación se representa la varianza única y compartida entre las variables independientes correlacionadas X_1 y X_2 (Fig. 4.18).

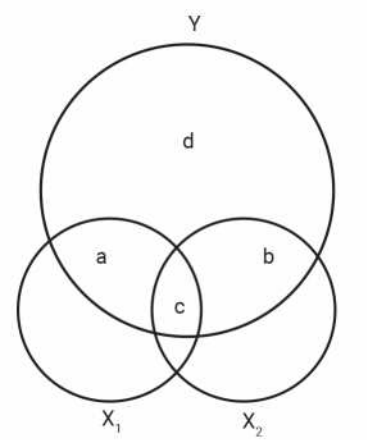


Fig. 4.18. Representación gráfica de los tipos de varianza: única y compartida entre la variable dependiente Y y las independientes X_1 y X_2 correlacionadas.

donde:

a: representa la varianza de Y explicada únicamente por X_1.

b: representa la varianza de Y explicada únicamente por X_2.

c: representa la varianza de Y explicada o prevista conjuntamente por X_1 y X_2.

d: varianza de Y no explicada por X_1 o por X_2.

La varianza asociada con la correlación parcial de Y, X_2 controlada por X_1 puede representarse como b/(d + b), donde d + b representa la varianza sin explicar después de tener en cuenta X_1.

La parte de la correlación de X_2 controlada por X_1 es b/(a + b + c + d), donde (a + b + c + d) representa el total de la varianza de Y, siendo la cantidad únicamente explicada por X_2.

El investigador puede también determinar la varianza única y compartida de la variable dependiente Y según las variables independientes X_i a través de cálculos simples.

La correlación semiparcial entre la variable dependiente (Y) y una variable independiente (X_1), mientras se controla una segunda variable independiente (X_2) (expresando r el coeficiente de correlación lineal simple muestral que estima el poblacional de Pearson) se calcula mediante la formulación de estos términos:

- Correlación semiparcial de Y, X_1 dada X_2 se expresa mediante la formulación que emplea el coeficiente de correlación r:

  $$r_{Y,\, X_1(X_2)} = \frac{\text{Coor de Y, } X_1 - (\text{Coor de Y, } X_2 \cdot \text{Coor de } X_1,\, X_2)}{\sqrt{1 - (\text{Coor de } X_1,\, X_2)^2}}$$ donde la varianza única prevista o explicada por $X_1 = [r_{Y,\, X1\ (X2)}]^2$

- Correlación semiparcial de Y, X_2 dada X_1 se expresa mediante la formulación que emplea el coeficiente de correlación r como:

  $$r_{Y,\, X_2(X_1)} = \frac{\text{Coor de Y, } X_1 - (\text{Coor de Y, } X_{12} \cdot \text{Coor de } X_1,\, X_2)}{\sqrt{1 - (\text{Coor de } X_1,\, X_2)^2}}$$ donde la varianza única prevista o explicada por $X_2 = [r_{Y,\, X2\ (X1)}]^2$.

La varianza total explicada (R^2) para la variable dependiente Y a partir de dos variables independientes X_1 y X_2 correlacionadas con la misma, es estimada empleando los términos como:

- Varianza directa o correlación entre dos variables: Y, X_1; Y, X_2; X_1, X_2, que se calcula empleando la formulación para el coeficiente de correlación lineal simple de Pearson (r) elevado al cuadrado (r^2).
- Varianza única de Y explicada o prevista por X_1 dado $X_2 = [r_{Y,\, X1\ (X2)}]^2$.
- Varianza única de Y explicada o prevista por X_2 dado $X_1 = [r_{Y,\, X2\ (X1)}]^2$.
- Varianza compartida de Y explicada por X_1 y $X_2 = [r^2_{Y,\, X1} - (r_{Y,\, X1\ (X2)})^2]$ o $[r^2_{Y,\, X2} - (r_{Y,\, X2\ (X1)})^2]$.
- Varianza total de Y explicada por X_1 y $X_2 = [r_{Y,\, X1\ (X2)}]^2 + [r_{Y,\, X2\ (X1)}]^2 + [r^2_{Y,\, X1} - (r_{Y,\, X1\ (X2)})^2]$ o $[r_{Y,\, X1\ (X2)}]^2 + [r_{Y,\, X2\ (X1)}]^2 + [r^2_{Y,\, X2} - (r_{Y,\, X2\ (X1)})^2]$.

Para recordar

La base para la estimación de todas las relaciones de regresión es la correlación que mide la asociación entre dos variables. Tanto para la regresión como para la correlación entre variables se requiere emplear coeficientes que permitan evaluar el grado, fuerza, intensidad o valor predictivo entre las variables. La tabla 4.70 resume aspectos conceptuales de los tipos de coeficientes de regresión y correlación más utilizados en la bioestadística actual en los proyectos investigativos.

Tabla 4.70. Representación gráfica resumen de los distintos tipos de coeficientes empleados en la correlación y la regresión lineal con su respectiva simbología e interpretación

Tipo de coeficiente	Símbolo	Interpretación del coeficiente
De correlación lineal simple de Pearson	r	Expresa el grado, intensidad o fuerza de asociación entre dos variables cuantitativas, es decir, entre una variable independiente y la otra dependiente (valores oscilan entre 0 y 1)
De Correlación lineal simple al cuadrado (de determinación).¿	r^2	Expresa el porcentaje que representa la variación de una variable dependiente debido a la variación de la independiente respecto a la variación total de la variable dependiente, es decir, de la varianza de factores comunes (valores oscilan entre 0 y 1 o entre 0 % y 100 %)
De correlación lineal múltiple al cuadrado, de determinación o de correlación parcial	R^2	Expresa el porcentaje de la variación de la variable dependiente debido a la variación de una variable independiente ajustando o controlando por el resto de las variables independientes consideradas (valores oscilan entre 0 % y 100 %)
De regresión lineal simple	B	Expresa el peso, influencia o cuanto predice la variable independiente a la dependiente, es decir, cuanto varía la dependiente por cada unidad que lo hace la independiente (valores oscilan entre –1 y 1)
De regresión lineal múltiple.	B_k	Expresa el peso, magnitud e influencia que tiene cada variable independiente sobre la dependiente, al controlar la varianza o remover los efectos de todas las demás variables independientes (valores oscilan entre –1 y 1)

Otro tipo de método de regresión multivariada entre los modelos log lineales lo constituye la regresión multivariada de Poisson.

El modelo estimado o ecuación de la regresión multivariada de Poisson se expresa como:

$\ln (DI) = a + b_1 X_1 + b_2 X_2 + \ldots\ldots + b_p X_p$ y se cumple que $DI = e^{(bi)}$

donde:

DI: razón de densidades de incidencia, es decir, de dos tasas de incidencia y se corresponde con la razón expresada mediante el cociente de los casos nuevos aparecidos de un evento no frecuente, según el comportamiento de dos categorías en la variable independiente con respecto al total de personas-tiempo de seguimiento en cada categoría de la variable independiente;

$b_1, b_2, \ldots, b_p$: coeficiente de regresión parcial b con los valores i de la ecuación o modelo de Poisson, que se estima por la formulación:

$$b_1, b_2, \ldots, b_p = \frac{\dfrac{\text{Número de casos nuevos con la categoría seleccionada de la variable independiente específica}}{\text{Total de personas con categoría analizada x el tiempo de seguimiento}}}{\dfrac{\text{Número de casos nuevos con la categoría de referencia de la variable independiente específica}}{\text{Total de personas con la categoría de referencia x el tiempo de seguimiento}}}$$

$X_1, X_2, \ldots, X_p$: variables independientes X con valores i del modelo estimado;
a: se corresponde con el valor constante que toma el modelo para el ln DI cuando el valor teórico de la ecuación a su derecha vale 0, es decir, cuando las X_p tienen valor nulo.

En este modelo la variable dependiente que es el número de casos o sujetos con un evento no frecuente según el número total de personas con cierta característica o factor relacionado (variable independiente) por el tiempo de seguimiento y se expresa como en otros modelos log lineales como una razón de tasas de incidencias.

El modelo de Poisson como todo modelo de regresión permite hacer predicción de riesgos, si es interpretado como razón de tasas de incidencia de un evento. Si se desea predecir la densidad de incidencia para un cierto evento en un sujeto basta con sustituir los valores de sus variables independientes X.

En los paquetes estadísticos actuales, de forma general se muestran los elementos para el análisis del modelo de regresión multivariada de Poisson como se representan tabularmente (Tabla 4.71).

Tabla 4.71. Elementos a tomar en cuenta en el análisis del modelo teórico estimado de la regresión multivariada de Poisson

Variable independientes (con la categoría seleccionada)	Coeficiente de Regresión (b_i)	Error estándar de b_i ($EE\ b_i$)	Estadístico de contraste $Z = [(b_i)/(EEb_i)$	p-valor (P > IZI)	IC (b_i)
X_1					
X_2					
X_p					
Valor de la constante de la ecuación (a)					
Total de personas con la categoría seleccionada /tiempo de seguimiento					

Método multivariado de análisis discriminante

El análisis discriminante o de separación-diferenciación en grupos de sujetos u objetos es la técnica estadística multivariada de dependencia apropiada cuando la variable dependiente es cualitativa (categórica o nominal u ordinal convertida en categórica) y las variables independientes son cuantitativas (métricas). Entre los aspectos que lo caracterizan:

- Tiene la capacidad de tratar tanto dos grupos como grupos múltiples (tres o más). Cuando se incluyen dos clasificaciones o categorías en la variable dependiente, la técnica es conocida como análisis discriminante de dos grupos, mientras que cuando se identifican tres o más clasificaciones la técnica es conocida como análisis discriminante múltiple.
- El análisis discriminante implica obtener un valor teórico, es decir, una combinación lineal de dos o más variables independientes que discrimine mejor entre los grupos definidos *a priori* para la variable cualitativa dependiente, es decir, se puede considerar como un análisis de regresión donde la variable dependiente es categórica y tiene como categorías la etiqueta de cada uno de los grupos, donde las variables independientes son continuas y determinan a qué grupos pertenecen los sujetos u objetos.
- Se pretende encontrar relaciones lineales entre las variables continuas que mejor discriminen en los grupos dados a los sujetos u objetos y construir una regla de decisión que asigne un objeto nuevo, que no se sabe clasificar previamente a uno de los grupos prefijados con un cierto grado de riesgo.
- La aplicación e interpretación del análisis discriminante es parecida al análisis de regresión múltiple, es decir, la función discriminante es una combinación lineal (valor teórico) de medidas métricas o cuantitativas de dos o más variables independientes y se usa para describir o predecir una única variable dependiente.
- En el análisis discriminante donde la variable dependiente (única) es categórica, y las variables independientes son cuantitativas (métricas) es el caso opuesto al MANOVA, donde se incluyen variables dependientes: cuantitativas (métricas) y variables independientes cualitativas categóricas.
- Se puede considerar al análisis discriminante tanto un tipo de análisis de perfil como una técnica predictiva analítica. En el análisis de perfil se proporciona una valoración objetiva de las diferencias entre grupos sobre un conjunto de variables independientes. En este caso el análisis discriminante es bastante similar al análisis multivariante de la varianza.
- Para comprender las diferencias de grupos el análisis discriminante tiene en cuenta tanto el papel de las variables independientes, como las combinaciones que se cons-

truyen con estas variables que representan dimensiones de discriminación entre los grupos. Estas dimensiones son los efectos conjuntos de varias variables que trabajan unidas para diferenciar los grupos.

El análisis discriminante puede tratar cualquiera de estos objetivos de investigación:

- Determinar si existen diferencias estadísticamente significativas entre los perfiles delas puntuaciones medias sobre un conjunto de variables de dos o más grupos definidos *a priori*.
- Determinar cuál de las variables independientes cuantifica mejor de las diferencias en los perfiles de las puntuaciones medias de dos o más grupos.
- Establecer los procedimientos para clasificar objetos (individuos, empresas, productos, entre otros), dentro de los grupos, en base a sus puntuaciones sobre un conjunto de variables independientes.
- Establecer el número y la composición de las dimensiones de la discriminación entre los grupos formados a partir del conjunto de variables independientes.

Pasos a tener en cuenta al realizar el análisis discriminante

Al igual que otros tipos de métodos multivariados el análisis discriminante requiere seguir un grupo de pasos en su ejecución como procedimiento para obtener los resultados esperados:

- Fijar los objetivos del análisis discriminante como pueden ser:
  - Determinar si existen diferencias estadísticamente significativas entre los perfiles de las puntuaciones medias de dos o más grupos definidos *a priori* para un conjunto de variables independientes empleando pruebas de hipótesis para el análisis de varianzas.
  - Determinar cuál de las variables independientes cuantifica mejor, de las diferencias en los perfiles de las puntuaciones medias de dos o más grupos empleando estadísticos específicos.
  - Establecer los procedimientos para clasificar objetos (individuos, empresas, productos, entre otros) dentro de los grupos, en base a sus puntuaciones sobre un conjunto de variables independientes.
  - Establecer el número y la composición de las dimensiones de la discriminación entre los grupos formados, a partir del conjunto de variables independientes.
- Conducir las cuestiones de diseño específicas y estar seguro de que se cumplen los supuestos básicos para poder ejecutarlo.
- Comprobar la obtención de la función discriminante y la determinación de su significación estadística en la separación de los dos o más grupos.

- Interpretación de la función discriminante con la determinación de cuál de las variables independientes es la que más contribuye a discriminar entre los grupos.
- Evaluar los resultados discriminantes para determinar la capacidad predictiva de la función discriminante construyendo una matriz de clasificación.
- Validación de la función discriminante mediante una ampliación de la muestra.
- Empleo del modelo estimado validado, en este caso la función discriminante, en la predicción de grupos discriminatorios de sujetos relacionando sus variables independientes con la dependiente.

Algunas claves a tener en cuenta al realizar el análisis discriminante

Al realizar este método multivariante no se debe dejar de tener en cuenta orientaciones y adecuaciones precisas para ejecutarlo de la manera más eficaz:

- El éxito en la aplicación del análisis discriminante requiere tener en cuenta, fundamentalmente, la selección correcta tanto de la variable dependiente como de las independientes, el tamaño muestral necesario para la estimación de la función discriminante y la división de la muestra con fines de validación.
- Los supuestos clave para obtener la función discriminante son el de normalidad en la distribución de los valores de las variables independientes y el de estructuras (matrices) de varianza y de covarianza desconocidas (pero iguales) para los grupos a formar. En la mayoría de los programas estadísticos se ofrecen dos o más contrastes para el supuesto de igualdad de las matrices de covarianzas o de dispersión. El contraste más común es el de la M de Box (proviene del término *Box's M*).
- Se hace necesario para la aplicación del análisis discriminante que existan al menos dos grupos y para cada grupo se necesitan dos o más casos.
- El número de variables discriminantes debe ser menor que el número de objetos menos dos: $x_1, ..., x_p$, donde $p < (n - 2)$ y n es el número de sujetos u objetos.
- El número máximo de funciones discriminantes es igual al mínimo entre el número de variables y el número de grupos menos 1 con q grupos $(q - 1)$ funciones discriminantes.
- Ninguna variable discriminante puede ser combinación lineal de otras variables independientes discriminantes, así como no debe estar presente la multicolinealida entre esas variables. La multicolinealidad consiste en que dos o más variables independientes están altamente correlacionadas, por lo que una variable puede venir muy bien explicada o predicha por otras variables y añadir poca capacidad explicativa al conjunto completo. Las relaciones no lineales no están reflejadas en la función discriminante a menos que se realicen transformaciones específicas de la variable para representar los efectos no lineales.

- El análisis discriminante es bastante sensible a la razón (*ratio*) entre el tamaño muestral y el número de variables predictoras o independientes. Muchos estudios sugieren un *ratio* o razón de 20 observaciones por cada variable predictora o factor. El tamaño mínimo recomendado es de cinco observaciones por variable independiente o predictora. El tamaño del grupo más pequeño debe ser mayor que el número de variables independientes. Como una regla práctica cada grupo debe tener al menos 20 observaciones, pero incluso, aunque todos los grupos excedan las 20 observaciones, el investigador debe también considerar los tamaños relativos de los grupos. En la etapa de clasificación, los grupos más grandes tienen una posibilidad desproporcionadamente más grande de clasificación. Si los tamaños de los grupos varían de forma importante puede que el investigador quiera muestrear aleatoriamente desde el grupo más grande y reducir su tamaño a un nivel comparable con el grupo más pequeño.
- El procedimiento habitual al trabajar con la muestra de estudio para ejecutar el análisis discriminante es dividir la misma en su totalidad en dos submuestras: una utilizada para la estimación de la función discriminante denominada *muestra de análisis* y otra con fines de validación denominada *ampliación de la muestra*. Este método de validación de la función se denomina *división de la muestra* o *enfoque de validación cruzada*. No obstante, no se ha establecido ninguna regla fiable en su forma práctica de ejecución, algunos investigadores prefieren una división 60 a 40 o 75 a 25 entre los grupos de análisis y ampliación respectivamente para un tamaño muestral de 100.
- Para la evaluación de la exactitud predictiva del modelo discriminante es necesario contar con las matrices de clasificación se determina la puntuación de corte óptima para cada variable independiente y se especifica el criterio para valorar la razón o el *ratio* de aciertos y la significación estadística de la exactitud predictiva.
- Se aconseja al investigador examinar todos los resultados por la presencia de casos atípicos y emplear técnicas específicas de evaluación y control para dicho inconveniente en el análisis discriminante.

Aspectos teóricos-matemáticos involucrados en el análisis discriminante

La comprensión teórica del método multivariante de análisis discriminante requiere tener en cuenta los principios matemáticos involucrados desde su origen y que constituyen el soporte para efectuar su aplicación por los sistemas computacionales. A continuación se muestran algunos de los más importantes involucrados en su génesis.

Modelo matemático

A partir de q grupos donde se asignan a una serie de sujetos observados u objetos y de p variables medidas sobre ellos (x_1,..., x_p), se trata de obtener para cada objeto una serie de puntuaciones que indican el grupo al que pertenecen (y_1,..., y_m), de modo que sean funciones lineales de x_1,..., x_p,

$y_1 = a_{11}x_1 + \ldots + a_{1p}x_p + a_{10}$ $\cdot y_m = a_{m1}x_1 + \ldots + a_{mp}x_p + a_{m0}$,

donde m = mín (q − 1, p), tales que discriminen o separen lo máximo posible a los q grupos. Estas combinaciones lineales de las p variables deben maximizar la varianza entre los grupos y minimizar la varianza dentro de los grupos.

Descomposición de la varianza

Se puede descomponer la variabilidad total de la muestra en variabilidad dentro de los grupos y entre los grupos. Si partimos de:

$$\mathrm{Cov}(x_j, x_{j'}) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})$$

Se puede considerar la media de la variable x_j en cada uno de los grupos I_1,..., I_q, es decir:

$$\bar{x}_{kj} = \frac{1}{nk}\sum_{i \in I_k} x_{ij} \text{ , para k = 1,..., q.}$$

De este modo la media total de la variable x_j se puede expresar como función de las medias dentro de cada grupo. Así tenemos que si:

$$\sum_{i \in I_k} x_{ij} = n_k \bar{x}_{kj} \text{ , entonces:}$$

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = \frac{1}{n}\sum_{k=1}^{q}\sum_{i \in I_k} x_{ij} = \frac{1}{n}\sum_{k=1}^{q} n_k \bar{x}_{kj} = \sum_{k=1}^{q}\frac{n_k}{n}\bar{x}_{kj}$$

De esta forma tenemos a:

$$\mathrm{Cov}(x_j, x_{j'}) = \frac{1}{n}\sum_{k=1}^{q}\sum_{i \in I_k}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})$$

Si en cada uno de los términos se pone:

$$(x_{ij} - \bar{x}_j) = (x_{ij} - \bar{x}_{kj}) + (\bar{x}_{kj} - \bar{x}_j)$$

$$(x_{ij'} - \bar{x}_{j'}) = (x_{ij'} - \bar{x}_{kj'}) + (\bar{x}_{kj'} - \bar{x}_{j'})$$

al simplificar se obtiene que:

$$\mathrm{Cov}(x_j, x_{j'}) = \frac{1}{n}\sum_{k=1}^{q}\sum_{i \in I_k}(x_{ij} - \bar{x}_{kj})(x_{ij'} - \bar{x}_{kj'}) + \sum_{k=1}^{q}\frac{n_k}{n}(\bar{x}_{kj} - \bar{x}_j)(\bar{x}_{kj'} - \bar{x}_{j'}) = d(x_j, x_{j'}) + e(x_j, x_{j'})$$

Es decir, la covarianza total es igual a la covarianza dentro de grupos (d) más la covarianza entre grupos (e). Si se denomina como t (x_j, $x_{j'}$) a la covarianza total entre x_j y $x_{j'}$ (sin distinguir grupos), entonces lo anterior se puede expresar como:

$$t(x_j, x_{j'}) = d(x_j, x_{j'}) + e(x_j, x_{j'})$$

En notación matricial esto es equivalente a:

$T = E + D$

donde:

T: matriz de covarianzas total;

D: matriz de covarianzas dentro de grupos;

E: matriz de covarianzas entre grupos.

Extracción de las funciones discriminantes

La idea básica del análisis discriminante consiste en extraer a partir de $x_1,..., x_p$ variables observadas en k grupos, m funciones $y_1,..., y_m$ de forma:

$y_i = a_{i1}x_1 + ... + a_{ip}x_p + a_{i0}$, donde el número de funciones m = mín (q − 1, p) y q expresa el número de grupos con p variables, tales que la correlación corr $(y_i, y_j) = 0$ para todo $i \neq j$.

Si las variables $x_1,..., x_p$ están tipificadas o expresadas en forma estandarizada, entonces las funciones: $y_i = a_{i1}x_1 + ... + a_{ip}x_p$ para i = 1,..., m, se denominan *funciones discriminantes canónicas*.

Las funciones $y_1, . . ., y_m$ se extraen de modo que:

- y_1: sea la combinación lineal de $x_1,..., x_p$ que proporciona la mayor discriminación posible entre los grupos.
- y_2: sea la combinación lineal de $x_1,..., x_p$ que proporciona la mayor discriminación posible entre los grupos, después de y_1, tal que corr $(y_i, y_2) = 0$.

En general y_i es la combinación lineal de $x_1,..., x_p$ que proporciona la mayor discriminación posible entre los grupos después de y_i −1 y tal que:

corr $(y_i, y_j) = 0$ paraj = 1,..., (i − 1)

Procedimiento matricial

Se sigue un método parecido al análisis factorial, así se busca una función lineal de sujetos u objetos observados con las mediciones x para las variables p: x1,..., xp expresados como: y = a0x, de modo que:

Var(y) = a0Ta = a0Ea + a0Da, es decir, la variabilidad total corresponde a la presente entre grupos (término con E) más la variabilidad dentro de grupos (término con D).

Se quiere maximizar la variabilidad entre los grupos respecto a la variabilidad total para discriminarlos mejor y esto equivale a hacer:

$\text{máx}\left(\frac{a'Ea}{a'Ta}\right)$, es decir, maximizar la varianza entre grupos en relación al total de la varianza.

Si se considera la función: $f(a) = \frac{a'Ea}{a'Ta}$, se observa que f es una función homogénea, es decir, f(a) = f(μa) para todo $\mu \in R$.

El hecho de que la función sea homogénea implica que calcular:

$\text{máx}\left(\frac{a'Ea}{a'Ta}\right)$ equivale a calcular: máx (a´Ea), tal que a´Ta = 1.

Como este es el esquema habitual de los multiplicadores de Lagrange, se define: L = a´Ea − λ(a′T − 1) y se calcula su derivada: $\frac{\partial L}{\partial a} = 0$ que puede plantearse como: $\frac{\partial L}{\partial a} = 2Ea - 2\lambda Ta = 0$, si se conoce que Ea = λTa y se despeja λa en la fórmula anterior, se tiene que:

$(T^{-1}E)a = \lambda a$.

Por tanto, el autovector asociado a la primera función discriminante lo es de la matriz $T^{-1}E$ (que no es simétrica en general).

Como: Ea = λTa, a0Ea = λa0Ta = λ, luego si se toma el vector asociado al máximo autovalor, se obtiene la función que recoge el máximo poder discriminante.

El autovalor λ_i asociado a la función discriminante indica la proporción de la varianza total, que es explicada por las m funciones discriminantes que recoge la variable y_i.

Para obtener más funciones discriminantes, se siguen sacando los autovectores de la matriz $(T^{-1}E)$ asociados a los autovalores elegidos en orden decreciente:

$$\begin{array}{lcl} a'_2 & \Rightarrow & a'_2x = y_2 \\ \vdots & \ldots & \vdots \\ a'_m & \Rightarrow & a'_mx = y_m \end{array}$$

Donde m = mín (q − 1, p). Estos vectores son linealmente independientes y dan lugar a funciones incorreladas entre sí. La suma de todos los autovalores $\sum_{i=1}^{m}\lambda_i$ es la proporción de varianza total que queda explicada o se conserva, al considerar solo los ejes o funciones discriminantes.

Como consecuencia, el porcentaje explicado por y_i del total de varianza explicada por $y_1,\ldots, y_m$ es: $\frac{\lambda_i}{\sum_{i=1}^{m}\lambda_i}\cdot 100\ \%$.

Por lo que se puede resumir que el autovalor es el cociente de la variación debida a las diferencias entre los grupos (medida mediante la suma de cuadrados intergrupos) y la variación que se da dentro de cada grupo combinada en una única cantidad (medida mediante la suma de cuadrados intragrupos). Este estadístico se diferencia de la F del análisis de varianza multivariante, en que no intervienen los grados de libertad. Su interés principal está en que permite comparar como se distribuye la dispersión o variación intergrupos cuando existe más de una función. Aunque un autovalor tiene un mínimo de cero, no tiene un máximo, lo que lo hace difícilmente interpretable por sí solo. Por esta razón se acostumbra a utilizar el estadístico lambda de Wilks (λ), que se encuentra estrechamente relacionado con los autovalores.

Comprobación de los supuestos paramétricos del análisis discriminante

En sentido estricto, la función discriminante minimiza la probabilidad de equivocarse al clasificar los individuos en cada grupo. Para eso las variables originales se deben distribuir como una normal multivariante y las matrices o estructuras de covarianzas deben ser iguales en todos los grupos. En la práctica es una técnica robusta y funciona bien, aunque las dos restricciones anteriores no se cumplan.

Respecto a la primera restricción para llevar a cabo el análisis discriminante, si el conjunto de variables independientes seleccionadas se distribuye como una normal multivariante, entonces cualquier combinación lineal de ellas se distribuye como una normal univariante. Por eso si alguna de las variables originales no se distribuye como una normal, entonces es seguro que todas las variables conjuntamente no se distribuirán como una normal multivariante.

La segunda restricción es la que se refiere a la igualdad entre las matrices de covarianzas de los grupos, para comprobar esto se puede usar la prueba M de Box, esta prueba tiene como hipótesis nula que las matrices de covarianzas de los grupos son iguales, lo que equivale a decir que los grupos proceden de la misma población y más concretamente que las matrices de varianzas-covarianzas poblacionales correspondientes a cada grupo son iguales entre sí. Se basa en el cálculo de los determinantes de las matrices de covarianzas de cada grupo y de los grupos de forma combinada.

El estadístico M de Box se formula de esta forma:

$$M = (n - g)\log|S| - \sum_{j=1}^{g}(n_j - 1)\log|S^{(j)}|$$

donde:

S: matriz de varianzas-covarianzas combinada (de todas las sumas de cuadrados intragrupos según los grados de libertad al unísono);

$S^{(j)}$: matriz de varianzas-covarianzas del j-ésimo grupo (de las sumas de cuadrados de cada grupo según grados de libertad hasta el j-ésimo grupo);

n: número total de casos de la muestra;

n_j: número de casos en el j-ésimo grupo de la muestra;

g: número de grupos formados en la muestra.

El estadístico M de Box carece de un tipo de distribución muestral conocida, pero su valor obtenido puede combinarse en formulaciones matemáticas con algunos elementos implicados en la formulación anterior, cuyos valores resultantes se pueden aproximar por la distribución teórica de probabilidades de estadísticos como chi-cuadrado o F de Snedecor. Para la toma de decisión respecto al contraste de hipótesis para la igualdad de matrices de covarianzas de los grupos o muestras que estiman las poblaciones, se toma en cuenta que si el p-valor estimado (donde se cumpliría H_0) según los estadísticos de contraste antes enunciados (que aparecen indistintamente en los paquetes estadísticos)

es menor que el valor alfa (generalmente tomado como 0,05) se rechaza la igualdad entre las matrices de covarianzas para los q grupos seleccionados.

La prueba de la M de Box es sensible a la falta de normalidad multivariante, es decir, matrices iguales pueden aparecer como significativamente diferentes si no existe normalidad en la distribución de las variables. Así mismo, esta prueba tiene poco poder si los tamaños de las muestras son pequeños, dado que en este caso un resultado que equivale a pensar en no significativo según el p-valor obtenido, no necesariamente indica igualdad en las matrices de covarianzas entre los grupos, por otra parte, si las muestras son grandes pierde efectividad esta prueba, pues es más fácil rechazar la hipótesis nula de la misma. En ocasiones en que no se cumplen los supuestos anteriores para efectuar la prueba M de Box se puede usar alternativamente para comprobar la homogeneidad de varianzas entre grupos a la prueba de Bartlett, cuyo procedimiento matemático se deriva de la prueba anterior o también se puede tomar en cuenta la prueba de Levene para el caso de no normalidad en la distribución de los valores de las variables independientes empleadas en el estudio.

Selección de las variables discriminantes

Primero se puede realizar un análisis descriptivo univariante calculando las medias y las desviaciones estándar de las variables independientes originales para cada uno de los grupos por separado. Si para alguna variable las medias de los grupos son diferentes y la variabilidad es pequeña, se considera que esa variable es importante a la hora de discriminar a los grupos, por lo que debe ser incluida inicialmente en el método que se escoja para identificar las funciones discriminantes en el procedimiento del análisis discriminante.

El proceso de discriminación entre grupos con empleo de la muestra de análisis requiere inicialmente estimar las medias de las puntuaciones discriminantes Z de los grupos para las variables independientes con el respectivo análisis de significación para incluirlas en el modelo. Esto último tomando en cuenta los valores aportados por el estadístico λ de Wilks univariante y la ratio F univariante con su respectivo p-valor por medio del ANOVA de una vía, donde:

- F univariante: se estima al comparar para cada variable, mediante un cociente las desviaciones de las medias de cada uno de los grupos a la media total y las desviaciones de las observaciones a la media dentro de cada grupo. Si F es grande para cada variable, entonces las medias de cada grupo están muy separadas y la variable discrimina bien. Si F es pequeña, la variable discrimina poco, ya que habrá poca homogeneidad en los grupos y estos estarán muy próximos en sus variabilidades para cada variable analizada.
- λ de Wilks: cuando se considera a las variables de modo individual, la λ es igual al cociente de la suma de cuadrados dentro de los grupos entre la suma de cuadrados total (sin distinguir grupos), es decir, equivale a las desviaciones de las observaciones a la media dentro de cada grupo, entre las desviaciones a la media total sin distinguir

grupos para cada variable. Si su valor es pequeño, la variable discrimina mucho: la variabilidad total se debe proporcionalmente más a las diferencias entre grupos, no a las diferencias dentro de los grupos. El estadístico λ de Wilks expresa la proporción de la variabilidad total debida a la variabilidad intragrupos de forma combinada y no debida a las diferencias entre los grupos, permite contrastar la hipótesis nula de que las medias multivariantes de los grupos (los centroides) son iguales, basándose en el principio de la razón de verosimilitud generalizada (según la cual la varianza de un espacio multivariante puede ser calculada mediante el determinante de la matriz de dispersión) y por ende su formulación se expresa como:

$$\lambda = \frac{\text{Suma de cuadrados intragrupos}}{\text{Suma de cuadrados toral}} = \frac{|S|}{|T|}$$

donde:

S: matriz de varianzas-covarianzas combinada, calculada a partir de las matrices de varianzas-covarianzas de cada grupo (intragrupos);

T: matriz de varianzas-covarianzas total, calculada sobre todos los sujetos como si pertenecieran a un solo grupo.

Los valores próximos a 1 indican un gran parecido entre los grupos, mientras que los valores próximos a 0 indican una gran diferencia entre ellos. Si se relaciona el estadístico λ de Wils con el estadístico eta (al cuadrado este último), se evidencia por la expresión:

$eta^2 + \lambda = 1$.

Como parte de este procedimiento de selección de las variables a analizar para estimar el modelo de la función discriminante se observan las relaciones entre las variables. Se calculan matrices de correlaciones en lugar de matrices de covarianzas por ser más fácilmente interpretables. Además de analizar la correlación entre pares de variables sin distinguir grupos se debe analizar las correlaciones de las variables dentro de cada grupo.

Variables originales a considerar en el modelo estimado (para la función discriminante)

La idea del análisis discriminante es construir funciones lineales de las variables independientes originales que discriminen a los sujetos u objetos observados entre los distintos grupos. Sin embargo, no todas las variables discriminan de la misma forma o tienen los mismos valores de significación de la F de Snedecor o de la λ de Wilks asociada univariantes empleando el ANOVA de una vía. Por lo que a la hora de construir las funciones lineales no es necesario incluir a todas las variables iniciales en la función, más bien se debe tener en cuenta los valores de estos estadísticos como criterio de selección.

Para la estimación de la función lineal discriminante de Fisher (D) como expresión de la combinación de las variables independientes clasificadoras o explicativas (X_k) elegidas por el investigador, se hace necesario estimar también los coeficientes de ponderación

(U_k) expresados como valores que acompañan a las variables incluidas en esa función y que consiguen hacer que los sujetos de uno de los grupos obtengan puntuaciones máximas y los sujetos del otro grupo puntuaciones mínimas, a partir de la sustitución de los valores respectivos de cada sujeto observado en las variables elegidas. Considerando que existen n observaciones, la función discriminante se puede expresar para los coeficientes de ponderación y las variables explicativas como:

$D_i = u_1X_{1i} + u_2X_{2i} + \ldots + u_kX_{ki} \qquad i = 1, 2,\ldots, n$

donde:

D_i: puntuación discriminante correspondiente a la observación i-ésima;

X_{ki}: representa las variables independientes explicativas para la observación i;

u_k: representa los coeficientes de ponderación de cada variable independiente X.

La estimación de los coeficientes de ponderación respecto a cada variable explicativa o discriminadora se realiza en base a un procedimiento matemático derivativo descrito por Fisher en 1936, mediante el cual se maximiza la razón de la variabilidad entre grupos respecto a la variación intragrupos, incorporada su versión informatizada en los programas estadísticos modernos como el SPSS.

Se pueden emplear fundamentalmente dos métodos de selección de las posibles variables independientes a incluir en la estimación del modelo o función discriminante: el método simultáneo o directo y el método por etapas o *stepwise*.

En el método simultáneo de selección de las variables se consideran todas las variables originales de forma directa que verifiquen un criterio de selección a la vez. Aquí la función discriminante se calcula basándose en el conjunto completo de variables independientes, sin considerar la capacidad discriminante de cada variable independiente. El método simultáneo es apropiado cuando, por razones teóricas, el investigador quiere introducir todas las variables independientes en el análisis y no está interesado en observar resultados intermedios basados solamente en las variables que discriminan mejor.

Por el contrario, el método por etapas o *stepwise* es una alternativa al enfoque simultáneo. En el mismo la estrategia de inclusión por pasos conlleva a que las variables independientes van siendo incorporadas paso a paso a la función discriminante tras evaluar su grado de contribución individual a la diferenciación entre grupos, evidenciando su capacidad discriminatoria. Se comienza eligiendo la variable que mejor discrimina según los valores de F univariante. La variable inicial se empareja entonces con cada una de las variables independientes (de una en una), y se elige la variable que más consigue incrementar la capacidad discriminante de la función en combinación con la primera variable, esto a partir de diferentes estadísticos donde los valores de λ de Wilks o F de cambio correspondiente con su respectivo p-valor, son los más empleados en la mayoría de los paquetes estadísticos. La tercera y posteriores variables se seleccionan de una manera similar.

A continuación se presentan algunos estadísticos de los que se pueden utilizar en el análisis de la selección de las variables a incluir en el modelo o función discriminante definitiva mediante el método por etapas o *stepwise*:

– Lambda de Wilks (λ): cada variable independiente candidata a ser incluida en el modelo se evalúa mediante un estadístico F_{cambio} (conocida como R de Rao) que mide el cambio que se produce en el valor de λ al incorporar cada una de las variables al modelo. Una vez obtenido el valor del estadístico F_{cambio} para cada variable se incorpora al modelo la variable a la que le corresponde el mayor valor F_{cambio}, es decir, la que produce el mayor cambio en λ. La formulación de la F_{cambio} es:

$$F_{cambio} = \left(\frac{n-g-p}{g-1}\right)\left(\frac{1-\lambda_{p+1}/\lambda_p}{\lambda_{p+1}}\right)$$

donde:

n: número de sujetos o casos válidos;

g: número de grupos;

λ_p: lambda de Wilks que corresponde al modelo antes de incluir la variable que se está evaluando;

λ_{p+1}: lambda de Wilks que corresponde al modelo después de incluir la variable que se está evaluando.

– Distancia de Mahalanobis (H^2): se incorpora en cada paso la variable que maximiza la distancia H^2 entre los dos grupos más próximos. La distancia multivariante entre los grupos a y b se define como:

$$H^2_{ab} = (n-g)\sum_{i=1}^{p}\sum_{j=1}^{p} w_{ij}^{*}\left(\overline{X}_i^{(a)} - \overline{X}_i^{(b)}\right)\left(\overline{X}_j^{(a)} - \overline{X}_j^{(b)}\right)$$

donde:

n: número de casos o sujetos válidos;

g: número de grupos;

$\overline{X}_i^{(a)}$: media del grupo a en la i-ésima variable independiente;

$\overline{X}_i^{(b)}$: media del grupo b en la i-ésima variable independiente;

w_{ij}^{*} : elemento de la inversa de la matriz de varianzas-covarianzas intragrupos.

– V de Rao: es una transformación de la traza de Lawley-Hotelling, que es directamente proporcional a la distancia entre los grupos. Al utilizar este criterio, la variable que se incorpora al modelo es la que produce un mayor incremento en el valor de V y se expresa como:

$$V = (n_k - g)\sum_{i=1}^{p}\sum_{j=1}^{p} w_{ij}^{*}\sum_{k=1}^{g}\left(\overline{X}_i^{(k)} - \overline{X}_i\right)\left(\overline{X}_j^{(k)} - \overline{X}_j\right)$$

donde:

p: número de variables en el modelo;

g: es el número de grupos;

n_k: número de casos válidos del grupo k;

$\bar{X}_i^{(k)}$: media del grupo k en la i-ésima variable independiente;

$\bar{X}_j^{(k)}$: media del grupo k en la j-ésima variable independiente;

$\bar{X}_i$: media de todos los grupos en la i-ésima variable independiente;

$\bar{X}_i$: media de todos los grupos en la j-ésima variable independiente;

w_{ij}^* : elemento de la inversa de la matriz de varianzas-covarianzas intragrupos.

De forma resumida los pasos a dar en este método, que es el más frecuentemente empleado son:

- Se incluye en el análisis la variable independiente que tenga el mayor valor aceptable para el criterio de selección o de entrada (según el valor de F parcial-univariante o el valor λ de Wilks, con mayor significación en la discriminación entre los grupos).
- Se evalúa el criterio de selección para las variables no seleccionadas. La variable que presenta el valor más alto para el criterio o sea de F univariante se selecciona (siempre que esté dentro de un límite).
- Se examinan dos variables seleccionadas según un criterio de salida (mediante la F multivariante de salida) y se examinan también las variables no seleccionadas, para ver si cumplen el criterio de entrada. Se excluyen o se incluyen variables según cumplan los criterios de entrada y de salida (determinados según F multivariante de entrada y de salida, respectivamente).
- Se repite el paso anterior hasta que ninguna variable más pueda ser seleccionada o eliminada.

Conjuntamente con todo lo anterior se puede considerar un número máximo de pasos, dado que una variable puede ser incluida y eliminada en más de una ocasión. Se toma el doble del número de variables originales como número máximo de pasos del método *stepwise*.

Otro aspecto a tener en cuenta en la selección de posibles variables para estimar la función discriminante m es la tolerancia. Para definirla se requiere acudir al concepto del coeficiente de correlación múltiple R_i, que se define para un conjunto de p variables y que expresa el porcentaje de variabilidad de la variable x_i (i = 1,..., p) recogida por el resto de (p − 1) variables. Si se eleva al cuadrado R^2_i se obtiene el coeficiente de determinación. Entonces la tolerancia se define como $1 - R^2_i$. Así, cuanto mayor sea la tolerancia de una variable, más información independiente del resto de las variables recogerá. De este modo si en una iteración dada del procedimiento *stepwise* la variable seleccionada verifica que su tolerancia respecto a las variables ya incluidas en la función discriminante es muy pequeña, esto impide obtener mayor información de las otras variables en conjunto, entonces esa variable no se debe incluir en esta correspondiente etapa para evitar la redundancia de información.

La F y la λ de Wilks multivariantes adoptan el término de entrada o salida en relación a que sus valores estimados para las p variables en la función discriminante, determinan si una nueva variable incorporada en la función discriminante, permite mejor o peor dis-

criminación del modelo por lo que la misma debe permanecer o ser retirada respectivamente según corresponda. En los paquetes estadísticos como el SPSS aparecen como:

- F de entrada (F *to enter*): el valor de F se relaciona con la disminución en la λ de Wilks que se produce si se incluye una variable dada de las que no están dentro de la función discriminante. Si el valor de F es pequeño la disminución de la λ de Wilks es inapreciable y la variable no entra en la función.
- F de salida (F *to remove*): el valor de F se relaciona con el incremento que se produce en la λ de Wilks si se elimina de la función discriminante una variable dada. Si el valor de la F es pequeño el incremento no es significativo en la λ de Wilks y la variable se elimina del análisis.

Generalmente se utiliza el criterio de que una variable pasa a formar parte de la función discriminante si su inclusión se asocia con que el valor del estadístico F es mayor que 3,84 (valor de entrada), mientras que es expulsada de la función si se corresponde con un valor menor que 2,71 (valor de salida).

Otra medida empleada en este procedimiento de selección de variables para estimar la función discriminante como modelo del análisis discriminante es la correlación canónica, que expresa una medida del grado de asociación entre las puntuaciones discriminantes de cada uno de los sujetos u objetos observados en la muestra de estudio y el grupo concreto de pertenencia, es decir, viene expresada esta relación de asociación a partir de la proporción de la variabilidad total explicada mediante la diferencia en el comportamiento para las variables empleadas entre los grupos para las funciones discriminantes.

Cuando solo se tienen dos grupos la correlación canónica es igual al coeficiente de correlación entre la puntuación discriminante y el grupo de pertenencia determinado previamente según la clasificación del sujeto u objeto observado para la variable cualitativa dependiente dicotómica en este caso, que se representa como una variable codificada en 0 o 1 según el criterio empleado para el grupo correspondiente. Las puntuaciones discriminantes se corresponden con los valores obtenidos al proyectar cada punto del espacio k dimensional de las variables sobre el eje discriminante.

Con más de dos grupos la correlación canónica es equivalente al estadístico eta (R) = raíz cuadrada del cociente entre la suma de cuadrados intergrupos y la suma de cuadrados total, que es utilizado en el análisis de varianza de un factor.

Evaluación de la significación de la función discriminante

De la relación lineal seleccionada de variables independientes incluidas en la ecuación de discriminación: cuando no existen diferencias entre los grupos, las funciones discriminantes solo indican variabilidad aleatoria (ruido). En caso contrario se puede usar la λ de Wilks para realizar una prueba en la cual la hipótesis nula es que las medias de las funciones discriminantes en cada grupo son iguales.

En la ocasión en que se tienen varios grupos y varias funciones se calcula una λ de Wilks total mediante el producto de las λ de Wilks de cada función. Esta se puede aproximar por una χ^2, usando esta transformación:

$V = -\left(n - 1 - \frac{p+k}{2}\right)\ln(\lambda)$, de modo que $V \sim \chi^2 p\,(k-1)$, con k igual al número de grupos en la variable de respuesta aproximadamente. De este modo si λ es pequeño entonces V es grande y se rechaza la hipótesis nula referida a que no existen diferencias entre las medias de las funciones discriminantes estimadas para los grupos correspondientes.

Si la significación asociada al valor de la χ^2 es menor que 0,05 (o bien otro valor prefijado) se rechaza la hipótesis nula (a este nivel de confianza).

En general la interpretación de las funciones discriminantes conlleva su evaluación mediante:

Para el caso de una función discriminante se emplea en su evaluación los aspectos como:

- Ponderaciones discriminantes: denominados como *coeficientes discriminantes* β en algunos paquetes estadísticos, representan la contribución relativa de cada variable independiente asociada a una función lineal discriminante; una ponderación pequeña puede indicar, bien que su correspondiente variable es irrelevante para determinar una relación, o bien que se ha apartado de la relación debido a un alto grado de multicolinealidad.
- Cargas discriminantes: denominadas también *correlaciones de estructura*, miden la correlación lineal simple entre cada variable independiente y la función discriminante, reflejan la varianza que las variables independientes comparten con la función discriminante y pueden ser interpretadas como cargas de los factores para valorar la contribución relativa de cada variable independiente a la función discriminante.
- Valores del estadístico F parciales o univariante: indican los niveles de significación asociados a cada variable independiente discriminatoria, valores de F con p-valor asociado menor que el alfa predeterminado, expresa significación estadística en la contribución de la variable independiente en la discriminación de los grupos según la función de discriminación estimada.

Para el caso de dos o más funciones discriminantes se emplea en su evaluación:

- Si son funciones separadas se realiza por medio de:
  - Ponderaciones discriminantes.
  - Cargas discriminantes.
  - Valores del estadístico F parciales.
- Si son funciones combinadas, entonces se realiza por medio de:
  - Rotación de las funciones: después de haber construido las funciones discriminantes, estas pueden ser "rotadas" para redistribuir la varianza.
  - Índice de potencia: es una medida relativa entre todas las variables que señala la capacidad discriminante de cada variable independiente para cada

función discriminante significativa; cuando se obtienen dos o más funciones discriminantes es una medida resumen o compuesta que es útil para describir las contribuciones de la variable entre todas las funciones significativas.

- Representación gráfica de las cargas: se emplea el procedimiento de extensión del vector siendo la longitud de cada vector la que indica la importancia relativa de cada variable para discriminar entre los grupos. Para extender un vector, el investigador multiplica la carga discriminante (preferiblemente después de la rotación) por su respectivo valor univariante de la F, los vectores señalan a los grupos que tienen las medias más altas sobre el respectivo predictor y están lejos de los grupos que tienen las puntuaciones medias más bajas.
- Representación gráfica de los centroides de grupo: los centroides de grupo son también extendidos en este procedimiento multiplicándolos por el valor aproximado de la F asociado a cada función discriminante. Si las cargas son extendidas, los centroides deben ser también extendidos para representarlas exactamente en el mismo gráfico; los valores de la F apropiados para cada función discriminante se obtienen a partir de la expresión:

$$\text{Valor F}_{\text{de la función}} = \text{Autovalor}_{\text{función}} \left[\frac{\text{Tamaño muestral utilizado en la estimación} - \text{Número de grupos}}{\text{Número de grupos} - 1} \right]$$

Elementos que expresan la contribución de las variables independientes a la función discriminante

Si usamos variables originales tipificadas, se obtienen los coeficientes a_{ij} que relacionan las variables con las funciones discriminantes como se muestra a continuación y constituyen los coeficientes discriminantes:

	y_1	$\cdots$	y_m
x_1	a_{11}	$\cdots$	a_{m1}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_p	a_{1p}	$\cdots$	a_{mp}

Los coeficientes discriminantes se pueden interpretar como las magnitudes de la importancia relativa, ponderaciones o el peso de cada variable en la respectiva función discriminante. Así, si a_{ij} es grande en valor absoluto, entonces hay una fuerte asociación entre la variable x_j y la función y_i, en relación al resto de variables. Aún así, al existir en general correlaciones significativas entre las variables originales se debe tener cuidado al hacer interpretaciones precipitadas.

Estas ponderaciones del valor teórico para cada variable independiente se estiman de tal forma que maximicen la varianza entre grupos (formados según el número de categorías de la variable dependiente) frente a la varianza dentro o intragrupos y conllevan el uso de la medida del estadístico H^2 o D^2de Mahalanobis con su respectivo nivel de

significación en el procedimiento por etapas para determinar las variables con mayor capacidad discriminatoria.

La combinación lineal de las variables con sus respectivas ponderaciones para el análisis discriminante, también conocida como *función discriminante*, se deriva de una ecuación o modelo que adopta la expresión:

$Z_{jk} = a + W_1X_{1k} + W_2X_{2k} + W_3X_{3k} + \ldots + W_nX_{nk}$

donde:

Z_{jk}: puntuación Z de la función discriminante j para el objeto o sujeto k;

a: constante del modelo;

W_i: ponderación o peso discriminante para la variable independiente con orden numérico i.

X_{ik}: variable independiente con orden numérico i para el objeto o sujeto k.

Otra forma de calcular la contribución de cada variable a una función discriminante es examinar las correlaciones entre los valores de las variables independientes y los valores de las puntuaciones de la función discriminante mediante una matriz de estructura. Se calculan dentro de cada grupo las correlaciones entre las variables y las puntuaciones, luego se combinan en una matriz *pooled within-groups correlation matrix*. Los valores obtenidos dan una medida de las contribuciones de cada variable independiente al modelo estimado en el análisis discriminante.

Evaluación de la capacidad predictiva discriminante de la función discriminante

La evaluación de la capacidad predictiva discriminante de la función discriminante se realiza mediante la clasificación de los sujetos u objetos observados en los diferentes grupos.

Una vez estimadas las funciones discriminantes, es decir, las combinaciones lineales de las variables originales definitorias, a cada sujeto u objeto se le puede asignar una puntuación o valor dado en la función discriminante mediante la sustitución de los valores observados X de cada unod e los sujetos u objetos para las variables independientes j incluidas definitivamente en el modelo estimado.

Esto equivale al valor que se recoge en una ecuación de regresión. Siendo así, si x_{ij} es el valor que alcanza el sujeto objeto i-ésimo en la variable j-ésima, entonces la puntuación o valor alcanzado en la función discriminante k será:

$y_{ik} = a_{k1}x_{i1} + \ldots + a_{kp}x_{ip} + a_{k0}$

Estas puntuaciones discriminantes obtenidas y_{ik} de emplean como parte del procedimiento para evaluar la capacidad predictiva de clasificación de los sujetos u objetos de la función discriminante en los grupos de forma predicha. Una de las formas es mediante el empleo de un enfoque frecuentista de esta manera: se lleva a cabo la construcción de matrices de clasificación con la determinación de una puntuación de corte óptima

o valor de la puntuación Z crítica. Esta puntuación es el criterio numérico frente al cual cada puntuación discriminante individual es comparada para determinar dentro de que grupo debe ser clasificado o ubicado cada objeto o sujeto observado en la muestra estudiada. La misma difiere dependiendo de si los tamaños o número de sujetos u objetos observados de los grupos son iguales o distintos. Si los grupos son de igual tamaño la puntuación Z crítica estará a mitad de camino entre los valores centroides (media de las puntuaciones para las variables en cada grupo). El punto de corte óptimo (valor calculado que es empleado como referencia para poder ubicar a los sujetos u objetos en los grupos predichos según las puntuaciones discriminantes calculadas por medio de la función discriminante) para dos grupos a discriminar de igual tamaño se obtiene según la fórmula:

$$Z_{CE} = \frac{Z_A + Z_B}{2}$$

donde:

Z_{CE}: es el valor de la puntuación de corte crítica para grupos de igual tamaño;

Z_A: es el valor del centroide del grupo A;

Z_B: es el valor del centroide del grupo B.

Si los grupos no son de igual tamaño y se supone que son representativos de las proporciones de la población, una media ponderada de los valores centroides de los grupos proporciona una puntuación de corte óptima para una función discriminante y se calcula como:

$$Z_{CU} = \frac{N_A Z_B + N_B Z_A}{N_A + N_B}$$

donde:

Z_{CU}: es el valor de la puntuación de corte crítica para tamaños de grupos distintos;

Z_A: es el valor del centroide del grupo A;

Z_B: es el valor del centroide del grupo B;

N_A: es el número de sujetos del grupo A;

N_B: es el número de sujetos del grupo B.

A modo de ejemplo, si el grupo A es mucho más pequeño que el grupo B, el punto de corte óptimo estará más cercano al centroide del grupo A que al centroide del grupo B. Además, si se utiliza una puntuación de corte no ponderada, ninguno de los objetos del grupo A estaría mal clasificado, pero una parte importante de los que están en el grupo B si lo estaría.

Un estadístico de amplio uso en el análisis discriminante que permite revelar lo correctamente que la función discriminante clasificó a los sujetos u objetos observados en cada grupo es el ratio o razón de aciertos expresado en por ciento el cual es análogo al estadígrafo R^2 empleado en la regresión múltiple. El contraste F para evaluar

la significación estadística del R^2 en la regresión múltiple es análogo al contraste χ^2 empleando D^2 con el mismo objetivo en relación a la razón de aciertos en el análisis discriminante.

Si el número de grupos es de tres o más, entonces el investigador debe decidir no solamente si la discriminación entre el total de los grupos es estadísticamente significativa, sino también si cada una de las funciones discriminantes estimadas es estadísticamente significativa.

Se dispone de un contraste de hipótesis para determinar la significación estadística de la precisión discriminatoria del modelo en el análisis discriminante, empleando el estadístico t-student. La expresión para el análisis de dos grupos con tamaño muestral igual sería

$$t = \frac{p - 0{,}5}{\sqrt{\frac{0{,}5(1{,}0 - 0{,}5)}{N}}}$$

donde:

p: proporción de sujetos clasificados correctamente de ambos grupos o razón de aciertos;

N: tamaño muestral.

Existen otros métodos para la medición de la capacidad predictiva aparte del correspondiente al ratio o razón de aciertos como es el método de la aleatoriedad (por el criterio de máxima aleatoriedad y el de aleatoriedad proporcional).

Si se desea emplear medidas de precisión clasificatoria fundamentadas estadísticamente y relacionadas con la aleatoriedad, como parte del análisis de la significación estadística de la capacidad predictiva de la función discriminante se puede emplear un contraste estadístico a partir de la matriz de clasificación cuando se compara con un modelo de aleatoriedad mediante el estadístico de contraste o de prueba:

$$\text{Q de Press} = \frac{\left[N - (nK)\right]^2}{N(K-1)}$$, que sigue una distribución χ^2 con un grado de libertad (χ^2 teórico – χ^2 esperado = 3,84)

donde:

N: tamaño muestral total;

n: número de observaciones correctamente clasificadas;

K: número de grupos.

Valor de p (p-valor) bajo H_0 = Pr ($\chi^2_{1,\,1-\alpha}$ teórico-esperado ≥ Q de Press (observado-calculado).

Regla de decisión

– Si Q de Press (observado-calculado) > $\chi^2_{1,\,1-\alpha}$ (teórico-esperado), entonces la matriz de clasificación (por la razón de aciertos) puede considerarse estadísticamente su-

perior o mejor a la que emplea la aleatoriedad al evaluar la capacidad predictiva de la función discriminante para el nivel de confianza $(1 - \alpha)$ seleccionado por el investigador.

También la evaluación de la capacidad predictiva de la función discriminante como modelo estimado en el análisis discriminante para la clasificación de los sujetos u objetos observados en los diferentes grupos a partir de las respectivas puntuaciones discriminantes de cada uno se puede llevar a cabo con el empleo de un método bayesiano mediante la regla de Bayes.

La probabilidad de que un sujeto u objeto j, con una puntuación discriminante: $D = (y_{j1}, ..., y_{jm})$, pertenezca al grupo i-ésimo se puede estimar a partir de que:

$$P(G_i|D) = \frac{P(D|G_i)P(G_i)}{\sum_{i=1}^{k} P(D|G_i)P(G_i)}$$

donde:

$P(G_i)$: probabilidad *a priori* y es una estimación de la confianza de que un objeto pertenezca a un grupo si no se tiene información previa, por ejemplo, si 70 de 113 personas sobreviven en la muestra, la probabilidad de sobrevivir se aproxima por 70/113, lo que representa al grupo que sobrevive, mientras 43/113 complementa la probabilidad para el grupo que no sobrevive. Las probabilidades *a priori* se pueden determinar de distintos modos. Si la muestra se considera representativa de la población se pueden usar las proporciones de los casos en cada grupo como estimadores de esas probabilidades. Cuando todos los grupos tienen el mismo número de sujetos u objetos y no se tiene ningún tipo de información previa se asignan probabilidades *a priori* iguales para todos los grupos.

$P(D|G_i)$: expresa la probabilidad de obtener la puntuación D estando en el grupo i-ésimo. Como las puntuaciones discriminantes D se calculan a partir de combinaciones lineales de p variables distribuidas según una normal se distribuyen a su vez como una normal, cuya media y varianza se estiman a partir de todas las puntuaciones que se recogen en el grupo i-ésimo.

$P(G_i|D)$: probabilidad *a posteriori* que se estima a través de $P(G_i)$ y de $P(D|G_i)$. En realidad mide lo mismo que la $P(G_i)$, pero refina la medida de incertidumbre al tener en cuenta la información que recogen las puntuaciones discriminantes D, es decir, lo que interesa es calcular la probabilidad de que un sujeto u objeto pertenezca al grupo G_i, dado que presenta la puntuación D. Se asigna un sujeto u objeto al grupo Gi cuya probabilidad *a posteriori* sea máxima, es decir, dado que presenta la puntuación D. Como se conoce el grupo al que pertenece cada sujeto u objeto observado de la muestra de estudio, según la clasificación empleada para la variable cualitativa dependiente al inicio del procesamiento de la información para el análisis discriminante, se puede comprobar la efectividad del método de clasificación usando la máxima probabilidad *a posteriori*,

cuando se observa el porcentaje de casos bien clasificados. No obstante, se tiene que tener en cuenta también la tasa de clasificaciones incorrectas esperadas según las probabilidades *a priori*.

Validación de la función discriminante

La validación de los resultados discriminantes se realiza mediante la validación cruzada empleando la ampliación de la muestra (submuestra de validación) para poder perfilar las diferencias de grupos.

Para validar la función discriminante por medio de matrices de clasificación la muestra debe dividirse aleatoriamente en dos grupos: uno de los grupos (la muestra de análisis) se utiliza para calcular la función discriminante empleada para el desarrollo de los aspectos anteriores y el otro grupo (la muestra de validación o ampliación de la muestra) se usa para la elaboración de la matriz de clasificación. El proceso consiste en multiplicar las ponderaciones o coeficientes generados para las variables independientes por la muestra de análisis por las medidas de las variables de la ampliación de la muestra en cada sujeto u objeto, obteniéndose las puntuaciones discriminantes respectivas. Estas puntuaciones se comparan con el valor de la puntuación de corte crítica y se clasifican los sujetos de forma de esta forma: pertenece al grupo A si $Z_n < Z_{ct}$ y pertenece al grupo B Si $Z_n > Z_{ct}$, donde:

Z_n: puntuación Z discriminante para el individuo n-ésimo;

Z_{ct}: valor de la puntuación de corte crítica.

Los resultados del proceso de clasificación de los sujetos u objetos de forma predicha por el modelo estimado empleado (función discriminante) se confrontan con los resultados reales asumidos de los sujetos u objetos observados que se habían clasificado anteriormente según la variable cualitativa dependiente mediante una tabla de contingencia, adoptando una forma matricial al confrontar los resultados observados (reales) en la variable dependiente y los predichos a partir del modelo estimado o función discriminante (sujetos u objetos clasificados en ambos grupos, de acuerdo a si la puntuación discriminante es mayor o menor que el valor del punto de corte calculado) (Tabla 4.72).

Tabla 4.72. Representación gráfica para la matriz de clasificación

Grupo real	Grupo predicho		Tamaño del grupo real	Porcentaje bien clasificado
	1	2		
1	-	-	-	-
2	-	-	-	-
Tamaño del grupo predico	-	-	-	-

Donde la precisión total de la clasificación por el modelo estimado (ratio o razón de aciertos %) = Porcentaje correctamente clasificado en el grupo 1 + Porcentaje correctamente clasificado en el grupo 2/Número total de observaciones) · 100 o lo que equivale decir que:

(Sumatoria del número de sujetos en que coincide lo observado y lo predicho en cada grupo) /Número total de observaciones) · 100.

Para recordar

A continuación, de forma resumida, se presentan algunas ideas a tomar en cuenta siempre que se intente realizar un análisis discriminante:

- El enfoque de las cargas es algo más válido que el uso de las ponderaciones y debe ser utilizado siempre que fuera posible.
- El uso de valores de la F multivariantes y univariantes junto a los respectivos valores λ de Wilks, así como las ponderaciones, las cargas discriminantes y la correlación canónica, entre otras, permiten al investigador la utilización de varias medidas y la búsqueda de algo de consistencia en las evaluaciones de las variables independientes.
- Si se estiman dos o más funciones, el investigador puede emplear varias herramientas gráficas y el índice de potencia, que ayuda a interpretar la solución multidimensional.
- Lo más importante es que el investigador emplee todos los métodos disponibles para llegar a la interpretación más precisa.
- Una vez que se han identificado las funciones discriminantes significativas, la atención se desplaza a averiguar el ajuste global de las funciones discriminantes consideradas. Esta valoración conlleva tres tareas:
  - Calcular la puntuación Z discriminante para cada observación.
  - Evaluar diferencias de grupo sobre las puntuaciones Z discriminantes.
  - Valorar la precisión en la predicción de pertenencia al grupo.
- La puntuación Z discriminante de cualquier función discriminante puede calcularse para cada observación mediante la fórmula:

  $Z_{jk} = a + W_1X_{1k} + W_2X_{2k} + W_3X_{3k} + \ldots + W_nX_{nk}$

A partir de esta combinación lineal de las variables independientes se obtienen una serie de puntuaciones discriminantes Z para cada objeto o sujeto en cada grupo. Las puntuaciones discriminantes se calculan de acuerdo con la regla estadística de maximizar la varianza entre los grupos y minimizar la varianza intragrupos. Si la varianza entre los grupos es grande con relación a la varianza intragrupos, se dice que la función discriminante separa bien los grupos.

- La comparación de las medias de los grupos sobre las puntuaciones Z ofrece una medida de la discriminación entre grupos.
- Una medida resumen de las diferencias entre grupos es una comparación de los centroides de grupo, o sea, la puntuación Z discriminante media para todos los miem-

bros del grupo. Una medida de éxito del análisis discriminante es su capacidad para definir funciones discriminantes que den lugar a centroides de grupo significativamente diferentes. Las diferencias entre centroides se miden en términos de la medida del estadístico H^2 o D^2 de Mahalanobis, para la cual se dispone de contrastes que determinan si las diferencias son significativamente distintas. El investigador debe asegurarse de que incluso con funciones discriminantes significativas, existen diferencias significativas entre cada uno de los grupos. Los centroides de grupo entre cada función discriminante pueden también dibujarse para mostrar los resultados desde una perspectiva global.

- El análisis discriminante también cuenta con otros medios de discriminación: la estimación y el uso de valores teóricos múltiples en casos donde hay tres o más grupos. En estos casos, esas funciones discriminantes llegan a ser dimensiones de discriminación, pues cada dimensión se separa y se diferencia de las otras, por lo que además de mejorar la explicación de la pertenencia a un grupo, estas funciones discriminantes adicionales aportan ideas entre las varias combinaciones de variables independientes que discriminan entre los grupos.
- La estimación de más de una función discriminante (para más de dos grupos a discriminar), proporciona al investigador tanto una mejora en la discriminación como perspectivas adicionales sobre las características y combinaciones que mejor discriminan entre los grupos.
- La capacidad en la predicción se valora por el número de observaciones clasificadas dentro de los grupos adecuados. El investigador dispone de varios métodos, criterios o enfoques para llevar a cabo el proceso de clasificación de los sujetos u objetos, aunque todos están encaminados a facilitar la evaluación en cuanto a la significación estadística y práctica del ajuste que la función discriminante puede alcanzar en un análisis discriminante.
- El empleo del modelo estimado y validado para el análisis discriminante en la práctica se realiza sustituyendo el valor observado de cada sujeto u objeto para cada variable independiente y multiplicando este por el respectivo coeficiente o ponderación discriminante de la misma, con lo cual los valores calculados se suman y se obtiene como resultado final numérico un valor que se hace corresponder con el grupo específico discriminado de pertenencia de cada sujeto u objeto.

Método multivariado de análisis multivariante de la varianza (MANOVA)

El análisis multivariante o multivariado de la varianza (MANOVA) es un modelo para analizar la relación entre una o más variables independientes de tipo cualitativas categóricas y dos o más variables dependientes de tipo cuantitativas (medidas por intervalos o razón). De forma paralela cuando se intenta un control estadístico midiendo variables

relacionadas con las dependientes antes de que se analicen las relaciones de estas últimas con las variables independientes se emplea el método de análisis multivariado de covarianzas (MANCOVA).

Los usos del MANCOVA pueden ser para:

- Evaluar diferencias entre grupos a través de múltiples variables dependientes con el poder de evaluar no solamente las diferencias totales entre las variables dependientes, sino diferencias entre las combinaciones de las mismas, representa una extensión del análisis de varianza (ANOVA) para cubrir casos donde hay más de una variable dependiente o cuando las variables dependientes simplemente no pueden ser combinadas, por lo que reconoce si los cambios en las variables independientes tienen un efecto significativo en las dependientes, es decir, señala qué grupos difieren en una variable o en el conjunto de variables dependientes.
- Identificar las interacciones entre las variables independientes y la asociación entre las dependientes lo que se puede analizar mediante las tres clases principales del MANOVA:
  - Hotelling's T: es parecida a la prueba t-student (para dos grupos), pero con más variables dependientes, es decir, una variable independiente dicotómica y varias dependientes.
  - MANOVA unidireccional: análogo al ANOVA de una sola vía, pero con más variables dependientes, es decir: una variable independiente multicategórica y varias dependientes.
  - MANOVA factorial: similar al ANOVA factorial, pero con dos o más variables dependientes, es decir, varias variables independientes cualitativas categóricas y varias dependientes.
- Los modelos del MANOVA tienen en común que forman combinaciones lineales de las variables dependientes cuantitativas (métricas) que discriminan (que diferencian o separan) mejor entre los grupos en un experimento o una situación no experimental. Es una prueba de significación de las diferencias en los grupos en un espacio multidimensional donde cada dimensión está definida por combinaciones lineales del conjunto de variables dependientes.
- El MANOVA no puede ser sustituido por ANOVA individuales para cada variable dependiente a partir de su relación con una variable independiente cualitativa multicategórica debido a que las variables dependientes están correlacionadas muy frecuentemente, por lo que los resultados de varios ANOVA pueden ser redundantes y difíciles de integrar.
- Al incluir dos o más variables dependientes simultáneamente no se consideran las diferencias entre las medias en cada variable, sino las diferencias en variables canónicas. El interés no solo es saber si los grupos definidos por las variables independientes difieren en las variables canónicas, sino conocer la naturaleza de estas últimas.

- En el MANOVA una variable canónica es una variable artificial generada a partir de los datos. Representa constructos y se compone de variables reales, que deben ser descritas en términos de variables dependientes. Lo anterior se efectúa por medio de las cargas de los coeficientes de correlación entre una variable dependiente y una variable canónica. Si una carga entre la variable canónica y la dependiente es positiva y elevada, significa que altos valores en la dependiente se asocian con altos valores en la canónica.
- El número de variables canónicas no puede exceder el número de variables dependientes, pero es común que el número de dependientes sea mayor que el de variables canónicas estadísticamente significativas o los grados de libertad.
- Tanto el ANOVA como el MANOVA son particularmente útiles cuando se utilizan conjuntamente con diseños de experimentos en los que el investigador controla o manipula directamente una o más variables independientes para determinar el efecto sobre una ANOVA o más MANOVA variables dependientes. El ANOVA y el MANOVA proporcionan las herramientas necesarias para juzgar la fiabilidad de cualquier efecto observado, por ejemplo, si una diferencia observada es debida a un efecto del tratamiento o a la variabilidad aleatoria del muestreo.
- Los supuestos necesarios para realizar el MANOVA incluyen:
  - Emplear tamaños muestrales adecuados de los grupos, por ejemplo, 98 encuestados por grupo, que permitan la potencia estadística adecuada (0,80) en los contrastes de diferencia de grupos.
  - Presencia de homogeneidad de las matrices de varianzas/covarianzas en los datos de los grupos, evaluadas por la M de Box en que debe existir grupos con diferencias no significativas.
  - Presencia de la correlación entre las variables dependientes cuantitativas evaluadas mediante el contraste de Bartlett de esfericidad en que debe obtenerse correlación significativa entre estas variables.
- El contraste de hipótesis general en el MANOVA postula que:
  - H_0: las medias de los grupos o para las categorías de las variables independientes no difieren entre sí en las variables canónicas. También puede ser planteado como los vectores simples de las puntuaciones medias de las variables dependientes según las categorías de la variable independiente categórica son iguales.
  - H_1: las medias de los grupos o para las categorías de las variables independientes difieren entre sí en las variables canónicas o los vectores simples de las puntuaciones medias de las variables dependientes según las categorías de la variable independiente categórica no son iguales.
- El MANOVA proporciona varios criterios con los que valorar las diferencias multivariantes de los grupos. Los estadísticos de prueba empleados en el contraste de hipótesis pueden ser:

- Hotelling's T^2 o traza de Hotelling: puede ser expresado su contraste a través de una ratio o razón F, cuando hay dos grupos formados por las variables independientes dicotómicas y al menos dos variables dependientes cuantitativas.
- Cuando el estadístico T^2 (observado-calculado) excede a T^2 (critico-teórico-esperado) para un nivel de significación alfa dado, se concluye que los vectores de las puntuaciones medias para las variables dependientes son diferentes.
- Wilks' lambda o estadístico U: expresado a través de una ratio o razón F, cuando hay más de dos grupos formados por las variables Independientes cualitativas multicategóricas es el contraste más comúnmente empleado para la significación global del MANOVA.
- Criterio de Pillai-Bartlett: expresado a través de una ratio o razón F, que es el más robusto, debe ser empleado si los tamaños muestrales disminuyen o si se incumple la homogeneidad de las covarianzas y es de utilidad con el uso de coeficientes canónicos.
- Mayor raíz característica o canónica (MRC) de Roy: mide las diferencias solamente sobre la primera raíz canónica o función discriminante entre las variables dependientes y es el más adecuado cuando las variables dependientes están fuertemente interrelacionadas en una sola dimensión, pero también es la medida que es más probable que se vea afectada gravemente por incumplimientos de los supuestos básicos.

Cuando alguna de las pruebas de significación que emplea a estos estadígrafos muestra significación estadística para la diferencia de las medias de los grupos esto indica que hay, por lo menos, una variable canónica significativa (aunque puede haber varias). Si diversas variables canónicas son significativas, esto muestra que se presentan diferencias en las variables canónicas en cuestión, entre los grupos formados a partir de las categorías de las variables independientes en relación con los datos cuantitativos de las variables dependientes.

– Los paquetes estadísticos que contiene el MANOVA suelen posicionar a los grupos de las variables independientes por puntuaciones discriminantes, estas son calculadas con una función discriminante, que es una ecuación de regresión para un compuesto de variables dependientes. A cada grupo se le asigna una puntuación discriminante en cada variable canónica. Las puntuaciones discriminantes que se corresponden con una variable independiente pueden ser cero o tener un valor positivo o negativo. Una puntuación discriminante positiva y elevada para un grupo indica que este se coloca por encima de los demás en la respectiva variable canónica y deben considerarse las cargas como positivas o negativas. Las puntuaciones discriminantes son utilizadas para interpretar las separaciones de los grupos en las variables canónicas, en tanto que las cargas se usan para evaluar y ligar los resultados de las variables dependientes. La función discriminante informa de que combinación lineal de las variables dependientes se produce la diferencia entre los grupos más fiable.

Evaluación del valor teórico dependiente del MANOVA

La evaluación del valor teórico dependiente del MANOVA según variables métricas dependientes para diferenciar grupos en las independientes se utiliza para evaluar cuál de las variables dependientes contribuye a las diferencias globales señaladas en los contrastes estadísticos es esencial, ya que se puede identificar un conjunto de variables que o acentúa las diferencias mientras otras variables no son significativas u oculta los efectos significativos de las restantes.

Uno de los procedimientos a emplear se denomina contraste *post hoc* o de comparaciones múltiples. Entre los más habituales se considera el contraste de Scheffe, el método de la diferencia en verdad significativa de Tukey, la extensión del enfoque de la diferencia menos significativa de Tukey, el contraste de rango múltiple de Duncan y el contraste de Newman-Kuels. Cada método identifica que comparaciones entre los grupos, por ejemplo, grupo 1 frente a grupo 2 y 3 presentan diferencias significativas. Proporcionan al investigador los contrastes para cada combinación de grupos, simplificando el proceso interpretativo. De estos el contraste de Scheffe es el más prudente con respecto al error de tipo I. Los restantes contrastes son clasificados en este orden: diferencia en verdad significativa de Tukey, diferencia menos significativa de Tukey, Newman-Kuels y Duncan.

Se puede realizar además un procedimiento conocido como *análisis de reducción* para valorar individualmente las diferencias entre las variables dependientes. Este procedimiento comprende el cálculo de un estadístico univariante F para una variable dependiente después de eliminarlos efectos de otras variables dependientes que la preceden en el análisis. El procedimiento es de alguna forma similar a la regresión por etapas, pero se examina si una determinada variable dependiente proporciona información única (incorrelacionada) sobre las diferencias de los grupos. Sus resultados son exactamente los mismos que los que se obtendrían si se realiza un análisis de covarianzas con las otras variables dependientes precedentes empleadas como covariables. Un supuesto decisivo del análisis de reducción es que el investigador conoce el orden en que se deben introducir las variables dependientes, ya que la interpretación puede variar fuertemente según diferentes órdenes para entrar. Las variables identificadas como no significativas son "redundantes" con las anteriores variables significativas, con lo que añaden más información referente a las diferencias entre los grupos. Se puede cambiar el orden de las variables dependientes para comprobar si los efectos de las variables son redundantes o únicos, pero el proceso es bastante complicado conforme se incrementa el número de variables.

Otros procedimientos incluyen más análisis de las funciones discriminantes, en particular sobre la primera función discriminante, para obtener información adicional sobre que variables diferencian mejor entre los grupos. Todos estos análisis están pensados

para hacer comprender al investigador cuál de las variables dependientes contribuye a las diferencias en el valor teórico dependiente entre los tratamientos.

Una vez que se ha evaluado la significación estadística de la diferencia entre las medias de los grupos formados para el valor teórico dependiente, el investigador puede examinar los resultados por medio de cualquiera de estos tres métodos:

- Interpretar los efectos de las covariables si estas se han empleado.
- Valorar que variables dependientes presentan diferencias entre los grupos.
- Identificar que grupos difieren en una sola variable dependiente o en el valor teórico dependiente completo.

Análisis múltiple de covarianza (MANCOVA)

En cualquier diseño del ANOVA se pueden incluir variables independientes métricas denominadas *covariables*, que permiten estimar la covarianza para los grupos formados a partir de las categorías de las variables independientes para los datos de la variable dependiente y el diseño es entonces calificado como análisis de la covarianza (ANCOVA).

Las covariables o covarianzas (variables cuantitativas o métricas) son incluidas normalmente en un diseño experimental para eliminar influencias extrañas de la variable dependiente, que incrementen la varianza dentro de los grupos. Esta situación es similar al uso de un factor de bloqueo, solo que esta vez la variable es métrica. Se emplean procedimientos similares a la regresión lineal para eliminar la variación de la variable dependiente asociada con una o más covariables. Después se lleva a cabo un ANOVA convencional sobre la variable dependiente ajustada. El análisis multivariante de covarianza (MANCOVA) es una extensión simple de los principios de ANCOVA para el análisis multivariante (variables dependientes múltiples), es decir, se puede considerar MANCOVA como MANOVA de los residuos de regresión o varianza en las variables dependientes que no son justificados por la participación de covariables. El MANCOVA es apropiado para lograr dos objetivos específicos: eliminar cualquier error sistemático fuera del control del investigador que pueda sesgar los resultados y tener en cuenta las diferencias en las respuestas debidas a características propias de los encuestados.

Una cuestión habitual al emplear el MANCOVA es determinar cuántas covariables se añaden en su análisis. Aunque el investigador quiera tener en cuenta tantos efectos extraños como sea posible, un número demasiado grande reduce la eficiencia estadística de los procedimientos. El número de covariables debe ser menor que (0,10 · tamaño muestral) – (número de grupos – 1).

Por ejemplo, para un tamaño muestral de 100 encuestados y cinco grupos, el número de covariables debe ser menor que 6, esto es (0,10 · 100) – (5 – 1). El investigador siempre debe intentar minimizar el número de covariables, mientras que también asegura las que no van a ser eliminadas, ya que en muchos casos, particularmente cuando hay tamaños muestrales pequeños, estas covariables pueden mejorar de manera importante la sensibilidad de los contrastes estadísticos.

Existen dos requisitos para el empleo del análisis de la covarianza: las covariables deben tener algún tipo de relación con las medidas dependientes y las covariables deben presentar una homogeneidad del efecto de la regresión, significando esto que las covariables tienen efectos iguales sobre la variable dependiente entre los grupos. Existen contrastes estadísticos para valorar si este supuesto básico es cierto para cada covariable empleada. Si alguno de estos requisitos no se cumple da entonces el empleo de la covarianza es inapropiado.

Validación de los resultados obtenidos en el MANOVA

Las técnicas del análisis de la varianza: ANOVA y MANOVA, se han desarrollado según la tradición de la experimentación, donde la replicaciónes el principal medio de validación. La especificidad de los tratamientos experimentales permite un amplio empleo del mismo experimento en múltiples poblaciones para evaluar la generalidad de los resultados. En términos de validación, el investigador debe ser precavido al analizar múltiples poblaciones y comparar los resultados como una mera prueba de validez. En las ciencias sociales y en la investigación económica, la experimentación verdadera es muchas veces sustituida por contrastes estadísticos en situaciones no experimentales como un estudio con encuesta. El investigador debe considerar de manera importante el empleo de covariables con el respectivo análisis de covarianzas para controlar otras características que podrían estar relacionadas con los grupos para las variables independientes como edad, género, procedencia, escolaridad, entre otras y que pueden afectar a las variables dependientes, pero que no están incluidas en el análisis.

Método multivariado de análisis factorial

El método multivariado de análisis factorial no establece dependencias entre variables, pues no se denominan en este método variables independientes y dependientes, ni hace contrastes de hipótesis, simplemente su propósito es identificar una serie de factores subyacentes (muy pocos) conformados a partir de relaciones lineales entre diferentes variables en una matriz de datos procedentes de esas variables. El análisis de las interrelaciones lineales de variables cuanitativas puede conducir a un número menor de variables denominadas factores (si no son observables) o componentes principales (si son observables).

Entre los aspectos que caracterizan el análisis factorial se pueden mencionar:

– Con el análisis factorial se aspira a reducir una información de múltiples variables originales a la que pueden proporcionar solo dos o tres factores o “componentes principales” que sean combinaciones lineales de ellas y constituyan resúmenes, dándoles a unas más peso que a otras y puede representar una dimensión específica subyacente.

– El objetivo del análisis factorial es conseguir un modelo con parsimonia, o sea, maximizar la cantidad de varianza de las variables analizadas que pueda ser explicada por el mínimo número de factores o componentes subyacentes.
– El tipo de análisis factorial exploratorio es utilizado cuando el investigador desconoce cuántos factores son necesarios para explicar las interrelaciones entre un grupo de características, indicadores o variables, es el más común en la investigación del área biomédica, mientras que el análisis factorial confirmatorio es utilizado para determinar el grado en el que un grupo de factores identificados *a priori* son capaces de representar los datos de una matriz. Este método es empleado cuando el investigador tiene cierto conocimiento sobre la estructura subyacente a una serie de datos. Este tipo de análisis requiere un conocimiento exhaustivo de las estructuras de covarianzas, por lo que es necesaria la utilización de *software* estadísticos más complejos.

Esto son los pasos a seguir en el análisis factorial con un enfoque exploratorio siguiendo el método para obtener componentes principales:

– Estandarizar o tipificar las variables originales restándole a cada dato su media y dividiéndolo por la desviación estándar (valores Z). Este nuevo valor Z indica a cuántas desviaciones estándar está el sujeto de la media del total de sujetos en la muestra. Esta estandarización de cada variable conduce a la obtención de una nueva variable Z con media 0 y varianza 1. Esto tiene la ventaja de que cada una de las variables tiene una varianza de 1 y por lo tanto la varianza total (s^2_{total}) del conjunto de variables es igual al número de variables.
– Extracción de los factores que es un proceso iterativo, que por lo trabajoso se debe realizar en procesadores informáticos casi exclusivamente, que consiste en ir refinando la matriz de correlaciones entre variables de forma repetida obteniendo los diferentes factores subyacentes a los datos. Cada factor viene a ser una suma ponderada de los valores que para cada variable presenta un sujeto. A esas ponderaciones o pesos se les llama *coeficientes*. Los criterios para determinar el número de factores a seleccionar pueden ser:
  • Seleccionar los factores con autovalores superiores a la unidad si existen menos de 40 variables, el tamaño muestral es elevado y el número de factores esperados oscila entre número de variables/5 y número de variables/3 siguiendo el criterio de Kaiser.
  • Determinar el porcentaje acumulado de varianza extraída por diferentes factores, se termina la extracción de factores cuando alcanza un porcentaje acumulado de varianza explicada del 75 % a 80 %.
  • Representar gráficamente los factores extraídos (colocados en el eje de abscisas) frente a sus autovalores (colocados en el eje de ordenadas) para establecer un punto de inflexión en la gráfica (prueba de la pendiente de Cattell).

- Emplear dos conceptos claves en el análisis factorial:
  - La colectividad o comunalidad (*communalities*) (c) de una variable estandarizada es la cantidad de varianza de una variable explicada por los factores extraídos y puede oscilar entre 0 y 1, donde un valor de 0 para la colectividad de la variable Z(e) se traduce en que ninguna cantidad de la varianza de esa variable es explicada por los factores extraídos. Un valor 1 en e indica que toda la varianza de Z es explicada por el conjunto de factores extraídos en el análisis.
  - El autovalor (*eigenvalue*) de un factor (λ_i) se define como la cantidad de varianza total de la muestra que puede ser explicada por un determinado factor, generada a partir de los datos de las variables incluidas en el análisis factorial. La suma de los autovalores de todos los factores posibles es igual a la varianza total (s^2_{total}) del conjunto de variables estandarizadas.
- Seleccionar los factores con un autovalor superior o igual a 1 ya que son los únicos capaces de explicar más varianza que la que explica cada una de las variables originales y la finalidad es reducir el número de variables.
- La interpretación de los factores extraídos se realiza mediante la valoración de las correlaciones (coeficientes de Pearson) entre cada uno de los nuevos factores obtenidos y las variables originales expresado como a_i. El cuadrado de este coeficiente equivale a la comunalidad de la variable analizada: $c_i = a_i^2$. Los mayores valores de las correlaciones obtenidas en cada factor expresan las combinaciones lineales o sumas ponderadas de variables en cada factor.
- El análisis de estos factores o componentes principales también conlleva que si en cada factor se asigna un peso a cada variable por procedimientos matemáticos de correlación y a estos pesos o ponderaciones se les llama también coeficientes, a las sumas ponderadas de las variables que forman cada factor se les puede llamar *puntuaciones* y de esto se deriva que el comportamiento de los factores extraídos pueden ser perfectamente estimados con el empleo de las variables originales incluidas en el análisis factorial para cada sujeto de la muestra:

  $$\text{Factor } j = \sum_{i=1} W_{ij} \times Z_{i;}$$

  donde:

  i: representa cada variable en el análisis;

  Z_i: valor estandarizado de cada variable = (cada valor de la variable X_i – media muestral)/desviación estándar muestral;

  j: representa cada factor o componente principal;

  $W_{i,j}$: peso de la variable i sobre el factor j, también se denomina la carga y representa la correlación de la variable i con el factor j en la matriz factorial o de componentes.

  Por ejemplo, para el factor 1 con empleo de 10 variables originales se expresa:

  $$\text{Factor 1} = W_1Z_1 + W_2Z_2 + W_3Z_3 + W_4Z_4 + W_5Z_5 + W_6Z_6 + W_7Z_7 + W_8Z_8 + W_9Z_9 + W_{10}Z_{10}$$

Esto se repite para los otros factores que se hayan generado en el análisis inicial de los autovalores conformados en cualquier paquete estadístico.

Finalmente se tiene una reducción de variables a factores o componentes principales y se puede estimar la varianza total de las variables en la muestra explicada a partir de los factores generados y también contando con los pesos y valores de las variables para cada sujeto para determinar el comportamiento de cada factor en el mismo.

Nota: La correlación entre los factores es igual a 0, por lo que son ortogonales y no admiten análisis de asociaciones entre ellos.

Algunos cálculos numéricos según sus expresiones matemáticas a emplear:

– Cálculo del valor de una variable a través de los factores extraídos:

$$\text{Factor } Z_i = \sum_{j=1} (a_{ij}\text{Factor j})$$

$$Z \text{ variable } i = (a_{i1} \cdot \text{Factor 1} + a_{i2} \cdot \text{Factor 2} + \ldots\ldots + a_{ij} \cdot \text{Factor j})$$

donde:

i: cada variable;

$Z_{i:}$ valor estandarizado de cada variable;

j: cada factor;

a_{ij}: coeficiente de correlación de la variable i con el factor j. Es representado en la matriz factorial o matriz de componentes;

$a_{11\ldots} a_{ij}$: coeficientes de correlación de las i variables con los j factores de la ecuación.

– Correlación entre dos variables (r): la correlación para dos variables (i e i') es la suma de los productos de los pesos para estas dos variables (a_i) x ($a_{i'}$) a lo largo de los j factores. Esta ecuación queda expresada a través de la fórmula:

$$r_{i\,i'} = \sum_{j=1} a_i \cdot a_{i'}$$

$$r_{\text{variable}}\ 1,2 = (a_{11} \cdot a_{21}) + (a_{12} \cdot a_{22}) + (a_{13} \cdot a_{23}) + \ldots + (a_{1j} \cdot a_{2j})$$

donde:

r_{ii}': coeficiente de correlación de Pearson entre las variables i e i'.

$a_{1i,}\ a_{2i} \ldots a_{1j'};\ a_{2j}$: coeficientes de correlación de la variables 1 y 2 con los j factores.

– Colectividad de una variable (C): una vez llevado a cabo el análisis factorial. si calculamos la suma de los coeficientes de una variable i con los j factores al cuadrado, se obtiene el valor de su colectividad siguiendo la fórmula:

$$C_i = \sum_{i=1} (a_i)j^2$$

$$C_i = (a_{11})^2 + (a_{12})^2 + (a_{13})^2 + \ldots + (a_{1j})^2 = 1$$

donde:

i: cada variable;

j: cada factor;

$a_{11\ldots} a_{1j}$: coeficiente de correlación de la variable 1 sobre los j factores.

– Autovalor de un factor (λ): el autovalor asociado a un factor j puede calcularse a través de la suma de los cuadrados de los pesos de dicho factor j en cada variable i:

$$\lambda_j = \sum_{i=1} (a_j)^2 i$$

$\lambda_j = (a_{1i})^2 + (a_{2i})^2 + (a_{3i})^2 + \ldots + (a_{ii})^2 = 1$

donde:

i: cada variable;

j: cada factor;

a_{11}…. a_{i1}: pesos de las i variables sobre el factor 1.

Las condiciones necesarias para la aplicación del análisis factorial y extracción de componentes principales:

– Examinar la matriz de correlación de las variables implicadas en el análisis y si no existe un suficiente número de correlaciones significativas superiores a 0,30, la aplicación de un análisis factorial es probablemente inapropiado, así como si las variables están muy correlacionadas (r > 0,80), ya que pueden dar problemas de multicolinealidad conduciendo a un sistema inestable.
– Examinar la adecuación de la muestra, mediante las pruebas de esfericidad de Bartlet con el criterio donde no se rechace la hipótesis nula de no correlación entre las variables y la prueba de Kaiser-Meyer-Olkin que compara los coeficientes de correlación de Pearson entre cada par de variables con sus respectivos coeficientes de correlación parciales con el criterio de valores mayores o iguales a 0,70.
– Examinar el tamaño muestral donde el número de sujetos necesario para llevar a cabo un análisis factorial depende del número de variables que se vayan a analizar, se sugiere la inclusión de diez sujetos por variable incluida en el análisis.
– Examinar la rotación de los factores para tratar de mejorar el significado e interpretación de los factores obtenidos originariamente. La rotación ortogonal asume que los factores generados son independientes unos de otros (no están correlacionados), en la rotación oblicua se asume que los factores no son independientes, existiendo correlación entre dos o más de los factores.

En teoría, cada variable tiene una alta correlación con un único factor tras la rotación de los factores, mientras en la práctica, después de la rotación, suele ocurrir que algunas variables tengan bien una baja correlación con todos los factores, bien una alta correlación con múltiples factores, por lo cual:

– Para variables con bajos coeficientes de correlación con todos los factores se sugiere:
  • Eliminar las variables que no posean correlaciones lo suficientemente altas con alguno de los factores.
  • Evaluar la colectividad de las variables con bajas correlaciones con los factores y su contribución en la muestra objeto de valoración. Si la colectividad de esta

variable es pequeña y su importancia relativa en la base de datos es mínima, la variable es eliminada.

- Para variables con altos coeficientes de correlación con múltiples factores, se sugiere:
  - Eliminar la variable con altos valores, ya que su inclusión dificulta la interpretación de los diferentes factores.
  - Decidir sobre cuál de los posibles factores debe incluirse dicha variable. Generalmente, la variable suele incluirse dentro de aquel factor con el que se encuentra más relacionada desde el punto de vista conceptual.
    - Para variables importantes en el análisis con baja correlación con los factores, se sugiere, aunque no existe consenso entre los diferentes expertos, el mantenimiento de las mismas en la solución final.

Condiciones que facilitan el proceso de interpretación de los factores:

- Cuanto mayor es el coeficiente de correlación entre la variable y el factor, mayor es el grado de solapamiento entre la variabilidad de la variable y el factor y más parecido es el factor a la variable en cuestión.
- Cuanto mayor es el número de variables con altos coeficientes de correlación con un factor, más sencillo resulta aislar lo que el factor probablemente representa.
- Cuanto más pura es una variable que define a un factor, más sencillo resulta hacer inferencias sobre la naturaleza del factor.

Diferencias entre el análisis factorial común y el análisis de componentes principales según el tratamiento realizado para la varianza total del conjunto de variables:

- Análisis de componentes principales:
  - Esta primera aproximación establece que la varianza total de las variables es explicada por los factores extraídos.
  - Estima que la colectividad inicial para cada variable es 1.
  - Pueden obtenerse tantos factores como variables se incluyen en el análisis.
- Análisis factorial común:
  - Esta segunda aproximación separa la varianza total de las variables en dos componentes: uno de los componentes denominado *varianza común*, es la cantidad de varianza que las variables comparten en común y que es reflejada (explicada) por uno o más factores; el otro componente, denominado *varianza única* es exclusiva a cada variable y no participa en la identificación del factor.
  - Estima que la colectividad inicial para cada variable es siempre inferior a la unidad (establece diferentes métodos para calcularla entre los que destaca el coeficiente de correlación múltiple R^2).
  - Pueden obtenerse siempre un número de factores inferior al número de variables incluidas en el análisis.

Validación del análisis factorial

El método más directo de validación de los resultados consiste en adoptar una perspectiva de confirmación, valorando la replicabilidad de los resultados, bien dividiendo la muestra con los datos originales, bien con una muestra adicional. La comparación de los resultados de dos o más modelos factoriales siempre ha sido problemática. Sin embargo, existen varias alterativas para realizar una comparación objetiva. El auge del análisis factorial confirmatorio a través de la modelización de ecuaciones estructurales supone una alterativa, pero generalmente es más complicado y requiere *software* adicional como LISREL o EQS. Además del análisis factorial confirmatorio se han propuesto otros métodos que van desde un simple índice de adecuación a programas (FMATCH) diseñados especialmente para valorar la correspondencia entre matrices de factores. Estos métodos cuentan con un uso ocasional, debido, en parte, a la percepción de ausencia de sofisticación y a la no disponibilidad de *software* o programas analíticos que automaticen las comparaciones. Por eso cuando el análisis factorial confirmatorio n oes apropiado estos métodos facilitan una base objetiva para la comparación.

Otro aspecto de la generalización es la estabilidad de los resultados del modelo factorial. La estabilidad factorial depende primeramente del tamaño muestral y del número de casos por variable.

El investigador siempre está obsesionado por contar con el mayor tamaño muestral posible y desarrollar modelos parsimoniosos que incrementen la razón: casos/variable. Si el tamaño muestral lo permite, el investigador puede dividir aleatoriamente la muestra en dos subconjuntos y estimar los modelos factoriales de cada uno. La comparación de las dos matrices factoriales resultantes suministra una valoración de la robustez de la solución a lo largo de la muestra.

Además de la generalidad, otro aspecto de importancia para la validación del análisis factoriales la detección de observaciones influyentes. El investigador procura estimar el modelo con y sin observaciones identificadas como atípicas para valorar su influencia sobre los resultados. También son de aplicación al análisis factorial varias medidas de influencia que reflejan la posición relativa de una observación respecto a las otras, por ejemplo, la razón de la covarianza. Finalmente se han propuesto métodos específicos de análisis factorial para identificar observaciones influyentes, pero su complejidad ha restringido su aplicación.

Consideraciones finales

Dependiendo de los objetivos de partida al aplicar el análisis factorial, el investigador puede detenerse en la interpretación de los factores o proseguir con uno de los métodos de reducción de datos.

Si el objetivo simplemente consiste en identificar combinaciones lógicas de variables y una mejor comprensión de las relaciones entre variables, entonces basta con la interpretación de los factores.

Esta facilita una base empírica para enjuiciar la estructura de las variables y la influencia de esta estructura cuando se interpretan los resultados a partir de otras técnicas multivariantes. Si el objetivo, sin embargo, es identificar variables apropiadas para aplicaciones subsiguientes de otras técnicas estadísticas se emplea alguna forma de reducción de datos. Las alterativas consisten en examinar la matriz de factores y seleccionar la variable con mayor carga factorial como un representante de una dimensión factorial particular o reemplazar el conjunto original de variables por uno totalmente nuevo con menos variables creado a partir de escalas aditivas o de la puntuación de factores. Cada alterativa suministra nuevas variables para ser usadas como variables independientes en una regresión o en el análisis discriminante o como variables dependientes en el análisis multivariante de la varianza, incluso como las variables agrupadas en el análisis clúster.

Capítulo 5

Empleo del método o paradigma bayesiano en la inferencia bioestadística

Aunque las bases del método o enfoque bayesiano para el análisis estadístico de la información datan de hace más de dos siglos, con los primeros trabajos del reverendo Thomas Bayes, quien inició la solución cuantitativa al problema de determinar cuál de varias hipótesis es más probable sobre la base de los datos, solo recientemente se asiste a un uso apreciable de este enfoque en la investigación biomédica. Las razones que explican tal realidad y que a la vez augura un cada vez más prominente futuro es que algunos de los problemas de cierta complejidad, que posee este método, exigen el uso de recursos computacionales accesibles solo muy recientemente para el común de los investigadores. Este enfoque trata de una aproximación metodológica exenta de casi todas las críticas que se les hacen a las pruebas de significación aportadas por el método frecuentista y goza del atractivo de incorporar las evidencias aportadas por experiencias previas dentro del proceso analítico y las contempla en las conclusiones. En este capítulo se exponen las características fundamentales de este enfoque para la estadística inferencial, diferenciándolo del método clásico frecuentista en su metodología y bases de interpretación de los resultados, así como se brindan algunas de las posibilidades o aplicaciones en el manejo con muestras y variables, empleando los contrastes de hipótesis con sus respectivas particularidades en la bioestadística inferencial actual.

Diferencias entre el método frecuentista (clásico) y el bayesiano

El método frecuentista fue desarrollado a partir de los trabajos de Ronald Fisher, Jerzy Neyman y Egon Pearson en la década de los años treinta del siglo xx, mientras que en el paradigma bayesiano intervinieron el reverendo inglés Thomas Bayes (1702-1761), Richard Price y Pierre Laplace.

El método frecuentista emplea la prueba de significación en la evaluación del resultado estimado, englobando un índice para medir la fuerza de la evidencia llamado valor p y un procedimiento de elección de hipótesis, denominado *prueba de hipótesis* para este

resultado, mientras el paradigma bayesiano se apoya en las probabilidades *a priori* de un parámetro, valores puntuales y por intervalos de confianza previamente conocidos o de referencia llamado *componente subjetivo*, la verosimilitud que se refiere al valor numérico de probabilidad, la cual expresa cuan correspondientes son los datos observados de la muestra de estudio suponiendo que el valor del parámetro especifico fuera el empleado de referencia previa, llamada *componente objetivo*, y la combinación multiplicatoria de ambas en un nuevo componente denominado *las probabilidades a posteriori*, que sintetiza el nuevo grado de convicción del investigador para el resultado respecto al parámetro en estudio, siendo expresadas puntualmente y para las que se hace referencia también a sus intervalos de probabilidad, correspondiendo con los diferentes percentiles que dan su ubicación en la distribución de las mismas.

La diferencia esencial entre ambos métodos o paradigmas radica en que el método frecuentista se pronuncia sobre los datos a partir de supuestos (distribuciones de probabilidades), o sea, expresa como la probabilidad de obtener ciertos datos supuesta la veracidad de H_0 [$p = Pr(D \geq d_0 \mid H_0)$], mientras el paradigma bayesiano se pronuncia sobre los supuestos (valores de los parámetros de referencia) partiendo de los datos y se expresa como la probabilidad de que sea válida H_0, suponiendo que se observaron los datos que dan lugar a diferencias[$Pr(H_0 \mid D = d_0)$].

El método frecuentista obliga a una decisión dicotómica (rechazo o no) y sus resultados dependen crucialmente de un elemento exógeno a los datos que es el tamaño de muestra, a un efecto pequeño observado con un tamaño de muestra grande puede corresponder el mismo valor p que a un gran efecto registrado por medio de una muestra pequeña, incluso una diferencia mínima quizás "intrascendente" de ese valor puede ser declarado como estadísticamente significativa sobre todo en muestras grandes.

El paradigma bayesiano no trata de "elegir" entre una hipótesis y otra, en rechazar o no la hipótesis nula (H), sino identificar en qué medida o probabilidad una hipótesis H es más razonable que la otra alternativa (K), apoyando su análisis en el empleo del factor de Bayes (a favor y en contra de H), las probabilidades *a posteriori* de la veracidad de la hipótesis H y el intervalo de probabilidades *a posteriori* para la máxima distribución del parámetro según corresponda al planteamiento de la hipótesis.

No obstante, el paradigma bayesiano puede guardar cierta relación con el método frecuentista, al emplear un elemento del método frecuentista solo en su punto de partida como son los intervalos de confianza del parámetro, adoptado como probabilidad *a priori* en ocasiones por el paradigma bayesiano, elemento que en ocasiones sustituye a las pruebas de hipótesis en el método frecuentista para apartarse de la interpretación automática del valor de p en ese método.

El paradigma bayesiano se relaciona más con el tipo de situaciones en que se encuentra el científico habitualmente, o sea, tiene los datos (parámetros estimados incluyendo sus intervalos de confianza) y lo que quiere es descubrir que circunstancias o condiciones determinaron que los datos se comportaran así y no de otra forma.

El paradigma bayesiano constituye una aproximación metodológica exenta de casi todas las críticas que se hacen a las pruebas de significación y goza del atractivo de incorporar las evidencias aportadas por experiencias previas dentro del proceso analítico y las contempla de seguro en las conclusiones.

El paradigma bayesiano emplea el teorema de Bayes que constituye el puente para pasar de una probabilidad *a priori* o inicial [Pr(H)], de una hipótesis H a una probabilidad *a posteriori* o actualizada [Pr(H|D)] basado en una nueva observación D (de los datos muestrales).

El paradigma bayesiano puede proveer un nivel de protección mucho mayor que el método frecuentista contra la posibilidad de ver significación en hallazgos procedentes de la investigación científica que son enteramente espurios.

En el paradigma bayesiano a medida que aumentan las probabilidades *a priori* (q) de la validez para hipótesis H (de nulidad), tiende a aumentar la probabilidad *a posteriori* de la veracidad de la hipótesis H, así mismo, las muestras pequeñas no pueden afectar el resultado de las probabilidades *a posteriori* para la veracidad de la hipótesis H como si puede ocurrir con el caso del p-valor en el método frecuentista.

El axioma del paradigma bayesiano o teorema de Bayes se deriva de la definición de probabilidad condicional como: si se tienen dos sucesos A y B, donde A y B son ambos sucesos posibles, con probabilidad no nula, entonces la probabilidad condicional de A dado B se define como:

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$$ de forma análoga se puede plantear:

$$Pr(B|A) = \frac{Pr(A \cap B)}{Pr(A)}$$, entonces si se sustituye:

$Pr(A \cap B) = Pr(BA)Pr(A)$

La regla o axioma de Bayes queda expresada como:

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)}$$

Para el caso en que A_1, A_2, … A_K son sucesos mutuamente excluyentes, uno de estos debe de ocurrir necesariamente la ley de probabilidad total se expresa como:

$$Pr(B) = \sum_{i=1}^{k} Pr(B|A_i)Pr(A_i)$$

De modo que tomando el suceso A_j en lugar de A y aplicando al denominador dicha ley se tiene, otra forma de expresar la regla de Bayes como:

$$Pr(A_j|B) = \frac{Pr(B|A_j)Pr(A_j)}{\sum_{i=1}^{k} Pr(B|A_j)Pr(A_i)}$$, siendo para el caso particular en que k = 2:

$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}$ donde $\bar{A}$ representa el complemento de A, para dar lugar a la forma más conocida del teorema de Bayes expresada anteriormente.

La regla de Bayes produce probabilidades inversas en el sentido de que se expresa Pr(A|B) en términos de Pr(B|A).

La terminología convencional para Pr(A|B) es la probabilidad *a posteriori* de A dado B y para Pr(A) es la probabilidad *a priori* de A, dado que se aplica antes sin estar condicionada por la información de que B ocurrió.

De forma general el teorema de Bayes al aplicarlo a las probabilidades *a posteriori* se expresa como:

$$p_{ji}^{*} = P\left(E_j \middle| x_i\right) = \frac{P\left(x_i \middle| E_j\right)P_j}{\sum_{h=1}^{k} P\left(x_i \middle| E_h\right)P_h} \text{ , } j = 1, \ldots k; \; i = 1,\ldots, r$$

donde:

$E_1, E_2,\ldots, E_k$: k condiciones mutuamente excluyentes;

$P_j = P(E_j)$: probabilidad *a priori* de la condición E_j, j = 1,..., k con $\sum_{j=1}^{k} P_j = 1$;

$x_1, x_2,\ldots, x_r$: r observaciones posibles;

$p_{ij} = P(x_i E_j)$: a verosimilitud de x_i bajo la condición E_j, i = 1,...r, j = 1,....,k.

Una forma alternativa de expresar el teorema de Bayes en el caso que se trabaje con un evento y su complemento es especialmente útil cuando se trabaja con *Odds*, es decir, la razón entre la probabilidad de que un suceso o evento ocurra y la probabilidad de que no ocurra, lo que se expresa como:

$\frac{Pr(H|datos)}{Pr(K|datos)} = \frac{Pr(datos|H)Pr(H)}{Pr(datos|K)Pr(K)}$, donde H representa que una hipótesis dada es válida y K que lo es la hipótesis complementaria.

El apoyo relativo que dan los datos a cada una de estas dos hipótesis en términos del factor de Bayes (FB) se expresa como:

$$FB = \frac{Pr(datos|H)}{Pr(datos|K)}$$

El teorema de Bayes tiene una versión análoga para el caso en que, en lugar de estar ante un conjunto finito de sucesos, ocurriera que ese conjunto es infinito. En tal caso se opera con funciones de densidad para caracterizar la distribución *a priori* y el teorema "devuelve" una densidad *a posteriori*, luego de observar ciertos datos. La formulación de este caso se omite para no perder sencillez en este punto.

Lo cierto es que resulta útil, además de ser consistente con las demandas de la intuición, poder contar con un método que combine las evidencias subjetivamente

acumuladas con la información objetiva obtenida de un experimento en particular. Los métodos bayesianos ofrecen esta posibilidad. Por ejemplo, al estimar un parámetro, su valor, naturalmente, se desconoce, pero es absurdo suponer un vacío total de ideas o de presunciones en la cabeza de quienes tienen esa ignorancia, y una persona puede expresar su opinión acerca de su conocimiento e incertidumbre por medio de una distribución probabilística de un conjunto de valores posibles del parámetro. La distribución *a priori* es la distribución probabilística que refleja y sintetiza la idea acerca de lo que esa persona razonablemente piensa antes de observar los datos. Está claro que la realidad objetiva (muchas veces representada a través de lo que los estadísticos llaman "parámetros") es de determinada manera (ajena a las impresiones que uno tenga de ella) y también es obvio que no se comporta aleatoriamente, pero la visión sobre ella sí, y en gran medida depende de la comprensión del problema y, sobre todo, del conocimiento general prevaleciente en el momento del estudio.

Procede, sin embargo, aclarar que es posible (y ocasionalmente recomendable) emplear información empírica para conformar esta opinión "subjetiva". En esto se basa el análisis bayesiano empírico, consistente en la determinación de la distribución *a priori* a partir de la evidencia empírica y el uso de esta misma evidencia para los cálculos bayesianos formales.

La distribución probabilística *a priori* representa el estado de incertidumbre de un individuo particular acerca de ciertos parámetros. Por tanto, si dos individuos tienen diferentes creencias acerca de un parámetro y estas creencias son representadas honestamente por diferentes distribuciones probabilísticas *a priori*, entonces ambas distribuciones son, en cierto sentido y de momento, igualmente válidas.

Un elemento particular a utilizar en el enfoque bayesiano lo constituyen los intervalos de probabilidad. Para su comprensión se debe remitir a lo que se conoce en el marco frecuentista por un intervalo de confianza como preámbulo. Este se interpreta en el enfoque frecuentista en términos de repeticiones hipotéticas del estudio. Según este enfoque clásico y para una confianza del 95 %: "si se repitiera muchas veces el muestreo, se aplica el mismo procedimiento y se calcularan respectivos intervalos de confianza al 95 % según las fórmulas conocidas, 95 de cada 100 intervalos incluirían al verdadero parámetro que está siendo estimado". Esta afirmación es coherente con la definición de probabilidad como frecuencia "a la larga" que usa esta escuela.

La interpretación de los intervalos de confianza también se presta a confusión. Cuando un investigador calcula un intervalo de confianza al 95 %, suele pensar que este intervalo obtenido a partir de una muestra particular contiene al parámetro de interés con una alta probabilidad. Esta interpretación que es muy intuitiva no deja de ser errónea. Una vez seleccionada la muestra, el intervalo de confianza es el resultado de una experiencia aleatoria, lo que puede afirmarse es que aproximadamente para el

95 % de las muestras, el intervalo asociado contiene al parámetro desconocido. Con este enfoque solo se puede *confiar* en el procedimiento seguido: si este intervalo se calcula repetidamente, entonces se sabe que aproximadamente en el 95 % de los casos se obtienen intervalos que contienen el valor desconocido; si se hace una sola vez (que es lo que ocurre siempre), entonces solo puede "confiarse" en que este intervalo sea uno de los "exitosos".

La inferencia en el paradigma bayesiano admite que se considere a la probabilidad como un grado de creencia y el enfoque para su análisis parte de la curva que representa la función de densidad que se obtiene *a posteriori*, donde si el área bajo dicha curva entre los valores X e Y es igual a 95 %, entonces se puede afirmar de que el verdadero valor esté entre X e Y con una probabilidad del 95 %; por esto (X, Y) constituye un intervalo de probabilidad o de certidumbre al 95 %, un intervalo de confianza bayesiano o intervalo probabilístico.

Para recordar

Elementos básicos conceptuales a tomar en cuenta en el enfoque bayesiano:

- El modelo o enfoque bayesiano en el análisis de cierto fenómeno se acerca más que el enfoque frecuentista o clásico al método clínico-epidemiológico empleado por el médico en su práctica médica cotidiana, al tomar en cuenta la información con que se cuenta previamente de referencia sobre cierto fenómeno motivo de estudio, así mismo, el paradigma bayesiano es capaz de resolver cuantitativamente el problema de determinar cuál de varias hipótesis es más probable sobre la base de los datos procedentes de una muestra de estudio.
- Permite comprobar la probabilidad o cuan válido es el valor de algunos parámetros estimados poblacionalmente (de los que se tienen cierta información acumulada y se cuenta con una percepción subjetiva por el investigador), al observar objetivamente el comportamiento empírico de un grupo de datos muestrales y llevar a cabo el procesamiento estadístico de toda esa información. Para esto se combinan las probabilidades *a priori* de los valores del parámetro de referencia para el investigador con la verosimilitud relacionada con cada modelo o valor del parámetro en cuestión, todo esto para estimar *a posteriori* las probabilidades estimadas de cumplirse la hipótesis nula según el tipo de contraste empleado para esos parámetros.
- Lejos de operar en un vacío total de información anterior como sucede con el método o paradigma frecuentista, el modelo de análisis bayesiano exige contemplar formalmente el conocimiento previo en el proceso que conduce a las conclusiones, de manera que al obtenerlas el investigador no se circunscribe a la experiencia u observación puntual que ha realizado, sino que incorpora la información precedente explícitamente en el análisis, por lo que se busca la integración de los resultados estadísticos con el conocimiento biológico o epidemiológico.

– La inferencia bayesiana tiene la capacidad de proveer un nivel de protección mucho mayor que el enfoque frecuentista contra la posibilidad de "ver significación en hallazgos procedentes de la investigación científica que son enteramente espurios".
– En el caso de querer hacer estimaciones de ciertos parámetros poblacionales se debe contar con los conocimientos previos o tener idea de sus posibles valores, así como utilizar como apoyo el resultado de investigaciones que estimen los valores de ciertos parámetros poblacionales y concederles subjetivamente los valores de probabilidades *a priori* que vayan a ser utilizados como punto de partida para facilitar dicho enfoque.

Los elementos necesarios a incluir en el enfoque o paradigma bayesiano tanto para la estimación de un parámetro como para la valoración de hipótesis sobre el mismo se detallan a continuación.

En el caso de la estimación de un parámetro:

– Probabilidad *a priori*: es la probabilidad atribuida (Pr) para cierto valor de un parámetro (θ_i) subjetivamente por el investigador, basado en su experiencia personal o como fruto de resultados reportados por diferentes autores, lo que es utilizado como base en el análisis por este enfoque.

La sumatoria de los valores de probabilidad *a priori* atribuidos por el investigador para cada valor del parámetro en estudio es igual a 1:

$$\sum_{i=1}^{k} Pr(\theta_i) = 1$$

Se muestra un ejemplo de cómo se plantearía la probabilidad *a priori* de cierto parámetro como una proporción [$Pr(\theta_i)$] (Tabla 5.1).

Tabla 5.1. Representación gráfica para un ejemplo de las probabilidades *a priori* de una proporción para su estimación por el paradigma bayesiano

	Valores escogidos (cinco posibles) de la proporción por el investigador				
	0,05	**0,07**	**0,09**	**0,11**	**0,13**
Probabilidades *a priori* para una proporción (atribuidas por el investigador)	0,10	0,20	0,40	0,20	0,10

La sumatoria de cada valor de probabilidad atribuido al parámetro *a priori* [$Pr(\theta_i)$] multiplicado por el respectivo valor escogido por el investigador del parámetro en estudio (θ_i) permite estimar el valor esperado $E(\theta)$ para este parámetro *a priori*:

$$E(\theta) = \sum_{i=1}^{K} \theta_i \Pr(\theta_i)$$

- Verosimilitud: es el valor numérico de probabilidad que expresa cuan correspondientes son los datos observados de la muestra de estudio suponiendo que el valor del parámetro específico fuera el empleado en su formulación, es decir, expresa cuan correspondientes son los datos observados empíricamente en la muestra de estudio para el valor del parámetro respectivo de referencia empleado: Pr(datos θ_j), por ejemplo:
  Para el caso de una proporción (θ) como parámetro, la probabilidad de obtener esos datos, es decir, que se tengan e (éxitos) en n intentos independientes, si la probabilidad de cada intento es igual a θ, esto es el valor de una densidad binomial con parámetros n y θ evaluada en e (e: 0,1,...n):

$$\Pr(e|n,\theta) = \binom{n}{e} \theta^e (1-\theta)^{n-e} \text{ , donde } \binom{n}{e} = \frac{n!}{e!(n-e)!}$$

- Probabilidad *a posteriori*: es el valor de la probabilidad estimada en relación al valor o modelo escogido del parámetro, después de haber observado los datos de la muestra; es decir, es reconstruir nuestra visión inicial de los posibles valores del parámetro para cada modelo, luego de observar la información empírica (los datos muestrales empleados). Por esta vía se establece la probabilidad de que el verdadero valor del parámetro θ, escogido al inicio, o sea θ_j, una vez analizados y procesados los datos muestrales a partir de su verosimilitud y probabilidades *a priori* y se puede expresar como:

$$\Pr(\theta_j|\text{datos}) = \frac{\Pr(\text{datos}|\theta_j)\Pr(\theta_j)}{\sum_{i=1}^{k} \Pr(\text{datos}|\theta_j)\Pr(\theta_i)} \text{ , es decir,}$$

$$\Pr(\theta_j|\text{datos}) = \frac{(\text{Verosimilitud } \theta_j \cdot \text{Probabilidad a priori } \theta_j)}{\sum_{i=1}^{k} (\text{Verosimilitud } \theta_i \cdot \text{Probabilidad a priori } \theta_i)}$$

  La sumatoria de cada valor de probabilidad atribuido al paramétro *a posteriori* según datos [Pr(θ_j) | datos] multiplicado por el respectivo valor o modelo escogido por el investigador del parámetro en estudio (θ_i), permite estimar el valor esperado E(θ_j) *a posteriori* según los datos muestrales empíricos para este parámetro mediante la siguiente expresión:

$$E\left[(\theta_j|\text{datos}\right] = \sum_{i=1}^{k} \theta_i \Pr(\theta_j)$$

- Intervalo de probabilidad: también llamado *de credibilidad*, es un intervalo de valores de probabilidad *a posteriori* que se corresponde con dos valores (inferior y superior) del parámetro estudiado respectivamente, que acumula una cierta probabilidad, por ejemplo, del 95 % que es el más empleado, de obtener valores del parámetro estudiado:

- El intervalo de máxima densidad se define como aquel intervalo: inferior (I), superior (S), entre los que tienen probabilidad $1 - \alpha$, para el que la función de densidad f cumple la siguiente condición $f(x) \geq f(y)$ cualquiera sea $x \in (I, S)$ y cualquiera sea $y \in (I, S)$. Este intervalo coincide salvo situaciones excepcionales, con el intervalo más corto de entre los que cumplen la condición de tener bajo la curva un área de magnitud $(1 - \alpha)$ 100 %, es decir, el intervalo tal que el valor S sea mínimo, entre los que cumplen $P(I \leq \theta \leq S) = 1 - \alpha$.
- La diferencia que existe entre el intervalo de confianza calculado según los métodos tradicionales o del enfoque frecuentista y el intervalo probabilístico correspondiente al método bayesiano radica en que, por ejemplo, para el caso que se utiliza un nivel de confianza del 95 %, por el método clásico se dice que existe un 95 % de confianza en que la proporción de cierto evento de la comunidad se encuentra en el intervalo que tiene como límite inferior (LI) y como límite superior (LS), mientras que, según el método bayesiano, se tiene un 95 % de probabilidad de que el intervalo de probabilidades [LI; LS] contenga los valores del parámetro estudiado. Los intervalos son similares desde el punto de vista matemático, pero la interpretación es diferente. En el caso del enfoque frecuentista el intervalo en cuestión se ha obtenido por un método que el 95 % de las veces en que se repita el procedimiento, da lugar a intervalos que contienen al verdadero parámetro, sin embargo, no se puede hablar de la probabilidad de que el intervalo contenga o no al parámetro como es el caso del bayesiano.

– Representaciones gráficas de los resultados, donde se pueden encontrar a manera de ejemplos (Figs. 5.1., 5.2., 5.3., 5.4 y 5.5).

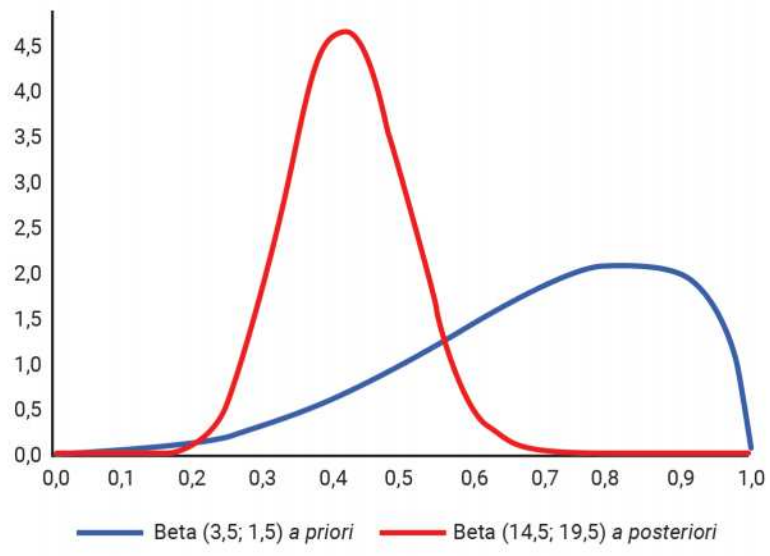


Fig. 5.1. Distribución beta *a priori* versus *a posteriori*, en la estimación de una proporción.

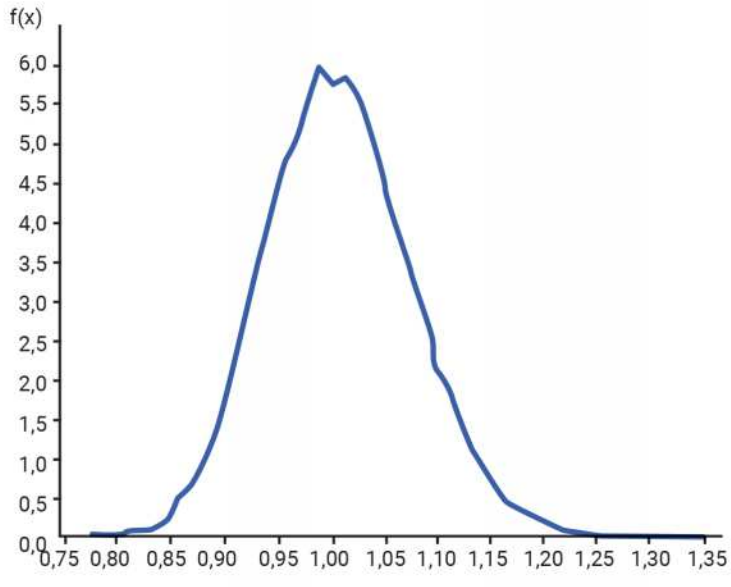


Fig. 5.2. Histograma para la estimación de la razón de proporciones (dos tasas).

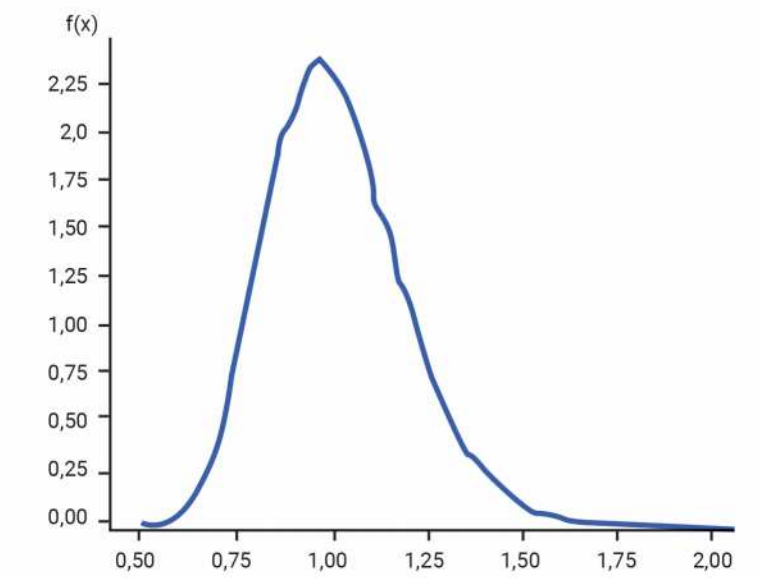


Fig. 5.3. Histograma para la estimación de una *odds ratio*.

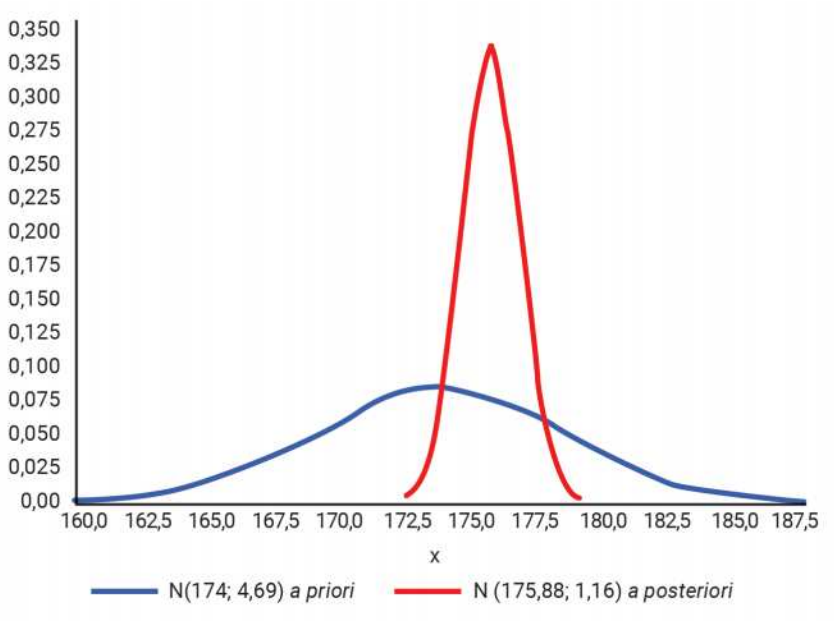


Fig. 5.4. Distribuciones normal *a priori* y *a posteriori* en la estimación de una media.

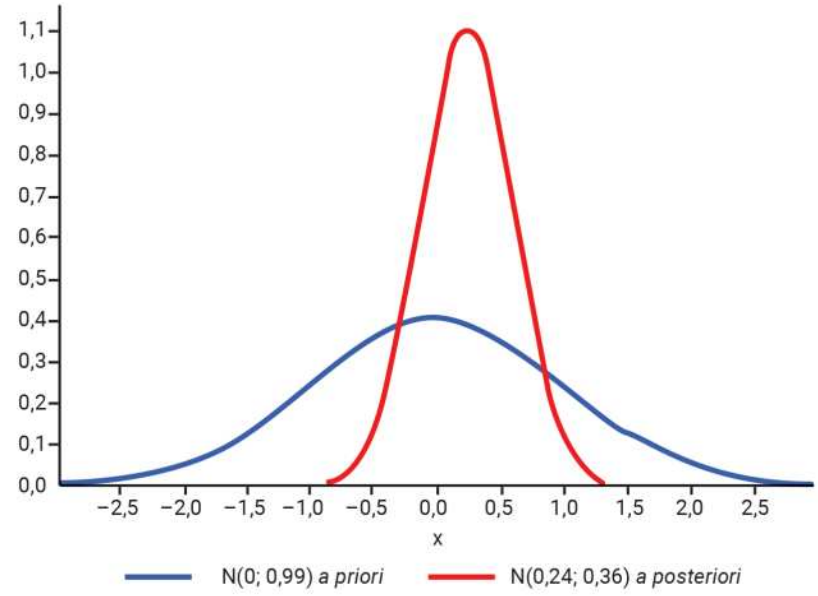


Fig. 5.5. Distribuciones normal *a priori* y *a posteriori* en la diferencia de dos medias.

En el caso de la valoración de hipótesis sobre un parámetro:

- Planteamiento de la hipótesis de trabajo (H) *versus* hipótesis alternativa (K), en torno al valor del parámetro.
- Probabilidad *a priori* de la validez de la c.
- Recolección y presentación de los datos observados (empíricos) para los parámetros muestrales.
- Estimación del factor de Bayes a favor de la veracidad de la hipótesis de trabajo [H(BF)] y en contra de la veracidad de la hipótesis de trabajo [H(BC)].
- Estimación de la probabilidad *a posteriori* (PP) de la veracidad de la hipótesis de trabajo (H).
- Representación gráfica de los resultados donde se pueden encontrar varios ejemplos (Figs. 5.6, 5.7 y 5.8).

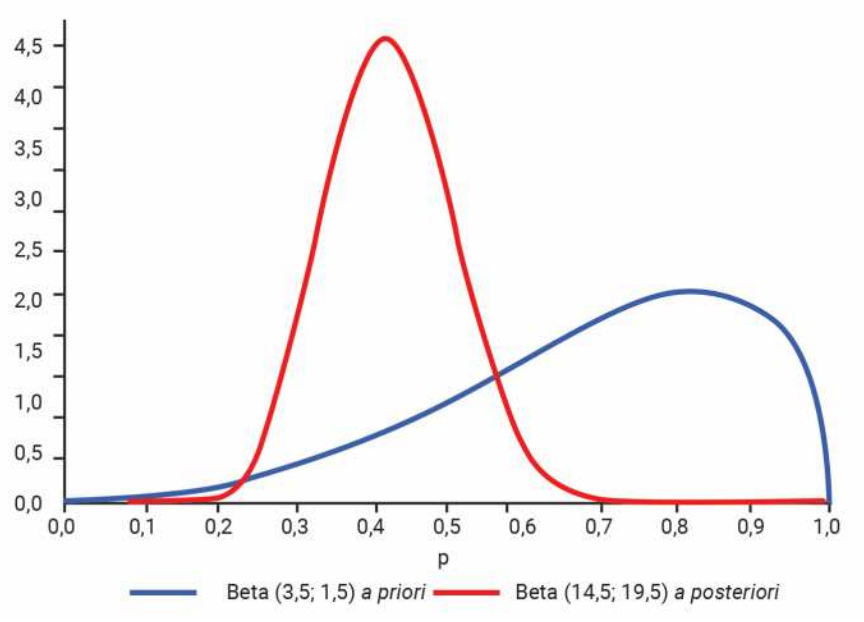


Fig. 5.6. Distribución beta *a priori* versus *a posteriori* del parámetro (una proporción).

Algunas de las aplicaciones del método bayesiano que pueden tener expresión en el procesamiento estadístico de la información en la vertiente inferencial se muestran a continuación, siguiendo la metodología expuesta en la ayuda del paquete estadístico epidemiológico EPIDAT versión 4.2 (Colectivo de autores, 2016).

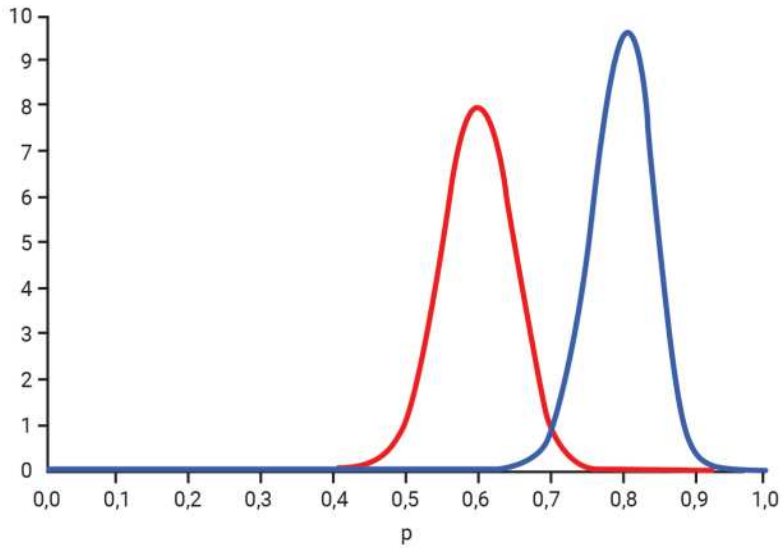


Fig. 5.7. Distribución beta *a priori* (diferencias de dos proporciones).

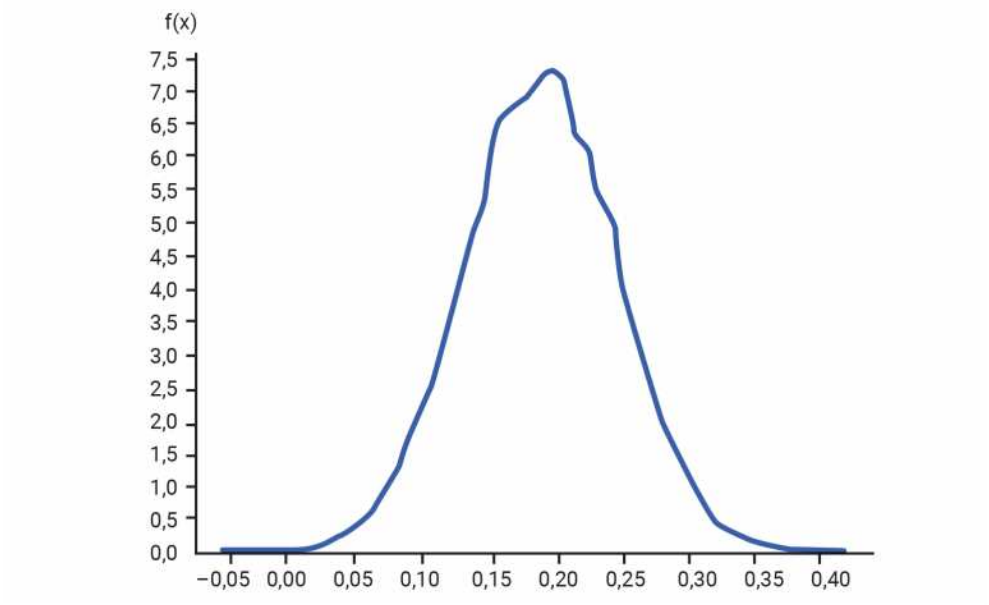


Fig. 5.8. Histograma de las diferencias (diferencia entre dos proporciones).

Aplicación del paradigma bayesiano en el análisis de un método diagnóstico

Conceptualmente hablando la sensibilidad (Sens) y la especificidad (Espec) de las pruebas diagnósticas son los dos indicadores clásicos para evaluar su capacidad demarcatoria, es decir, representan el grado de eficiencia inherente a una prueba diagnóstica. Si el enfermo (E) y sano (S) como posibles condiciones del estado patológico o no respectivamente del paciente; T+ como resultado positivo (patológico de la prueba) y T– como resultado negativo (no patológico de la prueba), entonces:

- La probabilidad condicional Sen = Pr(T+| E) es la capacidad de la prueba diagnóstica de detectar un sujeto enfermo, cuan sensible es la prueba para detectarlo. En tablas de contingencia 2x2 se calcula como:

  Sens = E con T+ /[(E con T+) + (E con T–)]
- La probabilidad condicional Espec = = Pr(T– | S) es la capacidad de la prueba de diagnosticar como sanos a los sujetos que verdaderamente lo son. En tablas de contingencia 2x2 se calcula como:

  Espec = S con T–/[(S con T+) + (S con T–)]
- La probabilidad condicional Pr(E | T+) denota el valor predictivo positivo (VPP), o sea, que si la prueba es positiva, cuan probable es que esté efectivamente enfermo el sujeto. En tablas de contingencia 2x2 se calcula como:

  VPP (de la prueba) = E con T+/[(E con T+) + (S con T+)]
- La probabilidad condiciona Pr(S | T–), denota el valor predictivo negativo (VPN), o sea, que si la prueba es negativa, cuan probable es que esté efectivamente sano el sujeto. En tablas de contingencia 2x2 se calcula como:

  VPN (de la prueba) = S con T–/ [(S con T–) + (E con T–)]

Al aplicar el teorema de Bayes asumiendo a p como la probabilidad *a priori* de que el sujeto esté enfermo y q = 1 – p como su complemento (que se tiene de referencia en la población igual al componente subjetivo), así como a partir de los datos de la muestra en relación al resultado de la prueba diagnóstica *versus* el estado del paciente (por tabla cruzada o contingencia 2x2) que permite determinar sensibilidad, especificidad igual al componente objetivo de los datos (verosimilitud) se determina la probabilidad *a posteriori* o actualizada como sigue:

- Para la sensibilidad (probabilidad *a posteriori* de estar enfermo teniendo la prueba diagnóstica positiva):

$$Pr(E|T+) = \frac{p \cdot Sens}{p \cdot Sens + q \cdot (1 - Espec)}$$

– Para la especificidad (probabilidad *a posteriori* de estar sano con la prueba diagnóstica negativa):

$$Pr(S|T-) = \frac{p \cdot Espec}{p \cdot Espec + q \cdot (1 - Sens)}$$

– Para los falsos positivos (probabilidad *a posteriori* de estar sano teniendo la prueba diagnóstica positiva):

$$Pr(S|T+) = Pr(E|T+)$$

– Para los falsos negativos (probabilidad *a posteriori* de estar enfermo teniendo la prueba diagnóstica negativa):

$$Pr(E|T-) = 1 - Pr(S|T-)$$

Aplicación bayesiana en la inferencia bioestadística en una población con el empleo del odds ratio

Se emplean los datos de:

– El valor puntual y por intervalos de probabilidad (al 95 %) de OR referidos en estudios previos a emplear por el investigador (OR *a priori*) (componente subjetivo de las probabilidades).

– El valor puntual y por intervalos de confianza de OR (al 95 %) (grado de error por la estimación realizada) por el método frecuentista habitual de los datos de la muestra estudiada (OR empírico) (componente objetivo de las probabilidades para la verosimilitud) siguiendo las fórmulas:

- Valor puntual OR:

$$OR = \frac{a \cdot d}{b \cdot c}$$

- Por intervalos de confianza (95 %):

$$\text{IC Inferior } OR = epx\left\{\ln(OR) - Z_{1-\alpha/2}\sqrt{Var\left[Ln(OR)\right]}\right\}$$

$$\text{IC Superior } OR = epx\left\{\ln(OR) + Z_{1-\alpha/2}\sqrt{Var\left[Ln(OR)\right]}\right\}$$

– Se estima el resultado puntual y por intervalos de probabilidad (al 95 %) del OR *a posteriori*, partiendo de los valores de OR (*a priori* y empíricos) y se determina su probabilidad de ser mayor a valores que representan puntos de corte comparativos (X) del OR, ejemplo, Pr (OR > X); X = 0,5; 0,8;...

- OR *a posteriori* (puntual):

$OR_{pt} = \exp(\mu_{pt})$

- OR *a posteriori* por intervalos de confianza (IC) al 95 %:

IC Inferior= $\exp(\mu_{pt} - 1{,}96\sigma_{pt})$

IC Superior= $\exp(\mu_{pt} + 1{,}96\sigma_{pt})$

– Probabilidad de que OR sea mayor que x:

$$\Pr(OR > x) = 1 - \Phi\left(\frac{\ln(x) - \mu_{pt})}{\sigma_{pt}}\right)$$

donde:

$$\mu_{pt} = \frac{\frac{\mu_{pr}}{\sigma^2_{pr}} + \frac{\mu_{dt}}{\sigma^2_{dt}}}{\frac{1}{\sigma^2_{pr}} + \frac{1}{\sigma^2_{dt}}} \; ;$$

$$\mu_{pr} = \ln(OR_{pr}) \; ;$$

$$\sigma_{pr} = \frac{\ln(Sup_{pr}) - \ln(Inf_{pr})}{3{,}92} \; ;$$

OR_{pr}: *odds ratio a priori* y ($Sup_{pr,}$ Inf_{pr}) su intervalo de probabilidad al 95 %;

$$\mu_{dt} = \ln(OR_{dt}) \; ;$$

$$\sigma_{dt} = \frac{\ln(Sup_{dt}) - \ln(Inf_{dt})}{3{,}92} \; ;$$

OR_{dt}: *odds ratio* empírico procedente de los datos y (Sup_{dt}, Inf_{dt}) su intervalo de confianza al 95 %.

Aplicación bayesiana en la inferencia bioestadística en una población, para la estimación de una proporción

En este caso se tratan las variantes: estimación de una proporción, valoración de hipótesis sobre el empleo de una proporción con especificación de esta para igualdad de la proporción con una constante y ubicación de una proporción dentro de un intervalo de referencia.

Estimación de una proporción

Conceptualmente hablando se asume que en la estimación de una proporción y en la valoración de hipótesis sobre ella el número de éxitos e en una muestra de tamaño n sigue una distribución binomial de parámetros n y p, donde p es la probabilidad de éxito: e ~Bin (n, p):

– La proporción p sigue, *a priori*, una distribución beta de parámetros a y b: p ~ beta (a, b).
– Se cumple que la distribución *a posteriori* de p, dado e, es una beta de parámetros a + e y b + f, donde: f = n − e, es el número de fracasos:
p| e ~ beta(a + e, b + f)

– La media y desviación estándar de la distribución beta se expresan como sigue:

$$\mu = \frac{a}{a+b}$$

$$\sigma = \sqrt{\frac{ab}{(a+b)^2(a+b+1)}}$$

– Los parámetros en función de la media, la desviación estándar y la función de densidad se expresan como:

$$a = \frac{(1-\mu)\mu^2 - \sigma^2\mu}{\sigma^2}$$

$$b = \frac{a(1-\mu)}{\mu}$$

$$f(p) = \frac{p^{a-1}(1-p)^{b-1}}{B(a,\ b)}\ ,\ 0 < p < 1$$

donde:

B (a, b) es la función beta:

$$B(a,b) = \int_0^1 u^{a-1}(1-u)^{b-1}du = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Γ(x): es la función gamma:

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$$

Pasos a seguir

– Se fijan las probabilidades *a priori* ($Pr\theta_1$) que corresponden a la proporción en cuestión de interés para el investigador de la cual se tiene referencias por estudios previos.

– Se conforma una distribución *a priori* para posibles valores de la proporción (θ) o de la prevalencia de una enfermedad, donde se pueda sospechar un intervalo de los mismos para la θ que van de 0 a 1.

– Se acepta que hay k posibles valores de θ_1, θ_2,..., θ_k a los que se atribuyen respectivas probabilidades que reflejan formal y cuantitativamente los puntos de vista del investigador o recogidos de referencia en estudios previos (probabilidades *a priori* igual a $Pr(\theta_1)$, $Pr(\theta_2)$,..., $Pr(\theta_k)$, donde:

$$\sum_{i=1}^{k} Pr(\theta_i) = 1$$

– Se calcula la verosimilitud del modelo para cada uno de los valores de proporciones referidos o tomados *a priori*, es decir, la probabilidad de que se hayan producido los datos observados suponiendo que es válido ese modelo, siguiendo la expresión:

$$Pr(e|n,\theta) = \binom{n}{e}\theta^e(1-\theta)^{n-e}$$

donde:

$\binom{n}{e} = \frac{n!}{e!(n-e)!}$ y siendo:

e: número de éxitos en la muestra;

n: el número de sujetos observados de la muestra;

$\theta = \frac{e}{n}$: valor puntual de la proporción para la muestra estudiada.

- Se multiplica la probabilidad *a priori*: Pr(θ), para cada valor de proporción empleado de referencia por la verosimilitud correspondiente: Pr(e|n, θ), se suman todos y posteriormente se divide cada resultado entre el valor de su suma obteniéndose así la probabilidad *a posteriori* (posterior) para cada valor correspondiente de probabilidad de proporción *a priori* según los datos muestrales empleados.
- Se estiman los intervalos de probabilidad o credibilidad para el parámetro (proporción poblacional) estudiado en cada modelo partiendo de la probabilidad *a priori* y la verosimilitud según el comportamiento de los datos empleados, donde el más empleado es el intervalo de máxima densidad.

Son varias las expresiones matemáticas o fórmulas a emplear en el caso de estimación de una proporción por el paradigma bayesiano:

- Área a la izquierda de x:
  $F(x) = Pr(p \leq x)$
- Percentil de orden k:
  $x = F^{-1}(k)$
- Intervalo de máxima densidad con probabiliadad $1 - \alpha$ para la proporción p = intervalos (I, S) de menor amplitud que cumple la condición:
  $F(S) - F(I) = 1 - \alpha$.
  donde:
  F: función de distribución beta *a posteriori* (a + e, b + f);
  a y b: parámetros de la distribución beta *a priori*;
  e: número de éxitos;
  f = n – e: número de fracasos;
  n: tamaño de la muestra;
  x: valor entre 0 y 1;
  k: recorre el conjunto de valores {0,025; 0,05; 0,1; 0,25; 0,5; 0,75; 0,9; 0,95; 0,975}.

Valoración de hipótesis sobre una proporción

En esta condición se puede realizar como parte de la valoración de hipótesis las variantes: probar la igualdad de una proporción con una constante de su ubicación dentro de un intervalo de referencia, lo cual se pasa a describir.

Probar que una proporción es igual a una constante

Planteamiento de hipótesis:

H: $P = P_0$ *vs.* K: $P \neq P_0$

Nota: Se debe establecer una probabilidad *a priori* (q) del grado de certeza o validez para hipótesis H a la hora de introducir los datos en procesadores estadísticos como, por ejemplo, Epidat 4.2.

Se debe contar con:

- Proporción de referencia que se quiere valorar (P_0).
- Número de éxitos (e) y fracasos (f) en n = e + f experiencias a partir de los datos de la muestra a emplear para estimar P.
- Emplear parámetros a y b de la distribución beta.
- Fórmula del factor de Bayes a favor de H (BF):

$$BF = \frac{p_0^e(1-p_0)^f B(a,b)}{B(a+e,b+f)}$$

- Fórmula del factor de Bayes en contra de H (BC):

$$BC = \frac{1}{BF}$$

- Fórmula de la probabilidad *a posteriori* de la veracidad de H:

$$P(H|\text{datos}) = \frac{qBF}{qBF+(1-q)}$$

donde:

a y b: parámetros de la distribución beta *a priori*;

e: número de éxitos;

f = n − e: número de fracasos;

n: tamaño de la muestra;

B(r, s) : función beta evaluada en r > 0 y s > 0;

q: probabilidad *a priori* de la validez de H.

Interpretación del resultado:

- La comparación del factor de Bayes (BF) *versus* factor de Bayes en contra de H (CB) respecto a la hipótesis H, permite determinar cuan más o menos veces probables es que sea cierta la hipótesis H respecto a la hipótesis complementaria (F) según sea mayor o menor respectivamente el factor de Bayes (BF) *versus* factor de Bayes en contra de H.
- La probabilidad *a posteriori* expresa la veracidad o cuan probable es que sea cierta la hipótesis H, es decir, del valor de la probabilidad del parámetro o estadígrafo de referencia con que se cuenta respecto a los datos empleados en la muestra de estudio al combinarlos para su análisis bayesiano.

Probar que una proporción se encuentra dentro de un intervalo de referencia

Planteamiento de hipótesis:

H: $P \in [P_1, P_2]$ *vs.* K: $P < P_1$ o $P > P_2$

Nota: Se debe establecer una probabilidad *a priori* (q) del grado de certeza o validez para hipótesis H al introducir los datos en procesadores estadísticos como, por ejemplo, Epidat 4.2.

Se debe contar con:

- Valores de referencia P_1 y P_2 (proporciones que constituyen los puntos extremos del intervalo de P_0).
- Número de éxitos (e) y fracasos (f) en n = e + f experiencias a partir de los datos de la muestra a emplear para estimar P.
- Emplear parámetros a y b de la distribución beta.
- Probabilidad de la hipótesis H:
  $Pr(H) = F(P_2) - F(P_1)$
- Factor de Bayes a favor de H (BF):
  $$BF = \frac{Pr(H)}{Pr(K)} \text{ o } BF = \frac{P(H)}{1-P(H)}$$
- Factor de Bayes en contra de H (BC):
  $$BC = \frac{1}{BF}$$
- Probabilidad *a posteriori* de la veracidad de H (PP):
  $$PP = \frac{qBF}{qBF + 1 - q}$$, en caso que q = 0,5 (probabilidad *a priori* para H = 0,5) la probabilidad *a posteriori* quedaría en función del factor de Bayes solamente y se expresa como:
  $$PP = \frac{BF}{BF + 1}$$
  donde:
  P(H) = F (p_2) − F (p_1): probabilidad de la hipótesis H;
  F: función de distribución *a posteriori*, beta (a + e, b + f);
  e: número de éxitos;
  f = n − e: número de fracasos;
  n: tamaño de la muestra;
  q: probabilidad *a priori* de la validez de H.

Interpretación del resultado:

- La comparación del factor de Bayes (BF) *versus* factor de Bayes en contra de H respecto a la hipótesis H, permite determinar cuan más o menos veces probables es que sea cierta la hipótesis H respecto a la hipótesis complementaria (F) según sea mayor o menor respectivamente el factor de Bayes (BF) *versus* factor de Bayes en contra de H.

– La probabilidad *a posteriori* expresa la veracidad o cuan probable es que sea cierta la hipótesis H, es decir, del valor de la probabilidad para el parámetro o estadígrafo de referencia con que se cuenta respecto a los datos empleados en la muestra de estudio al combinarlos en un análisis bayesiano.

Aplicación bayesiana en la inferencia bioestadística en dos poblaciones

En este caso se tratan las variantes: estimación de efectos basados en dos proporciones y la valoración de hipótesis sobre una diferencia de dos proporciones. En ambas situaciones se asume que:

– Se tienen dos poblaciones 1 y 2.
– El número de éxitos e_i en una muestra de tamaño ni de la población i sigue una distribución binomial de parámetros n_i y p_i, donde p_i es la probabilidad de éxito: $e_i \sim \text{Bin}(n_i, p_i)$, i = 1, 2.
– La proporción p_i sigue, *a priori*, una distribución beta de parámetros a_i y b_i:
$p_i \sim \text{beta}(a_i, b_i)$, i = 1,2.
Entonces se cumple que la distribución a posteriori de p_i, dado e_i, es una beta de parámetros $a_i + e_i$ y $b_i + f_i$, donde $f_i = n_i - e_i$ es el número de fracasos:
$P_i \mid e_i \sim \text{beta}(a_i + e_i, b_i + f_i)$, i = 1, 2.

Estimación de efectos basados en dos proporciones

La distribución *a posteriori* del efecto se estima empíricamente a partir de las distribuciones *a posteriori* de las dos proporciones:

– $p_{11}, p_{12}, \ldots, p_{1n}$ son n realizaciones de la primera distribución beta *a posteriori*: Beta($a_1 + e_1, b_1 + f_1$).
– $p_{21}, p_{22}, \ldots, p_{2n}$ son n realizaciones de la segunda distribución beta *a posteriori*: Beta($a_2 + e_2, b_2 + f_2$).
– θ_j: efecto estimado con el par de valores p_{1j}, p_{2j}, j = 1,..., n;
$\theta_j = p_{1j} - p_{2j}$: par una diferencia de proporciones;

$\theta_j = \frac{p_{1j}}{p_{2j}}$: para una razón de proporciones;

$\theta_j = \frac{p_{1j}(1-p_{2j})}{p_{2j}(1-p_{1j})}$: para una razón de odds.

A partir de la muestra de n valores ordenados $\theta_1, \theta_2, \ldots, \theta_n$, se tiene la estimación no paramétrica de la función de distribución (áreas a la izquierda de un valor x) y de los percentiles:

– Área a la izquierda de x:

$$Pr(\theta \leq x) = \frac{1}{n}\sum_{j=1}^{n} I(\theta_j \leq x)$$

– Percentil de orden k:

$P_k = (1 - f)\theta_r + f\theta_{r+1}$

donde:

a_i y b_i: parámetros de la distribución beta *a priori*, i = 1, 2;

e_i: número de éxitos;

$f_i = n_i - e_i$: número de fracasos, i = 1, 2;

n_i: tamaño de la muestra i = 1, 2;

n: número de simulaciones;

I(c): función indicadora;

I (c) = {valor 1, si la condición es cierta; valor 0, en otro caso};

k: recorre el conjunto de valores {0,025; 0,05; 0,1; 0,25; 0,5; 0,75; 0,9; 0,95; 0,975};

r = [R]: parte entera de R = (n + 1)k, $0 \leq r \leq n$;

f = R – r: parte freaccionaria de R;

$\theta_0 = \theta_1$;

θn + 1 = θn.

Valoración de hipótesis sobre una diferencia de proporciones

En esta condición se puede realizar como parte de la valoración de hipótesis las variantes: probar la igualdad de dos proporciones y que la diferencia de proporciones está dentro de un intervalo, lo cual se pasa a describir.

Probar una igualdad de dos proporciones

Se comienza realizando el planteamiento de hipótesis:

H: $p_1 = p_2$ *vs.* K: $p_1 \neq p_2$

Nota: Se debe establecer una probabilidad *a priori* (q) del grado de certeza o validez para hipótesis H a la hora de introducir los datos en procesadores estadísticos como, por ejemplo, Epidat 4.2.

Se debe contar con:

– Número de éxitos (e) y fracasos (f) en n = e + f experiencias a partir de los datos de las muestras a emplear para estimar P_1 y P_2.
– Emplear parámetros a y b de la distribución beta.
– Factor de Bayes a favor de H (BF):

$$BF = \frac{B(a_1 + e_1 + e_2, b_1 + f_1 + f_2)B(a_2, b_2)}{B(a_1 + e_1 + e_2, b_1 + f_1)B(a_2 + e_2, b_2 + f_2)}$$

– Factor de Bayes en contra de H (BC):

$$BC = \frac{1}{BF}$$

– Probabilidad *a posteriori* para la veracidad de H:

$$P(H|datos) = \frac{qBF}{qBF + (1-q)}$$

donde:

a_i y b_i: parámetros de la distribución beta *a priori*, i = 1, 2;

e_i: número de éxitos;

$f_i = n_i - e_i$: número de fracasos, i = 1, 2;

n_i: tamaño de la muestra, i = 1, 2;

B (r, s): función beta evaluada en r > 0 y s > 0;

q: es la probabilidad *a priori* de la validez de H.

Interpretación del resultado:

– La comparación del factor de Bayes (BF) *versus* factor de Bayes en contra de H (BC) respecto a la hipótesis H permite determinar cuan más o menos veces probables es que sea cierta la hipótesis H respecto a la hipótesis complementaria (F) según sea mayor o menor respectivamente factor de Bayes (BF) *versus* factor de Bayes en contra de H (BC).
– La probabilidad *a posteriori* expresa la veracidad o cuan probable es que sea cierta la hipótesis H, es decir, del valor de la probabilidad para el parámetro o estadígrafo de referencia con que se cuenta respecto a los datos empleados en las muestras de estudio al combinarlos en un análisis bayesiano.

Probar que la diferencia de proporciones está dentro de un intervalo

Se valora el planteamiento de hipótesis:

H: $p_1 - p_2 \in [p_3, p_4]$ *vs.* K: $p_1 - p_2 < p_3$ o $p_1 - p_2 > p_4$

Nota: Se debe establecer una probabilidad *a priori* (q) del grado de certeza o validez para hipótesis H a la hora de introducir los datos en procesadores estadísticos como, por ejemplo, Epidat 4.2.

Se debe contar con:

– Número de éxitos (e) y fracasos (f) en n = e + f experiencias a partir de los datos de las muestras a emplear para estimar P_1 y P_2.
– Emplear parámetros a y b de la distribución beta.
– Factor de Bayes a favor de H (BF):

$$BF = \frac{P(H)}{1-P(H)}$$

– Factor de Bayes en contra de H (BC):

$$BC = \frac{1}{BF}$$

– Probabilidad *a posteriori* para la veracidad de H:

$$P(H \mid datos) = \frac{qBF}{qBF + (1-q)}$$

donde:

P (H) = F (p_4) – F (p_3): probabilidad de la hipótesis H;

F: estimación no paramétrica de la función de distribución *a posteriori* de la diferencia de proporciones:

$$F(x) = Pr(d \leq x) = \frac{1}{n}\sum_{j=1}^{n} I(d_j \leq x)$$

n: número de simulaciones;

$d_j = p_{1j} - p_{2j}$, j = 1, 2,..., n;

p_{11}, p_{12},..., p_{1n}: n realizaciones de la primera distribución beta *a posteriori*, Beta(a_1 + e1, b_1 + f_1);

p_{21}, p_{22},..., p_{2n}: n realizaciones de la segunda distribución beta *a posteriori*, Beta(a_2 + e_2, b_2 + f_2);

a_i y b_i: parámetros de la distribución beta *a priori*, i = 1, 2;

e_i: número de éxitos;

$f_i = n_i - e_i$: número de fracasos, i = 1, 2;

n_i: tamaño de la muestra, i = 1, 2;

q: probabilidad *a priori* de la validez de H.

Interpretación del resultado:

– La comparación del factor de Bayes (BF) *versus* factor de Bayes en contra de H (BC) respecto a la hipótesis H permite determinar cuan más o menos veces probables es que sea cierta la hipótesis H respecto a la hipótesis complementaria (F) según sea mayor o menor respectivamente el factor de Bayes (BF) *versus* factor de Bayes en contra de H (BC).
– La probabilidad *a posteriori* expresa la veracidad o cuan probable es que sea cierta la hipótesis H, es decir, del valor de la probabilidad para el parámetro o estadígrafo de referencia con que se cuenta respecto a los datos empleados en las muestras de estudio al combinarlos en un análisis bayesiano.

Aplicación bayesiana en la inferencia bioestadística en una población, empleo de una media

En este caso se tratan las variantes: estimación de una media, valoración de hipótesis sobre el empleo de una media con especificación de esta para igualdad de la media

empleada con una media constante de referencia y la ubicación de la media empleada dentro de un intervalo de referencia.

Estimación de una media

Siendo la distribución *a posteriori* de la media normal de media (m_p) y desviación estándar (s_p):

- Si la distribución *a priori* de la media es la uniforme, entonces m_p y s_p se estiman como:

$$m_p = \bar{x} \ , \ s_p = \frac{s}{\sqrt{n}}\left(1+\frac{20}{n^2}\right)$$

- Si la distribución *a priori* de la media es la normal, entonces m_p y s_p se estiman empleando media (m_0) y desviación estándar (s_0) como:

$$m_p = \frac{c_0 m_0 + c\bar{x}}{c_1} \ , \ s_p = \frac{s}{\sqrt{c_1}}$$

donde:

$$c_0 = \frac{1}{s_0^2} \ , \ c = \frac{n}{s^2\left(1+\frac{20}{n^2}\right)^2} \ , \ c_1 = c_0 + c$$

con: $s_0 = \dfrac{x_1 - x_2}{\phi^{-1}(p_1) - \phi^{-1}(p_2)}$ y $m_0 = x_1 - s_0\varphi^{-1}(p_1)$, si se calculan a partir de dos valores de la distribución normal *a priori* x_1 y x_2 con probabilidades asociadas, respectivamente, p_1 y p_2.

Se realiza la estimación paramétrica como parte del análisis a:

- Área a la izquierda de x (estimación paramétrica de probabilidades):

$$F(x) = \int_{-\infty}^{x} \frac{1}{s_p\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x - m_p}{s_p}\right)^2\right\}$$

- Percentil de orden k (estimación paramétrica de percentiles de la distribución *a posteriori*):

$x = F^{-1}(k)$

- Intervalo de probabilidad ($x_{\alpha/2}$; $x_{1-\alpha/2}$) o de máxima densidad con probabilidad $1 - \alpha$: $[F^{-1}(\alpha/2), F^{-1}(1 - \alpha/2)]$

donde:

$\bar{x}$ y S: media y desviación estándar, respectivamente, de la muestra;

n: tamaño de la muestra;

ϕ: función de distribución normal estándar;

F: función de distribución normal *a posteriori*, N (m_p, S_P);

k: recorre el conjunto de valores [0,025; 0,05; 0,1; 0,25; 0,5; 0,75; 0,9; 0,95; 0,975]

Valoración de hipótesis sobre el empleo de una media

En esta condición se puede realizar como parte de la valoración de hipótesis las variantes: probar que la media empleada es igual a una media constante de referencia y probar que la media empleada se encuentra dentro de un determinado intervalo de referencia.

Probar que la media empleada es igual a una media constante de referencia

Planteamiento de hipótesis:

H: $m = m_0$ *vs.* K: $m \neq m_0$

Nota: Se debe establecer una probabilidad *a priori* (q) del grado de certeza o validez para hipótesis H al introducir los datos en procesadores estadísticos como, por ejemplo, Epidat 4.2.

Se debe contar con:

- Valor de la media de referencia o como constante a comparar (m_0).
- Desviación estándar (t) para la distribución de la media que rige hipótesis H.
- Desviación estándar (h) que rige cuando se supone que hipótesis H no es cierta.
- Datos muestrales (empíricos) observados: la media y el tamaño de la muestra.
- Factor de Bayes a favor de Hipótesis H (BF):

$$BF = \frac{\frac{1}{s_1}\exp\left[-\frac{1}{2s_1^2}(\bar{x}-m_0)^2\right]}{\frac{1}{s_1}\exp\left[-\frac{1}{2s_2^2}(\bar{x}-m_0)^2\right]}$$

- Factor de Bayes en contra de Hipótesis H (BC):

$$BC = \frac{1}{BF}$$

- Probabilidad *a posteriori* de la veracidad de hipótesis H:

$$P(H \mid datos) = \frac{qBF}{qBF + (1-q)}$$

donde:

$$s_1 = \frac{h}{\sqrt{n}}$$

$$s_2 = \sqrt{\frac{h^2}{n} + t^2}$$

$\bar{x}$: a media de la muestra;

n: tamaño de la muestra;

t: desviación estándar de la distribución *a priori* bajo la hipótesis H;

h: desviación estándar de la distribución *a priori* bajo la hipótesis (K);

q: probabilidad *a priori* de la validez de la hipótesis H.

Interpretación del resultado:

- La comparación del factor de Bayes (BF) *versus* factor de Bayes en contra de H (BC), permite determinar cuan más o menos veces probables es que sea cierta la hipótesis H respecto a la hipótesis complementaria (F) según sea mayor o menor respectivamente el factor de Bayes *versus* factor de Bayes en contra de H.
- La probabilidad *a posteriori* expresa la veracidad o cuan probable es que sea cierta la hipótesis H, es decir, del valor de la probabilidad para el parámetro o estadígrafo de referencia con que se cuenta respecto a los datos empleados en las muestras de estudio al combinarlos en un análisis bayesiano.

Probar que la media empleada se encuentra dentro de un determinado intervalo de referencia

Planteamiento de hipótesis:

H: $M \in [m_1, m_2]$ *vs.* K: $M < m_1$ o $M > m_2$

Nota: Se debe establecer una probabilidad *a priori* (q) del grado de certeza o validez para hipótesis H a la hora de introducir los datos en procesadores estadísticos como, por ejemplo, Epidat 4.2

Se debe contar con:

- Empleo de la distribución *a priori* uniforme.
- Empleo de la distribución *a posteriori* normal con media (m_p) y desviación estándar (s_p).
- -Valores de m_1 y m_2 que son los valores de referencia para el intervalo a probar de la media.
- Datos muestrales (empíricos) observados: la media, desviación estándar y el tamaño de la muestra.
- Factor de Bayes a favor de hipótesis H (BF):

$$BF = \frac{P(H)}{1 - P(H)}$$

- Factor de Bayes en contra de Hipótesis H (BC):

$$BC = \frac{1}{BF}$$

- Probabilidad *a posteriori* de la veracidad de hipótesis H:

$$P(H \mid datos) = \frac{qBF}{qBF + (1 - q)}$$

donde:

P (H) = F (m_2) – F (m_1): probabilidad de la hipótesis H;

F: función de distribución normal *a posteriori*, que tiene media m_p y desviación estándar S_p donde:

$m_p = \overline{x}$, $s_p = \frac{s}{\sqrt{n}}\left(1+\frac{20}{n^2}\right)$

$\overline{x}$: media de la muestra;

n: tamaño de la muestra;

S: desviación estándar de la muestra;

q: probabilidad *a priori* de la validez de la hipótesis H.

Interpretación del resultado:

- La comparación del factor de Bayes (BF) *versus* factor de Bayes en contra de H (BC) respecto a la hipótesis H, permite determinar cuan más o menos veces probables es que sea cierta la hipótesis H respecto a la hipótesis complementaria (F) según sea mayor o menor respectivamente el factor de Bayes *versus* factor de Bayes en contra de H.
- La probabilidad *a posteriori* expresa la veracidad o cuan probable es que sea cierta la hipótesis H, es decir, del valor de la probabilidad para el parámetro o estadígrafo de referencia con que se cuenta respecto a los datos empleados en las muestras de estudio al combinarlos en un análisis bayesiano.

Aplicación bayesiana en la inferencia bioestadística en dos poblaciones para la estimación de una diferencia de medias

En este caso se tratan las variantes: estimación de una diferencia de dos medias y valoración de hipótesis sobre la diferencia de dos medias. lo cual se pasa a describir:

Estimación de la diferencia de dos medias

Siendo la distribución *a posteriori* de la diferencia de medias normal de media (m_p) y desviación estándar (s_p):

$m_p = m_{p1} - m_{p2}$

$s_p = \sqrt{s_{p1}^2 + s_{p2}^2}$

- Si la distribución *a priori* de las dos medias es la uniforme:

$m_{pi} = \overline{x}_i$

$s_{pi} = \frac{s_i}{\sqrt{n_i}}\left(1+\frac{20}{n_i^2}\right)$, i = 1, 2.

- Si la distribución *a priori* de la media i-ésima (i = 1, 2) es mormal de media (m_{0i}) y desviación estándar (s_{0i}):

$$m_{pi} = \frac{c_{0i}m_{0i} + c_i\overline{x}_i}{c_{1i}}$$

$$s_{pi} = \frac{s_i}{\sqrt{c_{1i}}}$$

donde:

$$c_{0i} = \frac{1}{s_{0i}^2}$$

$$c_i = \frac{n_i}{s_i^2\left(1+\frac{20}{n_i^2}\right)^2}$$

$c_{1i} = c_{0i} + c_i$

con: $s_{0i} = \frac{x_{1i} - x_{2i}}{\phi^{-1}(p_{1i}) - \phi^{-1}(p_{2i})}$ y $m_{0i} = x_{1i} - s_{0i}\varphi^{-1}(p_{1i})$, si se calculan a partir de dos valores de la distribución normal *a priori* x_{1i} y x_{2i}, con probabilidades asociadas respectivamente, p_{1i} y p_{2i}.

Otros datos necesarios son los datos de ambas muestras que estiman los de dos poblaciones: $\overline{x}_1$, s_1, n_1 y $\overline{x}_2$, s_2, n_2.

Se realiza la estimación paramétrica como parte del análisis de:

- Área a la izquierda de x (estimación paramétrica de probabilidades):

$$F(x) = \int_{-\infty}^{x} \frac{1}{s_p\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x - m_p}{s_p}\right)^2\right\}$$

- Percentil de orden k (estimación paramétrica de percentiles de la distribución *a posteriori*):
  $X = F^{-1}(k)$
- Intervalo de máxima densidad ($x_{\alpha/2}$; $x_{1-\alpha/2}$) con probabilidad $1 - \alpha$:
  $[F^{-1}(\alpha/2), F^{-1}(1 - \alpha/2)]$
  donde:

  $\overline{x}_i$ y s_i: media y desviación estándar, respectivamente, de la muestra i, i = 1, 2;
  n: tamaño de la muestra i, i = 1, 2;
  ϕ: función de distribución normal estándar.
  F: función de distribución normal *a posteriori*, N (m_p, S_P);
  k: recorre el conjunto de valores [0,025; 0,05; 0,1; 0,25; 0,5; 0,75; 0,9; 0,95; 0,975].

Valoración de una hipótesis sobre la diferencia de dos medias

En este caso se prueba si la diferencia de medias se encuentra dentro de un intervalo de referencia. El planteamiento de hipótesis sería:

H: $D \in [d_1, d_2]$ *vs.* K: $D < d_1$ o $D > d_2$

Nota: Se debe establecer una probabilidad *a priori* (q) del grado de certeza o validez para hipótesis H a la hora de introducir los datos en procesadores estadísticos como, por ejemplo, Epidat 4.2.

Se debe contar con:

- Se asume una distribución *a priori* de tipo uniforme.
- Se asume las distribuciones *a posteriori* de tipo normal.
- Datos de ambas muestras (empíricos) que estiman los de dos poblaciones: $\overline{x}_1$, s_1, n_1 y $\overline{x}_2$, s_2, n_2.
- Valores de los extremos involucrados en la hipótesis H: d_1 y d_2.
- Factor de Bayes a favor de hipótesis H (BF):

$$BF = \frac{P(H)}{1-P(H)} \text{ o } BF = \frac{Pr(H)}{Pr(K)}$$

- Factor de Bayes en contra de hipótesis H (BC):

$$BC = \frac{1}{BF}$$

- Probabilidad *a posteriori* de la veracidad de hipótesis H:

$$p(H \mid datos) = \frac{qBF}{qBF + (1-q)}$$

donde:

$P(H) = F(d_2) - F(d_1)$: probabilidad de la hipótesis H;

F: estimación no paramétrica de la función de distribución *a posteriori* de la diferencia de medias:

$$F(x) = Pr(d \le x) = \frac{1}{n}\sum_{j=1}^{n} I(d_j \le x)$$

n: número de simulaciones;

$d_j = m_{1j} - m_{2j}$, j = 1, 2,..., n

$$m_{ij} = \sqrt{\frac{s_i^2 z_{ij}^2}{y_{ij} n_i}} + \overline{x}_i \text{ , i = 1, 2,..., n; j = 1, 2,..., n}$$

y_{i1}, y_{i2}, y_{in}: n realizaciones de la distribución chi-cuadrado con n – 1 grados de libertad, i = 1, 2;

z_{i1}, z_{i2}, z_{in}: n realizaciones de la distribución normal estándar, i = 1, 2;

$\overline{x}_i$ y s_i: la media y la desviación estándar, respectivamente, de la muestra i, i = 1, 2;

n_i: tamaño de la muestra, i = 1, 2;

q: es la probabilidad *a priori* de la validez de la hipótesis H.

Interpretación del resultado:

- La comparación del factor de Bayes (BF) *versus* factor de Bayes en contra de H (BC) respecto a la Hipótesis H, permite determinar cuan más o menos veces probables es

que sea cierta la hipótesis H respecto a la hipótesis complementaria (F) según sea mayor o menor respectivamente el factor de Bayes *versus* factor de Bayes en contra de H.

– La probabilidad *a posteriori* expresa la veracidad o cuan probable es que sea cierta la hipótesis H, es decir, del valor de la probabilidad para el parámetro o estadígrafo de referencia se cuenta respecto a los datos empleados en las muestras de estudio al combinarlos en un análisis bayesiano.

Aplicación bayesiana en la inferencia bioestadística en la valoración de pruebas convencionales

En este caso se tratan las variantes: prueba para la comparación de medias y prueba de chi-cuadrado, lo cual se pasa a describir.

Prueba de comparación de medias

Se emplea para el caso en que se valora una diferencia entre dos medias a través de la prueba de Z ~ N (0, 1), donde se estima una cota inferior para el factor de Bayes en función del valor observado o experimental Z a partir de la comparación de las dos medias en la muestra:

$$\exp\left(-\frac{Z^2}{2}\right) \leq FB$$

Esta expresión se incluye en la correspondiente cota inferior de probabilidad para nulidad en forma de desigualdad resultante, para la comparación de las medias como:

$$\left[1+\frac{1-Pr(H_0)}{Pr(H_0)}\exp\left(\frac{Z^2}{2}\right)\right]^{-1} \leq Pr(H_0 \mid D-d_0)$$

Lo que finalmente se estima con esta cota inferior es el menor valor de la probabilidad *a posteriori* para la veracidad de la hipótesis H de nulidad (P_{inf}) que equivale a la expresión: Pr (H_0 | D – d_0):

$$p_{inf} = \left[1+\frac{1-q}{q}\exp\left(\frac{Z^2}{2}\right)\right]^{-1}$$

donde:

Z: estadístico experimental de la prueba convencional de comparación de 2 medias para muestras grandes;

q o Pr(H_0): probabilidad *a priori* de la validez de la hipótesis H;

ϕ: expresa la función de distribución normal estándar;

p = 2[1 – φ(|Z|]: valor p de la prueba convencional para hipótesis H.

Prueba chi-cuadrado

Se emplea para el caso en que se valora una independencia entre dos variables a través de la prueba chi-cuadrado, donde se estima una cota inferior para el factor de Bayes en función del valor observado o experimental chi-cuadrado en la muestra:

$$\sqrt{\chi^2}\exp\left(\frac{1-\chi^2}{2}\right)\leq FB$$

Esta expresión se incluye en la correspondiente cota inferior de probabilidad para nulidad, en forma de desigualdad resultante, para la independencia de dos variables:

$$\left[1+\frac{1-Pr(H_0)}{Pr(H_0)\sqrt{\chi^2}\exp\left(\frac{1-\chi^2}{2}\right)}\right]^{-1}\leq Pr(H_0\,|\,D-d_0)$$

De esta forma se ha estimado la cota inferior o menor valor de la probabilidad *a posteriori* para la veracidad de la hipótesis H (P_{inf}) que equivale a Pr ($H_0|D - d_0$) a partir de la expresión:

$$p_{inf}=\left[1+\frac{1-q}{q\sqrt{\chi^2}\exp\left(\frac{1-\chi^2}{2}\right)}\right]^{-1}$$

donde:

χ^2: estadístico experimental de la prueba chi-cuadrado de Pearson;

k: número de grados de libertad;

q o Pr (H_0): probabilidad *a priori* de la validez de la hipótesis H;

F: función de distribución χ^2 con k grados de libertad;

$p = 1 - F(\chi^2)$: valor p de la prueba convencional.

Sección II

Propuestas algorítmicas para el manejo de muestras y variables en la inferencia bioestadística

Esta sección consta de dos capítulos: en el primer capítulo se presentan algunas propuestas algorítmicas esquematizadas a partir de la presentación de variadas figuras y tablas resumen, que ayudan a sintetizar de una manera interconectada los principales aspectos básicos involucrados en el manejo de muestras y variables para los diferentes contrastes de hipótesis por pruebas de significación paramétricas y no paramétricas, siguiendo el modelo o paradigma frecuentista, fundamentalmente, y de forma paralela en algunos ejemplos por el modelo bayesiano, ante las situaciones investigativas que más frecuentemente puede enfrentar el investigador joven. En el segundo capítulo se varía la forma de presentar la propuesta algorítmica relativa a este mismo tópico, en este caso se emplea una no esquematizada basada en preguntas y respuestas de forma ordenada y progresiva con tomas de decisión en el manejo de muestras y variables según el tipo de combinación, lo que representa los requerimientos o las bases necesarias que están en íntima relación con las respectivas opciones de técnicas estadísticas a utilizar para los contrastes de hipótesis lo que converge en el mismo objetivo final que en el capítulo anterior.

Capítulo 6

Propuestas algorítmicas esquematizadas para el manejo de muestras y variables

Al intentar la planificación de un proyecto de investigación con sus respectivas etapas, el investigador joven se apoya generalmente en sus tutores, asesores o personal de mayor experiencia en este sentido en su centro investigativo. Un punto álgido en este proceso consiste durante la etapa de planificación y ejecución del proyecto investigativo en cómo llevar a cabo el procesamiento estadístico eficaz de la información a obtener o recogida, lo que guarda relación en no pocas ocasiones con no contar con alguna herramienta o instrumento algorítmico facilitador a modo de guía a seguir, que articule las posibles opciones o posibilidades de como efectuarlo, sobre todo en la vertiente inferencial partiendo de la descriptiva en apego estricto a los objetivos trazados en la investigación.

Se presenta una propuesta de una guía práctica general algorítmica para el manejo de muestras y variables, basados en un modelo frecuentista en la inferencia bioestadística (Fig. 6.1), donde queda expuesto de forma general un posible camino a seguir para este empeño, tomando en cuenta el número y tipo de muestras a emplear, así como al unísono, el número y tipo de variables, con cuyas posibles combinaciones se relacionan de forma graficada un grupo de las posibles opcion es de técnicas estadísticas incluidas en los diversos tipos de contrastes de hipótesis o pruebas de significación correspondientes con la inferencia bioestadística, que permiten el análisis y la evaluación de los resultados obtenidos en función de cumplir con los objetivos planteados para resolver el problema investigativo identificado por el investigador joven.

Con la intención de dar continuidad y mostrar otra arista relacionada con la propuesta presentada en la figura 6.1 se muestra gráficamente otra manera didáctica de enfocar el manejo de muestras y variables y viceversa en las figuras 6.2 y 6.3, con similar contenido y simbología en ambas, pero con diferente punto de partida para llegar a las mismas opciones del análisis inferencial; son dos posibles flujogramas relacionados con las opciones más frecuentes a emplear en la inferencia bioestadística derivadas de las posibles combinaciones de variables y muestras según su número y tipo de las mismas.

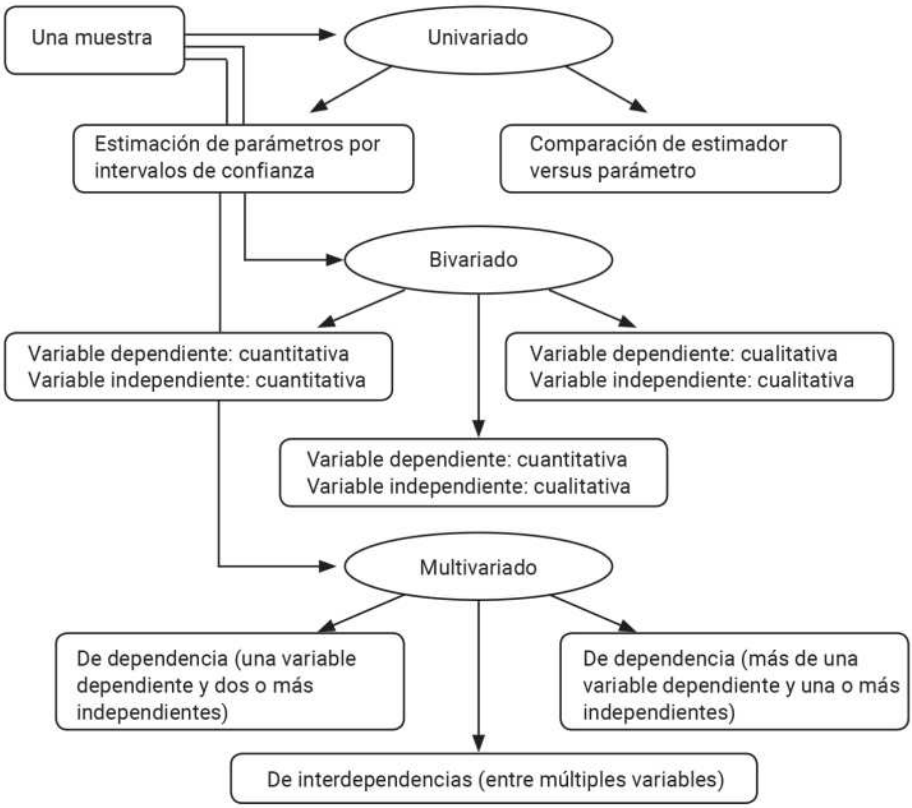

Una muestra
Univariado
Estimación de parámetros por intervalos de confianza
Comparación de estimador versus parámetro
Bivariado
Variable dependiente: cuantitativa
Variable independiente: cuantitativa
Variable dependiente: cualitativa
Variable independiente: cualitativa
Variable dependiente: cuantitativa
Variable independiente: cualitativa
Multivariado
De dependencia (una variable dependiente y dos o más independientes)
De dependencia (más de una variable dependiente y una o más independientes)
De interdependencias (entre múltiples variables)

Fig. 6.1.

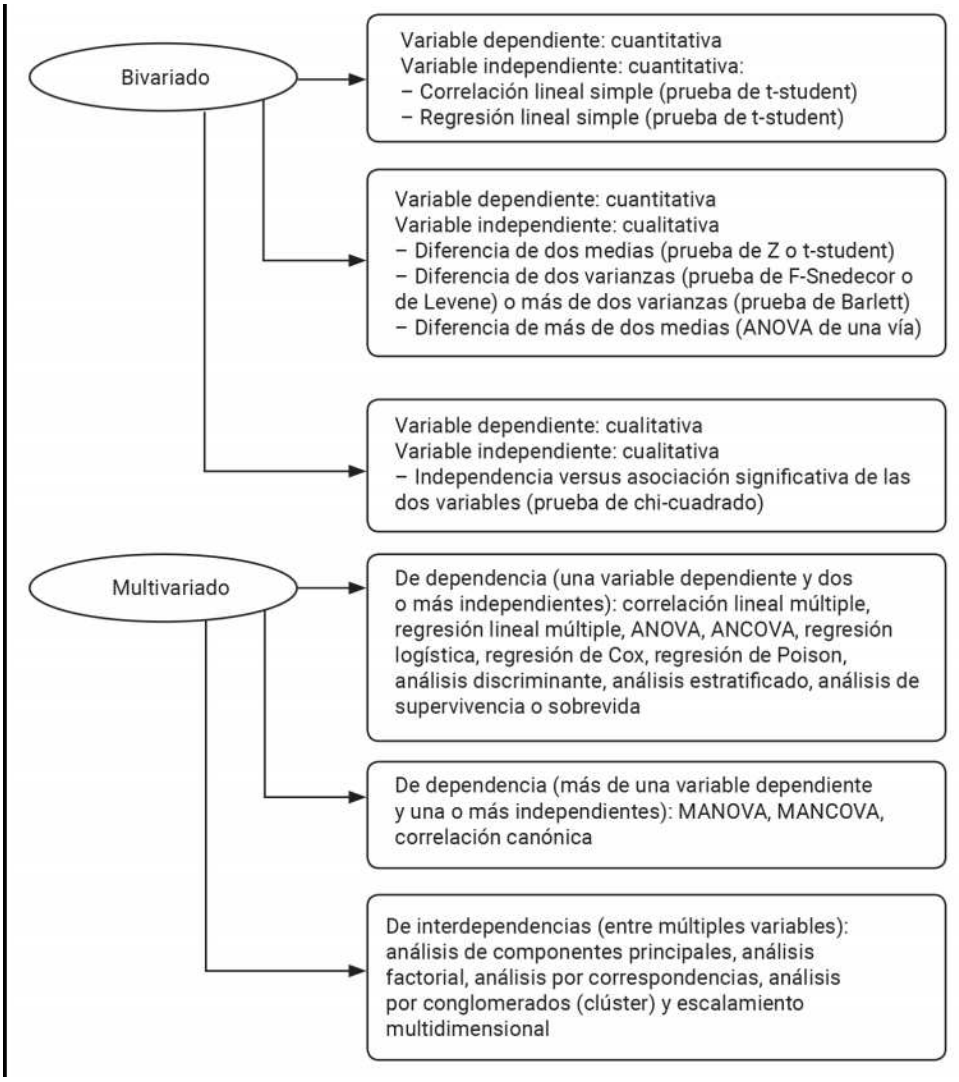

Bivariado
Variable dependiente: cuantitativa
Variable independiente: cuantitativa:
– Correlación lineal simple (prueba de t-student)
– Regresión lineal simple (prueba de t-student)
Variable dependiente: cuantitativa
Variable independiente: cualitativa
– Diferencia de dos medias (prueba de Z o t-student)
– Diferencia de dos varianzas (prueba de F-Snedecor o de Levene) o más de dos varianzas (prueba de Barlett)
– Diferencia de más de dos medias (ANOVA de una vía)
Variable dependiente: cualitativa
Variable independiente: cualitativa
– Independencia versus asociación significativa de las dos variables (prueba de chi-cuadrado)
Multivariado
De dependencia (una variable dependiente y dos o más independientes): correlación lineal múltiple, regresión lineal múltiple, ANOVA, ANCOVA, regresión logística, regresión de Cox, regresión de Poison, análisis discriminante, análisis estratificado, análisis de supervivencia o sobrevida
De dependencia (más de una variable dependiente y una o más independientes): MANOVA, MANCOVA, correlación canónica
De interdependencias (entre múltiples variables): análisis de componentes principales, análisis factorial, análisis por correspondencias, análisis por conglomerados (clúster) y escalamiento multidimensional

Fig. 6.1.

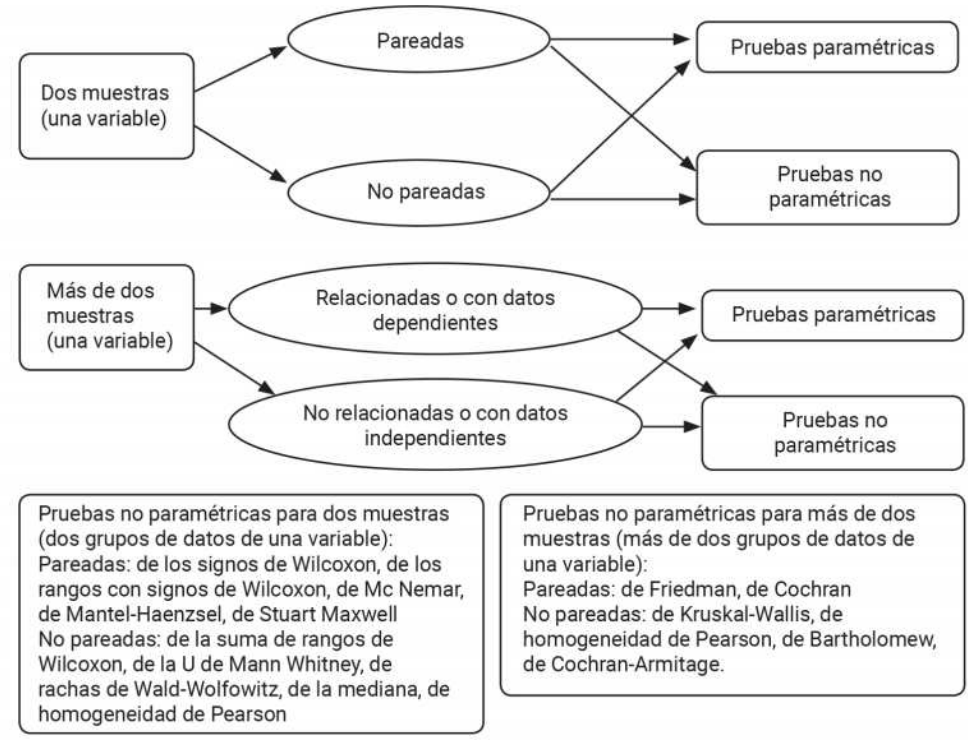


Fig. 6.1. Propuesta de una guía práctica general algorítmica para el manejo de muestras y variables, siguiendo un modelo frecuentista en la inferencia bioestadística.

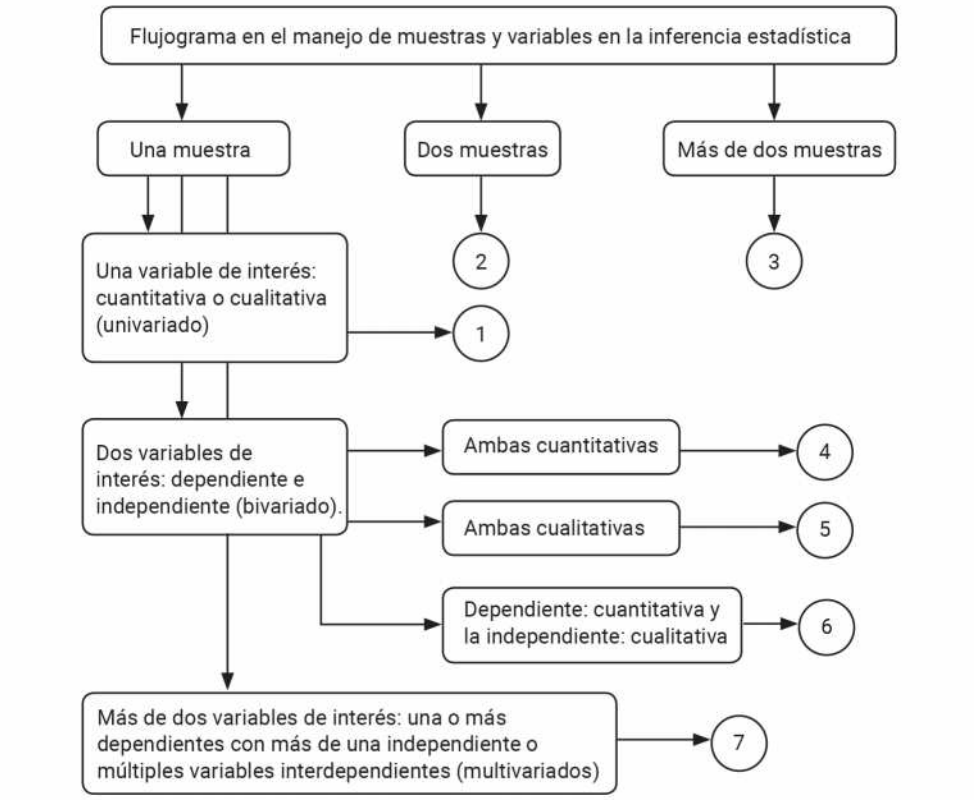


Fig. 6.2. Diagrama de flujo en el manejo bioestadístico de muestras y variables variante 1, partiendo del número de muestras y siguiendo un modelo o paradigma frecuentista en la inferencia bioestadística:

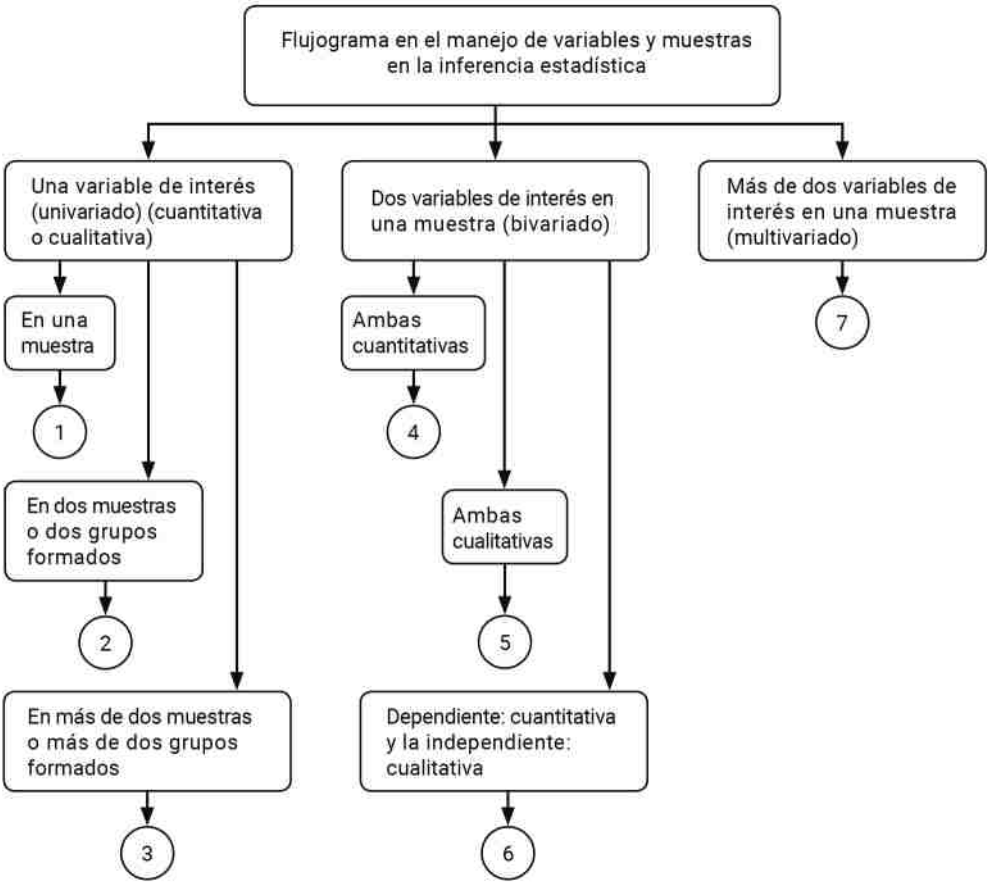
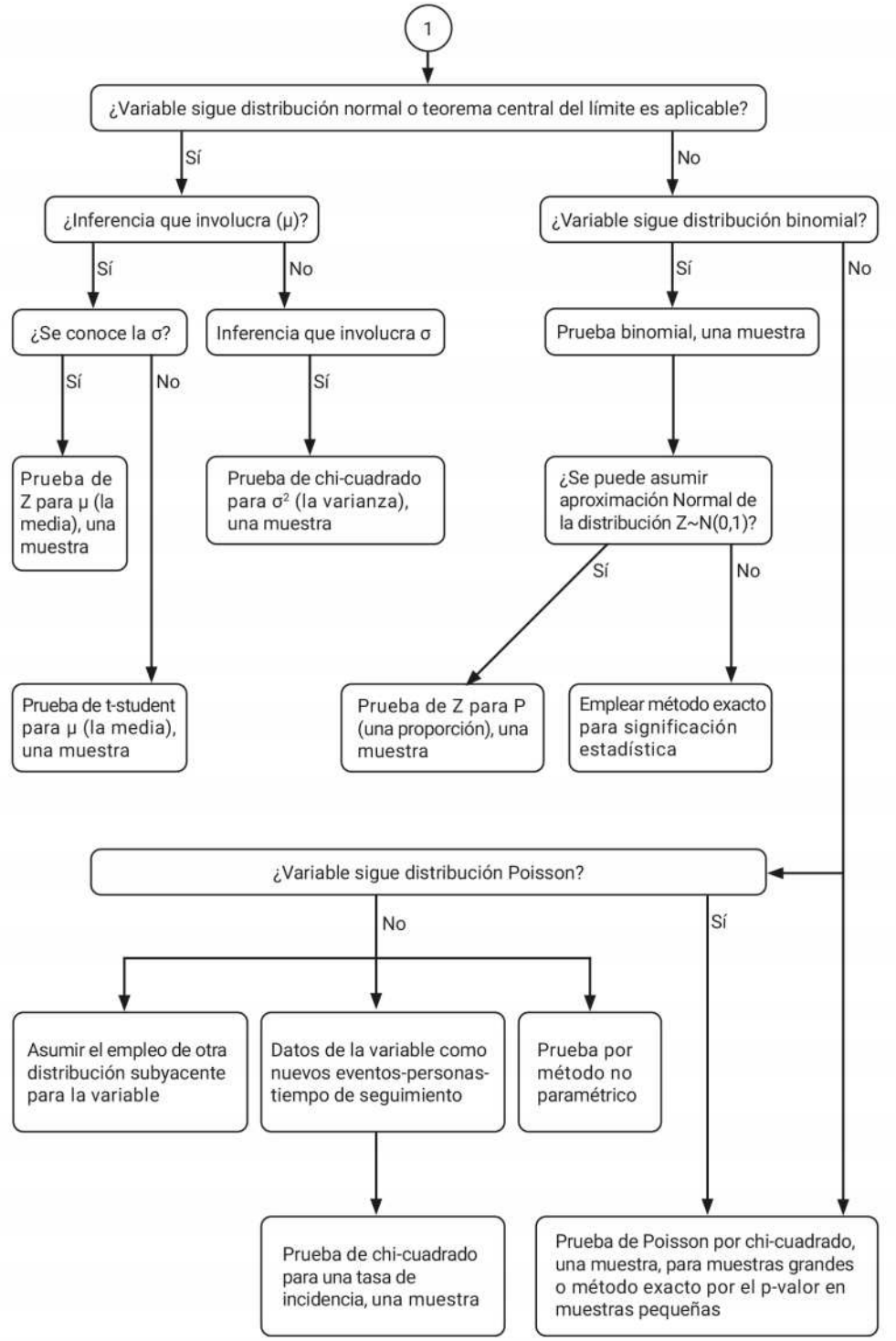
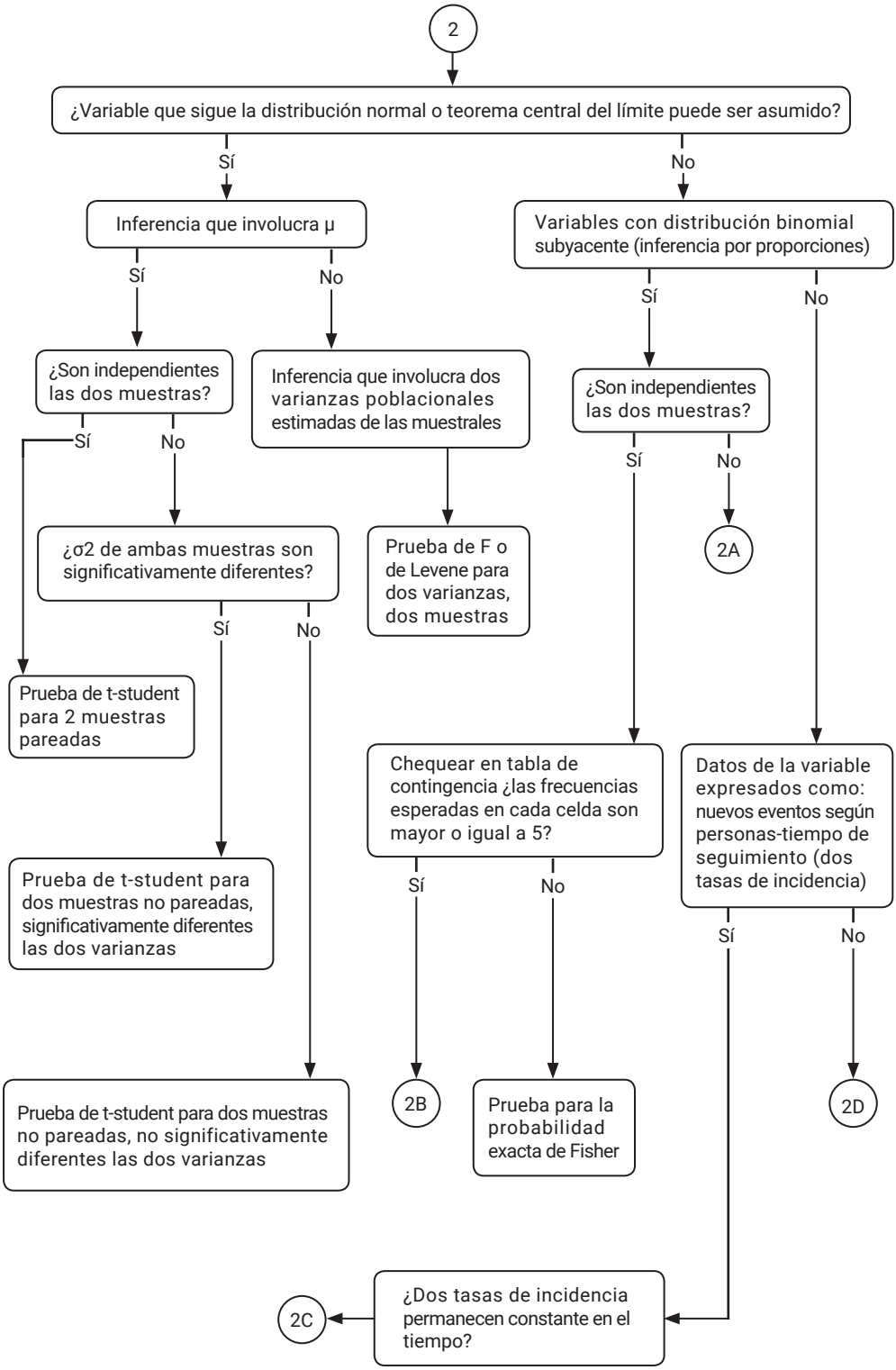
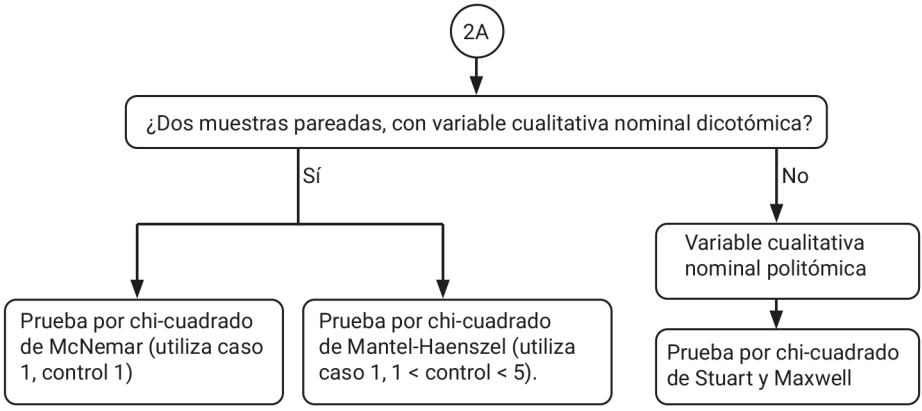


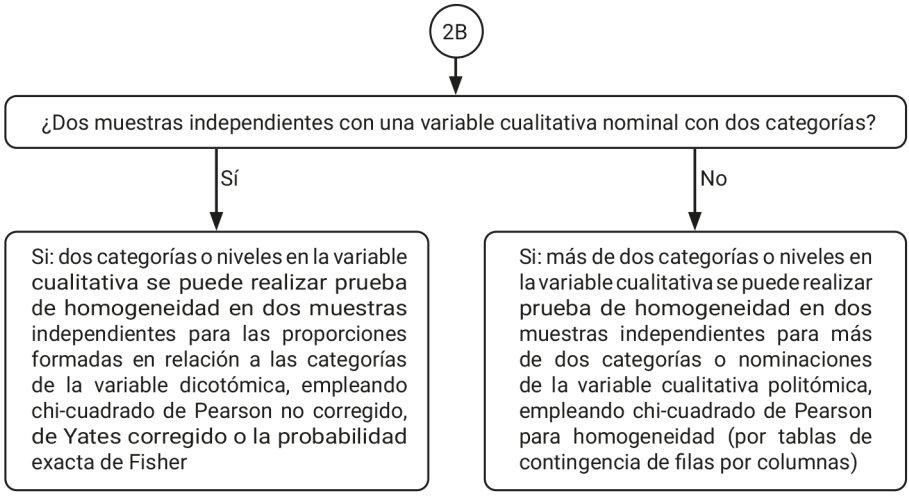
Fig. 6.3. Diagrama de flujo en el manejo bioestadístico de muestras y variables, variante 2, partiendo del número de variablesy siguiendo un modelo o paradigma frecuentista en la inferencia bioestadística.

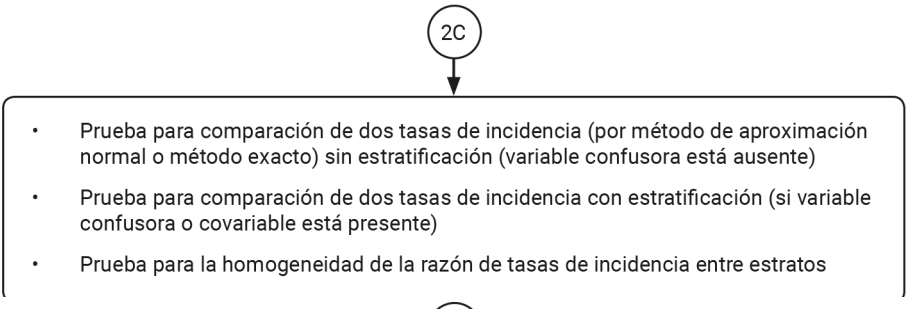
El desarrollo gráfico de estos esquemas algorítmicos con las posibles opciones de técnicas estadísticas a emplear como parte de la inferencia bioestadística y la consiguiente explicación de la simbología utilizada se presentan a continuación.

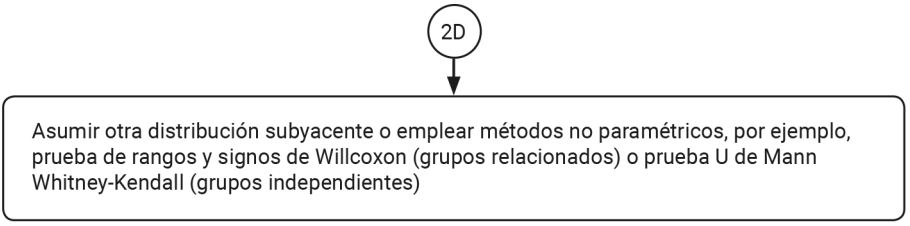

1
¿Variable sigue distribución normal o teorema central del límite es aplicable?
Sí
No
¿Inferencia que involucra (μ)?
¿Variable sigue distribución binomial?
Sí
No
Sí
No
¿Se conoce la σ?
Inferencia que involucra σ
Prueba binomial, una muestra
Sí
No
Sí
Prueba de Z para μ (la media), una muestra
Prueba de chi-cuadrado para σ² (la varianza), una muestra
¿Se puede asumir aproximación Normal de la distribución Z~N(0,1)?
Sí
No
Prueba de t-student para μ (la media), una muestra
Prueba de Z para P (una proporción), una muestra
Emplear método exacto para significación estadística
¿Variable sigue distribución Poisson?
No
Sí
Asumir el empleo de otra distribución subyacente para la variable
Datos de la variable como nuevos eventos-personas-tiempo de seguimiento
Prueba por método no paramétrico
Prueba de chi-cuadrado para una tasa de incidencia, una muestra
Prueba de Poisson por chi-cuadrado, una muestra, para muestras grandes o método exacto por el p-valor en muestras pequeñas

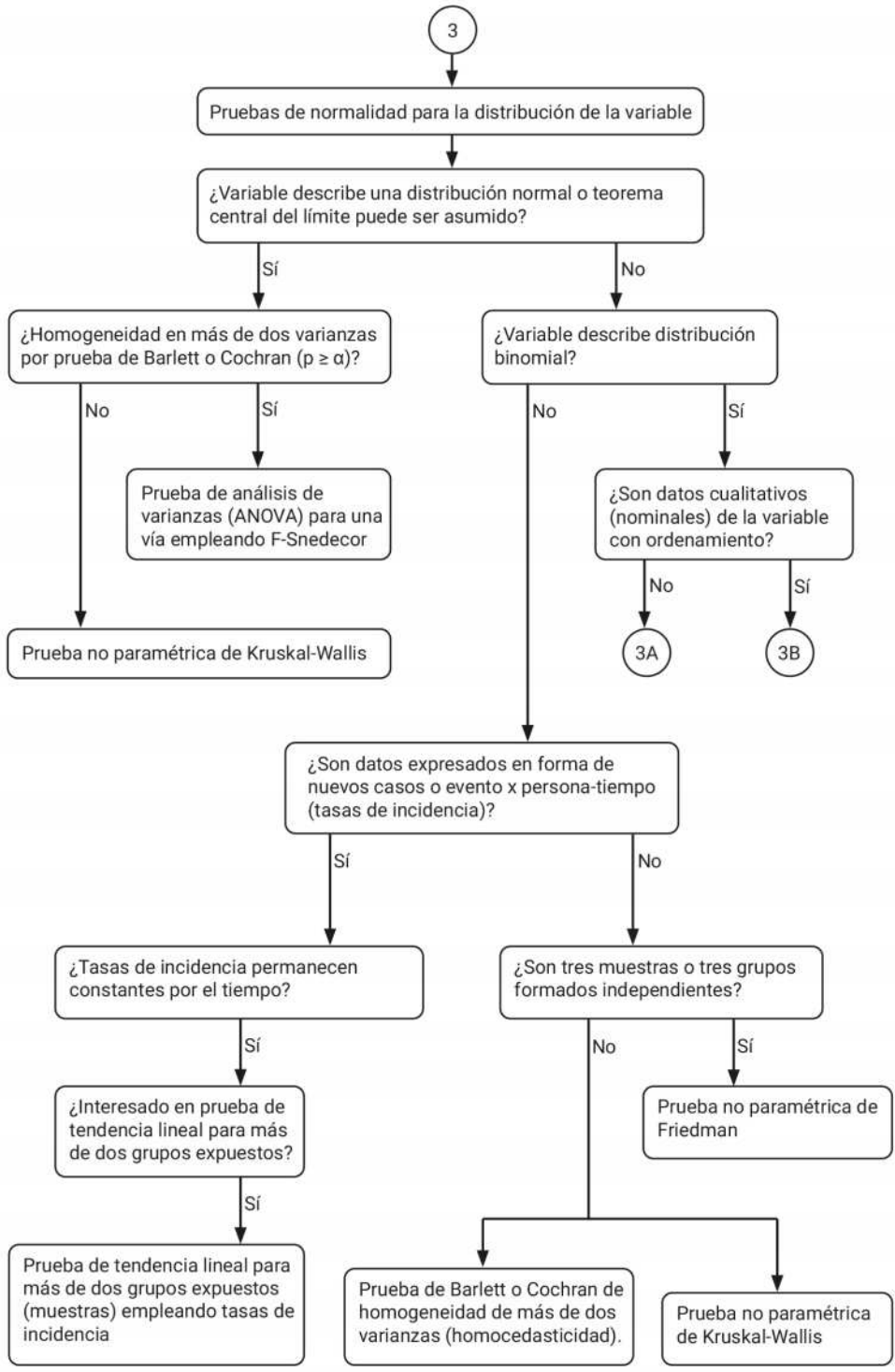

2
¿Variable que sigue la distribución normal o teorema central del límite puede ser asumido?
Sí
No
Inferencia que involucra μ
Variables con distribución binomial subyacente (inferencia por proporciones)
Sí
No
Sí
No
¿Son independientes las dos muestras?
Inferencia que involucra dos varianzas poblacionales estimadas de las muestrales
¿Son independientes las dos muestras?
Sí
No
Sí
No
¿σ2 de ambas muestras son significativamente diferentes?
Prueba de F o de Levene para dos varianzas, dos muestras
2A
Sí
No
Prueba de t-student para 2 muestras pareadas
Chequear en tabla de contingencia ¿las frecuencias esperadas en cada celda son mayor o igual a 5?
Datos de la variable expresados como: nuevos eventos según personas-tiempo de seguimiento (dos tasas de incidencia)
Prueba de t-student para dos muestras no pareadas, significativamente diferentes las dos varianzas
Sí
No
Sí
No
Prueba de t-student para dos muestras no pareadas, no significativamente diferentes las dos varianzas
2B
Prueba para la probabilidad exacta de Fisher
2D
¿Dos tasas de incidencia permanecen constante en el tiempo?
2C

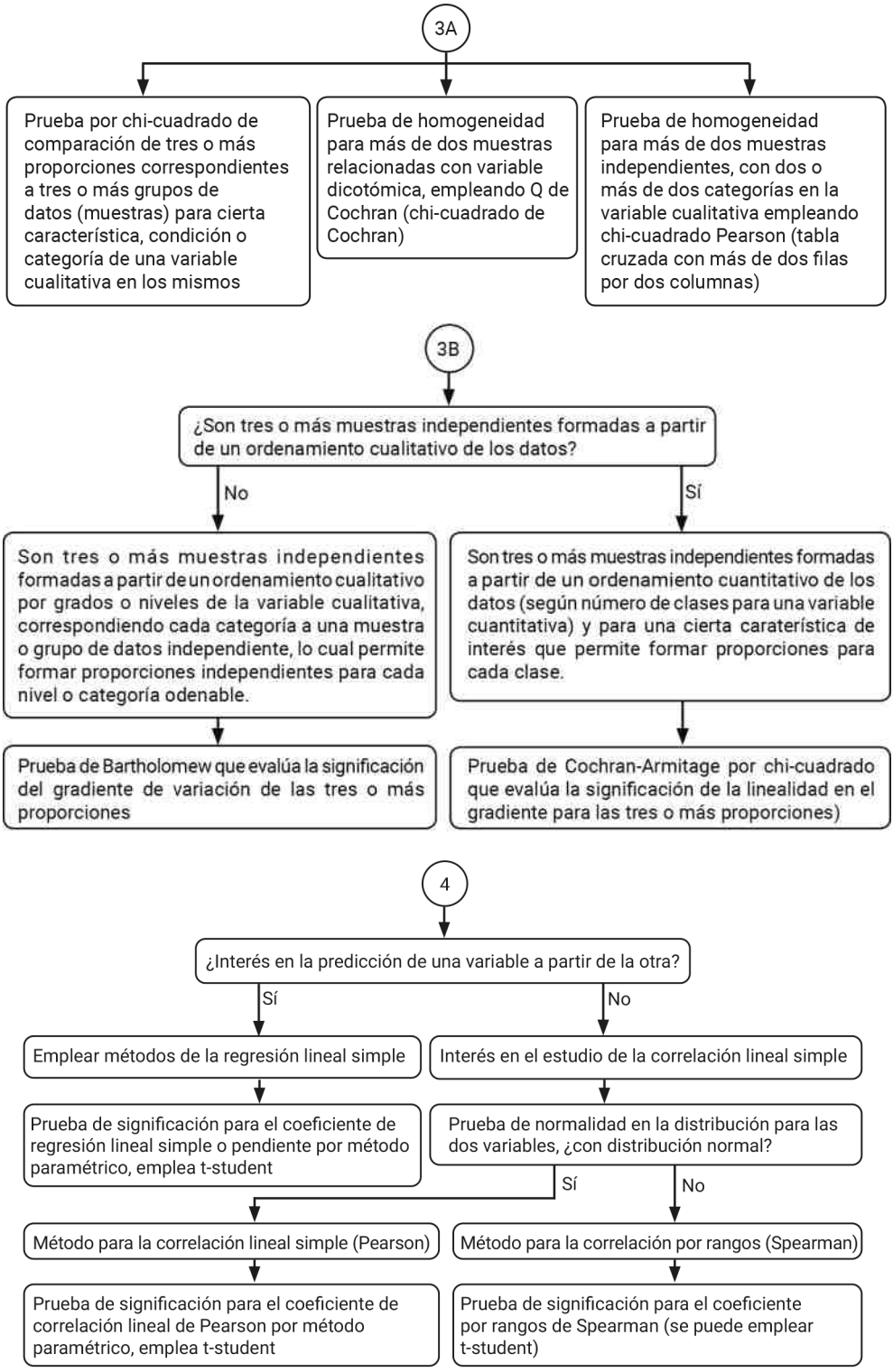

2A
¿Dos muestras pareadas, con variable cualitativa nominal dicotómica?
Sí
No
Prueba por chi-cuadrado de McNemar (utiliza caso 1, control 1)
Prueba por chi-cuadrado de Mantel-Haenszel (utiliza caso 1, 1 < control < 5).
Variable cualitativa nominal politómica
Prueba por chi-cuadrado de Stuart y Maxwell

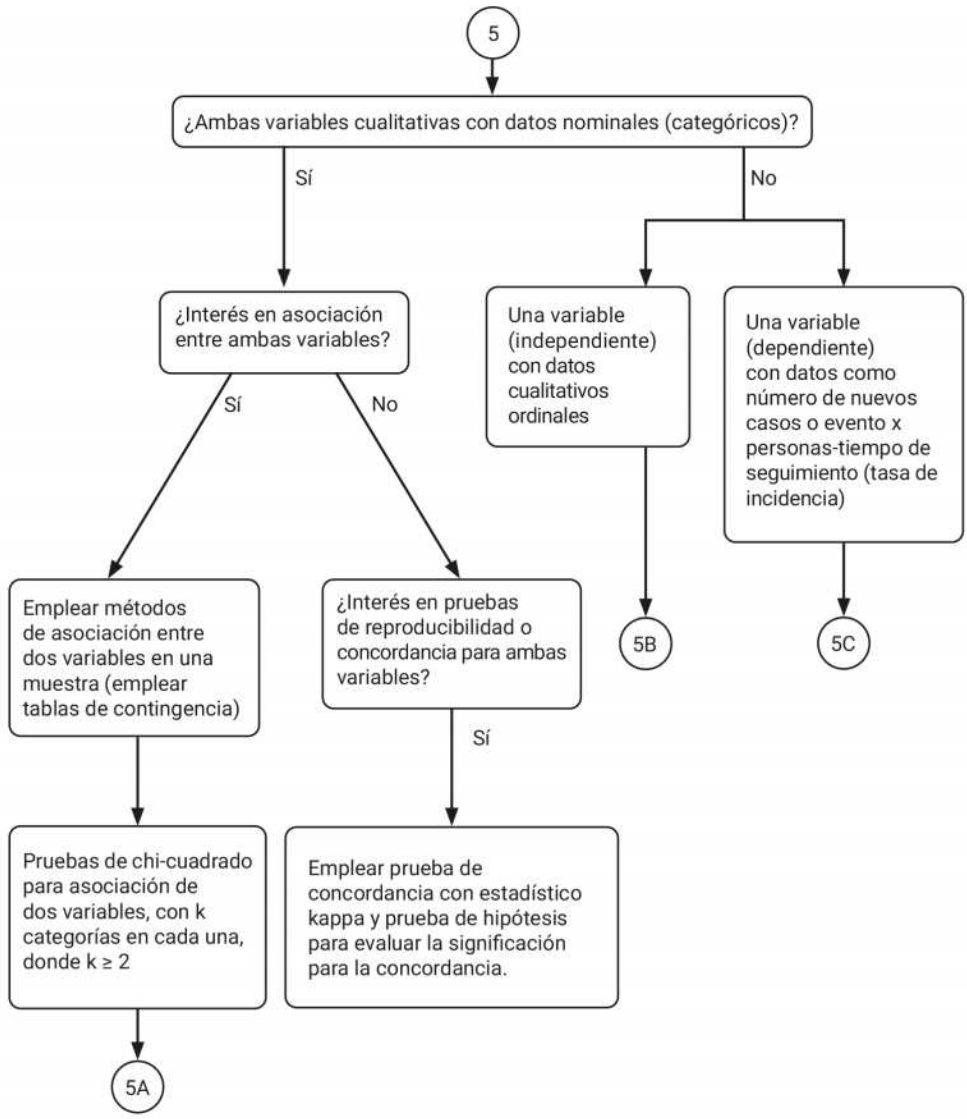

2B
¿Dos muestras independientes con una variable cualitativa nominal con dos categorías?
Sí
No
Si: dos categorías o niveles en la variable cualitativa se puede realizar prueba de homogeneidad en dos muestras independientes para las proporciones formadas en relación a las categorías de la variable dicotómica, empleando chi-cuadrado de Pearson no corregido, de Yates corregido o la probabilidad exacta de Fisher
Si: más de dos categorías o niveles en la variable cualitativa se puede realizar prueba de homogeneidad en dos muestras independientes para más de dos categorías o nominaciones de la variable cualitativa politómica, empleando chi-cuadrado de Pearson para homogeneidad (por tablas de contingencia de filas por columnas)


2C
Prueba para comparación de dos tasas de incidencia (por método de aproximación normal o método exacto) sin estratificación (variable confusora está ausente)
Prueba para comparación de dos tasas de incidencia con estratificación (si variable confusora o covariable está presente)
Prueba para la homogeneidad de la razón de tasas de incidencia entre estratos


2D
Asumir otra distribución subyacente o emplear métodos no paramétricos, por ejemplo, prueba de rangos y signos de Willcoxon (grupos relacionados) o prueba U de Mann Whitney-Kendall (grupos independientes)


3
Pruebas de normalidad para la distribución de la variable
¿Variable describe una distribución normal o teorema central del límite puede ser asumido?
Sí
No
¿Homogeneidad en más de dos varianzas por prueba de Barlett o Cochran (p ≥ α)?
¿Variable describe distribución binomial?
No
Sí
No
Sí
Prueba de análisis de varianzas (ANOVA) para una vía empleando F-Snedecor
¿Son datos cualitativos (nominales) de la variable con ordenamiento?
Prueba no paramétrica de Kruskal-Wallis
No
Sí
3A
3B
¿Son datos expresados en forma de nuevos casos o evento x persona-tiempo (tasas de incidencia)?
Sí
No
¿Tasas de incidencia permanecen constantes por el tiempo?
¿Son tres muestras o tres grupos formados independientes?
Sí
No
Sí
¿Interesado en prueba de tendencia lineal para más de dos grupos expuestos?
Prueba no paramétrica de Friedman
Sí
Prueba de tendencia lineal para más de dos grupos expuestos (muestras) empleando tasas de incidencia
Prueba de Barlett o Cochran de homogeneidad de más de dos varianzas (homocedasticidad).
Prueba no paramétrica de Kruskal-Wallis


3A
Prueba por chi-cuadrado de comparación de tres o más proporciones correspondientes a tres o más grupos de datos (muestras) para cierta característica, condición o categoría de una variable cualitativa en los mismos
Prueba de homogeneidad para más de dos muestras relacionadas con variable dicotómica, empleando Q de Cochran (chi-cuadrado de Cochran)
Prueba de homogeneidad para más de dos muestras independientes, con dos o más de dos categorías en la variable cualitativa empleando chi-cuadrado Pearson (tabla cruzada con más de dos filas por dos columnas)
3B
¿Son tres o más muestras independientes formadas a partir de un ordenamiento cualitativo de los datos?
No
Sí
Son tres o más muestras independientes formadas a partir de un ordenamiento cualitativo por grados o niveles de la variable cualitativa, correspondiendo cada categoría a una muestra o grupo de datos independiente, lo cual permite formar proporciones independientes para cada nivel o categoría odenable.
Son tres o más muestras independientes formadas a partir de un ordenamiento cuantitativo de los datos (según número de clases para una variable cuantitativa) y para una cierta caraterística de interés que permite formar proporciones para cada clase.
Prueba de Bartholomew que evalúa la significación del gradiente de variación de las tres o más proporciones
Prueba de Cochran-Armitage por chi-cuadrado que evalúa la significación de la linealidad en el gradiente para las tres o más proporciones)
4
¿Interés en la predicción de una variable a partir de la otra?
Sí
No
Emplear métodos de la regresión lineal simple
Interés en el estudio de la correlación lineal simple
Prueba de significación para el coeficiente de regresión lineal simple o pendiente por método paramétrico, emplea t-student
Prueba de normalidad en la distribución para las dos variables, ¿con distribución normal?
Sí
No
Método para la correlación lineal simple (Pearson)
Método para la correlación por rangos (Spearman)
Prueba de significación para el coeficiente de correlación lineal de Pearson por método paramétrico, emplea t-student
Prueba de significación para el coeficiente por rangos de Spearman (se puede emplear t-student)


5
¿Ambas variables cualitativas con datos nominales (categóricos)?
Sí
No
¿Interés en asociación entre ambas variables?
Una variable (independiente) con datos cualitativos ordinales
Una variable (dependiente) con datos como número de nuevos casos o evento x personas-tiempo de seguimiento (tasa de incidencia)
Sí
No
Emplear métodos de asociación entre dos variables en una muestra (emplear tablas de contingencia)
¿Interés en pruebas de reproducibilidad o concordancia para ambas variables?
5B
5C
Sí
Pruebas de chi-cuadrado para asociación de dos variables, con k categorías en cada una, donde k ≥ 2
Emplear prueba de concordancia con estadístico kappa y prueba de hipótesis para evaluar la significación para la concordancia.
5A

5A

Para tabla de contingencia de 2x2 simple: prueba de independencia vs. asociación significativa de dos variables dicotómicas por chi-cuadrado de Pearson o de Yates corregido si las frecuencias esperadas de las celdas son mayor o igual a cinco, en caso contrario realizar la prueba para la probabilidad exacta de Fisher.

para tabla de contingencia de 2x2 estratificada: prueba de independencia versus asociación entre las dos variables dicotómicas en una muestra con una variable de estratificación o covariable dicotómica, empleando chi-cuadrado de Mantel-Haenszel de asociación

Para tabla de contingencia de filas x columnas: prueba de independencia vs. asociación significativa de dos variables polinominales o politómicas (al menos una de ellas) en una muestra empleando chi-cuadrado de Pearson, según la comparación de los valores observados vs. esperados.

5B

Para tabla de contingencia 2xN simple:

- Prueba de tendencia lineal entre los N niveles de la variable independiente para el evento de la dependiente
- Prueba de Homogeneidad entre los N niveles de la variable independiente para el evento de la dependiente

Para tabla de contingencia 2xN estratificada: prueba de tendencia lineal entre los N niveles de la variable dependiente para el evento en la dependiente en cada estrato, según nominaciones del factor de estratificación o covariable

5C

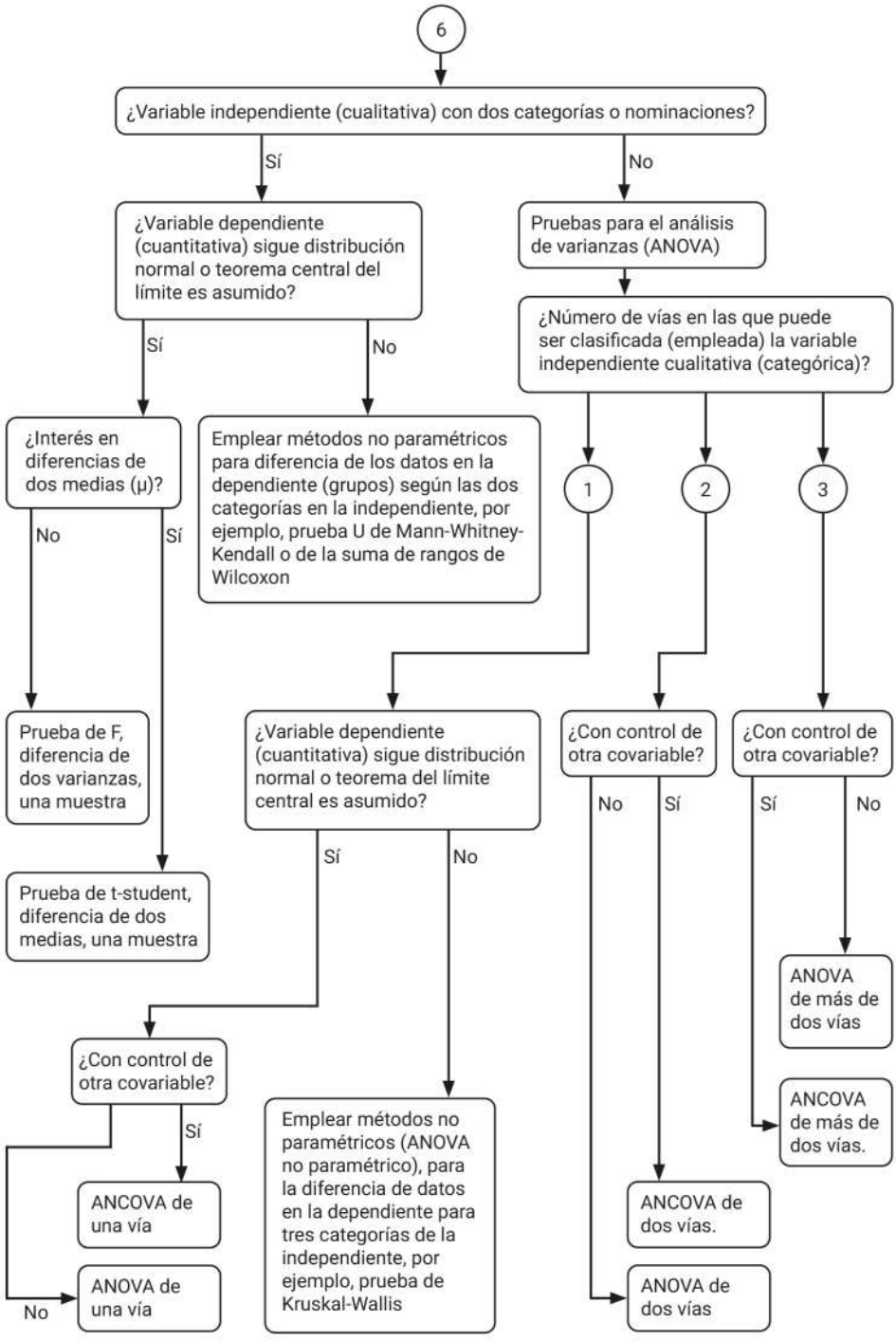
Para tabla de contingencia 2x2 simple: prueba de independencia vs. asociación significativa de dos variables en una muestra, empleando la prueba de Z por aproximación a la normal estándar: Z~N(0,1) o método exacto

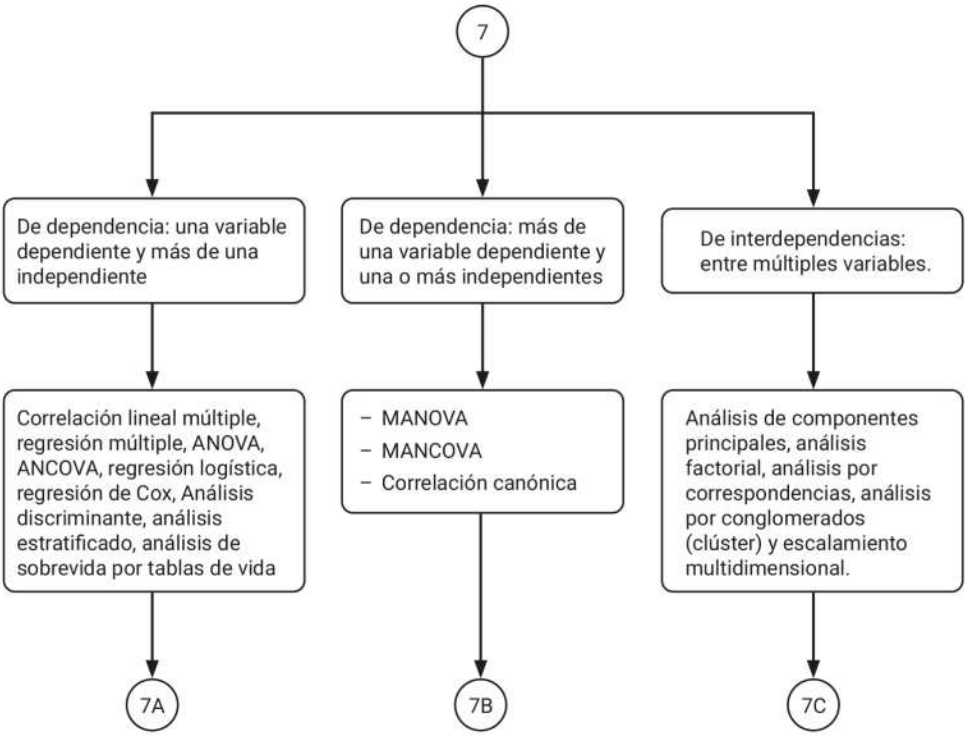
Para tabla de contingencia 2x2 estratificada: prueba de independencia vs. asociación entre las dos variables en una muestra con una variable de estratificación o covariable dicotómica, empleando al estadístico de contraste chi-cuadrado de Mantel-Haenszel de asociación

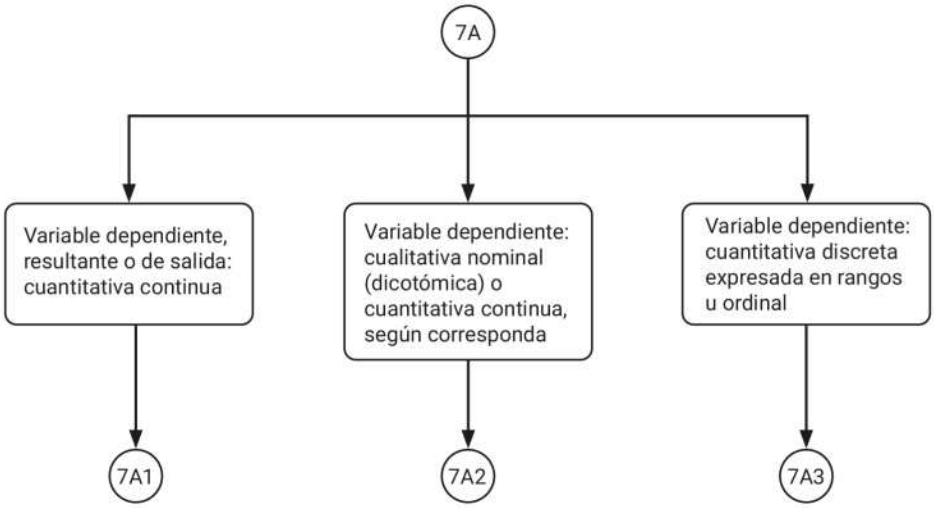
Para tabla de contingencia 2xN simple:

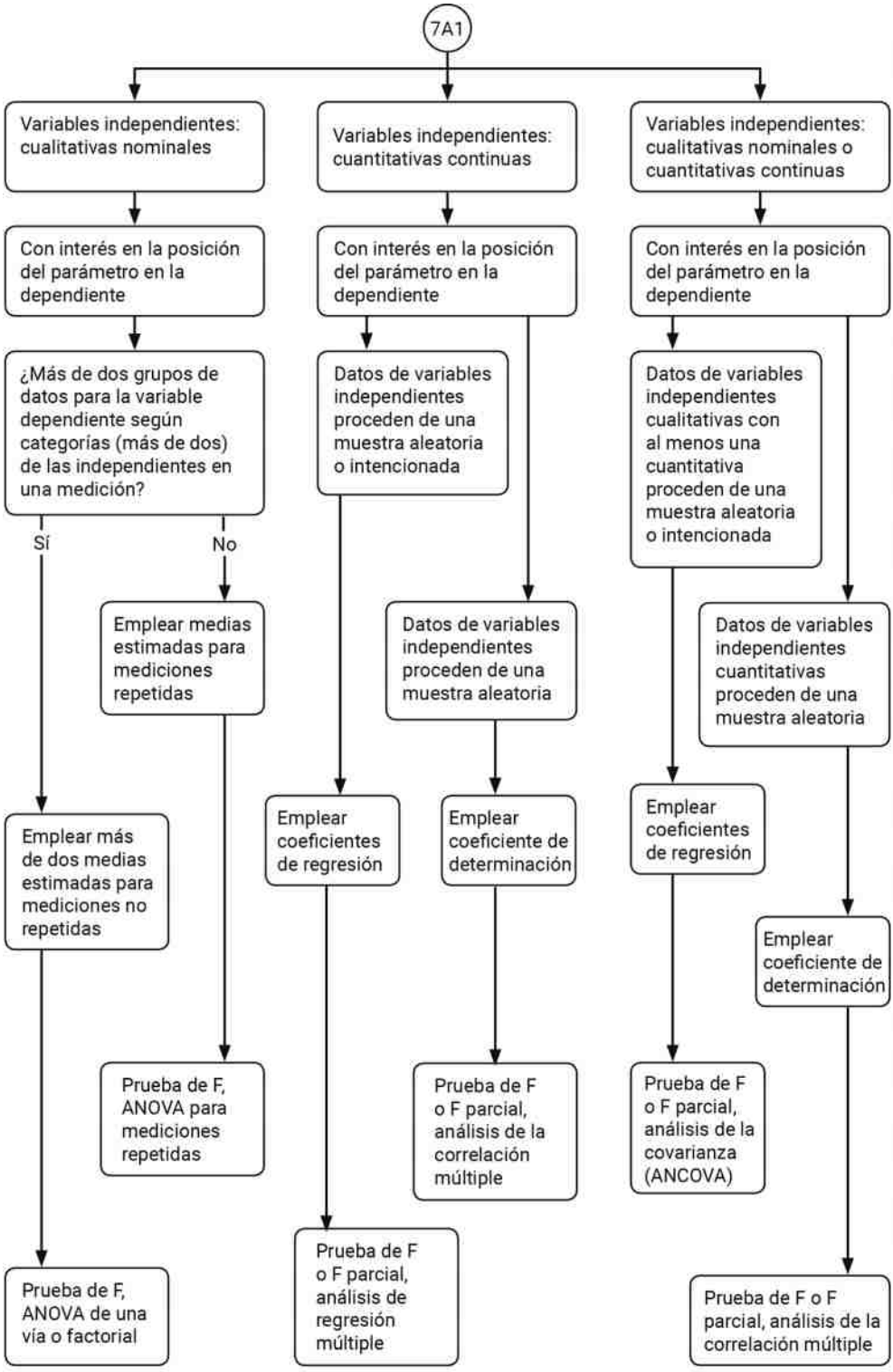
- Prueba de tendencia lineal entre los N niveles de la variable independiente para el evento de la dependiente
- Prueba de homogeneidad entre los N niveles de la variable independiente para el evento de la dependiente

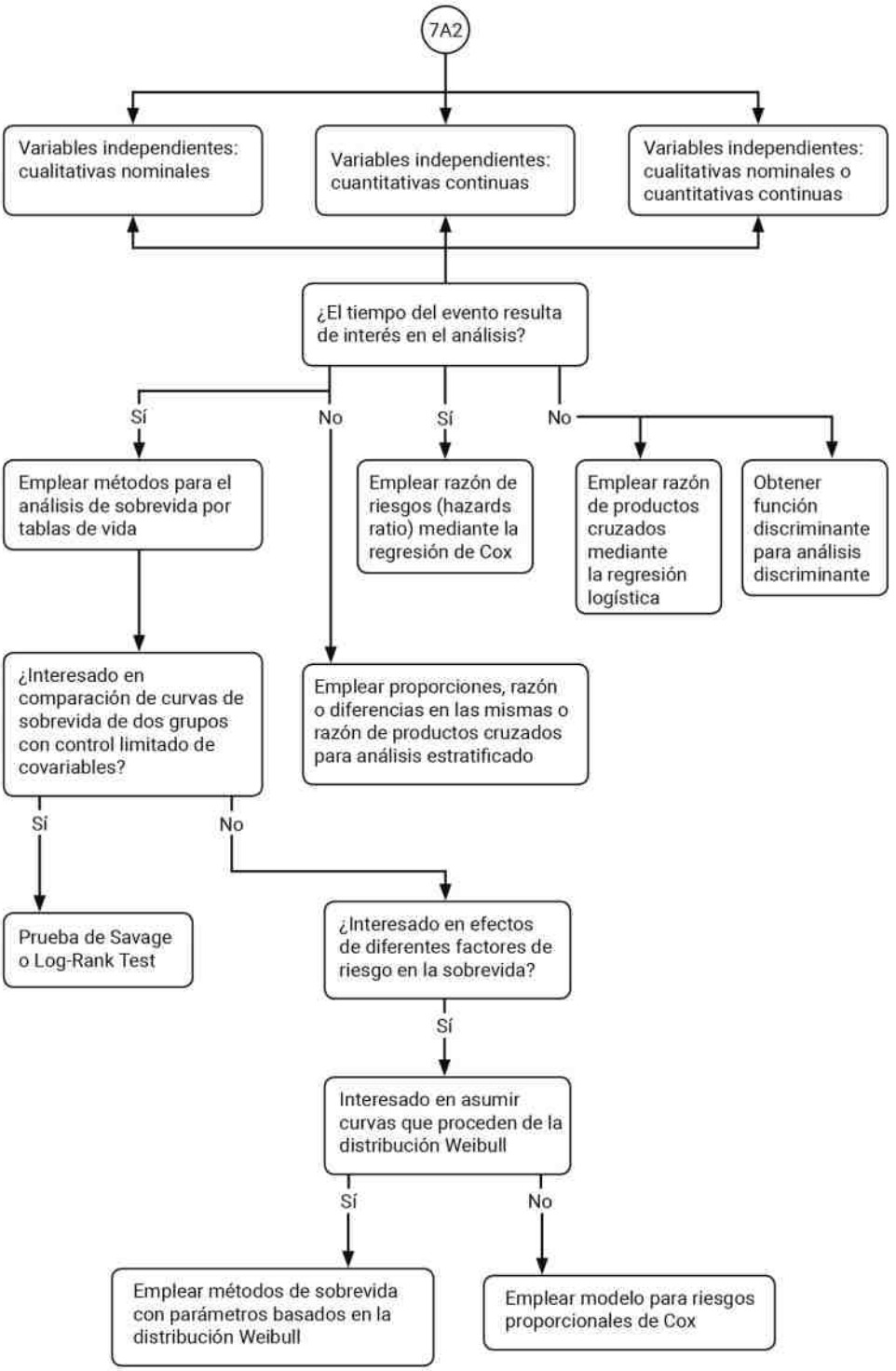
Para tabla de contingencia 2xN estratificada: prueba de tendencia lineal entre los N niveles de la variable independiente para el evento de la dependiente para cada estrato, según nominaciones del factor de estratificación o covariable

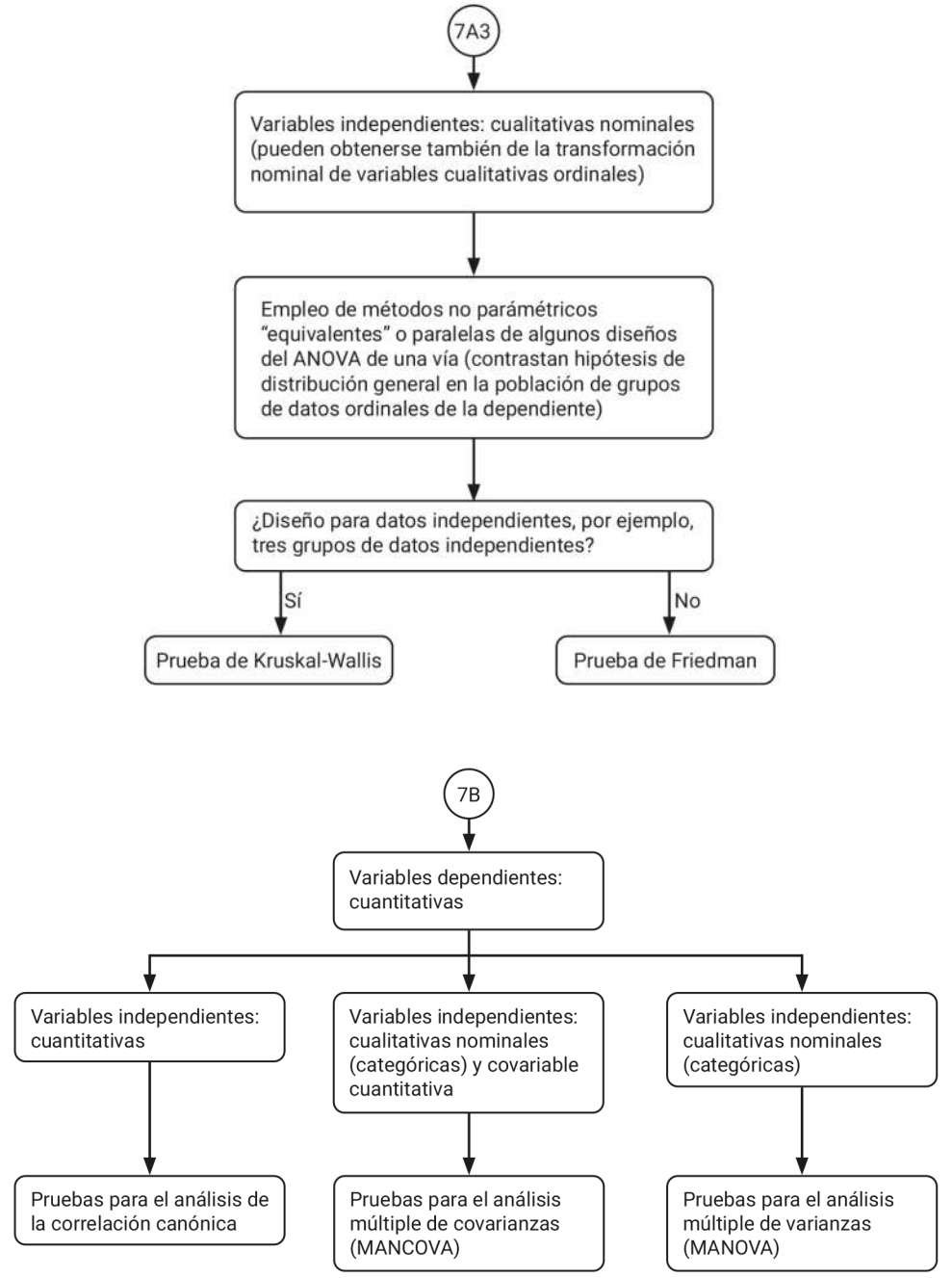

6
¿Variable independiente (cualitativa) con dos categorías o nominaciones?
Sí
No
¿Variable dependiente (cuantitativa) sigue distribución normal o teorema central del límite es asumido?
Pruebas para el análisis de varianzas (ANOVA)
Sí
No
¿Número de vías en las que puede ser clasificada (empleada) la variable independiente cualitativa (categórica)?
¿Interés en diferencias de dos medias (μ)?
Emplear métodos no paramétricos para diferencia de los datos en la dependiente (grupos) según las dos categorías en la independiente, por ejemplo, prueba U de Mann-Whitney-Kendall o de la suma de rangos de Wilcoxon
1
2
3
No
Sí
Prueba de F, diferencia de dos varianzas, una muestra
Prueba de t-student, diferencia de dos medias, una muestra
¿Variable dependiente (cuantitativa) sigue distribución normal o teorema del límite central es asumido?
¿Con control de otra covariable?
¿Con control de otra covariable?
No
Sí
Sí
No
Sí
No
ANOVA de más de dos vías
ANCOVA de más de dos vías.
¿Con control de otra covariable?
Emplear métodos no paramétricos (ANOVA no paramétrico), para la diferencia de datos en la dependiente para tres categorías de la independiente, por ejemplo, prueba de Kruskal-Wallis
Sí
ANCOVA de una vía
No
ANOVA de una vía
ANCOVA de dos vías.
ANOVA de dos vías

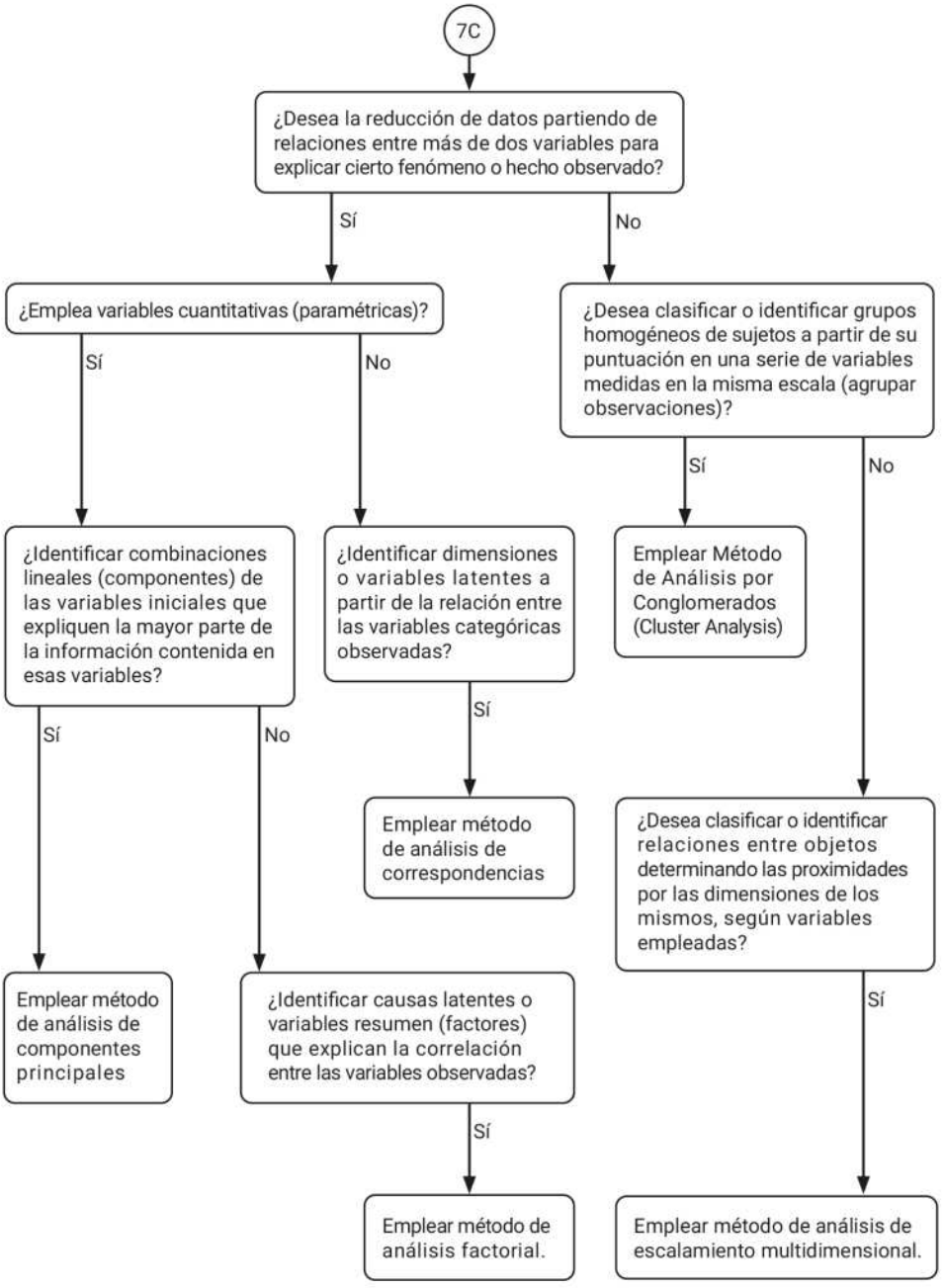

7
De dependencia: una variable dependiente y más de una independiente
De dependencia: más de una variable dependiente y una o más independientes
De interdependencias: entre múltiples variables.
Correlación lineal múltiple, regresión múltiple, ANOVA, ANCOVA, regresión logística, regresión de Cox, Análisis discriminante, análisis estratificado, análisis de sobrevida por tablas de vida
– MANOVA
– MANCOVA
– Correlación canónica
Análisis de componentes principales, análisis factorial, análisis por correspondencias, análisis por conglomerados (clúster) y escalamiento multidimensional.
7A
7B
7C


7A
Variable dependiente, resultante o de salida: cuantitativa continua
Variable dependiente: cualitativa nominal (dicotómica) o cuantitativa continua, según corresponda
Variable dependiente: cuantitativa discreta expresada en rangos u ordinal
7A1
7A2
7A3


7A1
Variables independientes: cualitativas nominales
Variables independientes: cuantitativas continuas
Variables independientes: cualitativas nominales o cuantitativas continuas
Con interés en la posición del parámetro en la dependiente
Con interés en la posición del parámetro en la dependiente
Con interés en la posición del parámetro en la dependiente
¿Más de dos grupos de datos para la variable dependiente según categorías (más de dos) de las independientes en una medición?
Datos de variables independientes proceden de una muestra aleatoria o intencionada
Datos de variables independientes cualitativas con al menos una cuantitativa proceden de una muestra aleatoria o intencionada
Sí
No
Emplear medias estimadas para mediciones repetidas
Datos de variables independientes proceden de una muestra aleatoria
Datos de variables independientes cuantitativas proceden de una muestra aleatoria
Emplear más de dos medias estimadas para mediciones no repetidas
Emplear coeficientes de regresión
Emplear coeficiente de determinación
Emplear coeficientes de regresión
Emplear coeficiente de determinación
Prueba de F, ANOVA para mediciones repetidas
Prueba de F o F parcial, análisis de la correlación múltiple
Prueba de F o F parcial, análisis de la covarianza (ANCOVA)
Prueba de F, ANOVA de una vía o factorial
Prueba de F o F parcial, análisis de regresión múltiple
Prueba de F o F parcial, análisis de la correlación múltiple


7A2
Variables independientes: cualitativas nominales
Variables independientes: cuantitativas continuas
Variables independientes: cualitativas nominales o cuantitativas continuas
¿El tiempo del evento resulta de interés en el análisis?
Sí
No
Sí
No
Emplear métodos para el análisis de sobrevida por tablas de vida
Emplear razón de riesgos (hazards ratio) mediante la regresión de Cox
Emplear razón de productos cruzados mediante la regresión logística
Obtener función discriminante para análisis discriminante
¿Interesado en comparación de curvas de sobrevida de dos grupos con control limitado de covariables?
Emplear proporciones, razón o diferencias en las mismas o razón de productos cruzados para análisis estratificado
Sí
No
Prueba de Savage o Log-Rank Test
¿Interesado en efectos de diferentes factores de riesgo en la sobrevida?
Sí
Interesado en asumir curvas que proceden de la distribución Weibull
Sí
No
Emplear métodos de sobrevida con parámetros basados en la distribución Weibull
Emplear modelo para riesgos proporcionales de Cox


7A3
Variables independientes: cualitativas nominales (pueden obtenerse también de la transformación nominal de variables cualitativas ordinales)
Empleo de métodos no paramétricos "equivalentes" o paralelas de algunos diseños del ANOVA de una vía (contrastan hipótesis de distribución general en la población de grupos de datos ordinales de la dependiente)
¿Diseño para datos independientes, por ejemplo, tres grupos de datos independientes?
Sí
No
Prueba de Kruskal-Wallis
Prueba de Friedman
7B
Variables dependientes: cuantitativas
Variables independientes: cuantitativas
Variables independientes: cualitativas nominales (categóricas) y covariable cuantitativa
Variables independientes: cualitativas nominales (categóricas)
Pruebas para el análisis de la correlación canónica
Pruebas para el análisis múltiple de covarianzas (MANCOVA)
Pruebas para el análisis múltiple de varianzas (MANOVA)


7C
¿Desea la reducción de datos partiendo de relaciones entre más de dos variables para explicar cierto fenómeno o hecho observado?
Sí
No
¿Emplea variables cuantitativas (paramétricas)?
Sí
No
¿Desea clasificar o identificar grupos homogéneos de sujetos a partir de su puntuación en una serie de variables medidas en la misma escala (agrupar observaciones)?
Sí
No
¿Identificar combinaciones lineales (componentes) de las variables iniciales que expliquen la mayor parte de la información contenida en esas variables?
¿Identificar dimensiones o variables latentes a partir de la relación entre las variables categóricas observadas?
Emplear Método de Análisis por Conglomerados (Cluster Analysis)
Sí
No
Sí
No
Sí
Emplear método de análisis de correspondencias
¿Desea clasificar o identificar relaciones entre objetos determinando las proximidades por las dimensiones de los mismos, según variables empleadas?
Emplear método de análisis de componentes principales
¿Identificar causas latentes o variables resumen (factores) que explican la correlación entre las variables observadas?
Sí
Sí
Emplear método de análisis factorial.
Emplear método de análisis de escalamiento multidimensional.

A continuación se presentan algunas situaciones en el orden del procesamiento estadístico de muestras y variables en que un mismo estadístico de prueba o contraste se puede emplear en diferentes tipos de contrastes de hipótesis según el rejuego que se haga entre muestras y variables en que al final se converge a igual cantidad de grupos de datos o muestras a comparar, todo depende del lado en que se sitúe el investigador según sus objetivos, a partir del número de muestras y el número y tipo de variables a emplear:

- Una muestra, dos variables (dependiente: cuantitativa con distribución normal, independiente: cualitativa dicotómica) o dos muestras, una variable cuantitativa con distribución normal (la misma en cada una), se emplea t-student para comparar dos grupos de datos cuantitativos.
- Una muestra, dos variables (dependiente: cuantitativa con distribución libre, independiente: cualitativa dicotómica) o dos muestras, una variable cuantitativa con distribución libre (la misma en cada una), se emplea la prueba de la U de Mann-Whitney o de la suma de rangos de Wilcoxon para comparar dos grupos de datos cuantitativos.
- Una muestra, dos variables (dependiente: cuantitativa con distribución normal, independiente: cualitativa politómica) o tres muestras independientes, una variable cuantitativa con distribución normal (la misma en cada una), se emplea la prueba F de Fisher-Snedecor por ANOVA de un factor o una vía para comparar más de dos grupos de datos cuantitativos.
- Una muestra, dos variables (dependiente: cuantitativa con distribución libre, Independiente: cualitativa politómica) o tres muestras independientes, una variable cuantitativa con distribución libre (la misma en cada una), se emplea la prueba de Kruskal-Wallis para comparar más de dos grupos de datos cuantitativos.
- Una muestra, dos variables cualitativas dicotómicas o dos muestras independientes, una variable cualitativa dicotómica (la misma en cada una), se emplea $Z \sim N(0, 1)$ para comparar dos proporciones.
- Una muestra, dos variables cualitativas dicotómicas o dos muestras independientes, una variable cualitativa dicotómica (la misma en cada una), se emplea la prueba chi-cuadrado de Pearson, en el primer caso para prueba de independencia de variables y en el segundo una prueba de homogeneidad para dos muestras.

Aunque en el capítulo 5 se describieron las características esenciales del modelo o paradigma bayesiano y se expusieron algunas de sus posibles aplicaciones en la inferencia bioestadística, la representación gráfica de sus elementos básicos de forma interrelacionada puede favorecer su comprensión y utilización adecuada en algunos tipos de investigaciones biomédicas a partir del manejo del par muestras y variables de forma paralela a como se abordó para el modelo frecuentista.

Se presenta a continuación un diagrama de flujo con algunas de las principales opciones y procedimientos a realizar en el procesamiento de la información, siguiendo el método o paradigma bayesiano en la inferencia bioestadística para una o dos pobla-

ciones y empleando parámetros como la media y la proporción como parámetros en correspondencia al tipo de variable seleccionada (Fig. 6.4).

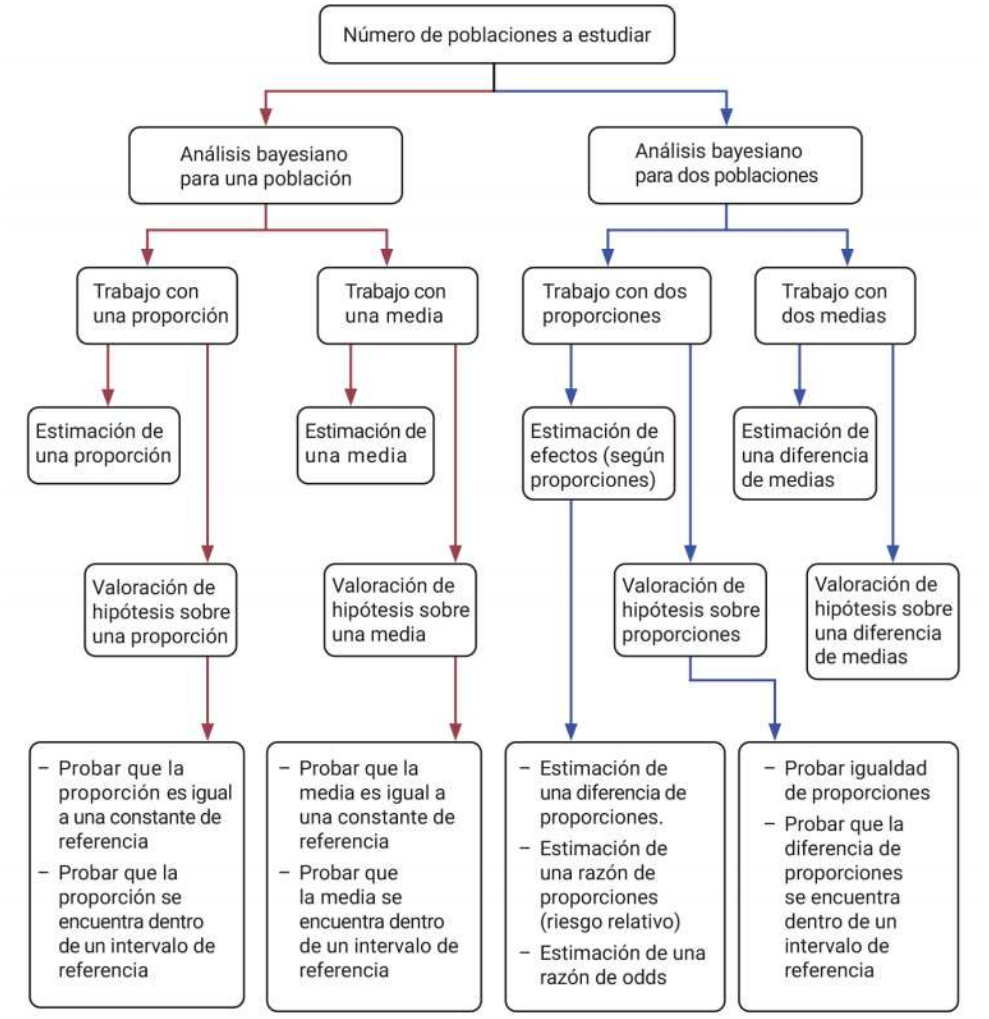


Fig. 6.4. Diagrama de flujo de los principales procedimientos bioestadísticos a realizar en el procesamiento de la información, siguiendo el método o paradigma bayesiano en la inferencia bioestadística.

Una vez que se cuenta con una visión ilustrada gráficamente de la manera de enfocarse cada modelo o paradigma en relación con el manejo de muestras y variables y las opciones de técnicas estadísticas a emplear en la inferencia bioestadística para cada uno, entonces se está en condiciones de mostrar de forma comparativa a través de un diagrama de flujo las posibilidades relacionadas con la inferencia bioestadística del enfoque frecuentista *versus* el bayesiano de forma general ante un problema de investigación determinado (Fig. 6.5).

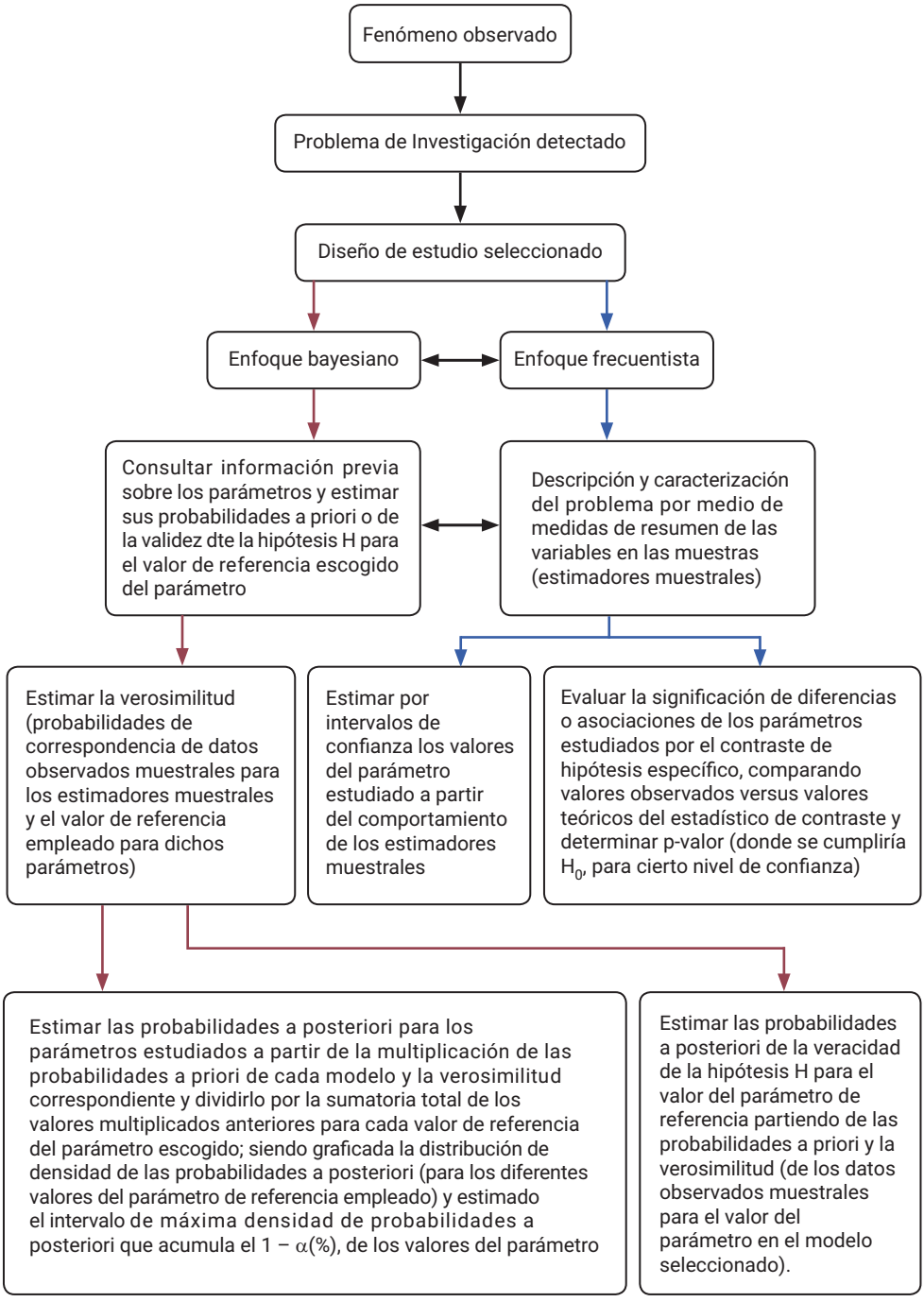


Fig. 6.5. Diagrama de flujo de las posibilidades en la inferencia bioestadística del enfoque frecuentista *versus* el bayesiano ante un problema de investigación.

Se representa la comparación de ambos modelos o paradigmas específicamente desde una perspectiva que comprende las opciones del procesamiento estadístico de los datos de una variable cuantitativa (métrica) o cualitativa (no métrica), según el número de muestras para ambas en la inferencia bioestadística (Figs. 6.6. y 6.7).

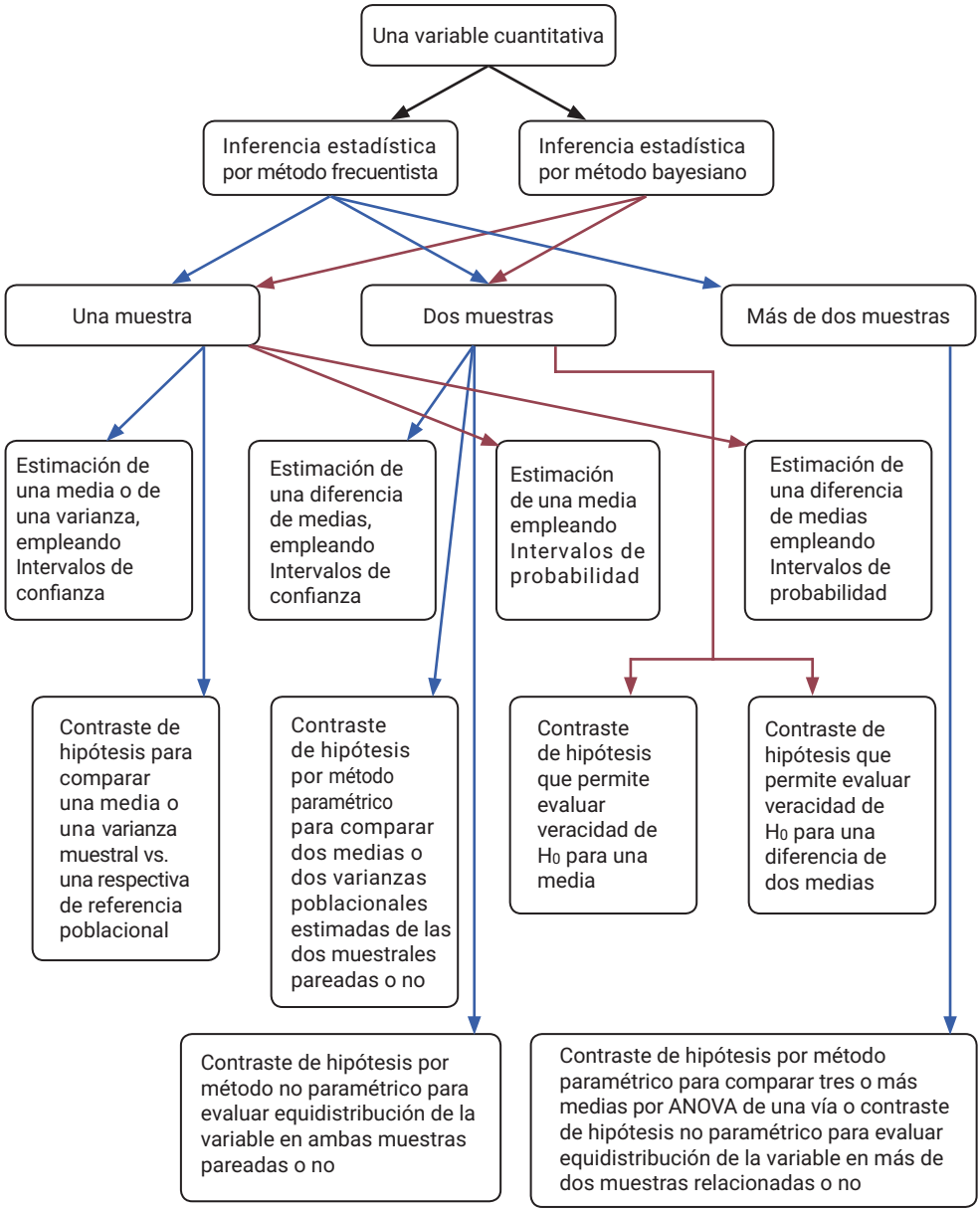


Fig. 6.6. Diagrama de las opciones en el procesamiento estadístico de una variable cuantitativa (métrica) según el número de muestras a partir de la comparación del enfoque frecuentista y el bayesiano.

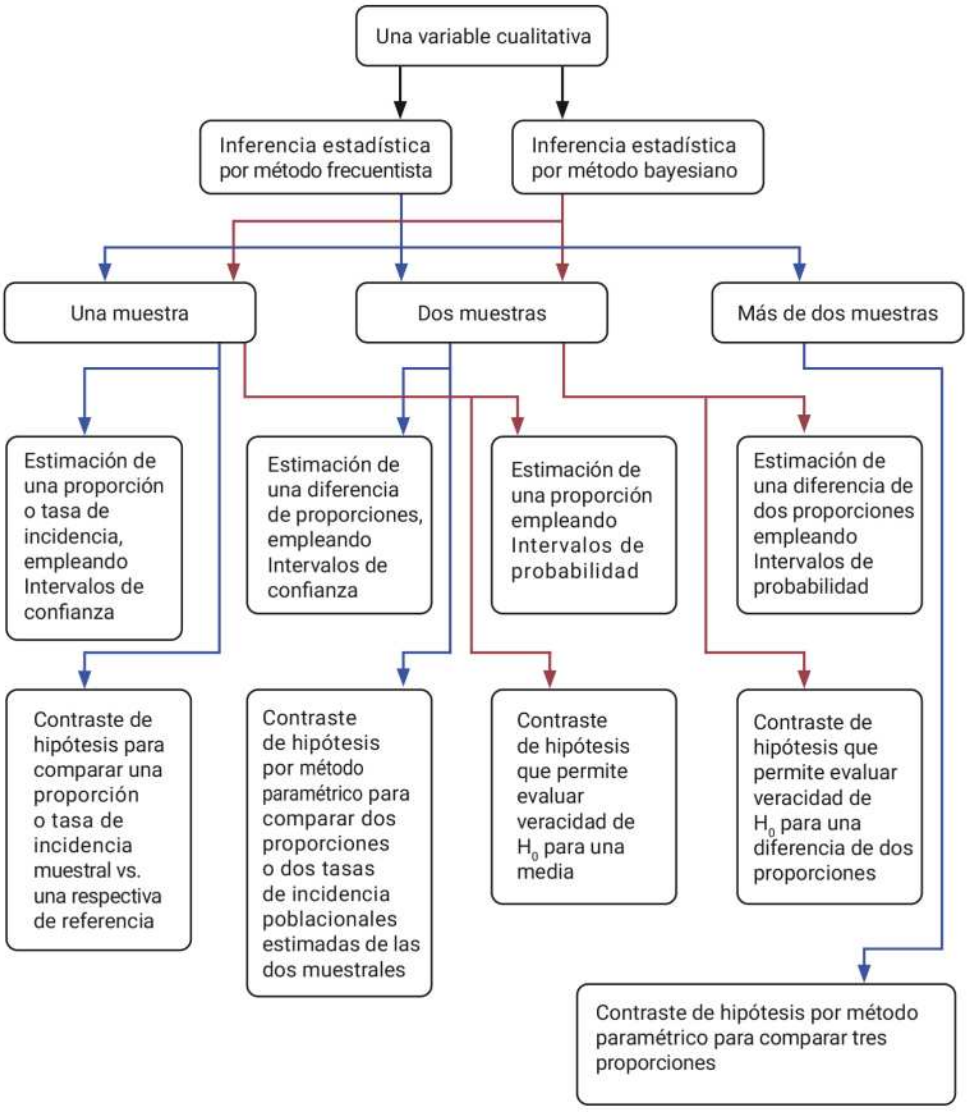


Fig. 6.7. Diagrama de las opciones en el procesamiento estadístico de una variable cualitativa (no métrica), según el número de muestras, a partir de la comparación del enfoque frecuentista y el bayesiano.

La utilización de medidas de tendencia central como la media aritmética de una variable cuantitativa, en diferentes números y tipos de muestras (relacionadas o no) es frecuentemente empleado en los proyectos investigativos epidemiológicos de tipo observacionales y experimentales. Algunas de las opciones a emplear como parte de la inferencia bioestadística para el parámetro: media poblacional en un número variable de muestras (Fig. 6.8).

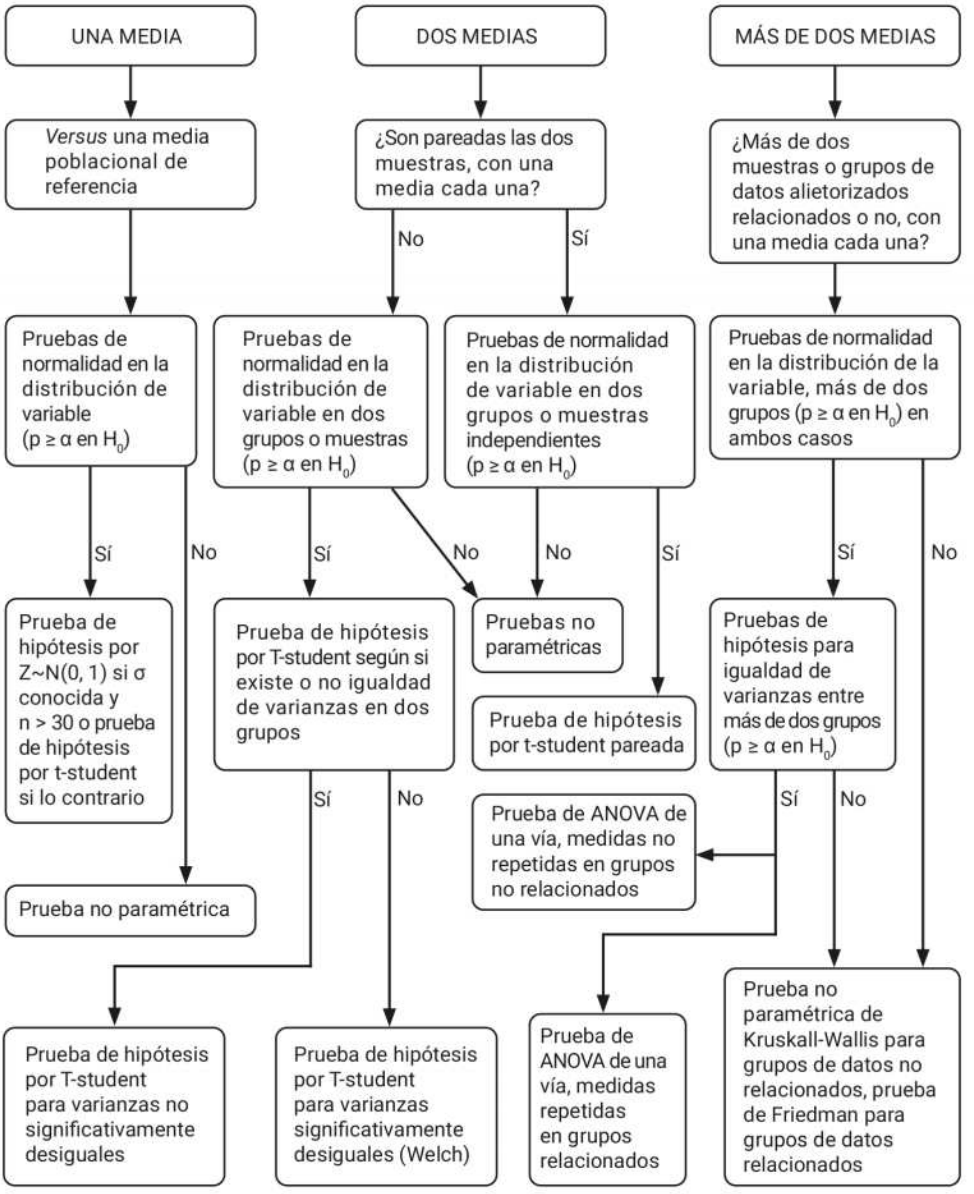


Fig. 6.8. Algoritmo a emplear al comparar medias aritméticas, entre grupos de datos determinados a partir de categorías de una variable o su equivalente en muestras.

A continuación se expresan algunas notas aclaratorias o ampliadas del algoritmo en la figura 6.8.

Para comprobar la asignación aleatoria de los individuos en la extracción de las muestras se emplea la prueba no paramétrica de rachas (tendencias) o prueba de independencia de las observaciones de una muestra.

Para comprobar la distribución normal de los datos en las muestras se emplean las pruebas estadísticas para el contraste de hipótesis relacionadas con la presencia de una distribución normal (normalidad) de las variables a estudiar, como son:

- De Shapiro-Wilk W.
- De Shapiro-Francia W.
- De D'Agostino.
- De Kolmogrov-Smirnov.
- De Anderson-Darling.
- De Lilliefors.

Para recordar

Cuando la variable analizada no sigue una distribución normal linealmente por los procedimientos de prueba de normalidad antes expresados se puede convertir los datos en su expresión logarítmica y en muchas ocasiones se logra la aproximación normal en su distribución y continuar el algoritmo propuesto, solo que los resultados se interpretan bajo esa condición logarítmica.

Para comprobar la homocedasticidad (similitud de varianzas) se emplean las pruebas estadísticas para el contraste de hipótesis respecto a las diferencias de varianzas:

- Para dos grupos de datos procedentes de dos muestras:
  - Prueba de F (Snedecor-Fisher) con estadístico:
    $F_{\text{observado-calculado}} = [S_1^2/S_2^2] \sim F_{(n-1,\, n-2;\, 1-\alpha)}$.
  - Prueba de Levene con estadístico:

$$W_{obs} = \frac{\dfrac{n_1(\bar{Z}_1 - \bar{Z})^2 + n_2(\bar{Z}_2 - \bar{Z})^2}{\sum_{j=1}^{n_1}(Z_{1j} - \bar{Z}_1)^2 + \sum_{j=1}^{n_1}(Z_{2j} - \bar{Z}_2)^2}}{n_1 + n_2 - 2} \sim F_{(1,\, n1 + n2 - 2;\, 1-\alpha).}$$

- Para más de dos grupos o muestras se tiene:
  - Prueba de Barlett con estadístico:

$$\chi^2_{obs} = \frac{1}{k}\left[(n-t)\ln \hat{S}^2_D - \sum_{i=1}^{t} \ln \hat{S}^2_i\right] \sim \chi^2_{t-1,1-\alpha}$$

  - Prueba de Cochran con estadístico:

$$R_{obs} = \frac{\text{máx}\left\{\hat{S}^2_i\right\}_{i=1}^{t}}{\sum_{i=1}^{t} \hat{S}^2_i} \sim R(\text{Cochran})_{n-1,t,1-\alpha}$$

En la comparación de medias de más de dos grupos de datos en la variable dependiente o muestras formadas se emplea el análisis de varianzas (ANOVA de una vía) con el estadístico para el cociente de varianzas entre grupos e intragrupos expresándose como:

F (observado-calculado) de Fisher-Snedecor = $S_E^2/S_D^2 \sim F_{(t-1\ o\ k-1,\, N-t\ o\ N-k).}$

A partir del número de categorías respectivas de cada variable cualitativa se pueden generar grupos de datos específicos para cada categoría, para los cuales se puede determinar

la homogeneidad o no en la distribución de los mismos desde el punto de vista inferencial mediante algunas opciones de pruebas de significación no paramétricas (Fig. 6.9).

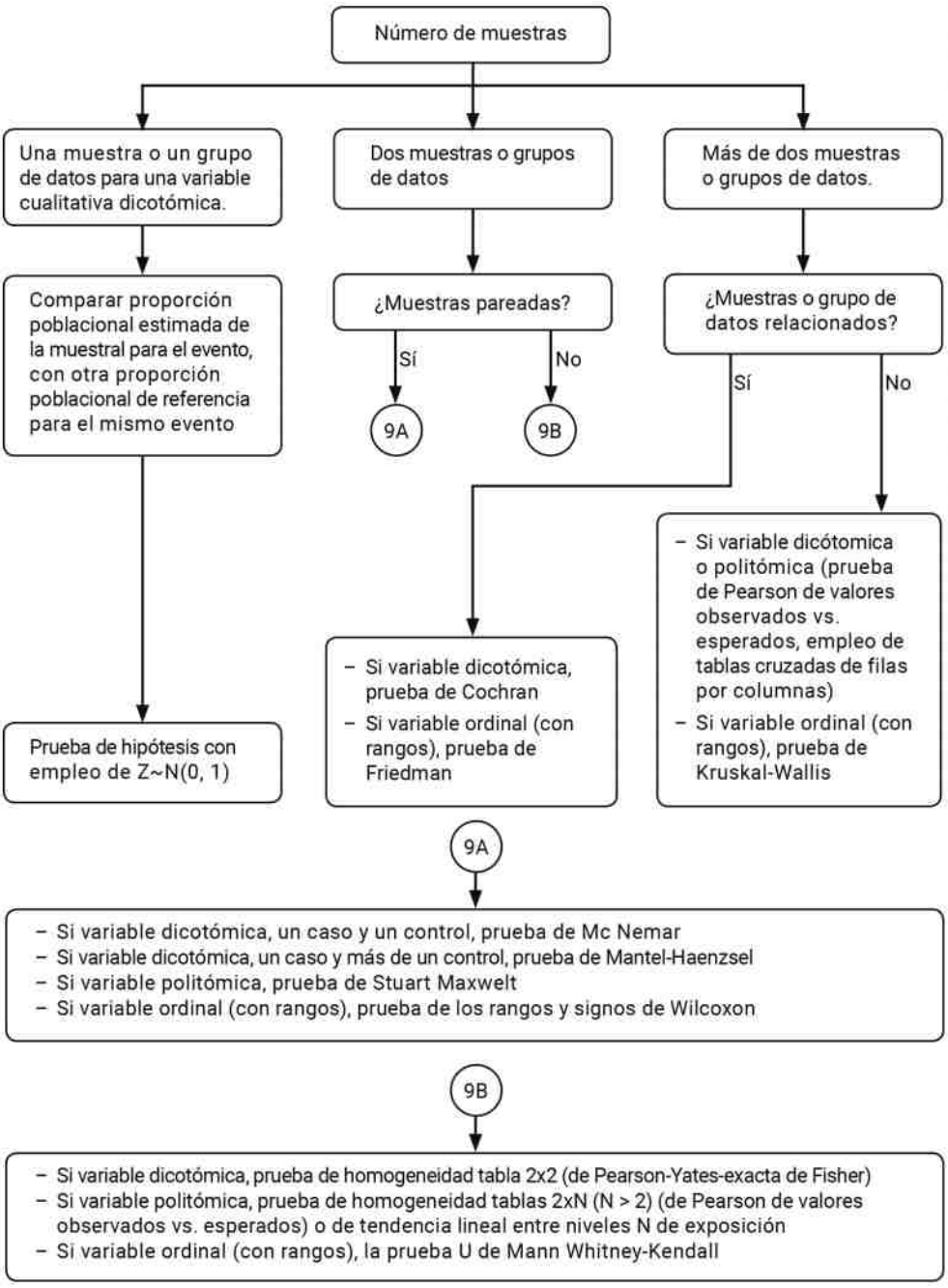


Fig. 6.9. Algoritmo para determinar la homogeneidad en la distribución de los datos entre los grupos o las muestras correspondientes para las distintas categorías, niveles u ordinalidad por rangos de los mismos, para diferentes tipos de variables.

Existen ocasiones en que se está con la necesidad investigativa de evaluar la relación entre dos variables de igual o diferente grupo clasificatorio en una muestra de estudio (métodos bivariados), con el uso de las respectivas opciones de técnicas estadísticas incluidas en las pruebas de significación paramétricas o no, que resultan específicas como parte de la inferencia bioestadística (Fig. 6.10).

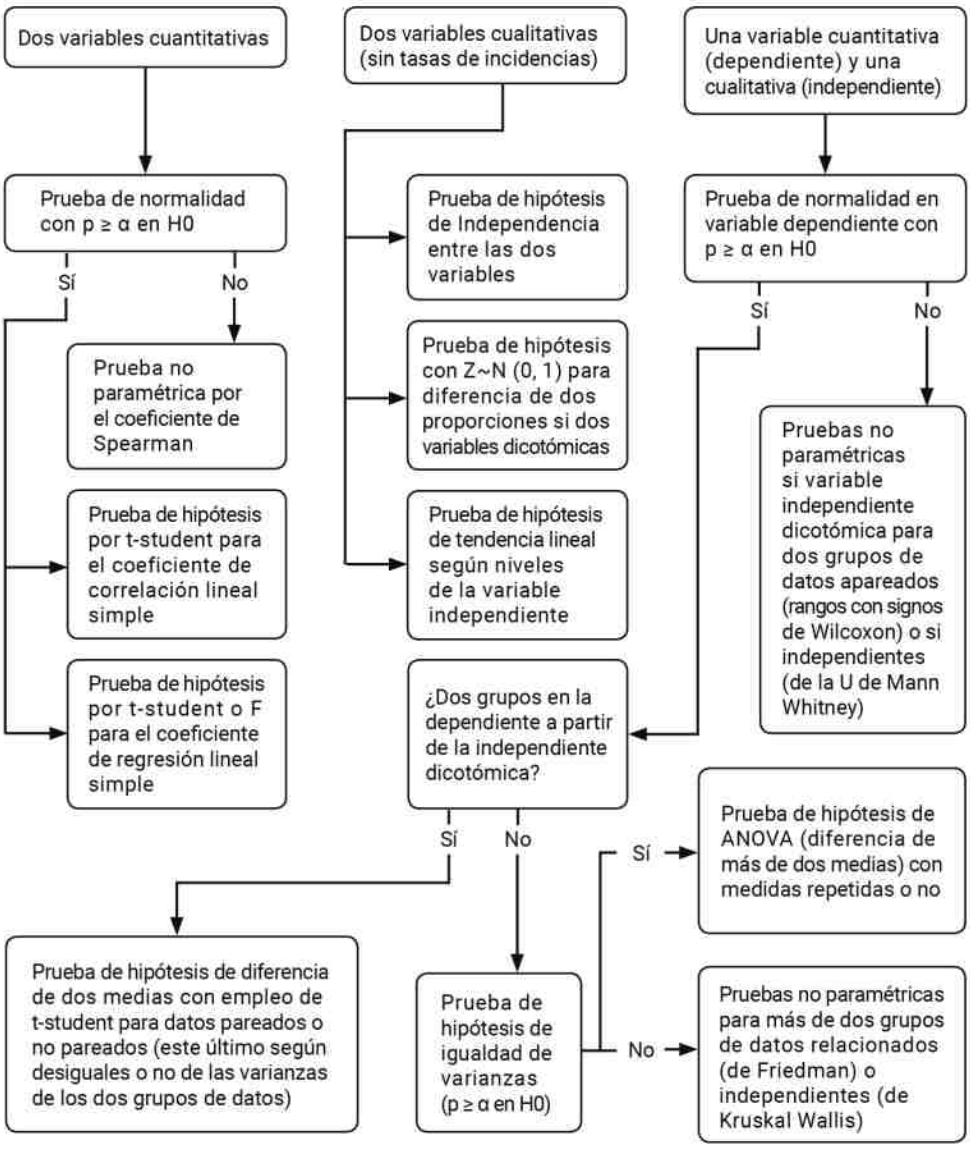


Fig. 6.10. Algoritmo para el manejo de dos variables en una muestra, con el empleo de pruebas de significación.

En algunas ocasiones se pueden encontrar en las investigaciones con variables que son de interés y que no siguen un tipo de distribución específica sus datos recogidos, es decir, que son de libre distribución en las muestras de estudio, lo que conlleva al empleo de pruebas no paramétricas solamente como parte del procesamiento estadístico en la inferencia bioestadística. Se representa un algoritmo que brinda algunas opciones en el manejo del número y tipo de muestras y variables, mediante técnicas no paramétricas (Fig. 6.11).

Los diseños de estudios epidemiológicos y entre estos específicamente los de tipo observacionales, cada vez más alcanzan mayor relevancia entre las investigaciones biomédicas en las épocas actuales en que se han puesto en evidencias verdaderas pandemias. Por lo que se debe tener en cuenta el enfoque de riesgo según exposición de esos diseños investigativos y su representación gráfica mediante tablas cruzadas o de contingencia simple o estratificada que emplean dos variables cualitativas sin y con covariable respectivamente, como parte del procesamiento estadístico de la información para la inferencia bioestadística (Fig. 6.12).

Al profundizar en algunos aspectos imprescindibles referentes a los diferentes tipos de diseños de estudios observacionales con un enfoque de riesgo según exposición, se debe recordar las respectivas medidas de resumen a emplear en íntima relación con su correspondiente tabla de contingencia a utilizar que favorece la presentación y análisis de los resultados (Fig. 6.13).

A pesar de que en el momento actual los procesadores estadísticos informáticos digitales son capaces de tener organizados internamente y soportar un gran número de procedimientos comprendidos o formando parte de la inferencia bioestadística que el investigador joven tiene a su alcance, en algunos procesadores se omiten algunos de ellos o en ocasiones se requiere tener bien agrupados y reconocidos los distintos tipos de contrastes de hipótesis o pruebas de significación (paramétricas y no paramétricas) con su respectiva aplicación para cada situación investigativa en particular, tomando en cuenta el objetivo investigativo y los requerimientos combinativos de muestras y variables, con las especificidades que ello conlleva implícito. Por esto nunca es ocioso presentar diversas formas de exponer la relación entre pruebas de significación y manejo de muestras y variables o viceversa como parte del procesamiento estadístico de la información (Fig. 6.20).

Como parte del procesamiento estadístico de la información, el investigador joven se puede enfrentar a la necesidad de emplear más de dos variables en una muestra de estudio, como parte del procesamiento estadístico de la información estadística a recoger, lo que se corresponde estadísticamente con el uso de los denominados *métodos multivariados* o *multivariantes en la inferencia bioestadística*. Las características generales de estos métodos y las bondades que brindan como parte de la inferencia bioestadística se describen en el capítulo 4. Para la exposición y ejemplificación gráfica de los diferentes tipos de métodos multivariados de dependencias y de interdependencias de las más de dos variables en la inferencia bioestadística, los investigadores han desarrollado múltiples esquemas algorítmicos con variantes y adaptaciones (Figs. 6.14 a la 6.18).

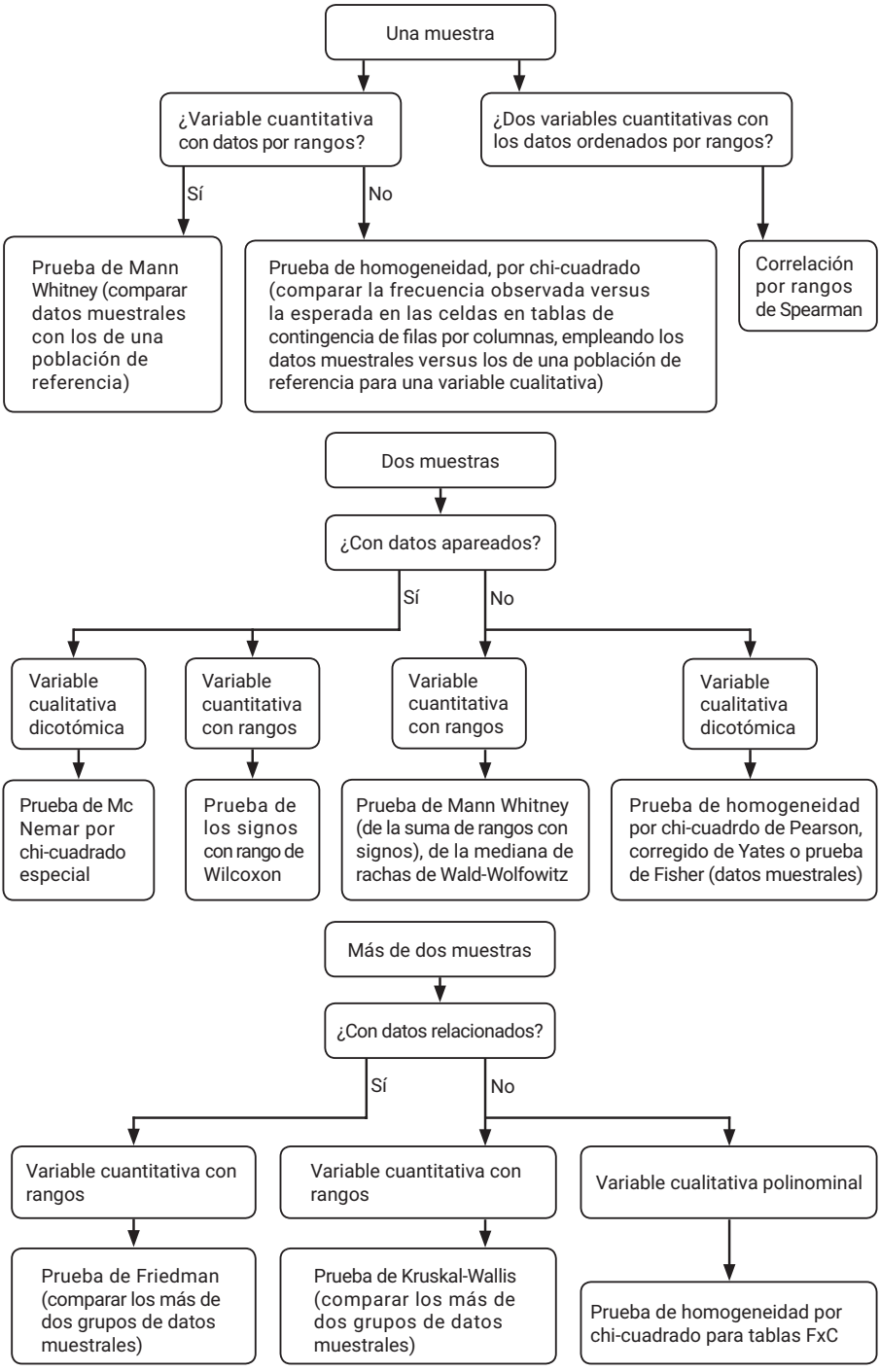


Fig. 6.11. Algoritmo para opciones en el manejo del número y tipo de muestras y variables, mediante técnicas no paramétricas.

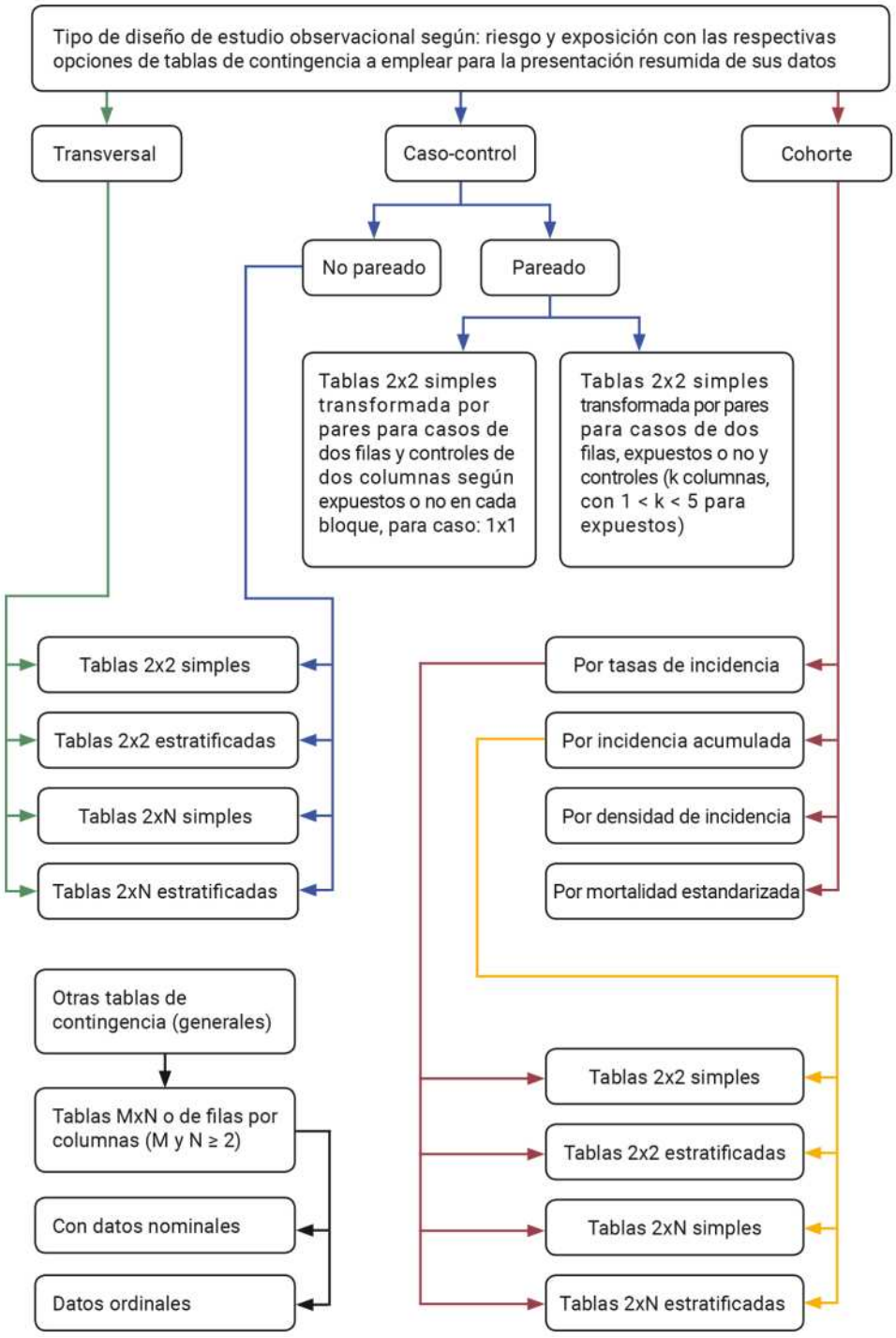


Fig. 6.12. Relación entre el tipo de diseño de estudio epidemiológico observacional, con un enfoque de riesgo según exposición y los tipos de tablas de contingencia a emplear.

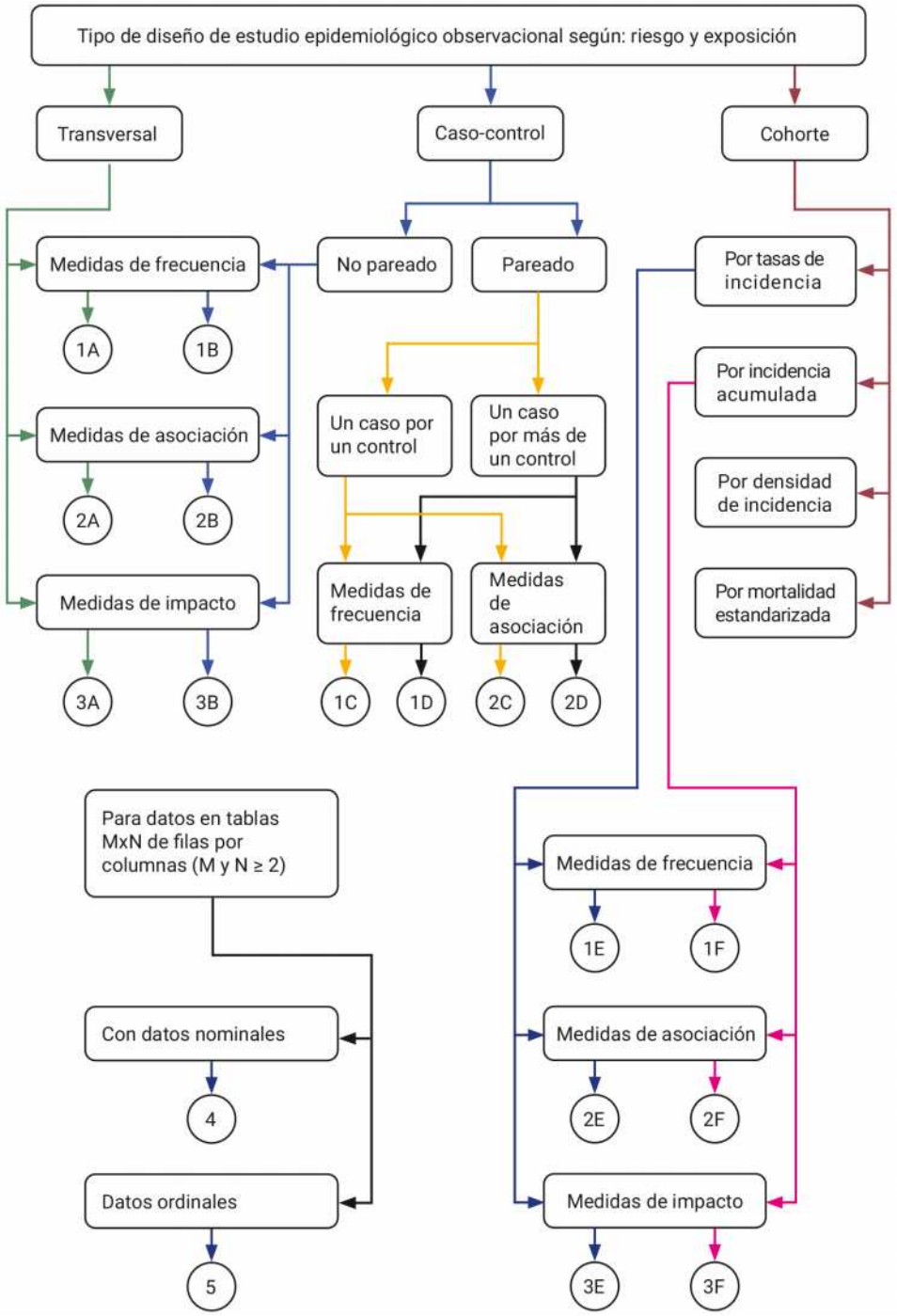


Fig. 6.13. Relación entre el tipo de diseño de estudio epidemiológico observacional con un enfoque de riesgo según exposición y las respectivas medidas de resumen, según sus tablas de contingencia.

1A

Si tabla de contingencia 2x2 simple:

Para la prevalencia de enfermedad:

- Prevalencia en expuestos y no expuestos,
- Razón de prevalencias: valor puntual y por intervalo de confianza.

Para la prevalencia de exposición:

- Prevalencia en enfermos y no enfermos,
- Razón de prevalencias: valor puntual y por intervalo de confianza.

Si tabla de contingencia 2x2 estratificada:

Razón de prevalencias de exposición: valor puntual y por intervalo de confianza:

- Para cada estrato.
- Cruda (tabla global).
- Combinada (de Mantel-Henszel).
- Ponderada.

Razón de prevalencias de enfermedad: valor puntual y por intervalo de confianza:

- Para cada estrato.
- Cruda (tabla global).
- Combinada (de Mantel-Haenszel).
- Ponderada.

Si tabla de contingencia 2xN simple:

- Prevalencia para cada nivel de exposición.
- Para cada nivel frente al de referencia: razón de prevalencias e intervalo de confianza.
- Razón de prevalencias general: valor puntual y por intervalo de confianza,

Si tabla de contingencia 2xN estratificada:

- Para cada nivel frente al de referencia: razón de prevalencias cruda: valor puntual y por intervalo de confianza.
- Razón de prevalencias general cruda: valor puntual y por intervalo de confianza.,
- Para cada nivel frente al de referencia, razón de prevalencias combinada (de Mantel-Haenzsel): valor puntual y por intervalo de confianza.
- Razón de prevalencias general ponderada: valor puntual y por intervalo de confianza.

2A

Si tabla de contingencia 2x2 simple:

Odds ratio: valor puntual y por intervalo de confianza.

Si tabla de contingencia 2x2 estratificada:

Odds ratio: valor puntual y por intervalo de confianza:

- Para cada estrato.
- Cruda (tabla global).
- Combinada (de Mantel-Haenzsel).
- Ponderada.

3A

Si tabla de contingencia 2x2 simple:

En relación a la prevalencia de enfermedad:

- Fracción atribuible o de prevención en los expuestos: valor puntual y por intervalo de confianza.
- Fracción atribuible o de prevención poblacional: valor puntual y por intervalo de confianza.

En relación a la prevalencia de exposición:

- Fracción atribuible o de prevención en los enfermos: valor puntual y por intervalo de confianza.
- Fracción atribuible o de prevención poblacional: valor puntual y por intervalo de confianza.

En estudios caso-control (no pareado)

1B

Si tabla de contingencia 2x2 simple:

Proporción de expuestos, en casos y en controles

2B

Si tabla de contingencia 2x2 simple:

Odds ratio: valor puntual y por intervalo de confianza.

Si tabla de contingencia 2x2 estratificada:

Odds ratio: valor puntual y por intervalo de confianza:

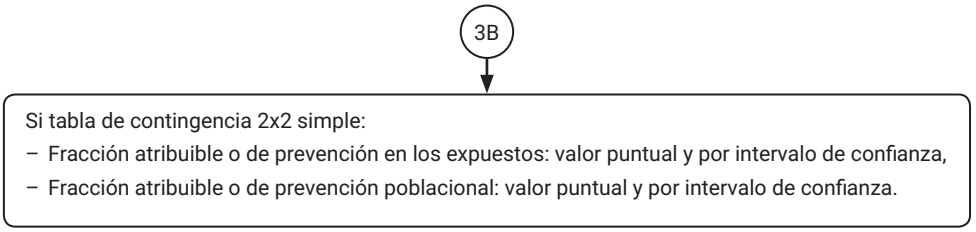
- Para cada estrato.
- Cruda (tabla global).
- Combinada (de Mantel-Haenzsel).
- Ponderada.

Si tabla de contingencia 2xN simple:

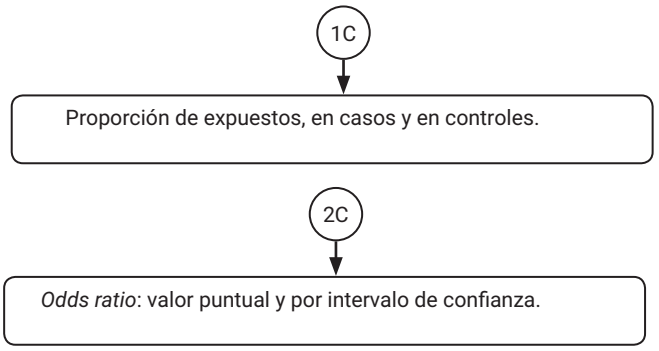
- *Odds* para cada nivel de exposición: valor puntual y por intervalo de confianza.
- Para cada nivel frente al de referencia: odds ratio, su valor puntual y por intervalo de confianza
- *Odds ratio* general: valor puntual y por intervalo de confianza,

Si tabla de contingencia 2xN estratificada:

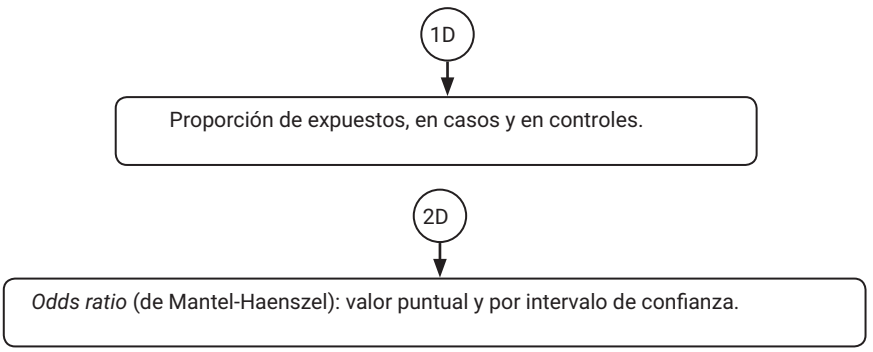
- Para cada nivel frente al de referencia: odds ratio cruda, su valor puntual y por intervalo de confianza.
- Odds ratio general cruda: valor puntual y por intervalo de confianza.
- Para cada nivel frente al de referencia, odds ratio combinada (de Mantel-Haenzsel), su valor puntual y por intervalo de confianza.
- Odds ratio general ponderada: valor puntual y por intervalo de confianza.



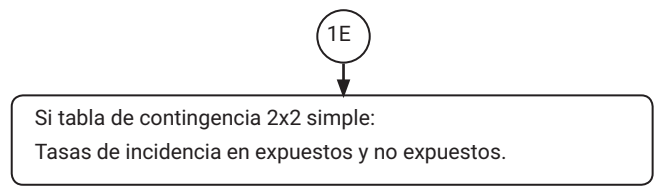
En estudios de caso-control (pareado), con 1 caso por 1 control



En estudios de caso-control (pareado), con 1 caso y más de 1 control



Estudios de cohorte (por datos de tasas de incidencia)



2E

Si tabla de contingencia 2x2 simple:

- Razón de tasas de incidencia: valor puntual y por intervalo de confianza.
- Diferencia de tasas de incidencia: valor puntual y por intervalo de confianza.

Si tabla de contingencia 2x2 estratificada:

Razón de tasas de incidencia, su valor puntual y por intervalo de confianza:

- Para cada estrato.
- Cruda (tabla global).
- Combinada (de Mantel-Haenzsel).

Diferencia de tasas de incidencia, su valor puntual y por intervalo de confianza:

- Para cada estrato.
- Cruda (tabla global).
- Combinada (de Mantel-Haenzsel).

Si tabla de contingencia 2xN simple:

- Tasa de incidencia para cada nivel de exposición.
- Para cada nivel frente al de referencia: razón de tasas de incidencia, su valor puntual y por intervalo de confianza.

Si tabla de contingencia 2xN estratificada:

- Para cada nivel frente al de referencia: razón de tasas de incidencia cruda, su valor puntual y por intervalo de confianza.
- Para cada nivel frente al de referencia: razón de tasas de incidencia combinada (de Mantel-Haenzsel), su valor puntual y por intervalo de confianza.

3E

Si tabla de contingencia 2x2 simple:

- Fracción atribuible en los expuestos, su valor puntual y por intervalo de confianza.
- Fracción atribuible y de prevención poblacional, su valor puntual y por intervalo de confianza.

En estudios de cohortes (por datos por Incidencia acumulada)

1F

Si tabla de contingencia 2x2 simple:

Proporción de nuevos enfermos en expuestos y no expuestos.

Si tabla de contingencia 2xN simple:

Riesgo para cada nivel de exposición

2F

Si tabla de contingencia 2x2 simple:

- Riesgo relativo: valor puntual y por intervalo de confianza.
- Diferencia de riesgos: valor puntual y por intervalo de confianza.
- *Odds ratio*: valor puntual y por intervalo de confianza.

Si tabla de contingencia 2x2 estratificada:

Riesgo relativo: valor puntual y por intervalo de confianza:

- Para cada estrato.
- Crudo (tabla global).
- Combinado (de Mantel-Haenzsel).
- Ponderado.

Diferencia de riesgos: valor puntual y por intervalo de confianza:

- Para cada estrato.
- Cruda (tabla global).
- Combinada (de Mantel-Haenzsel).
- Ponderada.

Odds ratio: valor puntual y por intervalo de confianza:

- Para cada estrato,
- Cruda (tabla global),
- Combinada (de Mantel-Haenzsel),
- Ponderada.

Si tabla de contingencia 2xN simple:

- Para cada nivel frente al de referencia: Riesgo relativo, su valor puntual y por intervalo de confianza.
- Riesgo relativo general: valor puntual y por intervalo de confianza.

Si tabla de contingencia 2xN estratificada:

- Para cada nivel frente al de referencia: riesgo relativo crudo, su valor puntual y por intervalo de confianza.
- Riesgo relativo general crudo: valor puntual y por intervalo de confianza.
- Para cada nivel frente al de referencia, Riesgo relativo combinado (de Mantel-Haenzsel), su valor puntual y por intervalo de confianza.
- Riesgo relativo general ponderado, su valor puntual y por intervalo de confianza.

3F

Si tabla de contingencia 2x2 simple:

- Fracción atribuible o de prevención en los expuestos e intervalo de confianza.
- Fracción atribuible o de prevención poblacional e intervalo de confianza.

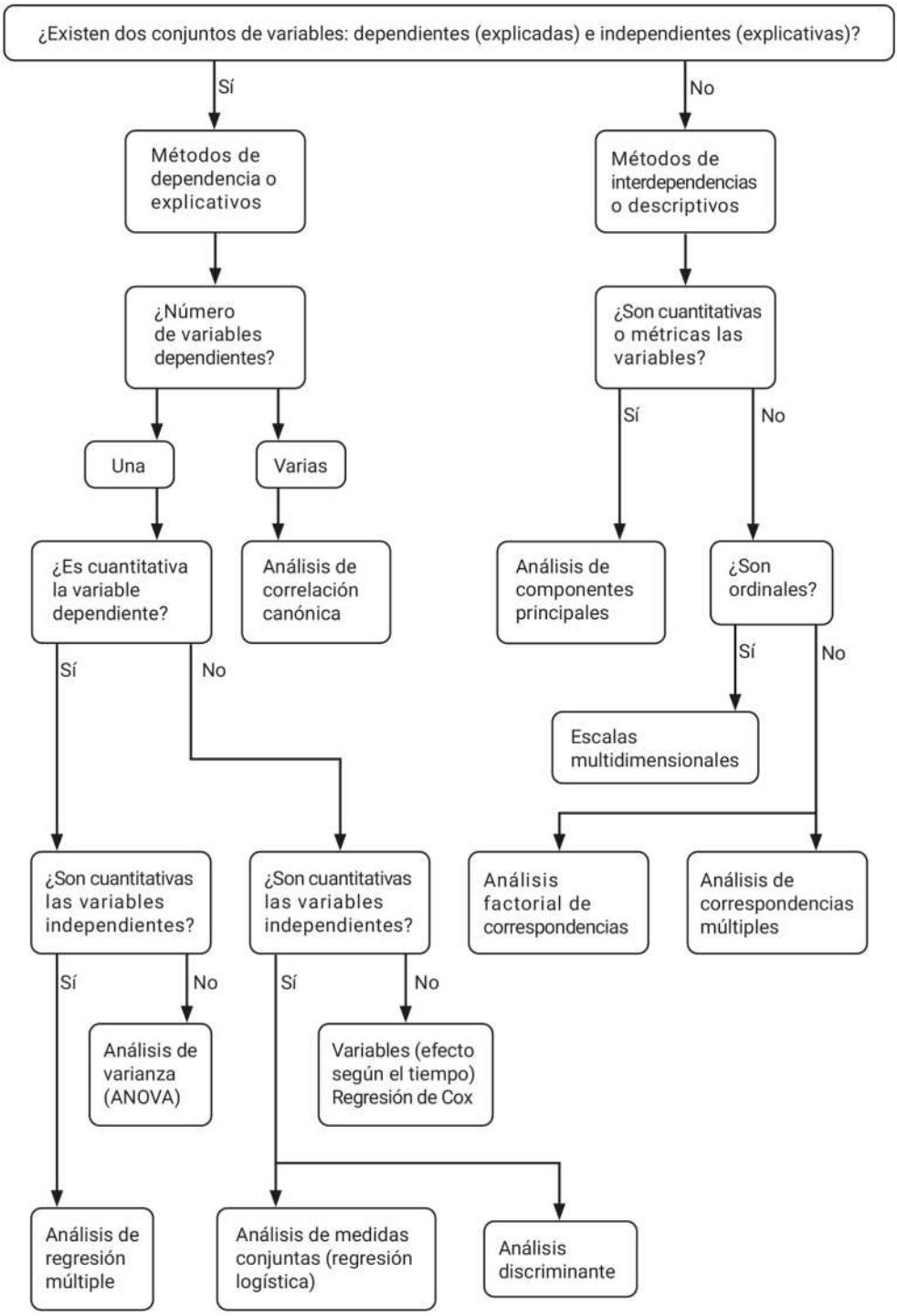


Fig. 6.14. Algoritmo para la selección de métodos o técnicas multivariantes (1).

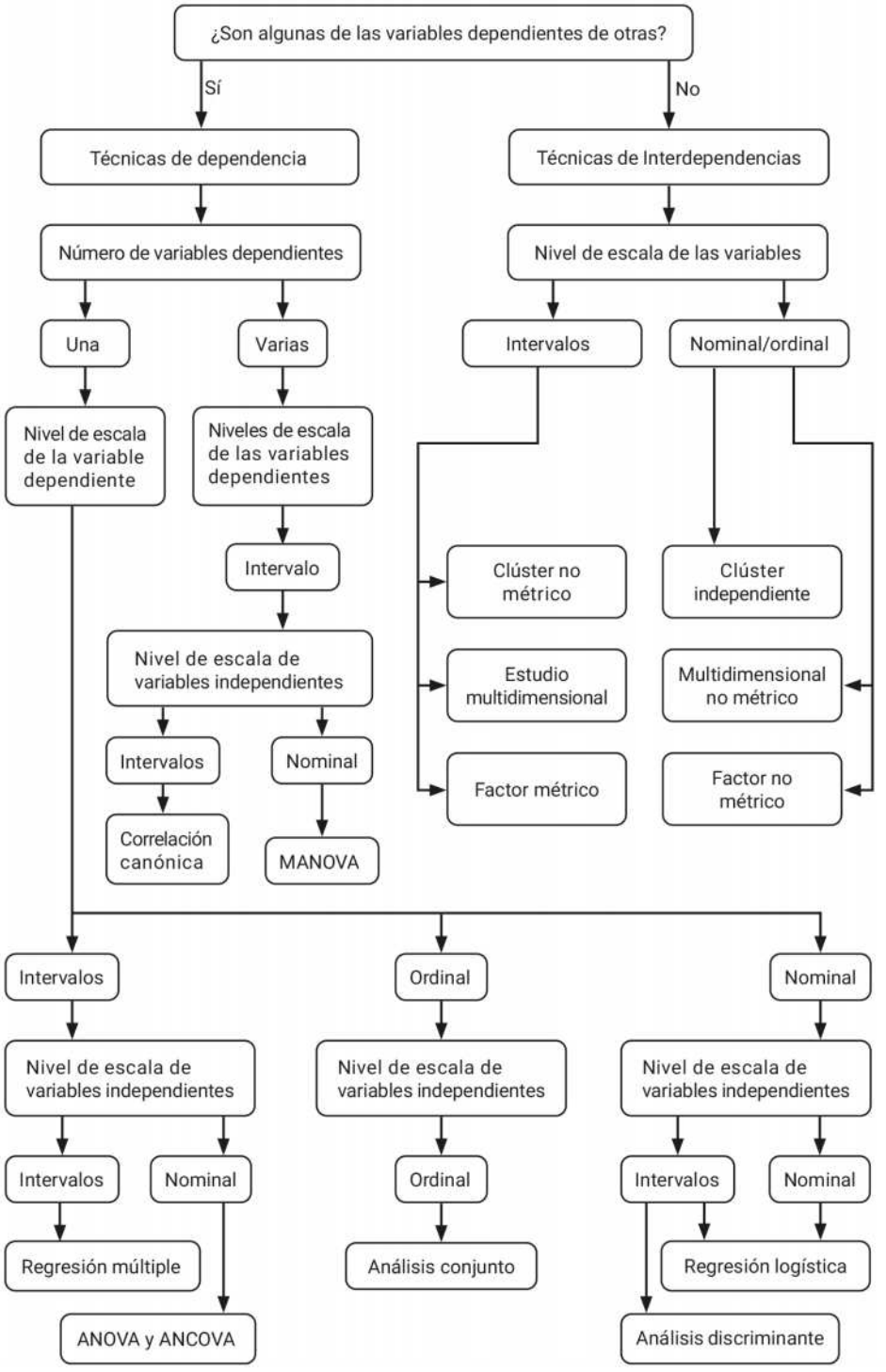


Fig. 6.15. Algoritmo para la selección de métodos o técnicas multivariantes (2).

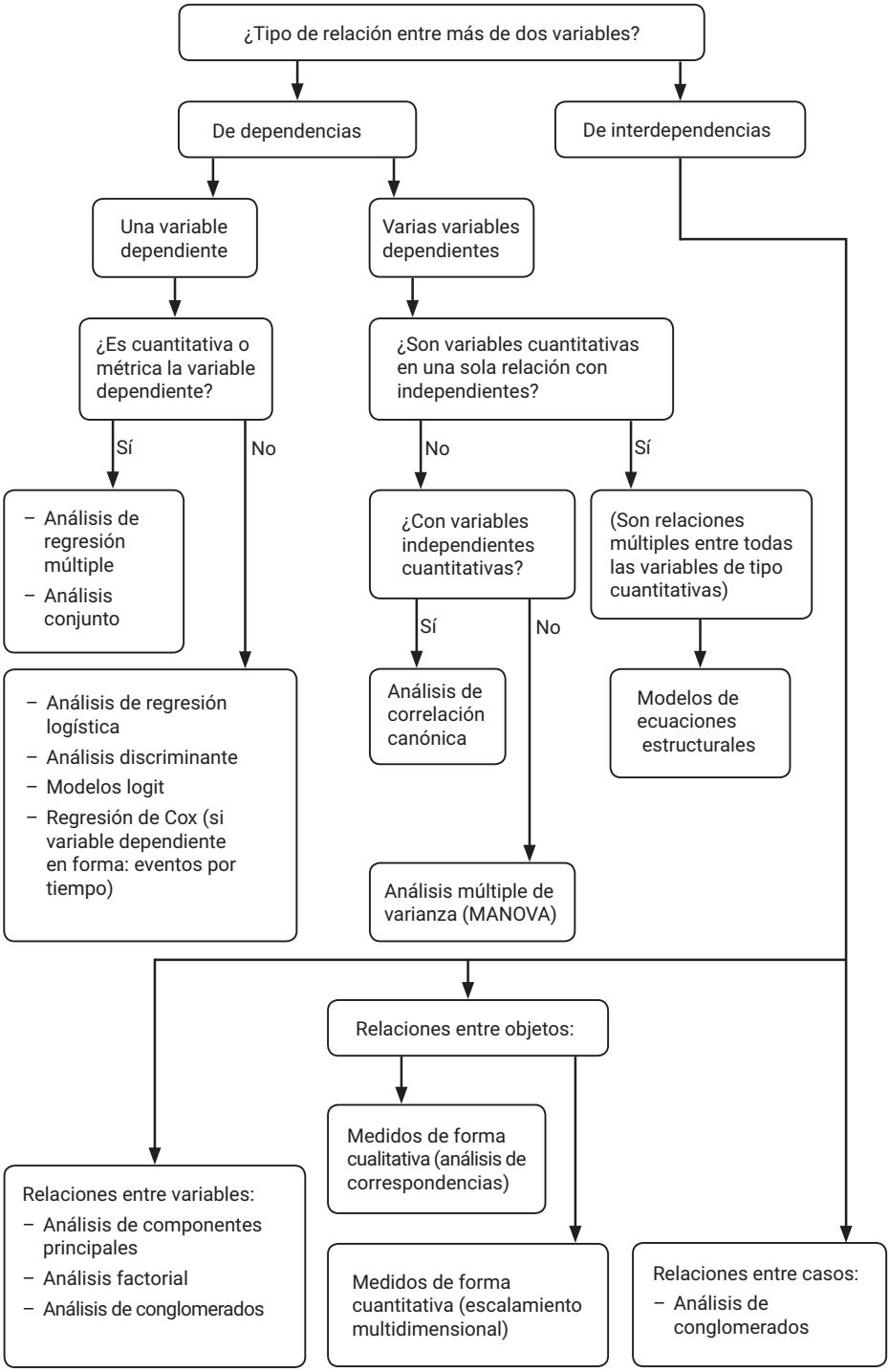


Fig. 6.16. Algoritmo para la selección de métodos o técnicas multivariantes (3).

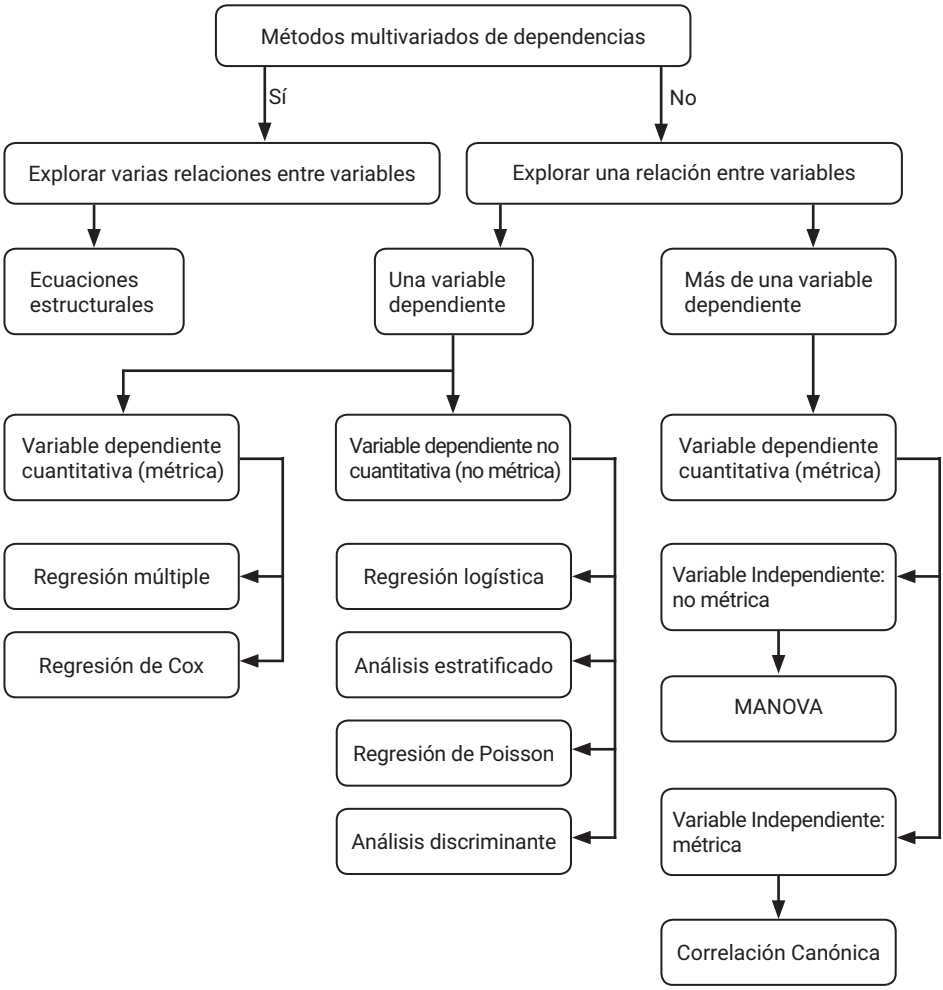


Fig. 6.17. Algoritmo para la selección de métodos o técnicas multivariantes (4).

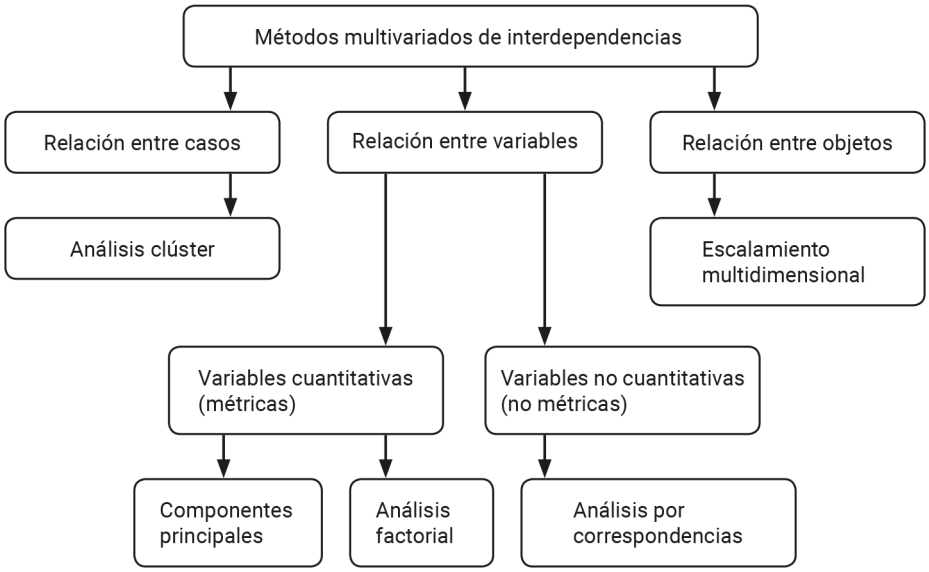
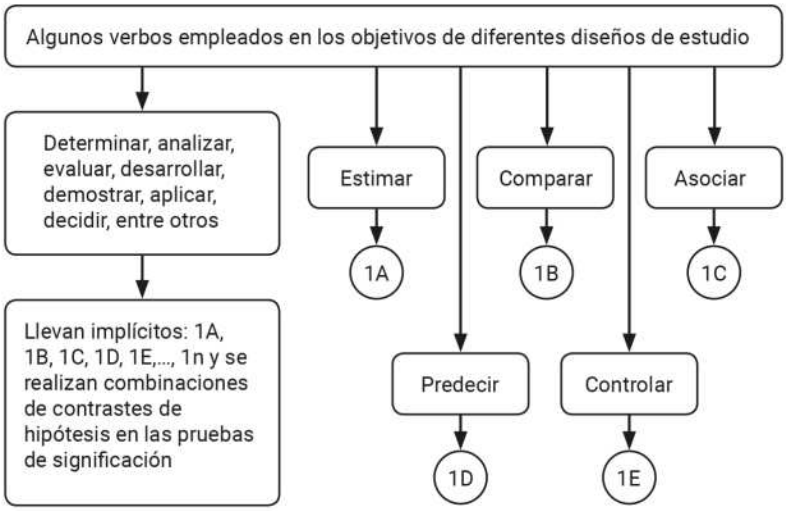
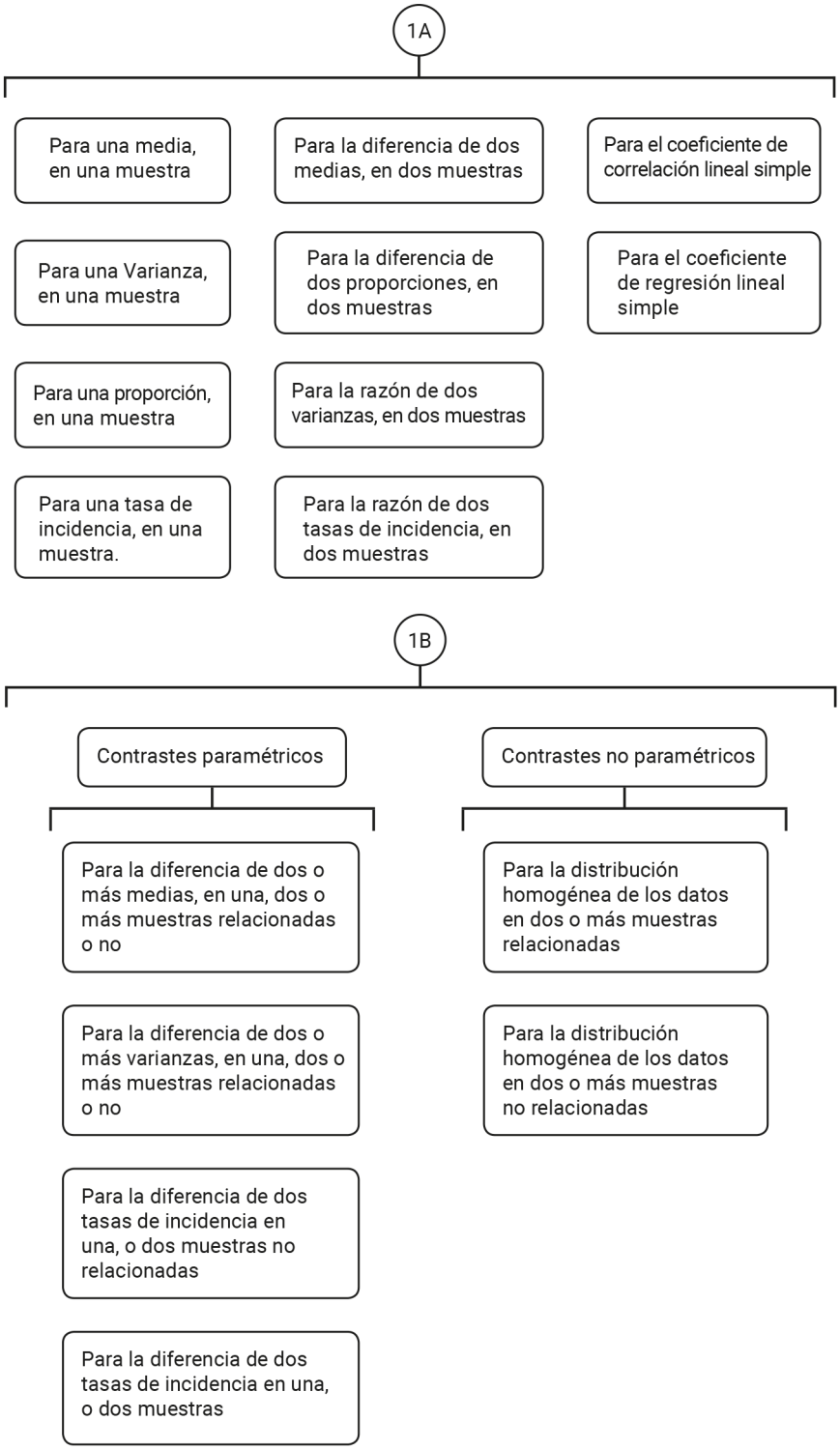
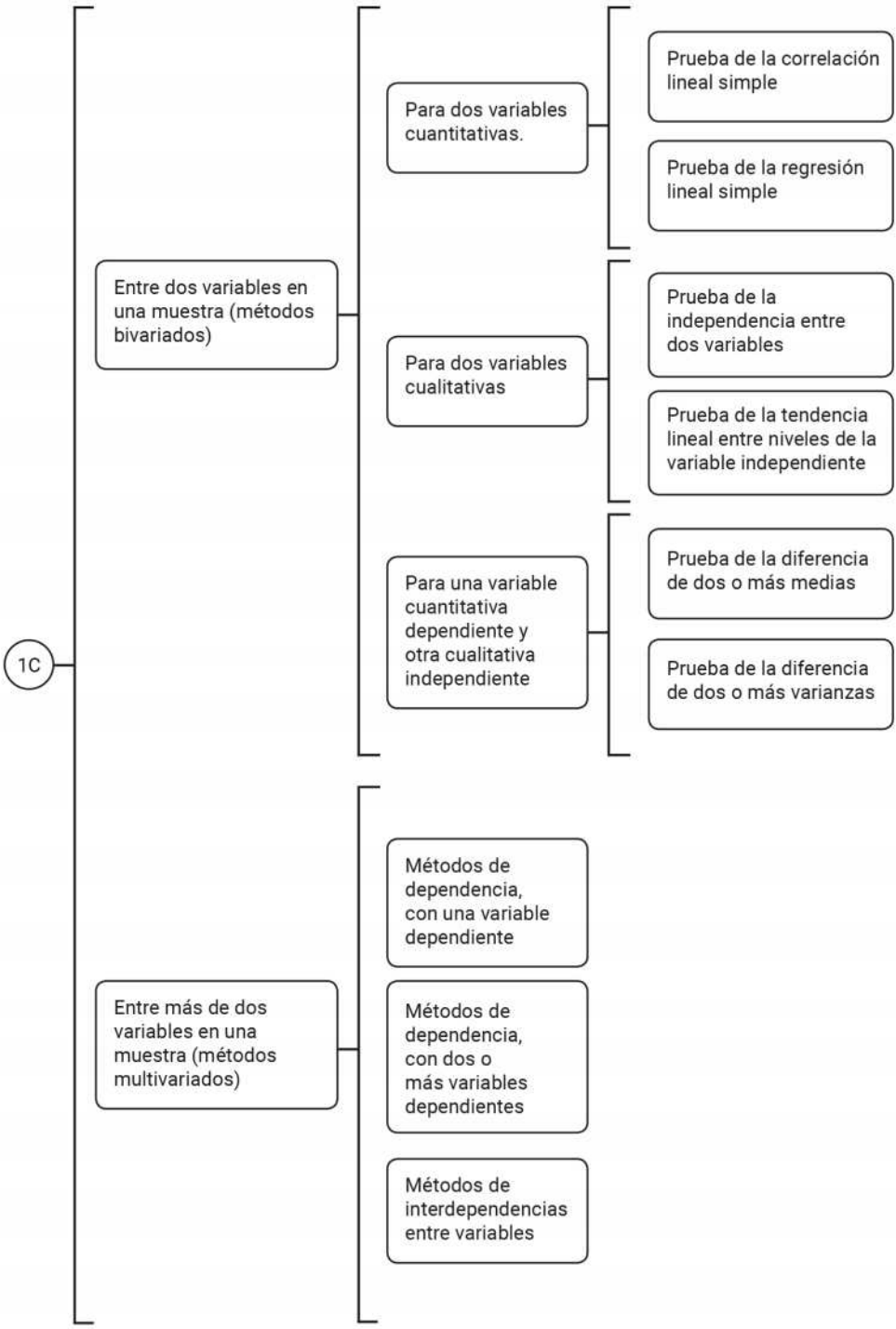
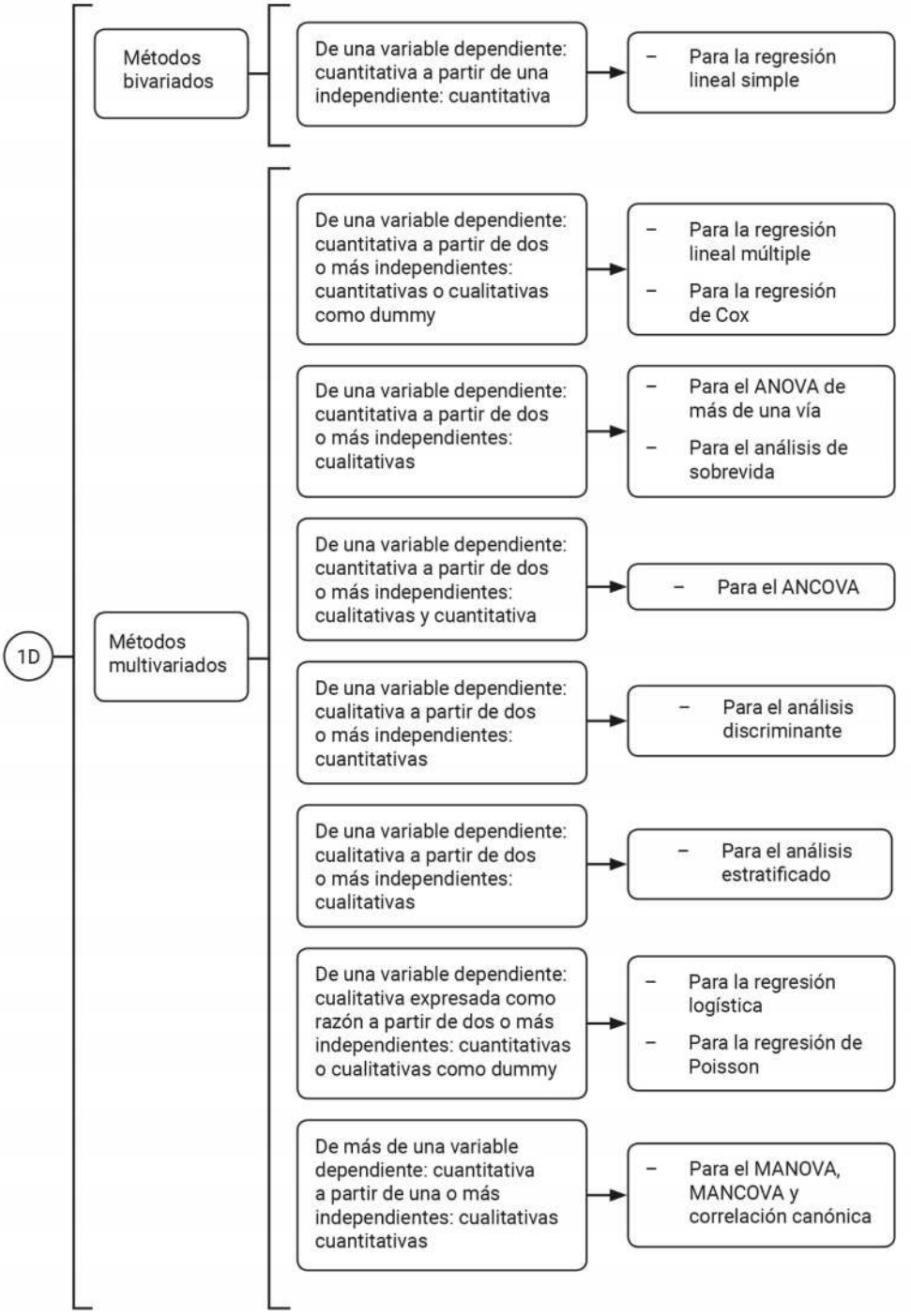


Fig. 6.18. Algoritmo para la selección de métodos o técnicas multivariantes (5).




1A
Para una media, en una muestra
Para la diferencia de dos medias, en dos muestras
Para el coeficiente de correlación lineal simple
Para una Varianza, en una muestra
Para la diferencia de dos proporciones, en dos muestras
Para el coeficiente de regresión lineal simple
Para una proporción, en una muestra
Para la razón de dos varianzas, en dos muestras
Para una tasa de incidencia, en una muestra.
Para la razón de dos tasas de incidencia, en dos muestras
1B
Contrastes paramétricos
Contrastes no paramétricos
Para la diferencia de dos o más medias, en una, dos o más muestras relacionadas o no
Para la distribución homogénea de los datos en dos o más muestras relacionadas
Para la diferencia de dos o más varianzas, en una, dos o más muestras relacionadas o no
Para la distribución homogénea de los datos en dos o más muestras no relacionadas
Para la diferencia de dos tasas de incidencia en una, o dos muestras no relacionadas
Para la diferencia de dos tasas de incidencia en una, o dos muestras


1C
Entre dos variables en una muestra (métodos bivariados)
Para dos variables cuantitativas.
Prueba de la correlación lineal simple
Prueba de la regresión lineal simple
Para dos variables cualitativas
Prueba de la independencia entre dos variables
Prueba de la tendencia lineal entre niveles de la variable independiente
Para una variable cuantitativa dependiente y otra cualitativa independiente
Prueba de la diferencia de dos o más medias
Prueba de la diferencia de dos o más varianzas
Entre más de dos variables en una muestra (métodos multivariados)
Métodos de dependencia, con una variable dependiente
Métodos de dependencia, con dos o más variables dependientes
Métodos de interdependencias entre variables


1D
Métodos bivariados
De una variable dependiente: cuantitativa a partir de una independiente: cuantitativa
– Para la regresión lineal simple
Métodos multivariados
De una variable dependiente: cuantitativa a partir de dos o más independientes: cuantitativas o cualitativas como dummy
– Para la regresión lineal múltiple
– Para la regresión de Cox
De una variable dependiente: cuantitativa a partir de dos o más independientes: cualitativas
– Para el ANOVA de más de una vía
– Para el análisis de sobrevida
De una variable dependiente: cuantitativa a partir de dos o más independientes: cualitativas y cuantitativa
– Para el ANCOVA
De una variable dependiente: cualitativa a partir de dos o más independientes: cuantitativas
– Para el análisis discriminante
De una variable dependiente: cualitativa a partir de dos o más independientes: cualitativas
– Para el análisis estratificado
De una variable dependiente: cualitativa expresada como razón a partir de dos o más independientes: cuantitativas o cualitativas como dummy
– Para la regresión logística
– Para la regresión de Poisson
De más de una variable dependiente: cuantitativa a partir de una o más independientes: cualitativas cuantitativas
– Para el MANOVA, MANCOVA y correlación canónica

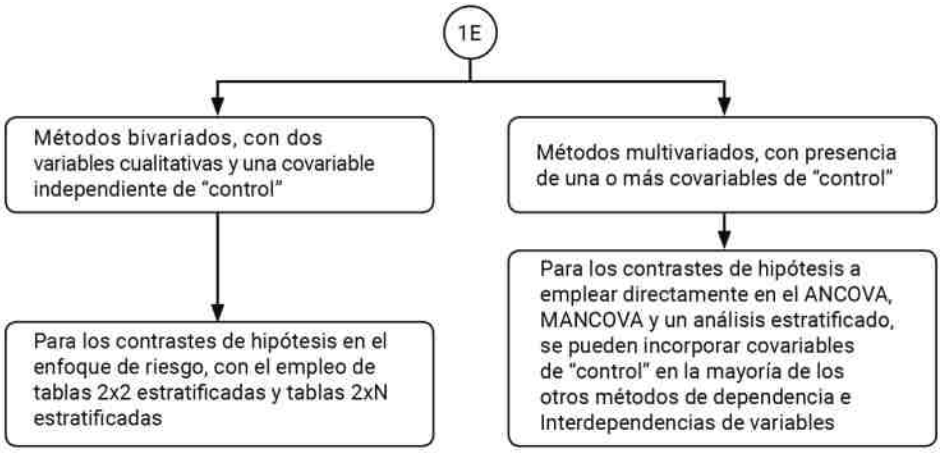
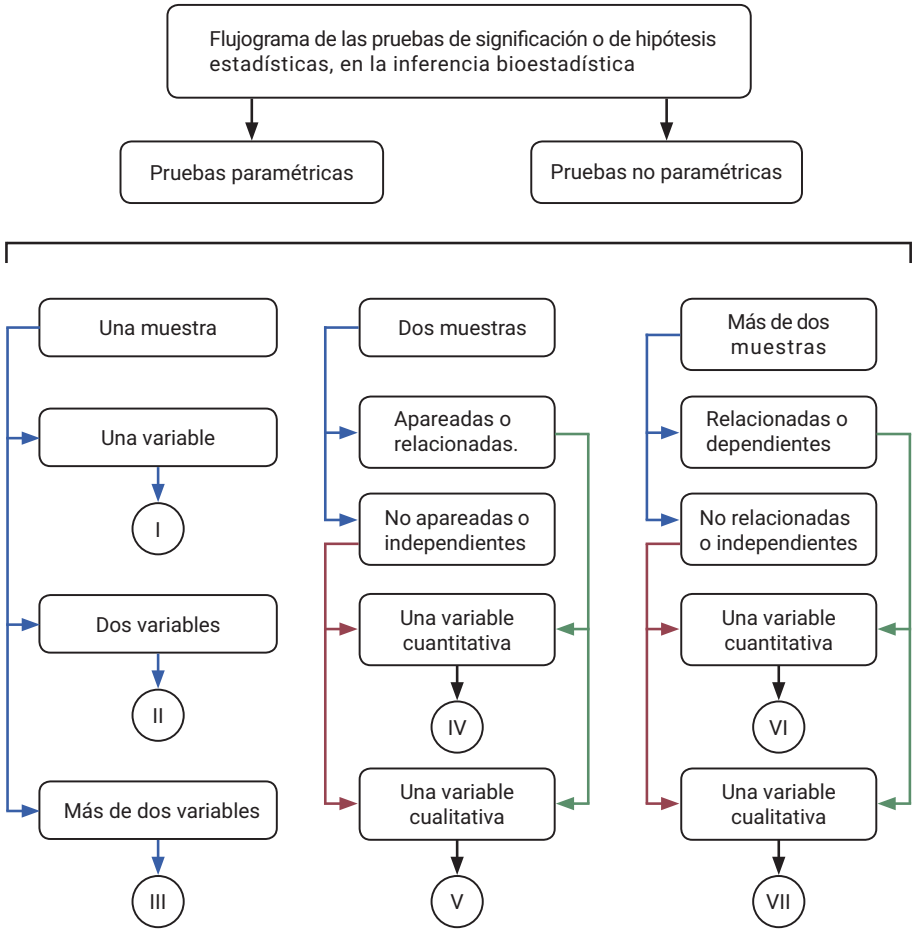
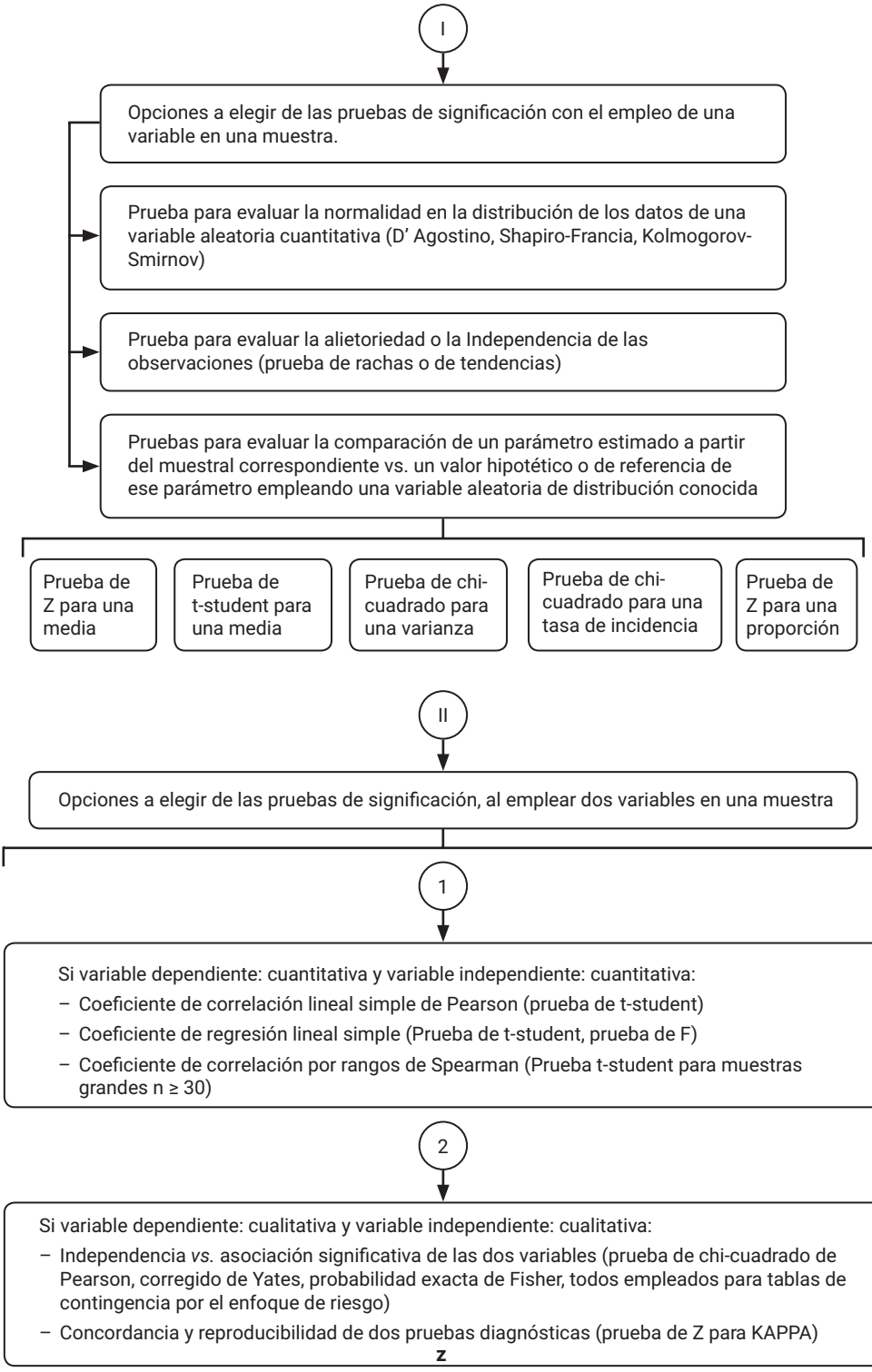
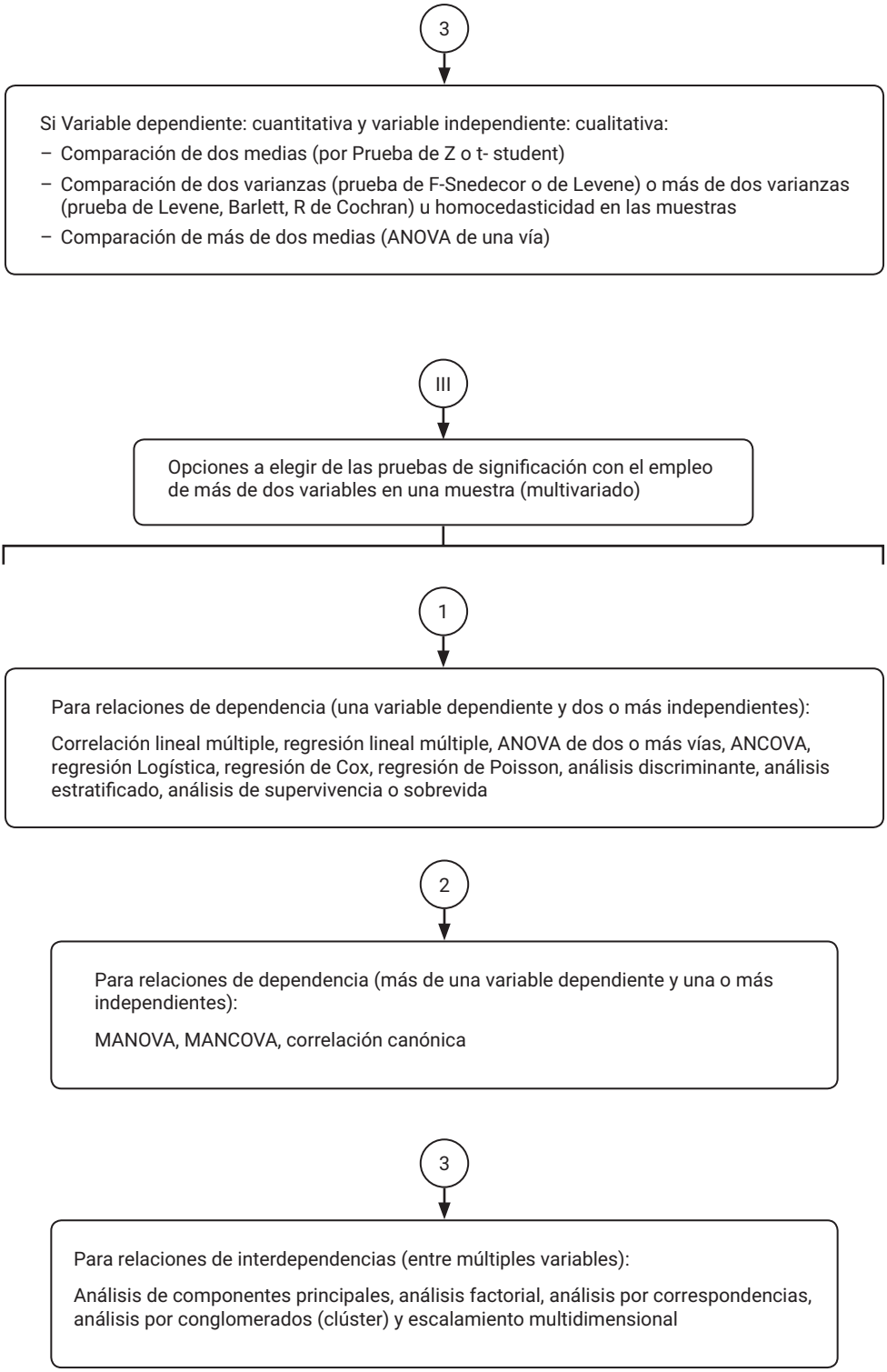


Fig. 6.19. Relación entre algunos verbos empleados en los enunciados de los objetivos de los diseños investigativos y las pruebas de contrastes de hipótesis a emplear en la inferencia bioestadística por el método frecuentista.




I
Opciones a elegir de las pruebas de significación con el empleo de una variable en una muestra.
Prueba para evaluar la normalidad en la distribución de los datos de una variable aleatoria cuantitativa (D' Agostino, Shapiro-Francia, Kolmogorov-Smirnov)
Prueba para evaluar la alietoriedad o la Independencia de las observaciones (prueba de rachas o de tendencias)
Pruebas para evaluar la comparación de un parámetro estimado a partir del muestral correspondiente vs. un valor hipotético o de referencia de ese parámetro empleando una variable aleatoria de distribución conocida
Prueba de Z para una media
Prueba de t-student para una media
Prueba de chi-cuadrado para una varianza
Prueba de chi-cuadrado para una tasa de incidencia
Prueba de Z para una proporción
II
Opciones a elegir de las pruebas de significación, al emplear dos variables en una muestra
1
Si variable dependiente: cuantitativa y variable independiente: cuantitativa:
– Coeficiente de correlación lineal simple de Pearson (prueba de t-student)
– Coeficiente de regresión lineal simple (Prueba de t-student, prueba de F)
– Coeficiente de correlación por rangos de Spearman (Prueba t-student para muestras grandes n ≥ 30)
2
Si variable dependiente: cualitativa y variable independiente: cualitativa:
– Independencia *vs.* asociación significativa de las dos variables (prueba de chi-cuadrado de Pearson, corregido de Yates, probabilidad exacta de Fisher, todos empleados para tablas de contingencia por el enfoque de riesgo)
– Concordancia y reproducibilidad de dos pruebas diagnósticas (prueba de Z para KAPPA)
z


3
Si Variable dependiente: cuantitativa y variable independiente: cualitativa:
– Comparación de dos medias (por Prueba de Z o t- student)
– Comparación de dos varianzas (prueba de F-Snedecor o de Levene) o más de dos varianzas (prueba de Levene, Barlett, R de Cochran) u homocedasticidad en las muestras
– Comparación de más de dos medias (ANOVA de una vía)
III
Opciones a elegir de las pruebas de significación con el empleo de más de dos variables en una muestra (multivariado)
1
Para relaciones de dependencia (una variable dependiente y dos o más independientes):
Correlación lineal múltiple, regresión lineal múltiple, ANOVA de dos o más vías, ANCOVA, regresión Logística, regresión de Cox, regresión de Poisson, análisis discriminante, análisis estratificado, análisis de supervivencia o sobrevida
2
Para relaciones de dependencia (más de una variable dependiente y una o más independientes):
MANOVA, MANCOVA, correlación canónica
3
Para relaciones de interdependencias (entre múltiples variables):
Análisis de componentes principales, análisis factorial, análisis por correspondencias, análisis por conglomerados (clúster) y escalamiento multidimensional

IV

Pruebas de significación para dos muestras pareadas o relacionadas con una variable cuantitativa de libre distribución

Prueba de los Signos de Wilcoxon, de los rangos con signos de Wilcoxon (prueba de Z para muestras grandes)

Prueba de significación para dos muestras pareadas o relacionadas con una variable cuantitativa que sigue la aproximación de la distribución normal estándar

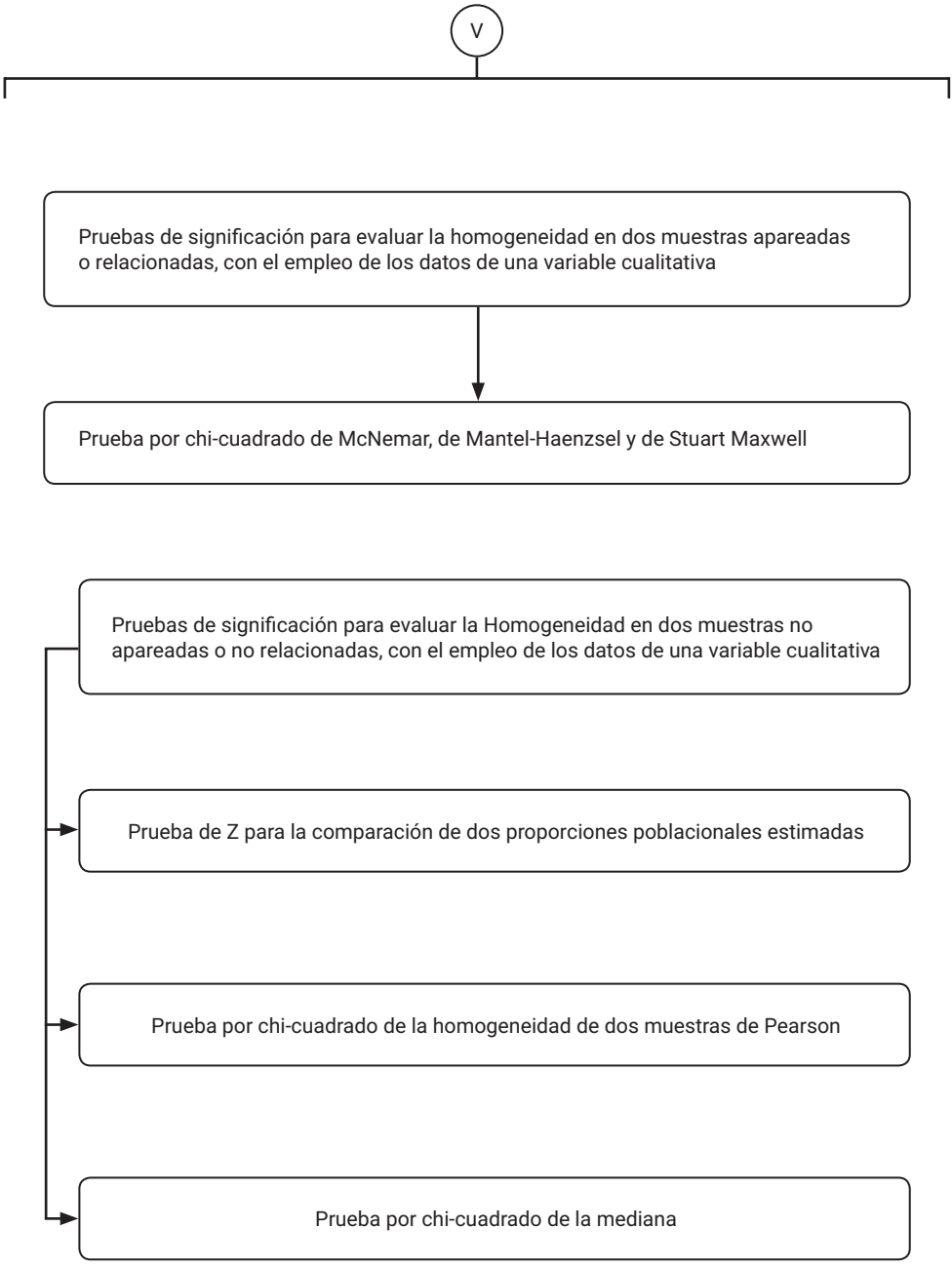
Prueba de t-studen pareada para la comparación de dos medias poblacionales estimadas pareadas

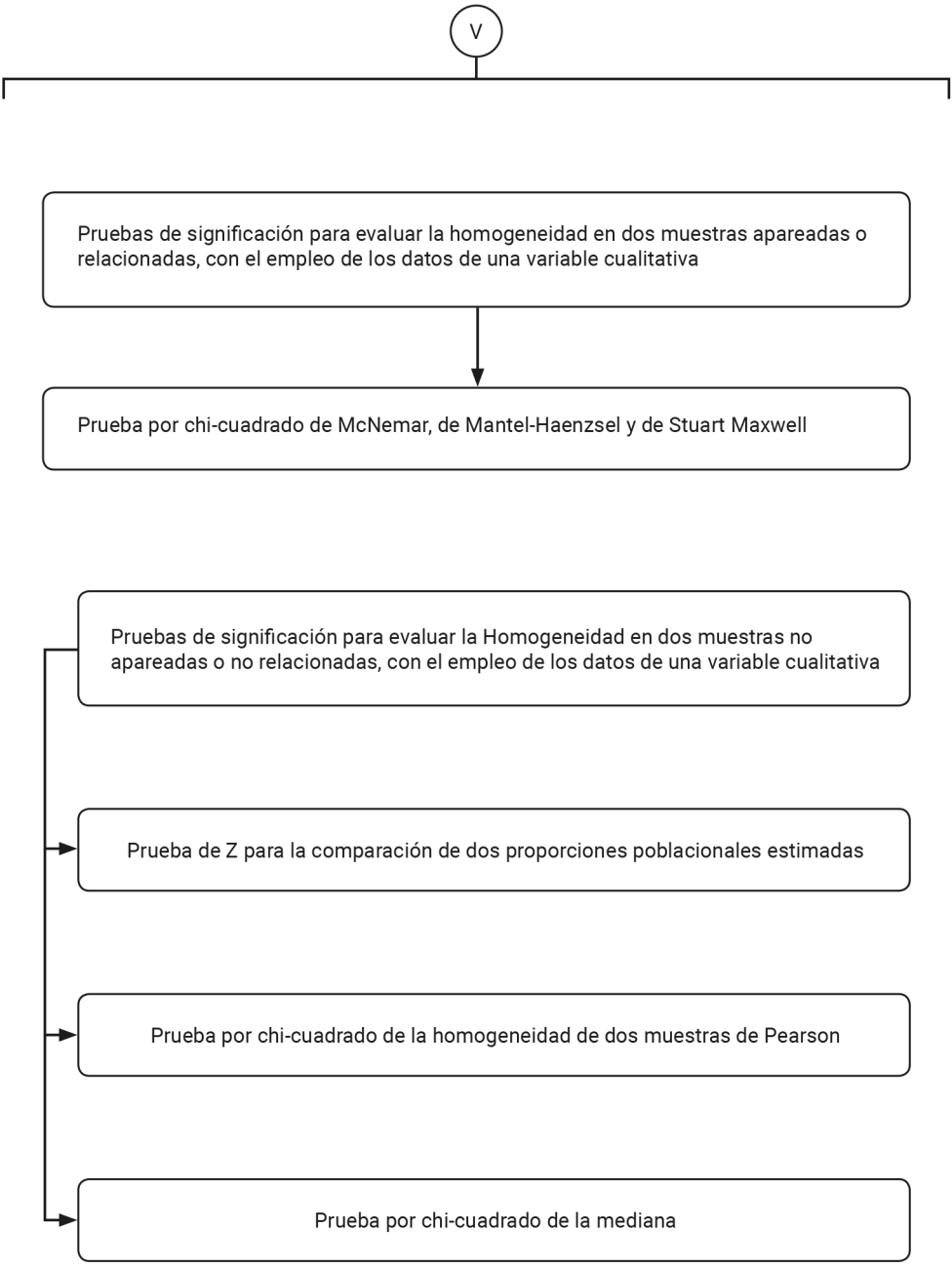
Pruebas de significación para dos muestras no pareadas o no relacionadas con una variable cuantitativa de libre distribución

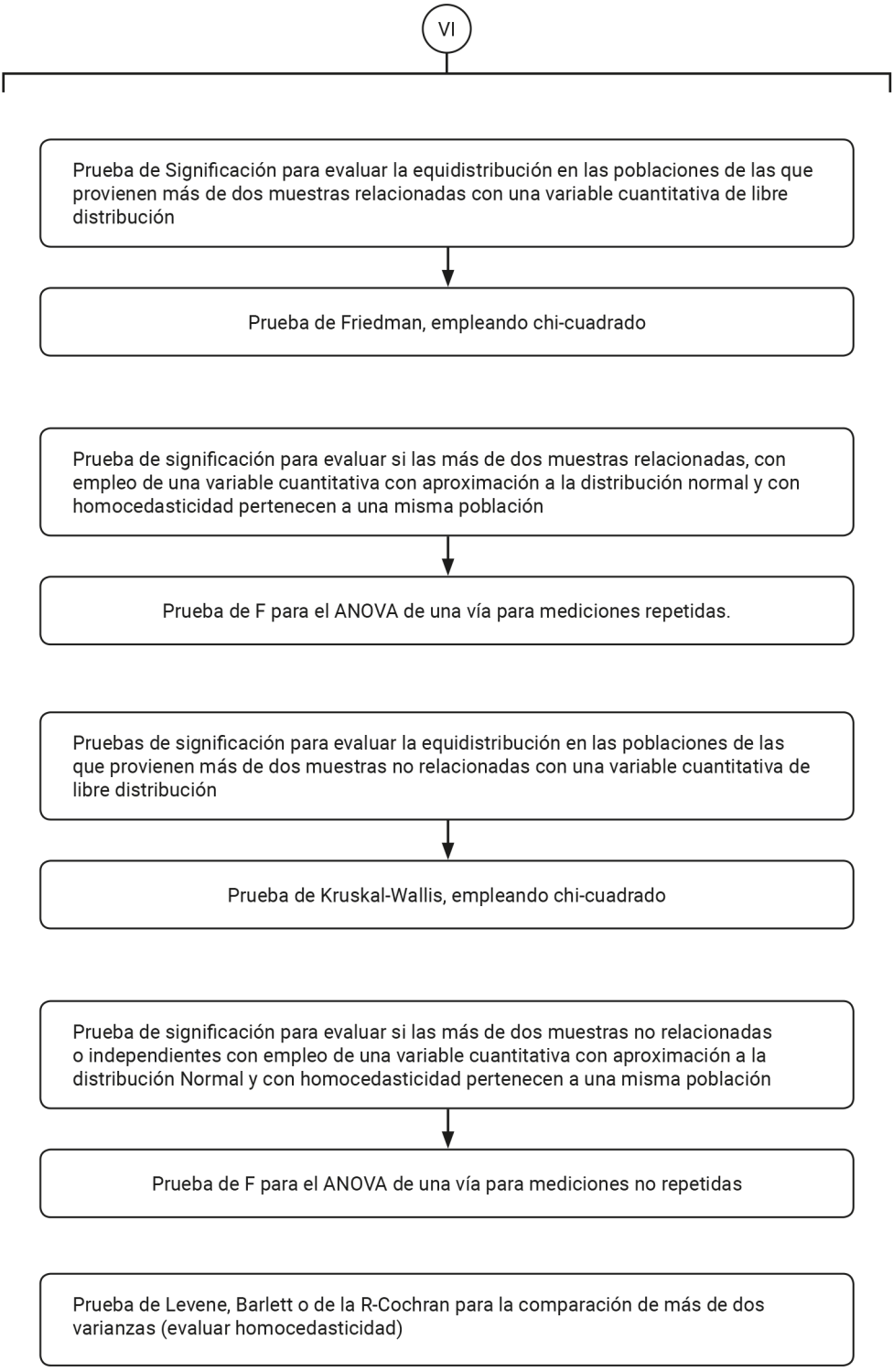
Prueba de la suma de rangos de Wilcoxon, de la U de Mann Whitney, de rachas de Wald-Wolfowitz (prueba de Z para muestras grandes)

Prueba de significación para dos muestras no pareadas o no relacionadas con una variable cuantitativa que sigue la aproximación de la distribución normal estándar

Prueba de t-studen no pareada para comparación de dos medias poblacionales estimadas


V
Pruebas de significación para evaluar la homogeneidad en dos muestras apareadas o relacionadas, con el empleo de los datos de una variable cualitativa
Prueba por chi-cuadrado de McNemar, de Mantel-Haenzsel y de Stuart Maxwell
Pruebas de significación para evaluar la Homogeneidad en dos muestras no apareadas o no relacionadas, con el empleo de los datos de una variable cualitativa
Prueba de Z para la comparación de dos proporciones poblacionales estimadas
Prueba por chi-cuadrado de la homogeneidad de dos muestras de Pearson
Prueba por chi-cuadrado de la mediana


V
Pruebas de significación para evaluar la homogeneidad en dos muestras apareadas o relacionadas, con el empleo de los datos de una variable cualitativa
Prueba por chi-cuadrado de McNemar, de Mantel-Haenzsel y de Stuart Maxwell
Pruebas de significación para evaluar la Homogeneidad en dos muestras no apareadas o no relacionadas, con el empleo de los datos de una variable cualitativa
Prueba de Z para la comparación de dos proporciones poblacionales estimadas
Prueba por chi-cuadrado de la homogeneidad de dos muestras de Pearson
Prueba por chi-cuadrado de la mediana


VI
Prueba de Significación para evaluar la equidistribución en las poblaciones de las que provienen más de dos muestras relacionadas con una variable cuantitativa de libre distribución
Prueba de Friedman, empleando chi-cuadrado
Prueba de significación para evaluar si las más de dos muestras relacionadas, con empleo de una variable cuantitativa con aproximación a la distribución normal y con homocedasticidad pertenecen a una misma población
Prueba de F para el ANOVA de una vía para mediciones repetidas.
Pruebas de significación para evaluar la equidistribución en las poblaciones de las que provienen más de dos muestras no relacionadas con una variable cuantitativa de libre distribución
Prueba de Kruskal-Wallis, empleando chi-cuadrado
Prueba de significación para evaluar si las más de dos muestras no relacionadas o independientes con empleo de una variable cuantitativa con aproximación a la distribución Normal y con homocedasticidad pertenecen a una misma población
Prueba de F para el ANOVA de una vía para mediciones no repetidas
Prueba de Levene, Barlett o de la R-Cochran para la comparación de más de dos varianzas (evaluar homocedasticidad)

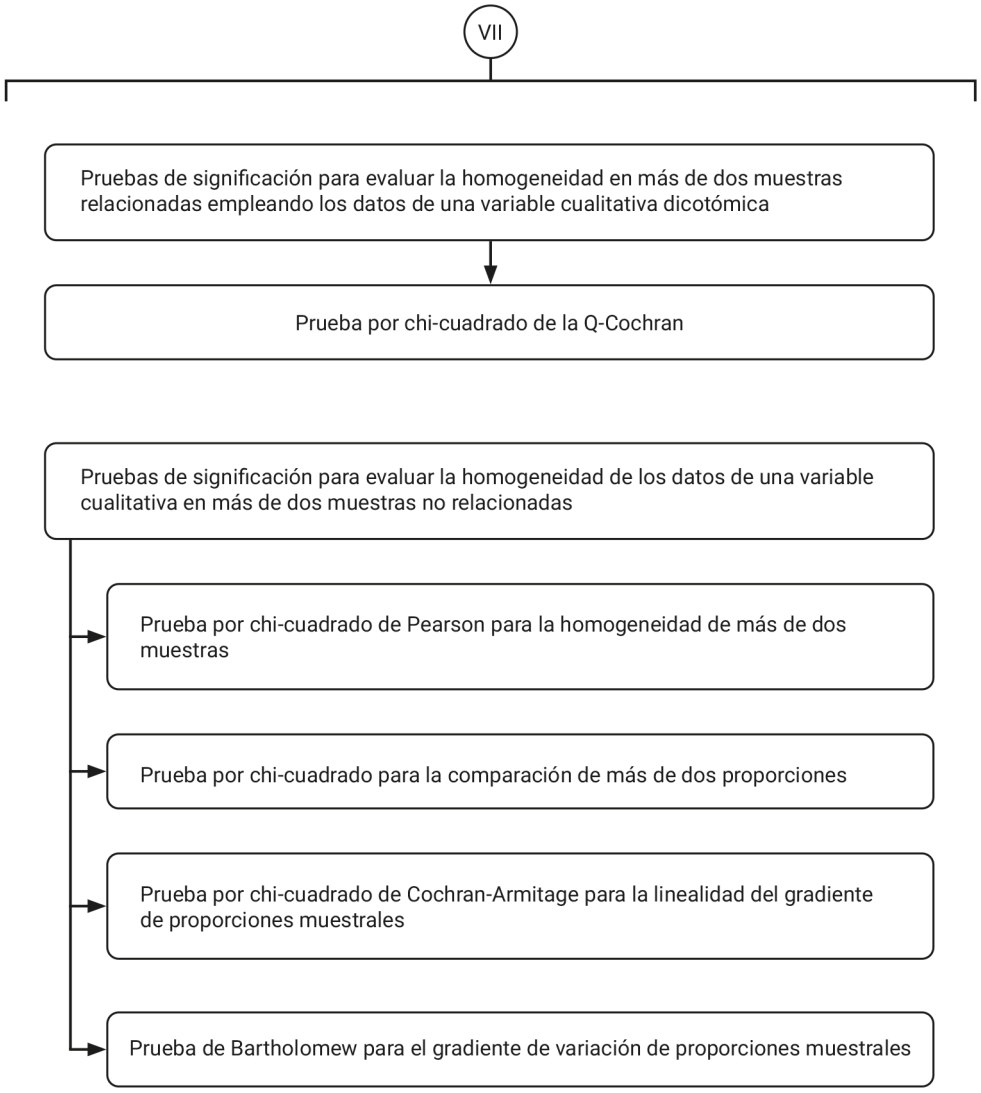


Fig. 6.20. Resumen de las posibles opciones de pruebas de significación relacionadas con las distintas combinaciones de muestras y variables en la inferencia bioestadística por el método frecuentista.

Además de los esquemas algorítmicos mostrados anteriormente y que permiten de una forma didáctica el aprendizaje acerca de la integración en la relación entre muestras-variables y aspectos de la inferencia bioestadística, las tablas resumen son capaces de mostrar una arista complementaria que puede facilitar el aprendizaje y aplicación de los elementos básicos involucrados y relacionados con la inferencia bioestadística para la realización del procesamiento estadístico eficaz de la información. En la tabla 6.1 se presenta un resumen de algunas opciones en el procesamiento estadístico para

el trabajo con dos variables en una muestra (método bivariado) y la interpretación de los resultados probables a obtener por medio de estos mediante la cual se complementan algunos aspectos referenciados en la figura 6.10.

Tabla 6.1. Resumen de algunas opciones en el procesamiento estadístico, para el trabajo con dos variables en una muestra

Tipo de combinación	Procedimiento estadístico para determinar	Interpretación
Dos variables cuantitativas	Correlación lineal simple entre dos variables	Expresa la fuerza, intensidad o grado de relación entre ambas variables. Cuanto de la variación de la variable dependiente es influenciado por la variación en la independiente. Se emplea una prueba de significación para el valor del coeficiente de correlación lineal simple de Pearson
Dos variables cuantitativas	Regresión lineal simple entre dos variables	Expresa cuanto es predecida o pronosticada la variable dependiente a partir de la variación de una unidad de la variable independiente. Se determina su valor puntual e intervalos de confianza y se emplea una prueba de significación para el valor del coeficiente de regresión lineal simple para cierto nivel de confianza.
Dos variables cualitativas	Diferencia entre proporciones:	Expresa las diferencias significativas o no de las proporciones formadas a partir del número de expuestos a cierta condición (o categoría 1 de la variable independiente) con el evento según el total de expuestos o con categoría 1 *versus* el número de no expuestos o de la categoría 2 de la independiente, con el evento según el total de no expuestos o con categoría 2
	Asociación o independencia entre dos variables	Explora la asociación significativa entre las dos variables cualitativas a partir de la homogeneidad o no de las proporciones formadas para ambas categorías en la variable independiente y el evento en la dependiente
	Estimando el riesgo relativo	Expresa la razón o cociente de probabilidades o riesgos de aparecer un efecto o evento a partir de dos condiciones diferentes. Se determinan su valor puntual e intervalos de confianza

Tipo de combinación	Procedimiento estadístico para determinar	Interpretación
	Estimando el *odds ratio*	Expresa la razón o cociente entre dos oportunidades diferentes para la aparición de un evento o no, cada oportunidad en relación a dos condiciones diferentes. Se determinan su valor puntual e intervalos de confianza
Una variable cuantitativa (dependiente) y una variable cualitativa (independiente)	Diferencia de medias	Expresa las diferencias de las medias de los grupos de datos formados a partir del número de categorías para la variable independiente cualitativa. Se determina su valor puntual y por intervalos de confianza y se emplea una prueba de significación en la evaluación de las diferencias de medias de los grupos para cierto nivel de confianza alfa prefijado
Una variable dependiente, expresada en forma del número de nuevos casos o eventos según el total de personas-tiempo de seguimiento en relación con la presencia de exposición o no (variable independiente)	Asociación o independencia entre dos variables cualitativas:	Explora la asociación significativa entre las dos variables cualitativas a partir de la homogeneidad o no de las proporciones formadas para ambas categorías en la variable independiente y la tasa de incidencia o nuevos eventos por persona-tiempo en la dependiente

El uso de la media aritmética como parámetro en dos o más poblaciones estimadas mediante las pruebas de significación paramétricas y el empleo de pruebas no paramétricas en dos o más muestras en relación con la misma, constituye un aspecto medular de los empleados en la inferencia bioestadística, lo que está con frecuencia en el centro de gran cantidad de procesamientos estadísticos, por ejemplo, de estudios epidemiológicos experimentales como los ensayos clínicos que emplean gran cantidad y variedades de variables cuantitativas. En la tabla 6.2 se presenta un resumen de algunas opciones empleadas en el estudio comparativo de medias aritméticas, según los tipos de pruebas de significación, mediante la cual se complementan algunos aspectos referenciados en la figura 6.8.

Tabla 6.2. Resumen de algunas opciones empleadas en el estudio comparativo de medias aritméticas según los tipos de pruebas de significación

	Comparación de dos medias		Comparación de más de dos medias	
	Prueba paramétrica	**Prueba no paramétrica**	**Prueba paramétrica**	**Prueba no paramétrica**
Datos independientes (no pareados o no relacionados tomados de forma transversal, en un momento del tiempo)	t-student para dos grupos de datos independientes (según igual o diferentes varianzas).	U de Mann-Whitney-Kendal	ANOVA para medidas no repetidas.	De Kruskall-Wallis
Datos no independientes (pareados o relacionados tomados de forma longitudinal, más de un momento del tiempo)	t-student para dos grupos de datos pareados	De los Rangos con signos de Wilcoxon	ANOVA para medidas repetidas	De Friedman

Otro aspecto que se explora con detenimiento en gran cantidad de proyectos investigativos es el comportamiento homogéneo o no de la distribución de los datos de diferentes variables para un número de muestras diferente. En la tabla 6.3 se presenta un resumen de algunas opciones para el estudio de la homogeneidad en la distribución de dos o más grupos de datos (equivalentes con muestras) empleando diferentes tipos de variables, mediante la cual se complementan algunos aspectos referenciados en la figura 6.9.

Tabla 6.3. Resumen de algunas opciones para el estudio de la Homogeneidad entre dos o más grupos de datos (muestras) empleando diferentes tipos de variables

Tipo de variable en las dos muestras	Prueba a realizar
Para dos muestras pareadas, dos grupos de datos no independientes (son pareados), (longitudinalmente, datos tomados en dos momentos en el tiempo)	
Nominal dicotómica	– De McNemar si: 1 caso por 1 control (no paramétrica). – De Mantel-Haenzsel si: 1 caso por más de 1 control (no paramétrica)
Nominal politómica	De Stuart-Maxwell (no paramétrica)
Cuantitativa discreta (distribución libre con valores por rangos igual al ordinal)	De los signos o de los rangos con signos de Wilcoxon (no paramétrica)
Cuantitativa continua (distribución aproximada de la normal)	De la t-student pareada (paramétrica)
Para más de dos muestras pareadas, más de dos grupos de datos no independientes (son relacionados) (longitudinalmente, datos tomados en más de dos momentos en el tiempo)	

Tipo de variable en las dos muestras	Prueba a realizar
Nominal dicotómica	De Cochran (no paramétrica)
Cuantitativa discreta (distribución libre con valores por rangos igual al ordinal)	De Friedman (no paramétrica)
Cuantitativa continua (distribución normal)	ANOVA (de una vía, para medidas repetidas) (paramétrica)
Para dos muestras no pareadas, dos grupos de datos independientes (transversalmente, datos tomados un solo momento en el tiempo)	
Nominal dicotómica	De Pearson, corregido de Yates, prueba exacta de Fisher por tablas 2x2 simples (no paramétrica)
Nominal politómica	– De Pearson por tablas cruzadas de filas por columnas (con filas o columnas mayor de 2) – De tendencia lineal entre niveles por tablas 2xN simple (con N niveles de exposición) (no paramétrica)
Cuantitativa discreta (distribución libre por rangos = ordinal)	De la U de Mann Whitney o de la suma de Rangos de Wilcoxon (no paramétrica)
Cuantitativa continua (distribución aproximada de la normal)	De la t-student no pareada (paramétrica)
Para más de dos muestras no pareadas, más de dos grupos de datos independientes (transversalmente, datos tomados un solo momento en el tiempo)	
Nominal dicotómica	De diferencia de más de dos proporciones por chi-cuadrado (una proporción por muestra independiente a partir de la categoría o característica de interés en cada grupo de datos o muestras) (no paramétrica)
Nominal dicotómica o politómica	De homogeneidad en las muestras por chi-cuadrado de Pearson para tablas de contingencia o de filas por columnas (filas igual o mayor de do, es el número de categorías de la variable; columnas mayor de dos, es el número de muestras) (no paramétrica)
Cualitativa ordinal	De Bartholomew, para la significación del gradiente de variación en las proporciones formadas para los grupos de datos (muestras) cualitativamente ordenados (correspondientes con las categorías o nominaciones ordinales (niveles) de la variable cualitativa) (no paramétrica)
Cuantitativa discreta (distribución libre y expresada por rangos = ordinal por rangos).	De Kruskall-Wallis (no paramétrica)

Tipo de variable en las dos muestras	Prueba a realizar
Cuantitativa continua (distribución normal)	ANOVA (una vía, para medidas no repetidas) (paramétrica)
Cuantitativa discreta (expresada en forma de clases)	De Cochran-Armitage para la significación del gradiente de variación en las proporciones formadas para los grupos de datos o muestras cuantitativamente ordenados correspondientes con las clases de la variable cuantitativa (no paramétrica)

El conocimiento de la relación e integración que existe entre los diferentes elementos y aspectos involucrados en el procesamiento estadístico de la información recogida, permite la convergencia hacia una correcta aplicación de las opciones incluidas en la inferencia bioestadística, lo que puede contribuir a lograr un mejor análisis y evaluación de los resultados obtenidos y con esto tratar de cumplir los objetivos declarados de la investigación en cuestión. En la tabla 6.4 se presenta un resumen que integra los elementos: variable -estimador-tipo de contraste de hipótesis-estadístico de prueba, aspectos que constituyen las bases de la vertiente inferencial del procesamiento estadístico de la información y son capaces de complementar la información facilitada para ese particular por las figuras 6.19 y 6.20.

Tabla 6.4. Resumen que integra los elementos variable-estimador-tipo de contraste de hipótesis-estadístico de prueba:

Tipo de variable	Estimador muestral empleado	Tipo de contraste de hipótesis	Estadístico de prueba o de contraste a emplear
Cuantitativa en una muestra	Una media aritmética muestral y una poblacional de referencia	Diferencia de medias (la muestral *versus* una de referencia)	– Z ~ N (0,1) normal estándar si n ≥ 30. – t-student si n < 30
Cuantitativa en dos muestras no pareadas	Dos medias aritméticas; una de cada muestra (muestras no pareadas)	Diferencia de dos medias muestrales no pareadas que estiman dos poblaciones	t-student no pareada
Cuantitativa en dos muestras pareadas	Dos medias aritméticas; una tomada en dos momentos diferentes (muestras pareadas)	Diferencia de dos medias muestrales pareadas que estima una diferencia poblacional	t-student pareada
Cuantitativa en más de dos muestras no relacionadas	Mas de dos medias aritméticas, cada una de muestras independientes	Diferencia de más de dos medias muestrales no relacionadas que estiman más de dos poblacionales, por ANOVA de una vía (medidas no repetidas)	F-Snedecor.

Tipo de variable	Estimador muestral empleado	Tipo de contraste de hipótesis	Estadístico de prueba o de contraste a emplear
Cuantitativa en más de dos muestras relacionadas	Más de dos medias aritméticas correspondientes a una misma muestra con más de dos mediciones (muestras relacionadas)	Diferencia de más de dos medias muestrales relacionadas que estiman más de dos poblacionales, por ANOVA de una vía (medidas repetidas)	F-Snedecor
Cuantitativa en dos muestras no relacionadas	Dos medianas, una de cada muestra independiente	Diferencia de dos medianas muestrales	Chi-cuadrado (prueba de Person para tablas de 2x2)
Cuantitativa en una muestra	Una varianza muestral y una de referencia poblacional	Homogeneidad de la varianza poblacional estimada de la muestral respecto a la poblacional de referencia (probar la homocedasticidad)	Chi-cuadrado
Cuantitativa en dos muestras no pareadas	Dos varianzas, una de cada muestra independiente.	Homogeneidad de dos varianzas poblacionales estimadas de cada muestral respectiva (probar homocedasticidad)	F-Snedecor
Cuantitativa en más de dos muestras no relacionadas	Más de dos varianzas, una en cada muestra	Homogeneidad de más de dos varianzas poblacionales estimadas de las muestrales (probar homocedasticidad)	– F-Snedecor (prueba de Levene) – Chi-cuadrado (prueba de Barlett) – R de Cochran (prueba de Cochran)
Cualitativa (dicotómica) en una muestra	Una proporción muestral y una poblacional de referencia	Diferencia de proporciones (la poblacional estimada de la muestral *versus* una poblacional de referencia)	$Z \sim N(0, 1)$ por aproximación de la normal-estándar preferiblemente para $pqn > 5$
Cualitativa (dicotómica) en dos muestras no pareadas.	Dos proporciones, una de cada muestra no pareada.	Diferencia de dos proporciones poblacionales estimadas de dos muestrales no pareadas	$Z \sim N(0,1)$ Normal estándar

Tipo de variable	Estimador muestral empleado	Tipo de contraste de hipótesis	Estadístico de prueba o de contraste a emplear
Cualitativa (politómica) en dos muestras no pareadas	Más de dos proporciones (una proporción para la categoría de interés en cada una de las dos muestras no pareadas).	Homogeneidad en las dos muestras no pareadas de las proporciones poblacionales estimadas formadas para la variable politómica	Chi-cuadrado de Pearson para Homogeneidad por tablas cruzadas de filas por columnas: con filas y columnas mayor de dos
Una cualitativa dicotómica en dos muestras no pareadas	Una proporción para la categoría de interés de la variable para cada muestra independiente	Homogeneidad de las dos muestras independientes para las proporciones formadas según variable dicotómica	Chi-cuadrado de homogeneidad de Pearson, Yates o probabilidad exacta de Fisher para tablas de 2x2 simples
Dos cualitativas (dicotómicas), una independiente y otra dependiente, en una muestra	Dos proporciones en una misma muestra en relación a la categoría presente o no de la variable independiente para el evento de la dependiente	Independencia *versus* asociación entre las dos variables dicotómicas	Chi-cuadrado de asociación (de Pearson, corregido de Yates o probabilidad exacta de Fisher para tablas de 2x2 simples
Dos cualitativas (dicotómica la dependiente y una ordinal la independiente, expresada en forma de niveles de exposición en una muestra)	Una proporción para cada categoría de la variable ordinal (niveles de exposición) en una muestra.	Tendencia lineal entre los niveles de exposición	Chi-cuadrado para tendencia lineal por tablas de 2xN simples
Una cualitativa ordinal ordenada por niveles que permite formar más de dos grupos o muestras no relacionadas.	Una proporción para cada categoría de la variable ordinal (corresponde a igual número de muestras independientes	Del gradiente de variación de las más de dos proporciones para las más de dos muestras independientes.	Estadístico de Bartholomew

Tipo de variable	Estimador muestral empleado	Tipo de contraste de hipótesis	Estadístico de prueba o de contraste a emplear
Una cuantitativa ordenada por clases que permite formar más de dos grupos o muestras no relacionadas en relación a una característica o categoría de una variable cualitativa	Una proporción para cada clase de la variable cuantitativa cuyas clases se corresponden a igual número de muestras independientes	De la asociación lineal o linealidad entre cada proporción y al punto medio de cada clase.	Estadístico chi-cuadrado de Cochran-Armitage
Una cualitativa dicotómica en dos muestras pareadas.	Una proporción para la categoría de interés para cada muestra pareada	Homogeneidad en las dos muestras pareadas de las proporciones formadas	Chi-cuadrado (prueba de McNemar)
Una cualitativa politómica en dos muestras no pareadas.	Una proporción para cada categoría de la variable politómica de cada muestra no pareada	Homogeneidad en las dos muestras para las proporciones.	Chi-cuadrado de Pearson para homogeneidad por tablas cruzadas de filas por columnas con filas y columnas mayor de dos
Una cualitativa politómica en dos muestras pareadas	Una proporción para la categoría de interés en cada muestra pareada	Homogeneidad en las dos muestras pareadas para las proporciones	Chi-cuadrado de Stuart-Maxwell
Una cualitativa dicotómica en más de dos muestras relacionadas	Una proporción para la categoría de interés en cada muestra relacionada	Homogeneidad en las más de dos muestras relacionadas para las proporciones	Chi-cuadrado de Cochran
Una cualitativa dicotómica en más de dos muestras no relacionadas	Una proporción para la categoría de interés en cada muestra no relacionada	Homogeneidad en las más de dos muestras relacionadas para las proporciones	Chi-cuadrado de Pearson para tablas cruzadas de filas por columnas con filas y columnas mayor de dos

Tipo de variable	Estimador muestral empleado	Tipo de contraste de hipótesis	Estadístico de prueba o de contraste a emplear
Una cualitativa expresada en forma de nuevos casos o eventos según personas por tiempo de seguimiento en una muestra.	Una tasa de incidencia muestral y una poblacional de referencia	Diferencia de tasas de incidencia (la poblacional estimada de la muestral *versus* la poblacional de referencia)	Chi-cuadrado para una tasa de incidencia
Una cualitativa expresada en forma de nuevos casos o eventos según personas por tiempo de seguimiento en dos muestras independientes	Una tasa de incidencia en dos muestras independientes	Diferencia de dos tasas de incidencia	Z~N (0,1) por aproximación de la normal estándar para dos tasas de incidencia
Una cualitativa expresada en forma de nuevos casos o eventos según personas-tiempo de seguimiento y una Cualitativa ordinal para más de dos categorías o niveles de exposición en una muestra.	Una tasa de incidencia correspondiente con los más de dos niveles de exposición en una muestra	Tendencia lineal en el incremento en las más de dos tasas de incidencias según niveles de exposición	Prueba de chi-cuadrado para tendencia lineal (según tasas de incidencia)
Dos cualitativas (dicotómicas) en una muestra	Dos proporciones en una muestra	Significación de la Concordancia entre las dos proporciones estimadas para la relación entre las dos variables dicotómicas empleando el estadístico KAPPA	Prueba de Z por la aproximación de la normal estándar: Z ~N (0, 1)

Nota: para evaluar la significación en el contraste de hipótesis se procede a comparar los estadísticos de prueba observado-calculado según el tipo de contraste específico *versus* los valores tabulados-esperados de la distribución teórica de frecuencias o probabilidades correspondiente al valor del percentil especifico (según alfa prefijado por el investigador) que representa el estadístico de prueba o contraste teórico-esperado-tabulado.

Regla de decisión

Si el valor de la probabilidad en que se cumpliría H_0 (donde el estadístico de prueba teórico fuera mayor o igual que el estadístico de prueba observado calculado en relación al valor calculado especifico de este último es igual a p-valor) fuese menor que el valor alfa prefijado, entonces se rechaza a la hipótesis nula y se acepta a la alternativa siendo significativa estadísticamente a condición contrastada en el estudio bajo un paradigma frecuentista.

En la tabla 6.5 se presenta un resumen de los diferentes tipos de diseños de estudios epidemiológicos observacionales por el enfoque de riesgo según exposición, los tipos de tablas de contingencia y las pruebas de significación correspondientes, mediante la cual se complementan algunos aspectos referenciados en las figuras 6.12 y 6.13.

Tabla 6.5. Resumen de los diferentes tipos de diseños de estudios epidemiológicos observacionales por el enfoque de riesgo según exposición, los tipos de tablas de contingencia y las pruebas de significación correspondientes

Tipo de estudio observacional	Tipo de tabla de contingencia	Prueba de significación a emplear
De corte o transversal	2x2 simple	– Prueba chi-cuadrado de asociación con y sin corrección – Prueba de la probabilidad exacta exacta de Fisher
De caso-control (no pareado)	2x2 simple	– Prueba chi-cuadrado de asociación con y sin corrección – Prueba de la probabilidad exacta exacta de Fisher
De caso-control (pareados, un control por caso)	2x2 simple (transformada de casos y control, ambos de expuestos y no expuestos, por pares)	Prueba chi-cuadrado de asociación de McNemar
De caso-control (pareados, k > 1 controles por caso)	2x2 simple (transformada de casos expuestos y no expuestos versus más de un control expuestos, por pares)	Prueba chi-cuadrado de asociación de Mantel-Haenszel
De cohortes (datos de tasas de incidencia)	2x2 simple	Prueba de asociación por aproximación a la normal Z~ N (0, 1) o por método exacto.
De cohortes (datos de incidencia acumulada)	2x2 simple	– Prueba chi-cuadrado de asociación con y sin corrección – Prueba de la probabilidad exacta exacta de Fisher

Tipo de estudio observacional	Tipo de tabla de contingencia	Prueba de significación a emplear
De corte o transversal.	2x2 estratificada	– Prueba de homogeneidad entre estratos para la razón de prevalencias de enfermedad – Prueba de homogeneidad entre estratos para la razón de prevalencias de exposición – Prueba de homogeneidad entre estratos para la *odds ratio* – Prueba de asociación de Mantel-Haenszel por chi-cuadrado combinado
De caso-control (no pareado)	2x2 estratificada	– Prueba de Homogeneidad entre estratos para la *odds ratio* – Prueba de asociación de Mantel-Haenszel por chi-cuadrado combinado
De cohortes, datos de tasas de incidencia	2x2 estratificada	– Prueba de homogeneidad entre estratos para la razón de tasas de incidencia – Prueba de homogeneidad entre estratos para la diferencia de tasas de incidencia – Prueba de asociación de Mantel-Haenszel por chi-cuadrado combinado
De cohortes, datos de incidencia acumulada.	2x2 estratificada	– Prueba de homogeneidad entre estratos para el Riesgo relativo – Prueba de homogeneidad entre estratos para la diferencia de riesgos – Prueba de asociación de Mantel-Haenszel por chi-cuadrado combinado
De corte o transversal	2xN simple	– Prueba de homogeneidad entre niveles de exposición – Prueba de tendencia lineal entre los niveles de exposición
De caso-control (no pareado)	2xN simple	– Prueba de homogeneidad entre niveles de exposición – Prueba de tendencia lineal entre los niveles de exposición

Tipo de estudio observacional	Tipo de tabla de contingencia	Prueba de significación a emplear
De cohortes, datos de tasas de incidencia	2xN simple	– Prueba de homogeneidad entre niveles de exposición – Prueba de tendencia lineal según los niveles de exposición
De cohortes, datos de incidencia acumulada	2xN simple	– Prueba de homogeneidad entre los niveles de exposición – Prueba de tendencia lineal entre los niveles de exposición
De caso-control (no pareado)	2xN estratificada	Prueba de tendencia lineal entre los niveles de exposición en cada estrato
De cohortes; datos de tasas de incidencia	2x2 estratificada	Prueba de tendencia lineal según los niveles de exposición en cada estrato
De cohortes; datos de incidencia acumulada.	2x2 estratificada	Prueba de tendencia lineal entre los niveles de exposición en cada estrato

Para el caso de tablas de contingencia (cruzadas) de MxN (filas por columnas), ambas mayores o igual a 2 si se emplean datos nominales en filas y columnas, entonces:

Estadígrafo utilizado:	Prueba de significación a emplear
Chi-cuadrado de Pearson (χ^2) ~ distribución (χ^2) para (F – 1) (C – 1) como grados de libertad y para el nivel de significación alfa	Prueba de Independencia entre las dos variables
Coeficiente de contingencia (C), que relaciona el valor de chi-cuadrado observado como: $C=\sqrt{\frac{\chi^2}{\chi^2+n}}$	No prueba especifica
Coeficiente V de Cramer (V), que relaciona el valor de chi-cuadrado observado como: $V=\sqrt{\frac{\chi^2}{n(m-1)}}$	No prueba específica
Taus de Goodman y Kruskal	No prueba específica

Para el caso de tablas de contingencia (cruzadas) de MxN (filas por columnas), ambas mayores o igual a 2 si se emplean datos ordinales en filas y columnas, entonces:

Estadígrafo utilizado	Prueba de significación a emplear
Chi-cuadrado de tendencia lineal (χ^2_t) ~ distribución (χ^2) para (F – 1) (C – 1) como grados de libertad y para el nivel de significación alfa	Prueba de tendencia lineal entre niveles
Tau-b de Kendall, Tau-c de Stuart, Gamma de Goodman y Kruskal, Ds de Sommers.	Prueba de asociación entre las dos variables empleando la aproximación a la normal estándar Z ~ N (0, 1) a partir de los valores estimados de dichos estadígrafos

En la tabla 6.6 se presenta un resumen de algunos aspectos de los tipos de métodos multivariados, mediante la cual se complementan algunos aspectos referenciados desde las figuras 6.14 a la 6.18.

Tabla 6.6. Resumen de algunos aspectos de los tipos de métodos multivariados:

Forma de relación entre las más de dos variables	Tipo de método multivariado	Tipo de variable dependientes	Tipo de variable independientes
	Correlación lineal múltiple	Cuantitativa	Cuantitativas
	Regresión lineal múltiple	Cuantitativa	Cuantitativas
	ANOVA	Cuantitativa	Cualitativas
	ANCOVA	Cuantitativa	Cualitativas y cuantitativa
	Análisis discriminante (no dependiente del tiempo)	Cualitativa	Cuantitativas
	Análisis estratificado (no dependiente del tiempo)	Cualitativa	Cualitativas
	Análisis de Sobrevida o Supervivencia (dependiente del tiempo, el número de casos sin el evento).	Cuantitativa	Cualitativas
De dependencias con una variable dependiente	Regresión logística (no dependiente del tiempo)	Cualitativa, expresada en forma de OR (razón de odds para el evento)	Cuantitativas o cualitativas. (transformadas a variables *dummy*)
	Regresión de Cox (dependiente del tiempo, es el tiempo hasta el evento)	Cuantitativa, expresada como *Hazard ratio* (razón de tasas de incidencias instantáneas o del tiempo para el evento)	Cuantitativas o cualitativas (transformadas a variables *dummy*)

Forma de relación entre las más de dos variables	Tipo de método multivariado	Tipo de variable dependientes	Tipo de variable independientes
	Regresión multivariada de Poisson (dependiente del tiempo; sujetos con el evento en el tiempo de seguimiento)	Cuantitativa expresada como razón de tasas de incidencia (casos con el evento según personas-tiempo de seguimiento)	Cuantitativas o cualitativas (transformadas a variables *dummy*)
De dependencias con más de una variable dependiente	MANOVA	Cuantitativas	Cualitativas
	MANCOVA	Cuantitativas	Cualitativas y covariable cuantitativa
	Correlación canónica	Cuantitativas	Cuantitativas
De interdependencias de variables	Componentes principales	Objetivo: reducción de datos, identificando combinaciones lineales de variables cuantitativas (componentes)	
	Análisis factorial	Objetivo: reducción de datos, identificando variables resumen (factores) a partir de correlaciones de variables cuantitativas	
	Análisis de correspondencias	Objetivo: identificar variables latentes o dimensiones a partir de la relación entre variables nominales (categóricas) para formar nuevas categorías en las variables	
	Análisis de conglomerados (*Cluster Analysis*)	Objetivo: identificar y clasificar grupos homogéneos de sujetos u objetos empleando variables cuantitativas y cualitativas	
	Escalamiento multidimensional	Objetivo: identificar relaciones entre objetos por medio de proximidades en sus dimensiones empleando variables cuantitativas y cualitativas	

Capítulo 7

Propuesta algorítmica no esquematizada para el manejo de muestras y variables con un enfoque frecuentista

En este capítulo se presenta la propuesta de una guía algorítmica no esquematizada a seguir, desarrollada en consulta con algunos expertos y con el apoyo en algunos elementos de los enfoques realizados por otros autores en la bibliografía consultada, de diferentes opciones de técnicas estadísticas a ejecutar empleando los distintos y respectivos contrastes de hipótesis paramétricos y no paramétricos, que forman parte y guardan relación estrecha con el manejo de muestras y variables que se vaya a realizar por el enfoque frecuentista, como parte del procesamiento estadístico de la información en la inferencia bioestadística, a partir de las situaciones en que más frecuentemente se puede hallar el estudiante o investigador joven al enfrentar su problema de investigación específico.

Inicialmente en esta guía algorítmica se propone partir del número de muestras seleccionadas por el investigador: una muestra, dos muestras o más de dos muestras. A su vez si se selecciona una muestra, puede necesitar trabajar con una variable (métodos univariados), dos variables (métodos bivariados) o más de dos variables (métodos multivariados):

- Si se emplean los métodos univariados (para una variable: cuantitativa o cualitativa), se comparan los estimadores muestrales que estiman los parámetros poblacionales *versus* los parámetros de otra población de referencia y se expresan sus valores por intervalo de confianza.
- Para el caso de los métodos bivariados si se emplean dos variables cuantitativas se evalúa la correlación o regresón lineal simple,si fuesen dos variables cualitativas, entonces se puede emplear la prueba de independencia entre estas y si se cuenta con una variable cuantitativa y una cualitativa, entonces se puede determinar la diferencia de medias de los grupos formados a partir de la variable independiente

cualitativa por medio de métodos paramétricos o no paramétricos y se expresan los valores de esta diferencia por intervalo de confianza en el caso de los métodos paramétricos.

- Si se eligen los métodos multivariados a emplear en una muestra se puede contar con opciones en el procesamiento estadístico de la información para el caso en que se desee evaluar: dependencia entre variables (para una variable dependiente o para más de una) o por otra parte se investiga la presencia de interdependencias entre las más de dos variables involucradas.
- Si el caso fuese de dos muestras emparejadas (pareadas o relacionadas) o no, para cada caso hay que determinar si se emplea una variable cuantitativa (diferencia de medias para muestras pareadas o no) en relación a cierta condición, categoría o característica de interés por el investigador, utilizando métodos o técnicas paramétricas o no paramétricas, así mismo, puede determinar diferencia de dos varianzas por métodos en las pruebas de significación paramétricas y en el caso de una variable cualitativa en ambas muestras se pueden comparar dos proporciones a partir de cierta categoría de dicha variable o realizar una prueba de homogeneidad en la distribución de los datos para una característica de una variable en las poblaciones estimadas de los grupos de datos o muestras, partiendo del comportamiento de las proporciones estimadas en relación a la categoría de la variable seleccionada para ambas muestras.
- En el caso de seleccionar más de dos muestras relacionadas (dependientes) o no, si se emplea una variable cuantitativa, se puede determinar diferencias de medias para muestras relacionadas (método del ANOVA de una vía) o la diferencias de más de dos varianzas por métodos no paramétricos. Si es necesario emplear una variable cualitativa se realiza de forma general la comparación de más de dos proporciones, así como una prueba de homogeneidad en las poblaciones estimadas a partir del comportamiento de las proporciones estimadas en relación a categorías de la variable en estudio por métodos no paramétricos.

Se reproduce la propuesta de la guía práctica general algorítmica para el manejo de muestras y variables, siguiendo un modelo frecuentista en la inferencia bioestadística (Fig. 7.1), la que constituye el hilo conductor y que resume toda la información antes brindada en el manejo de muestras y variables y donde se debe apoyar en lo adelante el investigador joven para comprender la propuesta de guía algorítmica con potencial para convertirse en una aplicación o *software* digital que se desarrollará en el este capítulo.

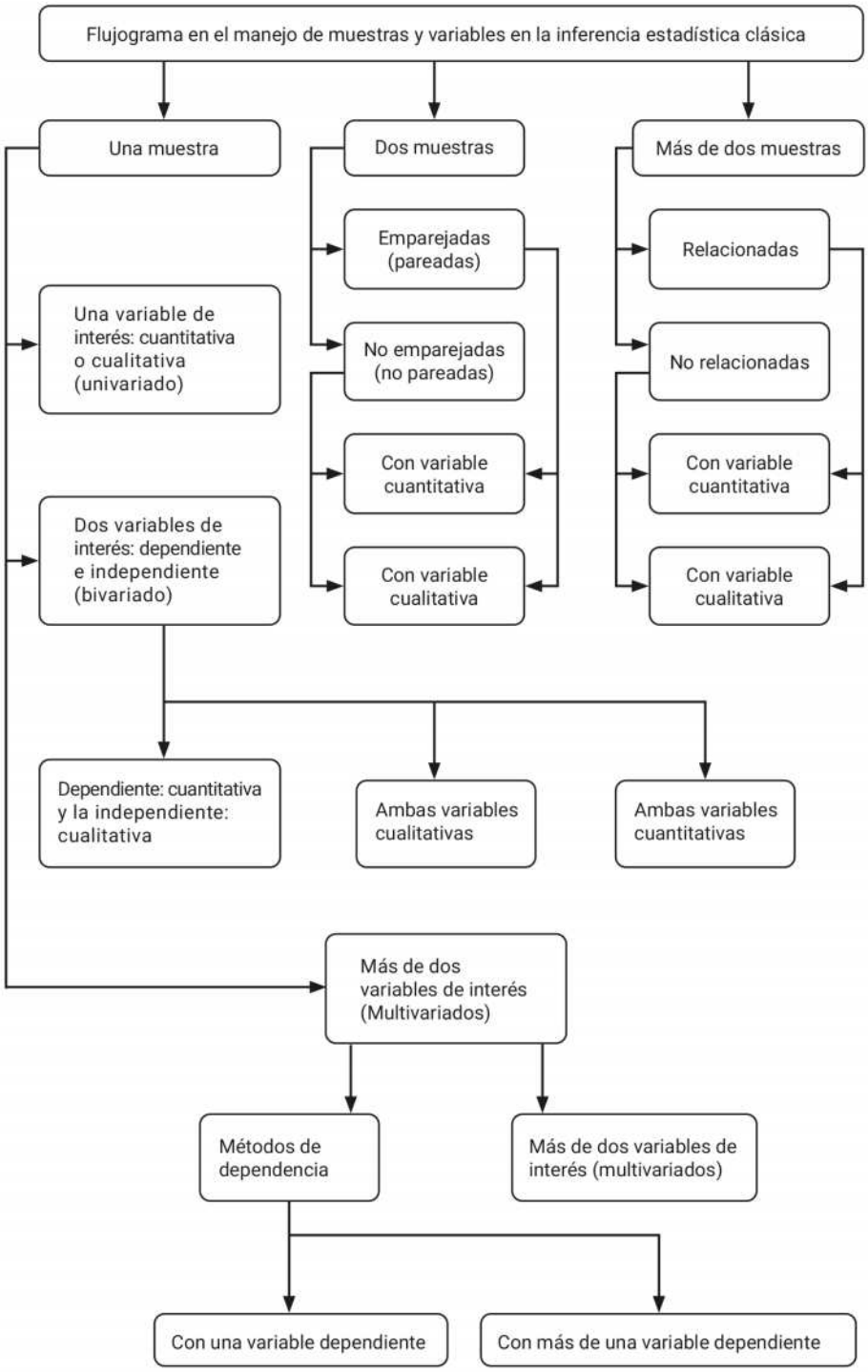


Fig. 7.1. Esquema empleado como base para desarrollar la propuesta algorítmica no esquematizada en el manejo de muestras y variables desde la inferencia bioestadística.

Los pasos propuestos a seguir por el investigador, como guía para el manejo de muestras y variables partiendo del anterior esquema algorítmico se desarrollan a continuación, aclarando que los números y las letras que puedan aparecer acompañando el fin de cada oración, representan a modo de síntesis cierta condición correspondiente con el enunciado realizado previamente y se emplea como elemento que articule los planteamientos para cada paso algorítmico en relación con la situación investigativa determinada que se esté abordando de la bioestadística inferencial preferiblemente.

Se comienza preguntando el investigador joven en el orden algorítmico:

I. ¿Número de muestras a emplear en el diseño de estudio?:
   1. Una muestra con una, dos o más de dos variables aleatorias.
   2. Dos muestras con la misma variable aleatoria cada una.
   3. Más de dos muestras con la misma variable aleatoria cada una.

   Si I.1 seguir con:

II. ¿Número de variables aleatorias a seleccionar para una muestra?:
   1. Una variable (univariado).
   2. Dos variables (bivariado).
   3. Más de dos variables (multivariado).

   Si II.1 una muestra con una variable aleatoria, entonces seguir con:

III. ¿Tipo de variable aleatoria a utilizar en una muestra?
   1. Cuantitativa o paramétrica.
   2. Cualitativa o categórica.

   Si III.1 una muestra con una variable aleatoria cuantitativa, entonces seguir con:

IV. ¿Sigue la variable aleatoria cuantitativa el modelo de una distribución normal o puede asumirse el teorema del límite central para la misma? (para su comprobación, realizar pruebas de normalidad, por ejemplo: prueba de D´ Agostino, Shapiro-Wilks, Shapiro-Francia)
   1. Variable aleatoria que sigue el modelo el modelo de distribución normal.
   2. Variable aleatoria que no sigue el modelo de distribución normal (tiene libre distribución).

   Si IV.1 una muestra con una variable aleatoria continua que sigue el modelo de distribución normal, entonces seguir con:

   Opciones de técnicas bioestadísticas investigativas de manejo para una muestra- una variable aleatoria cuantitativa con modelo de distribución de frecuencias modelada (normal):

   1. Estimación del parámetro poblacional con interés en la posición (media poblacional), a partir del estimador muestral por intervalo de confianza (IC) empleando la distribución normal estándar: $Z \sim N(0, 1)$ para un nivel de confianza preseleccionado.
   2. Estimación del parámetro poblacional con interés en la dispersión (varianza poblacional), a partir del estimador muestral por intervalo de confianza (IC) em-

pleando la distribución teórica de chi-cuadrado según los grados de libertad respectivos y el nivel de confianza preseleccionado.

3. Prueba de hipótesis (con interés en la posición del parámetro poblacional) para comparar el estimador muestral (media) que estimaría el poblacional versus el parámetro poblacional de referencia conocido o hipotético esperado, empleando en dependencia del tamaño muestral: la prueba de $Z \sim N(0, 1)$ para un nivel de confianza preseleccionado o la prueba de t-student según los grados de libertad respectivos y el nivel de confianza preseleccionado.
4. Prueba de hipótesis con interés en la dispersión del parámetro poblacional para comparar el estimador muestral (varianza) que estimaría el poblacional versus el parámetro (poblacional) de referencia conocido o hipotético esperado, empleando la prueba de chi-cuadrado según los grados de libertad respectivos y el nivel de confianza preseleccionado.

   Si IV.2 una muestra con una variable cuantitativa aleatoria que no sigue el modelo de distribución normal (tiene libre distribución):

   – Empleo de métodos no paramétricos con pruebas de hipótesis para comparar la distribución de los valores de la variable cuantitativa de la muestra expresados por rangos en relación a otro grupo de valores de dicha variable predeterminados o de referencia mediante:
   – Prueba de la U de Mann-Whitney-Kendall o de la suma de los rangos de Wilcoxon.

   Si III.2 una muestra con una variable aleatoria cualitativa, entonces seguir con:

V. ¿Modelo de la distribución de frecuencias que sigue la variable aleatoria cualitativa nominal a emplear?:

1. Binomial (para eventos con desenlace común).
2. De Poisson (para eventos con desenlace no común).
3. De aproximación a la Poisson para valores observados relacionados con datos expresados en forma de: eventos/personas-tiempo de seguimiento (para tasas de incidencia).

   Si V.1 una muestra con una variable aleatoria cualitativa que sigue el modelo de distribución de frecuencias binomial, entonces seguir con "opciones de técnicas bioestadísticas investigativas de manejo para una muestra una variable cualitativa que sigue una distribución binomial por aproximación de la distribución normal estándar: $Z \sim N(0,1)$:

   – Estimación por intervalo de confianza (IC) de un parámetro poblacional (proporción) a partir del estimados muestral empleando la aproximación de la distribución normal estándar: $Z \sim N(0,1)$ para un nivel de confianza preseleccionado, en muestras grandes ($npq > 5$).
   – Prueba de hipótesis con interés en la posición del parámetro para comparar el estimados muestral (proporción) que estima uno poblacional *versus* el parámetro poblacional de referencia conocido o hipotético esperado, empleando

la aproximación de la normal estándar Z ~ N (0, 1) para un nivel de confianza preseleccionado en muestras grandes (npq > 5).

Si V.2 una muestra con una variable cualitativa que sigue modelo de distribución de frecuencias tipo Poisson en sus valores (para eventos con desenlace no común y datos en forma de eventos/número de personas-tiempo de seguimiento), entonces seguir con:

- Prueba de hipótesis para comparar el número de eventos promedio esperados en una población estimados de una muestra *versus* el número de eventos promedio esperados en una población de referencia, hipotética o definida por un procedimiento teórico si asumimos la aproximación de la distribución normal estándar a la distribución de Poisson y si el número promedio de eventos poblacional de referencia ($\mu_0 > 10$), empleando para el contraste una prueba de chi-cuadrado con 1 grado de libertad para un nivel de confianza preseleccionado en muestras grandes.

  Si V.3 una muestra con una variable cualitativa con aproximación al modelo de distribución de Poisson para datos de esa variable expresados como eventos observados según el número de personas-tiempo de seguimiento, entonces seguir con:

- Prueba de hipótesis para comparar la densidad de incidencia o tasa de incidencia de una población estimada de una muestral *versus* la tasa de incidencia de una población de referencia conocida, hipotética esperada o definida por un procedimiento teórico asumiendo la aproximación de la distribución normal-estándar a la distribución de Poisson para ($\mu_0 = tID_0 > 10$), se emplea para el contraste un prueba de chi-cuadrado con 1 grado de libertad para un nivel de confianza predeterminado, en muestras grandes.

  Si I.1 y II.2 una muestra con dos variables, métodos bivariados, entonces seguir con:

VI. ¿Tipo de combinación bivariada que desea seleccionar en una muestra?:

1. Variable dependiente: cuantitativa continua, variable independiente: cuantitativa continua.
2. Variable dependiente: cuantitativa continua, variable independiente: cualitativa nominal.
3. Variable dependiente: cuantitativa continua, variable independiente: cualitativa ordinal.
4. Variable dependiente: cualitativa nominal, variable independiente: cualitativa nominal.

   Entonces seguir con:

   Si VI.1 una muestra con variable dependiente: cuantitativa continua; variable independiente: cuantitativa continua, entonces seguir con:

   A) ¿Tipo de modelos de distribución para los datos de las variables cuantitativas?:

1. Siguen el modelo de distribución normal o es aplicable el teorema del límite central.
2. No siguen el modelo de distribución normal o no es aplicable el teorema del límite central.

   Si VI.1, A)1 una muestra con dos variables cuantitativas aleatorias que siguen el modelo de distribución normal o es aplicable el teorema del límite central a sus datos, entonces seguir con:

Opciones de técnicas bioestadísticas investigativas de manejo para una muestra con dos variables aleatorias: variable dependiente: cuantitativa continua; variable independiente: cuantitativa continua, por inferencia paramétrica:

1. Estimación por intervalo de confianza (IC) del coeficiente de regresión lineal simple poblacional estimado a partir del muestral, empleando la distribución teórica de la t-student según los grados de libertad respectivos y el nivel de confianza preseleccionado.
2. Prueba de hipótesis para evaluar la significación de la correlación lineal simple entre dos variables cuantitativas por medio del coeficiente de correlación lineal simple de Pearson y la prueba de t-student según los grados de libertad respectivos y el nivel de confianza preseleccionado.
3. Prueba de hipótesis para evaluar la significación de la regresión lineal simple entre dos variables cuantitativas por medio del coeficiente de regresión lineal simple y la prueba de t-student según los grados de libertad respectivos y el nivel de confianza preseleccionado.

   Si VI.1, A)2 una muestra con dos variables aleatorias cuantitativas que no siguen el modelo de distribución normal o no es aplicable el teorema del límite central, entonces seguir con:

Opciones de técnicas bioestadísticas investigativas de manejo para una muestra con dos variables: variable dependiente e independiente: cuantitativas; por inferencia no paramétrica:

1. Prueba de hipótesis para evaluar la significación de la correlación por rangos entre dos variables cuantitativas de libre distribución empleando el coeficiente por rangos de Spearman y auxiliándonos del estadístico de contraste t-student según los grados de libertad respectivos y el nivel de confianza preseleccionado.

   Si VI.2 una muestra con una variable dependiente: cuantitativa continua; variable independiente: cualitativa nominal; entonces seguir con:

   A) ¿Tipo de modelo de distribución de frecuencias que sigue la variable aleatoria dependiente cuantitativa? para su determinación se debe realizar una prueba de normalidad y según el resultado obtenido de la distribución de frecuencias de la variable como sigue:

1. Modelo de distribución normal o teorema del límite central es aplicable; entonces realizar inferencia por pruebas paramétricas como sigue:

2. No modelo de distribución normal o teorema del límite central no es aplicable; entonces realizar inferencia por pruebas no paramétricas como sigue:
   Si VI.2, A)1 una muestra con una variable dependiente: cuantitativa continua, variable independiente: cualitativa nominal; inferencia por pruebas paramétricas, entonces seguir con:

Opciones de técnicas bioestadísticas investigativas de manejo de una muestra y dos variables; variable dependiente: cuantitativa continua, variable independiente: cualitativa nominal; inferencia por pruebas paramétricas:

1. Estimación por intervalo de confianza (IC) de la diferencia de dos medias poblacionales estimadas a partir de las dos medias calculadas para la variable cuantitativa a partir de las dos categorías en la variable independiente cualitativa nominal dicotómica, empleando la distribución teórica para la t-student según los grados de libertad respectivos y el nivel de confianza preseleccionado.
2. Estimación intervalo de confianza (IC) de la razón de dos varianzas poblacionales estimadas a partir de las dos muestrales calculadas para la variable dependiente cuantitativa según las dos categorías de la variable independiente: cualitativa nominal dicotómica, empleando la distribución teórica de la F-Snedecor según los grados de libertad respectivos y el nivel de confianza preseleccionado.
3. Prueba de hipótesis para evaluar la significación de la diferencia entre las dos medias poblacionales estimadas de las dos medias muestrales respectivas calculadas para la variable dependiente cuantitativa a partir de las dos categorías en la variable independiente cualitativa dicotómica (que permite formar dos grupos de datos), empleando el estadístico de contraste t-student según los grados de libertad respectivos y el nivel de confianza preseleccionado.
4. Prueba de hipótesis para evaluar la significación de la diferencia entre las más de dos medias poblacionales estimadas de las más de dos medias muestrales respectivas calculadas para la variable dependiente cuantitativa a partir de las más de dos categorías en la variable independiente cualitativa politómica (que permite formar más de dos grupos de datos), por medio del ANOVA de una vía, empleando el estadístico de contraste F-Snedecor según los grados de libertad respectivos y el nivel de confianza preseleccionado.
5. Prueba de hipótesis para evaluar la significación de la razón entre dos varianzas poblacionales, estimadas de las dos varianzas muestrales respectivas calculadas para la variable dependiente cuantitativa a partir de las dos categorías en la variable independiente: cualitativa nominal dicotómica (que permite formar dos grupos de datos), empleando el estadístico de contaste F-Snedecor según los grados de libertad respectivos y el nivel de confianza preseleccionado o también se puede emplear la prueba de Levene como alternativa en este propósito.

Si VI.2, A)2 una muestra una variable dependiente: cuantitativa continua, variable independiente: cualitativa nominal; inferencia por pruebas no paramétricas, entonces seguir con:

Opciones de técnicas bioestadísticas investigativas de manejo de una muestra y dos variables; variable dependiente: cuantitativa continua, variable independiente: cualitativa nominal; inferencia por pruebas no paramétricas:

1. Prueba de la U de Mann-Whitney-Kendall, para comprobar la equidistribucion de los datos de una variable dependiente cuantitativa con libre distribución empleando el estadístico U, en dos grupos independientes o muestras a partir de dos nominaciones o categorías en la variable independiente (la versión no paramétrica de la habitual prueba de t-student).
2. Prueba de la suma de rangos de Wilcoxon, para comprobar la equidistribución de los datos de una variable dependiente cuantitativa de libre distribución utilizando la sumatoria de sus rangos y auxiliándonos del estadístico T de Wilcoxon, según los dos grupos formados a partir de dos nominaciones en la variable Independiente.
3. Prueba no paramétrica de la mediana para comparar la equidistribución los datos de una variable dependiente cuantitativa con distribución libre según el valor de la mediana calculada para los mismos, a partir de dos nominaciones en la variable independiente, auxiliándonos de una tabla de contingencia 2x2 simple y empleo del estadístico de contraste chi-cuadrado para 1 grado de libertad y el nivel de confianza preseleccionado. Su principal uso es cuando existen valores atípicos (*outliers*) en la distribución de frecuencias de esos datos o la sospecha de sesgo en obtención de los mismos.
4. Prueba no paramétrica de Kruskal-Wallis para comparar la equidistribución de los datos de una variable dependiente cuantitativa continua con distribución libre, a partir de las más de dos nominaciones en la variable independiente (que permite formar más de dos grupos), empleando el estadístico K que sigue una distribución chi-cuadrado según los grados de libertad respectivos y el nivel de confianza preseleccionado.

   Si VI.3 una muestra con una variable dependiente: cuantitativa continua, variable independiente: cualitativa ordinal; entonces, seguir con:

   A) ¿Tipo de modelo de distribución de frecuencias que sigue la variable aleatoria dependiente cuantitativa? para su determinación se deberá realizar una prueba de normalidad y según el resultado obtenido de la distribución de frecuencias de la variable como sigue:

1. Modelo de distribución normal o teorema del límite central es aplicable, entonces realizar inferencia por pruebas paramétricas como sigue:

2. No modelo de distribución normal o teorema del límite central no es aplicable; entonces realizar inferencia por pruebas no paramétricas como sigue:
   Si VI.3, A)1 una muestra con una variable dependiente: cuantitativa continua, variable independiente: cualitativa ordinal, inferencia por pruebas paramétricas, entonces seguir con:

Opciones de técnicas bioestadísticas investigativas de manejo de una muestra y dos variables; variable dependiente: cuantitativa continua, variable independiente: cualitativa ordinal, inferencia por pruebas paramétricas:

1. Prueba de hipótesis para evaluar la significación de la diferencia entre las más de dos medias poblacionales estimadas de las más de dos medias muestrales respectivas calculadas para la variable dependiente cuantitativa a partir de las más de dos categorías en la variable independiente cualitativa ordinal (que permite formar más de dos grupos de datos), por medio del ANOVA de una vía, empleando el estadístico de contraste F-Snedecor según los grados de libertad respectivos y el nivel de confianza preseleccionado.
   Si VI.3, A)2 una muestra con una variable dependiente: cuantitativa continua, variable independiente: cualitativa ordinal; inferencia por pruebas no paramétricas, entonces seguir con:

Opciones de técnicas bioestadísticas investigativas de manejo de una muestra y dos variables; variable dependiente: cuantitativa continua, variable independiente: cualitativa ordinal, inferencia por pruebas no paramétricas:

1. Prueba no paramétrica de Kruskal-Wallis para comparar la equidistribución de los datos de una variable dependiente cuantitativa continua con distribución libre, a partir de las más de dos categorías ordinales en la variable independiente cualitativa ordinal (que permite formar más de dos grupos de datos), empleando el estadístico K que sigue una distribución chi-cuadrado según los grados de libertad respectivos y el nivel de confianza preseleccionado.
   Si VI.4 una muestra con variable dependiente: cualitativa nominal, variable independiente: cualitativa nominal, entonces seguir con:

Opciones de técnicas bioestadísticas investigativas de manejo de una muestra y dos variables; variable dependiente: cualitativa nominal, variable independiente: cualitativa nominal dicotómica:

1. Estimación por intervalo de confianza (IC) de la diferencia de dos proporciones poblacionales estimadas de las dos proporciones muestrales respectivas a partir de las dos categorías o nominaciones en la variable independiente y una categoría en la variable dependiente (la presencia del evento), empleando la aproximación de la distribución normal estándar $Z \sim N(0, 1)$ para un nivel de confianza preseleccionado.
2. Prueba de hipótesis de la diferencia de dos proporciones poblacionales estimadas de las dos proporciones muestrales respectivas formadas a partir de las dos categorías o nominaciones en la variable independiente y una categoría en la

variable dependiente (la presencia del evento), empleando la aproximación de la distribución normal estándar Z ~ N(0, 1) para un nivel de confianza preseleccionado.

3. Prueba de hipótesis para evaluar la significación en la asociación o independencia entre la variable independiente (presencia de exposición) y la variable dependiente (presencia de un evento), auxiliándonos de tablas de contingencia 2x2 simples y empleando la prueba de chi-cuadrado (de Pearson, corregido de Yates o prueba exacta de Fisher) para un nivel de confianza preseleccionado.
4. Prueba de hipótesis para evaluar la significación en la asociación entre la variable independiente (presencia de exposición) y la variable dependiente (presencia de un evento) con la presencia de covariable cualitativa dicotómica, auxiliándonos de tablas de contingencia 2x2 estratificadas y empleando la prueba de chi-cuadrado de Mantel-Haenszel de asociación según los grados de libertad respectivos y el nivel de confianza preseleccionado.
5. Prueba de hipótesis para evaluar la significación en la homogeneidad entre los estratos formados según una covariable cualitativa dicotómica, a partir del análisis de la asociación entre la variable independiente (presencia de exposición) y la variable dependiente (presencia de un evento), con el auxilio de tablas de contingencia 2x2 estratificadas y empleando la prueba de chi-cuadrado de homogeneidad entre estratos, según los grados de libertad respectivos y el nivel de confianza preseleccionado.
6. Prueba de hipótesis para evaluar la homogeneidad entre los niveles de la variable independiente cualitativa ordinal (presencia de exposición por niveles) y la variable dependiente (presencia de un evento), con el auxilio de las tablas de contingencia 2xn simple y empleando la prueba de chi-cuadrado de homogeneidad según los grados de libertad respectivos y el nivel de confianza preseleccionado.
7. Prueba de hipótesis para evaluar la significación en la tendencia lineal entre los niveles de la variable independiente cualitativa ordinal (presencia de exposición por niveles) y la variable dependiente (presencia de un evento), con el auxilio de las tablas de contingencia 2xn simple y empleando la prueba de chi-cuadrado para tendencia lineal según los grados de libertad respectivos y el nivel de confianza preseleccionado.
8. Prueba de hipótesis para evaluar la significación en la asociación entre la variable independiente (presencia de exposición) y la variable dependiente expresada como tasa de incidencia (número de casos o eventos según el número de personas por un tiempo de seguimiento), auxiliándonos de tablas de contingencia 2x2 simples y empleando la prueba de z por aproximación de la normal estándar N (0, 1) para un nivel de confianza preseleccionado.
9. Prueba de hipótesis para evaluar la significación en la asociación entre la variable independiente (presencia de exposición) y la variable dependiente: expresada como tasa de incidencia (número de casos o eventos según el número de personas por un tiempo de seguimiento), con presencia de covariable cualitativa dico-

tómica, auxiliándonos de tablas de contingencia 2x2 estratificadas y empleando la prueba de chi-cuadrado de Mantel-Haenszel de asociación según los grados de libertad respectivos y el nivel de confianza preseleccionado.

10. Prueba de hipótesis para evaluar la significación en la homogeneidad entre los estratos formados según una covariable cualitativa dicotómica, a partir del análisis de la asociación entre la variable independiente (presencia de exposición) y la variable dependiente expresada como tasa de incidencia (número de casos o eventos según el número de personas por un tiempo de seguimiento), auxiliándonos de tablas de contingencia 2x2 estratificadas y empleando la prueba de chi-cuadrado de homogeneidad entre estratos, según los grados de libertad respectivos y el nivel de confianza preseleccionado.
11. Prueba de hipótesis para evaluar la significación en la homogeneidad entre los niveles de la variable independiente (presencia de exposición por niveles) y la variable dependiente expresada como tasa de incidencia (número de casos o eventos según el número de personas por un tiempo de seguimiento), auxiliándonos de tablas de contingencia 2xn simples y empleando la prueba de chi-cuadrado de homogeneidad entre niveles, según los grados de libertad respectivos y el nivel de confianza preseleccionado.
12. Prueba de hipótesis para evaluar la significación en la tendencia lineal entre los niveles de la variable independiente cualitativa ordinal (presencia de exposición por niveles) y la variable dependiente expresada en forma de tasas de incidencia; auxiliándonos de las tablas de contingencia 2xn simple y empleando la prueba de chi-cuadrado para tendencia lineal según los grados de libertad respectivos y el nivel de confianza preseleccionado.

Para recordar

El empleo de tablas de contingencia es el método graficado para representar los datos de dos variables cualitativas y guardan relación con el tipo de prueba de hipótesis estadísticas o el estadístico de prueba a emplear para el análisis de las asociaciones significativas entre las mismas, así se observa:

- Para tablas de 2x2 simples (dos variables cualitativas nominales dicotómicas):
  - Prueba de asociación *versus* independencia entre las dos variables cualitativas dicotómicas (chi-cuadrado sin corrección de Pearson).
  - Prueba de asociación *versus* independencia entre las dos variables cualitativas dicotómicas (chi-cuadrado con corrección de Yates).
  - Prueba de la probabilidad exacta de Fisher para la asociación entre las dos variables cualitativas dicotómicas.
- Para tablas de 2x2 estratificadas (dos variables cualitativas nominales dicotómicas estratificadas por una tercera):

- Prueba de asociación de Mantel-Haenszel.
- Prueba de homogeneidad entre estratos.

– Para tablas de 2xN simples (una variable dependiente cualitativa nominal dicotómica y otra independiente cualitativa con ordinalidad por niveles N):
  - Prueba de homogeneidad entre los niveles.
  - Prueba de tendencia lineal entre los niveles de la variable independiente para el evento en la dependiente.

– Para tablas de 2xN estratificadas (lo mismo al anterior, pero estratificado por una tercera variable nominal o factor de estratificación): prueba de tendencia lineal entre los niveles de la variable independiente para el evento en la dependiente, según estratos.

– Para tablas de NxM (N número de filas mayor que dos nominaciones; M número de columnas mayor que dos nominaciones (para variables dependiente e independiente: nominales politómicas):
  - Prueba de asociación empleando dos variables cualitativas con dos o más categorías o niveles.
  - Prueba de la razón de verosimilitudes para datos nominales.
  - Prueba de tendencia lineal para datos ordinales.

13. Determinar la reproducibilidad, concordancia o acuerdo en el comportamiento de dos variables cualitativas dicotómicas (para dos pruebas diagnósticas con resultado probable dicotómico), con el auxilio para su representación gráfica de una tabla de contingencia 2x2 simple y se emplea el índice de acuerdo o concordancia KAPPA y se evalúa la significación de este índice mediante la prueba de Z por la aproximación de la distribución normal estándar $Z \sim N(0, 1)$ según el nivel de confianza preseleccionado en muestras grandes.

    Si I.1 y II.3 una muestra y más de dos variables (métodos multivariados) entonces seguir con:

VII. ¿Tipo de combinación multivariada a seleccionar?

1. De dependencia: una variable dependiente y dos o más variables independientes.
2. De dependencia: más de una variable dependiente y una o más variables independientes.
3. De interdependencias entre las variables.

    Si VII.1 de dependencia: una muestra con una variable dependiente y dos o más variables independientes; entonces, seguir con:

    ¿Forma de expresión de la variable dependiente?

    A) Variable dependiente expresada por un valor numérico o una categoría.

    B) Variable dependiente expresada mediante una razón de sus valores.

    Si VII.1, A) entonces seguir con:

1. ¿Tipo de combinación a utilizar en una muestra y más de dos variables?

    a) Variable dependiente: cuantitativa continua, variables independientes: cuantitativas continuas (que siguen modelo normal de distribución de frecuencias.

b) Variable dependiente: cuantitativa continua, variables independientes: cualitativas nominales.
c) Variable Dependiente: Cuantitativa continua, Variables Independientes: Cualitativas nominales y cuantitativa continua.
d) Variable dependiente: cualitativa nominal, variables independientes: cualitativas nominales.
e) Variable dependiente: cualitativa nominal, variables independientes: cuantitativas.

Si VII.1, A)1.a) manejo de una muestra, variable dependiente: cuantitativa continua, variables independientes: cuantitativas continuas (que siguen un modelo normal en su distribución de frecuencias), entonces seguir con:

Opciones de técnicas bioestadísticas investigativas de manejo de una muestra con una variable dependiente: cuantitativa continua, variables independientes: cuantitativas continuas (que siguen un modelo normal en su distribución de frecuencias):

1. Con interés en la posición del parámetro poblacional para una muestra aleatoria, realizar análisis de la correlación lineal múltiple utilizando el estadígrafo: coeficiente de correlación múltiple con empleo del estadístico de prueba o contraste F para el contraste de hipótesis general en la evaluación de la significación estadística del resultado obtenido de las variables independientes en conjunto en relación con la dependiente para los grados de libertad correspondientes y un nivel de confianza preseleccionado.

Para recordar

- Coeficiente de correlación múltiple (R): expresa el grado o intensidad de la asociación (correlación) de la variable dependiente y todas las variables independientes tomadas en conjunto. Sus valores pueden ser $0 < R < 1$, cuanto más alto sea su valor, las variables independientes estarán más correlacionadas con la variable dependiente y explican en mayor medida sus fluctuaciones (varianza). Sus posibles valores a interpretar son:
  - Valor 0: no hay correlación entre las variaciones de las variables independientes y la dependiente.
  - Valores que se acercan a 1 o –1: hay intensidad en la correlación de la dependiente con las independientes.
  - Valor 1: correlación perfecta positiva, cuando varían las variables independientes en su conjunto lo hace en la misma intensidad y en el mismo sentido la dependiente.
  - Valor –1: correlación perfecta negativa, cuando varían las variables independientes en su conjunto lo hace con la misma intensidad, pero en sentido inverso la dependiente.

    El signo determina que si es positivo (+) las variables independientes y dependiente aumentan o disminuyen a la vez y si el signo es negativo (–): significa que a medida que unas aumentan (independientes) la otra disminuye.

- Coeficiente de determinación múltiple o correlación parcial (R^2): expresa la varianza explicada de la variable dependiente por todas las independientes, es decir, el porcentaje de variación en la dependiente es debido a todas las independientes en conjunto consideradas en el análisis. Sus posibles valores son $0 < R^2 < 100$. Sus posibles valores a interpretar son si (R^2) > 0,50 o 50 %, más de la mitad de los cambios en la variable dependiente se debe a las variaciones en las independientes de conjunto.
- Coeficiente de regresión múltiple (β): expresa cuanto puede variar la variable dependiente producto al cambio de valor en una unidad de una variable independiente, si se controla o mantiene en un valor fijo el resto de las independientes.

2. Con interés en la posición del parámetro poblacional, para una muestra aleatoria o intencionada realizar análisis de regresión lineal múltiple o regresión múltiple con empleo del estadístico de prueba F para el contraste de hipótesis general y F parcial en la evaluación de la significación estadística del resultado obtenido para los coeficientes de regresión de cada variable independiente con los grados de libertad correspondientes y un valor de alfa prefijado.

Si VII.1, A)1.b) manejo de una muestra con una variable dependiente: cuantitativa continua (que sigue un modelo normal de distribución de frecuencias), variables independientes: cualitativas nominales, entonces seguir con:

Opciones de técnicas bioestadísticas investigativas de manejo de una muestra con una variable dependiente: cuantitativa continua (que sigue un modelo normal de distribución de frecuencias), variables independientes: cualitativas nominales:

1. Prueba de hipótesis con interés en la posición del parámetro poblacional para datos no apareados o no relacionados, realizar la comparación de más de dos medias poblacionales estimadas de las muestrales calculadas con los datos de la variable cuantitativa (dependiente) a partir de las nominaciones para cada una de las variables cualitativas independientes empleando el estadísticos de prueba o contraste F-Snedecor por ANOVA de una vía o factorial (para datos no repetidos) en la evaluación de la significación estadística del resultado obtenido con los grados de libertad correspondientes y un nivel de confianza preseleccionado.
2. Prueba de hipótesis con interés en la posición del parámetro poblacional para datos apareados o relacionados, realizar la comparación de más de dos medias poblacionales estimadas de las muestrales calculadas, con los datos repetidos de la variable cuantitativa (dependiente) a partir de una nominación de la variable cualitativa (independientes), empleando estadísticos de prueba o contraste F-Snedecor por ANOVA de una vía o factorial (para datos repetidos), en la evaluación de la significación estadística del resultado obtenido con los grados de libertad correspondientes y un nivel de confianza preseleccionado.

Si VII.1, A)1.c) manejo de una muestra con una variable dependiente: cuantitativa continua (que sigue un modelo normal de distribución de frecuencias), variables independientes: cualitativas nominales y cuantitativa continua; entonces seguir con opciones de técnicas bioestadísticas investigativas de manejo de una muestra con una variable dependiente: cuantitativa continua (que sigue un modelo normal de distribución de frecuencias), variables independientes: cualitativas nominales y cuantitativa continua:

1. Con interés en la posición del parámetro poblacional, realizar análisis de la covarianza (ANCOVA) a partir de variables independientes cualitativas nominales y una covariable cuantitativa continua apoyándonos en el coeficiente de regresión múltiple y empleando el estadístico de prueba F para el contraste de hipótesis general en la evaluación de la significación estadística del resultado obtenido con los grados de libertad correspondientes y un valor de alfa prefijado.

Si VII.1, A)1.d) Manejo de una muestra con una variable dependiente: cualitativa nominal, variables Independientes: cualitativas nominales, entonces seguir con opciones de técnicas bioestadísticas investigativas de manejo de una muestra con una variable dependiente: cualitativa nominal, variables independientes: cualitativas nominales:

1. Con interés en la posición del parámetro poblacional dependiente del tiempo, realizar prueba de hipótesis para evaluar la significación en la comparación de tasas de supervivencia acumulada para la variable dependiente según nominaciones de las variables independientes a través de tablas de vida con empleo de la prueba de significación para el *log-rank*.
2. Con interés en la posición del parámetro poblacional no dependiente del tiempo, realizar análisis estratificado para determinar asociaciones de las variables independientes con la variable dependiente según el uso de covariables y el empleo del estadístico de prueba o contraste chi- cuadrado (ajustado) de Mantel-Haenzsel en la evaluación de la significación estadística del resultado obtenido para los grados de libertad correspondientes y el nivel de confianza preseleccionado.

Si VII.1, A)1.d) manejo de una muestra con una variable dependiente: cualitativa nominal, variables independientes: cuantitativas, entonces seguir con opciones de técnicas bioestadísticas investigativas de manejo de una muestra con una variable dependiente: cualitativa nominal, variables independientes: cuantitativas:

1. Con interés en la posición del parámetro poblacional no dependiente del tiempo, realizar análisis discriminante, estimando la función de discriminación para los diferentes grupos de la variable dependiente empleando las diferencias en las puntuaciones discriminantes de los grupos según las variables independientes cuantitativas y apoyándonos en los coeficientes y cargas discriminantes respectivas de las variables predictoras independientes con la evaluación de su significación estadística por el estadístico de prueba F para su contraste de hipótesis y

emplear la prueba de razón de máxima probabilidad en la evaluación de la significación estadística de la predicción del modelo en la variable dependiente.

Si VII.1, B) manejo de una muestra con una variable dependiente: cualitativa nominal, variables independientes: cuantitativas continuas y cualitativas nominales (transformadas a *dummy* estas últimas), entonces seguir con opciones de técnicas bioestadísticas investigativas de manejo de una muestra con una variable dependiente: cualitativa nominal, variables independientes: cuantitativas continuas y cualitativas nominales (transformadas a *dummy* estas últimas):

1. Con interés en la posición del parámetro poblacional dependiente del tiempo, realizar regresión de Cox se emplea el método de riesgos proporcionales y estimando la razón de tasas instántaneas del evento (*hazards ratio*) en la variable dependiente, apoyados en la conformación de la ecuación de regresión en los coeficientes de regresión beta respectivos para las distintas variables independientes predictoras con la evaluación de la significación de los mismos por el estadístico de prueba F y con el auxilio de la prueba de razón de máxima probabilidad en la evaluación de la significación estadística de la predicción de la razón de riesgos (*hazard ratio*) por el modelo en la variable dependiente.
2. Con interés en la posición del parámetro poblacional no dependiente del tiempo, realizar análisis de regresión logística, estimando la razón de productos cruzados (OR) a partir de los coeficientes de regresión beta de las diferentes variables independientes predictoras y auxiliándonos de la prueba de razón de máxima probabilidad en la evaluación de la significación estadística de la predicción del modelo en la variable dependiente.

   Si VII.2 manejo de una muestra con más de una variable dependiente y una o más variables independientes, entonces seguir con opciones de técnicas bioestadísticas investigativas de manejo de una muestra con más de una variable dependiente y una o más variables independientes:

   A) Determinar diferencias entre grupos a través de múltiples variables dependientes (medidas por intervalos o razón) a partir de las variables independientes categóricas por análisis múltiple de varianzas (MANOVA).
   B) Determinar diferencias entre grupos a través de múltiples variables dependientes (medidas por intervalos o razón) a partir de las variables independientes categóricas y de covariable cuantitativa por análisis múltiple de covarianzas (MANCOVA).
   C) Determinación de la asociación lineal de dependencia entre un conjunto de variables independientes (cuantitativas) con uno respectivo de variables dependientes (cuantitativas) por análisis de correlación canónica.

   Si VII.3 relación de interdependencias entre las múltiples variables; entonces seguir con opciones de técnicas bioestadísticas investigativas multivariadas a partir de interdependencias entre más de dos variables:

A) Si se desea la reducción de los datos de un grupo de variables cuantitativas (paramétricas) para establecer que causas latentes (factores) o variables resumen causan la correlación entre las variables observadas se puede realizar un análisis de componentes factoriales, que es una técnica para formar nuevas variables.

B) Si se desea la reducción de datos de un grupo de variables cuantitativas (paramétricas) y construir combinaciones lineales de las variables iniciales que expliquen la mayor parte de la información contenida en esas variables, dichas combinaciones formadas son el objetivo del análisis por componentes principales, el cual es una técnica para formar nuevas variables.

C) Si se desea la reducción de datos de un grupo de variables cualitativas (no paramétricas) para identificar dimensiones o variables latentes a partir de la relación entre dos o más variables cualitativas observadas, se puede emplear el análisis de correspondencias, que es una técnica para formar nuevas categorías de variables.

D) Si desea utilizar las proximidades entre objetos para realizar una representación espacial de los mismos e identificar las dimensiones subyacentes que explican por ejemplo los criterios que utilizan los individuos para considerar que diferentes objetos son parecidos o distintos se puede emplear el análisis de escalamiento multidimensional, que es una técnica de clasificación y forma grupos de objetos que emplean variables cuanti-cualitativas.

E) Si desea identificar grupos homogéneos de sujetos u objetos a partir de su puntuación en una serie de variables (cuantitativas y cualitativas), siendo estos grupos a su vez, muy diferentes entre sí y con una alta homogeneidad interna, se puede realizar el análisis de conglomerados (clúster), que constituye una técnica de clasificación de datos que agrupa observaciones y emplea variables en la misma escala.

Si I.2 dos muestras con una misma variable aleatoria a estudiar en cada muestra; entonces seguir con:

VIII. ¿Tipo de muestras a emplear: pareadas o no pareadas?

1. Pareadas, emparejadas, relacionadas o dependientes (datos de la variable de la muestra, tomados en dos momentos diferentes consecutivos).
2. No pareadas, no emparejadas, no relacionadas o independientes (datos de la variable tomados en un único momento, en dos muestras diferentes).

   Si VIII.1 manejo de dos muestras pareadas con la misma variable, entonces seguir con:

   ¿Tipo de variable aleatoria presente en las muestras pareadas?

   A) Variable cuantitativa.

   B) Variable cualitativa.

Si VIII.1.A) dos muestras pareadas con la misma variable cuantitativa en cada muestra, entonces seguir con:

IX. ¿Tipo de modelo de distribución de frecuencias para los datos obtenidos en la variable cuantitativa? (para su comprobación, realizar pruebas de normalidad, por ejemplo: prueba de D´ Agostino, Shapiro-Wilks, Shapiro-Francia):

1. Variable aleatoria cuantitativa que sigue un modelo de distribución normal o teorema del límite central es aplicable para emplear parámetros (pruebas paramétricas).
2. Variable aleatoria cuantitativa que no sigue un modelo de distribución normal o teorema del límite central no es aplicable para emplear parámetros (pruebas no pramétricas).]

   Si IX.1 dos muestras pareadas con una variable aleatoria cuantitativa que sigue modelo de distribución normal; entonces seguir con opciones de técnicas bioestadísticas investigativas de manejo de dos muestras pareadas, una variable aleatoria cuantitativa por pruebas paramétricas:

   A) Prueba de hipótesis para evaluar la significación de la diferencia de dos medias poblacionales pareadas estimadas de las dos muestrales pareadas respectivas calculadas con los datos de la variable cuantitativa (para una categoría o característica dada), empleando la prueba de t-student pareada según los grados de libertad respectivos y el nivel de confianza preseleccionado.

   Si IX.2 dos muestras pareadas con una variable aleatoria cuantitativa que no sigue el modelo de distribución normal; entonces seguir con opciones de técnicas bioestadísticas investigativas de manejo de dos muestras pareadas con una variable aleatoria cuantitativa por pruebas no paramétricas:

   A) Prueba de hipótesis para evaluar si las observaciones en cada par provienen de distribuciones idénticas o si existe una equidistribución en las dos muestras pareadas de los datos de la variable cuantitativa de libre distribución empleando el procedimiento de la prueba del signo de Wilcoxon y auxiliándonos del estadístico S+ (número de diferencias con signo positivo de las n observaciones por pares).

   B) Prueba de hipótesis para evaluar si las observaciones en cada par provienen de distribuciones idénticas o si existe una equidistribución en las dos muestras pareadas de los datos de la variable cuantitativa de libre distribución empleando el procedimiento de la prueba de rangos y de signos de Wilcoxon y auxiliándonos del estadístico T+ (suma de los rangos positivos para los n pares de observaciones por pares).

   Si VIII.1.B) dos muestras pareadas con la misma variable cualitativa en cada muestra, entonces seguir con opciones de técnicas bioestadísticas investigativas de manejo de dos muestras pareadas con una variable cualitativa nominal dicotómica en cada muestra:

A) Prueba de hipótesis para evaluar la significación en la homogeneidad de las dos muestras pareadas (un caso por un control) para las proporciones formadas a partir de las categorías de la variable cualitativa dicotómica para ambas muestras, empleando la prueba de chi-cuadrado de McNemar según con los grados de libertad respectivos y el nivel de confianza preseleccionado.

B-) Prueba de hipótesis para evaluar la significación en la homogeneidad de las dos muestras pareadas (1 caso x 1 < control < 5) para las proporciones formadas a partir de las dos categorías de la variable cualitativa dicotómica para ambas muestras, empleando la prueba de chi-cuadrado de Mantel-Haenszel según los grados de libertad respectivos y el nivel de confianza preseleccionado.

C) Prueba de hipótesis para evaluar la significación en la homogeneidad de las muestras pareadas para las proporciones calculadas a partir de las más de dos categorías de la variable cualitativa politómica, empleando la prueba de Stuart-Maxwell.

Si VIII.2 dos muestras no pareadas con la misma variable en cada una; entonces seguir con:

X. ¿Tipo de variable aleatoria presente en las dos muestras no pareadas?]

1. Variable cuantitativa.
2. Variable cualitativa.

   Si X.1 dos muestras no pareadas con la misma variable cuantitativa en cada una, entonces seguir con:

XI. ¿Tipo de modelo de distribución de frecuencias para los datos obtenidos en la variable cuantitativa? (para su comprobación, realizar pruebas de normalidad, por ejemplo, prueba de D´ Agostino, Shapiro-Wilks, Shapiro-Francia).]

1. Variable aleatoria cuantitativa que sigue un modelo de distribución normal o teorema del límite central es aplicable para emplear parámetros (pruebas paramétricas).]
2. Variable aleatoria cuantitativa que no sigue un modelo de distribución normal o teorema del límite central no es aplicable para emplear parámetros (pruebas no pramétricas).

   Si XI.1 dos muestras pareadas con una variable aleatoria cuantitativa que sigue modelo de distribución normal, entonces seguir con opciones de técnicas bioestadísticas investigativas de manejo de dos muestras pareadas con una variable aleatoria cuantitativa:

   A) Prueba de hipótesis para evaluar la significación de la diferencia de dos medias poblacionales estimadas de las dos medias muestrales respectivas calculadas para los datos de la variable cuantitativa (para una categoría o característica dada), con desviaciones estándar poblacionales conocidas para muestras grandes ($n \geq 30$), empleando la prueba de z por la aproxima-

ción de la normal estándar Z~ N (0,1) según el nivel de confianza preseleccionado.

B) Prueba de Hipótesis para evaluar la significación de la diferencia de dos medias poblacionales estimadas de las dos medias muestrales respectivas calculadas para los datos de la variable cuantitativa (para una categoría o característica dada), con desviaciones estándar poblacionales desconocidas, pero supuestamente iguales para muestras pequeñas (n < 30) empleando la prueba de t-student según los grados de libertad respectivos y el nivel de confianza preseleccionado.

C) Prueba de hipótesis para evaluar la significación de la diferencia de dos medias poblacionales estimadas de las dos medias muestrales respectivas calculadas para los datos de la variable cuantitativa (para una categoría o característica dada), con desviaciones estándar poblacionales desconocidas y significativamente diferentes, para muestras pequeñas (n < 30) empleando la prueba de t-student según los grados de libertad respectivos y el nivel de confianza preseleccionado.

Si XI.2 dos muestras no pareadas con una misma variable cuantitativa que no sigue la distribución normal de frecuencias, entonces seguir con opciones de técnicas bioestadísticas investigativas de manejo de dos muestras no pareadas con una misma variable aleatoria cuantitativa en cada muestra, inferencia por pruebas no paramétricas":

A) Prueba de hipótesis para evaluar si las observaciones en cada muestra no pareada provienen de distribuciones idénticas o si existe una equidistribución en las dos muestras no pareadas de los datos de la variable cuantitativa de libre distribución empleando el procedimiento de la prueba de la U de Mann-Whitney-Kendall.

B) Prueba de hipótesis para evaluar si las observaciones en cada muestra no pareada provienen de distribuciones idénticas o si existe una equidistribución en las dos muestras no pareadas de los datos de la variable cuantitativa de libre distribución empleando el procedimiento de la prueba de la suma de los rangos de Wilcoxon.

C) Prueba de hipótesis para evaluar si las observaciones en cada muestra no pareada provienen de distribuciones idénticas o si existe una equidistribución en las dos muestras no pareadas de los datos de la variable cuantitativa de libre distribución empleando el procedimiento de la prueba ba de la mediana.

D) Prueba de hipótesis para evaluarsi las observaciones en cada muestra no pareada provienen de distribuciones idénticas o si existe una equidistribución en las dos muestras no pareadas de los datos de la variable cuantitativa de libre distribución empleando el procedimiento de la prueba de tendencia o de rachas de Wald-Wolfowitz.

Si X.2 dos muestras no pareadas con una misma variable cualitativa en cada muestra, entonces seguir con opciones de técnicas bioestadísticas investigativas de manejo de dos muestras no pareadas con una misma variable aleatoria cualitativa en cada muestra:

A) Prueba de hipótesis para evaluar la significación de la diferencia entre dos proporciones poblacionales estimadas de las dos muestrales respectivas, calculadas a partir de las categorías de la variable cualitativa nominal dicotómica para cada muestra prefiriendo el uso de muestras grandes (npq > 5) con el empleo de la prueba de Z por la aproximación de la distribución normal estándar Z ~ N (0, 1) según el nivel de confianza preseleccionado.

B) Prueba de hipótesis para evaluar la significación de la homogeneidad en las dos poblaciones estimadas de las dos muestrales no pareadas para las proporciones formadas a partir de las categorías de la variable cualitativa nominal dicotómica o politómica empleando la prueba de chi-cuadrado de Pearson con los grados de libertad respectivos y el nivel de confianza preseleccionado.

D) Prueba de hipótesis para evaluar la significación de la diferencia de dos tasas de incidencia poblacionales estimadas de las dos tasas de incidencia muestrales respectivas al ser comparadas (dos grupos diferentes expuestos) empleando la prueba de Z por la aproximación de la distribución normal a la binomial según el nivel de confianza preseleccionado en muestras grandes o el método exacto a partir del p-valor para muestras pequeñas o con número de eventos observados muy pequeños.

Si I.3 más de dos muestras con una misma variable; entonces seguir con:

XII. ¿Tipo de muestras (más de dos) con la misma variable en cada una?]

1. Más de dos muestras relacionadas o dependientes (son más de dos grupos de datos de la variable tomados en más de dos momentos diferentes consecutivos).
2. Más de dos muestras no relacionadas o independientes (son más de dos grupos de datos de la variable tomados en un mismo momento).]

Si XII.1 más de dos muestras relacionadas o dependientes, entonces elegir:

A) Más de dos muestras relacionadas con una misma variable aleatoria cuantitativa.)]

B) Más de dos muestras relacionadas con una misma variable aleatoria cualitativa.]

Si XII.1.A) más de dos muestras relacionadas o dependientes con una misma variable aleatoria cuantitativa, entonces seguir con:

XIII. ¿Tipo de modelo de distribución de frecuencias para los datos obtenidos en la variable cuantitativa? (para su comprobación, realizar pruebas de normalidad, por ejemplo: prueba de D´ Agostino, Shapiro-Wilks, Shapiro-Francia).

1. Variable aleatoria cuantitativa que sigue un modelo de distribución normal o teorema del límite central es aplicable para emplear parámetros (pruebas paramétricas).
2. Variable aleatoria cuantitativa que no sigue un modelo de distribución normal o teorema del límite central no es aplicable para emplear parámetros (pruebas no paramétricas).]

Si XIII.1 más de dos muestras relacionadas con una variable aleatoria cuantitativa que sigue modelo de distribución normal; entonces seguir con opciones de técnicas bioestadísticas investigativas de manejo de más de dos muestras relacionadas con una variable aleatoria cuantitativa:

A) Prueba de hipótesis para evaluar la significación de la diferencia de más de dos medias poblacionales relacionadas, estimadas de las más de dos medias muestrales respectivas correspondientes a los datos obtenidos en más de dos momentos repetidos de la variable cuantitativa empleando el procedimiento del ANOVA de una vía (para medidas repetidas) y con el auxilio del estadístico de contraste F-Snedecor según los grados de libertad respectivos y el nivel de confianza preseleccionado.

Si XIII.2 más de dos muestras relacionadas con una misma variable aleatoria cuantitativa que no sigue el modelo de distribución normal de frecuencias, entonces seguir con opciones de técnicas bioestadísticas investigativas de manejo de más de dos muestras relacionadas con una misma variable aleatoria cuantitativa que no sigue el modelo de distribución normal de frecuencias, por pruebas no paramétricas:

A) Prueba de hipótesis para evaluar la significación en la equidistribución de los datos de la variable cuantitativa de libre distribución entre las más de dos muestras relacionadas empleando la prueba de Friedman y con el auxilio del estadístico de Friedman que sigue una distribución de chi-cuadrado para muestras grandes con los grados de libertad respectivos y el nivel de confianza preseleccionado.

Si XII.1.B) más de dos muestras relacionadas o dependientes con una misma variable aleatoria cualitativa nominal dicotómica, entonces seguir con:

A) Prueba de hipótesis para evaluar la significación de la homogeneidad en las más de dos muestras relacionadas de las proporciones calculadas según las categorías de la variable cualitativa nominal dicotómica por la prueba de Cochran, con el auxilio del estadístico de contraste de Cochran que sigue una distribución chi-cuadrado según los grados de libertad y el nivel de confianza preseleccionado.

Si XII.2 más de dos muestras no relacionadas o independientes con una misma variable; entonces elegir:

XIV. ¿Tipo de variable aleatoria en las más de dos muestras no relacionadas?

1. Más de dos muestras no relacionadas con una misma variable aleatoria cuantitativa.
2. Más de dos muestras no relacionadas con una misma variable aleatoria cualitativa.

   Si XIV.1 más de dos muestras no relacionadas o independientes con una misma variable aleatoria cuantitativa, entonces seguir con:

XV. ¿Tipo de modelo de distribución de frecuencias para los datos obtenidos en la variable cuantitativa? (para su comprobación, realizar pruebas de normalidad, por ejemplo, prueba de D´ Agostino, Shapiro-Wilks, Shapiro-Francia).

1. Variable aleatoria cuantitativa que sigue un modelo de distribución normal o teorema del límite central es aplicable para emplear parámetros (pruebas paramétricas).
2. Variable aleatoria cuantitativa que no sigue un modelo de distribución normal o teorema del límite central no es aplicable para emplear parámetros (pruebas no paramétricas).

   Si XV.1 más de dos muestras no relacionadas con una variable aleatoria cuantitativa que sigue el modelo de distribución normal, entonces seguir con opciones de técnicas bioestadísticas investigativas de manejo de más de dos muestras no relacionadas con una variable aleatoria cuantitativa:

   A) Prueba de hipótesis para evaluar la significación de la diferencia de más de dos medias poblacionales relacionadas, estimadas de las más de dos medias muestrales respectivas correspondientes a los datos obtenidos al mismo tiempo en las más de dos muestras de la variable cuantitativa, empleando el procedimiento del ANOVA de una vía (para medidas no repetidas) y con el auxilio del estadístico de contraste F-Snedecor según los grados de libertad respectivos y el nivel de confianza preseleccionado.

   Si XV.2 más de dos muestras no relacionadas con una misma variable aleatoria cuantitativa que no sigue el modelo de distribución normal de frecuencias; entonces seguir con opciones de técnicas bioestadísticas investigativas de manejo de más de dos muestras no relacionadas con una misma variable aleatoria cuantitativa que no sigue el modelo de distribución normal de frecuencias, por pruebas no paramétricas:

   A) Prueba de hipótesis para evaluar la significación en la equidistribución de los datos de la variable cuantitativa de libre distribución entre las más de dos muestras no relacionadas, empleando la prueba de Kruskal-Wallis y auxiliándonos del estadístico de Kruskal que sigue una distribución de chi-cuadrado para muestras grandes con los grados de libertad respectivos y el nivel de confianza preseleccionado.

   Si XIV.2, más de dos muestras no relacionadas con una misma variable aleatoria cualitativa en cada muestra; entonces seguir con:

A) Prueba de hipótesis para evaluar la significación de la diferencia de mayor o igual a tres proporciones calculadas a partir de una cierta característica o categoría dada y que son correspondientes con igual número de muestras no relacionadas empleando la prueba de chi-cuadrado según los grados de libertad y el nivel de confianza preseleccionado.
B) Prueba de hipótesis para evaluar la significación de la homogeneidad en las más de dos muestras no relacionadas de las proporciones calculadas según las categorías de la variable cualitativa nominal dicotómica, politómica u ordinal, auxiliándonos de las tablas de contingencia de filas por columnas mayor o igual a dos y empleando la prueba de chi-cuadrado de Pearson de homogeneidad.
C) Prueba de hipótesis para evaluar la significación de la asociación lineal entre: los valores de las proporciones calculadas para una característica o categoría dada y el punto medio correspondiente a las mayores o iguales a las distintas clases o escalas de clasificación de una variable cuantitativa en las más de dos muestras no relacionadas empleando el procedimiento que utiliza el estadístico de contraste chi-cuadrado de Cochran-Armitage según los grados de libertad respectivos y el nivel de confianza preseleccionado en la respectiva prueba de chi-cuadrado.
D) Prueba de hipótesis para evaluar la significación del gradiente de variación de tre o más proporciones correspondientes a igual número de muestras no relacionadas, calculadas a partir de los datos para una variable cualitativa ordinal, empleando el procedimiento que utiliza el estadístico de contraste de Bartholomew.

Anexo 1

Tablas de valores críticos (teóricos-esperados) en diferentes distribuciones de estadísticos

Tabla para la distribución de probabilidad de la normal estándar tipificada Z ~ N (0, 1)

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
3,5	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002

Fuente: Martínez González, M. A., Sánchez Villegas, A. y Faulín Fajardo, F. J. (Eds.) (2006). *Bioestadística amigable*. 2da ed. España: Editorial Díaz de Santos, p. 911.

- Para +Z se hace referencia al área que deja una cola a la derecha o para –Z para el área que deja una cola a la izquierda.
- Dentro de la tabla se presenta el valor de p para + Z o para –Z (área de una sola cola), por lo que es necesario buscar los dos primeros dígitos de Z en la primera columna vertical y el último dígito de Z en la primera fila horizontal.
- En una distribución normal el 50 % de los individuos está por encima de la media (≥ 0,00 desviaciones estándar por encima de la media Z = 0,00, columna 2, fila 1, p = 0,500).
- La proporción de individuos que está al menos a 1,15 desviaciones típicas bajo la media es del 12,5 % (Z = –1,15; columna 7, fila 12, p = 0,1251).

Tabla para representar los valores teóricos de la distribución normal estándar tipificada: Z ~ N (0, 1), dependiendo del valor de p (área bajo la curva), bien en una sola cola ($\alpha/2$) o en cada una de las dos colas (α)

p (2 colas)	p (1 cola)	z
α	α/2	
0,50	0,25	0,6745
0,49	0,245	0,6903
0,48	0,24	0,7063
0,47	0,235	0,7225
0,46	0,23	0,7388
0,45	0,225	0,7554
0,44	0,22	0,7722
0,43	0,215	0,7892
0,42	0,21	0,8064
0,41	0,205	0,8239
0,40	0,2	0,8416
0,39	0,195	0,8596
0,38	0,19	0,8779
0,37	0,185	0,8965
0,36	0,18	0,9154
0,35	0,175	0,9346
0,34	0,17	0,9542
0,33	0,165	0,9741
0,32	0,16	0,9945
0,31	0,155	1,0152
0,30	0,15	1,0364
0,29	0,145	1,0581
0,28	0,14	1,0803
0,27	0,135	1,1031
0,26	0,13	1,1264
0,25	0,125	1,1503

p (2 colas)	p (1 cola)	z
α	α/2	
0,24	0,12	1,1750
0,23	0,115	1,2004
0,22	0,11	1,2265
0,21	0,105	1,2536
0,20	0,10	1,2816
0,19	0,095	1,3106
0,18	0,09	1,3408
0,17	0,085	1,3722
0,16	0,08	1,4051
0,15	0,075	1,4395
0,14	0,07	1,4758
0,13	0,065	1,5141
0,12	0,06	1,5548
0,11	0,055	1,5982
0,10	0,05	1,6449
0,09	0,045	1,6954
0,08	0,04	1,7507
0,07	0,035	1,8119
0,06	0,03	1,8808
0,05	0,025	1,9600
0,04	0,02	2,0537
0,03	0,015	2,1701
0,02	0,01	2,3263
0,01	0,005	2,5758
0,001	0,0005	3,2905
0,0001	0,00005	3,8906

Fuente: Martínez González, M. A., Sánchez Villegas, A. y Faulín Fajardo, F. J. (Eds.) (2006). *Bioestadística amigable.* 2da ed. España: Editorial Díaz de Santos, p. 912.

Nota: Dentro de la tabla se proporciona el valor de Z dependiendo del valor de p (área bajo la curva) bien en una sola cola o en cada una de las dos colas. Esta es la tabla donde se deben buscar los valores de Z para estimar intervalos de confianza o para calcular el tamaño muestral.

- En una distribución normal, hay una probabilidad p = 0,2 de encontrar a un individuo a 0,8416 desviaciones estándar o más por encima de la media (Z = 0,84 6) y una probabilidad de 0,4 de encontrar a alguien que se distancie 0,8416 desviaciones estándar o más de la media, sumando a los que están a igual o mayor de 0,8416 desviaciones estándar por encima y a los que están a menor o igual de 0,8416 desviaciones estándar por debajo de la media (columna 3, fila 11).
- La proporción de individuos que están al menos a 1,96 desviaciones típicas por encima de la media es del 2,5 %. Si se suman los que están por encima y por debajo de 1,96 desviaciones estándar hay un 5 % de individuos a una distancia superior o igual a 1,96 desviaciones estándar de la media (z = 1,96: columna 6, fila 20).

Tabla para la distribución de probabilidad acumulativa 1 – α de la normal estándar tipificada Z ~ N (0, 1)

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
−3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
−3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
−3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
−3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
−3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
−2,9	0,0019	0,0018	0,0017	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
−2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
−2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
−2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
−2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
−2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
−2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
−2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
−2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
−2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
−1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
−1,8	0,0359	0,0352	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
−1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
−1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
−1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
−1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0722	0,0708	0,0694	0,0681
−1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
−1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
−1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
−1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
−0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
−0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
−0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
−0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
−0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
−0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
−0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
−0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,33897	0,3859
−0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
−0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8949	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9278	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

Fuente: Álvarez Cáceres, R. (2007). *Estadística aplicada a las ciencias de la salud*. España: Editorial Díaz de Santos, p. 969.

Tabla para los valores críticos (teóricos-esperados) de la distribución t-student (t), según los grados de libertad (toda la primera columna con valores de v desde 1 hasta 100) y para los valores 1 – α (toda la primera fila desde $t_{0,001}$ hasta $t_{0,999}$)

v	$t_{0.001}$	$t_{0.005}$	$t_{0.010}$	$t_{0.025}$	$t_{0.050}$	$t_{0.100}$	$t_{0.200}$
1	−318.309	−63.657	−31.821	−12.706	−6.314	−3.078	−1.376
2	−22.327	−9.925	−6.965	−4.303	−2.920	−1.886	−1.061
3	−10.215	−5.841	−4.541	−3.182	−2.353	−1.638	−0.978
4	−7.173	−4.604	−3.747	−2.776	−2.132	−1.533	−0.941
5	−5.893	−4.032	−3.365	−2.571	−2.015	−1.476	−0.920
6	−5.208	−3.707	−3.143	−2.447	−1.943	−1.440	−0.906
7	−4.785	−3.499	−2.998	−2.365	−1.895	−1.415	−0.896
8	−4.501	−3.355	−2.896	−2.306	−1.860	−1.397	−0.889
9	−4.297	−3.250	−2.821	−2.262	−1.833	−1.383	−0.883
10	−4.144	−3.169	−2.764	−2.228	−1.812	−1.372	−0.879
11	−4.025	−3.106	−2.718	−2.201	−1.796	−1.363	−0.876
12	−3.930	−3.055	−2.681	−2.179	−1.782	−1.356	−0.873
13	−3.852	−3.012	−2.650	−2.160	−1.771	−1.350	−0.870
14	−3.787	−2.977	−2.624	−2.145	−1.761	−1.345	−0.868
15	−3.733	−2.947	−2.602	−2.131	−1.753	−1.341	−0.866
16	−3.686	−2.921	−2.583	−2.120	−1.746	−1.337	−0.865
17	−3.646	−2.898	−2.567	−2.110	−1.740	−1.333	−0.863
18	−3.610	−2.878	−2.552	−2.101	−1.734	−1.330	−0.862
19	−3.579	−2.861	−2.539	−2.093	−1.729	−1.328	−0.861
20	−3.552	−2.845	−2.528	−2.086	−1.725	−1.325	−0.860
21	−3.527	−2.831	−2.518	−2.080	−1.721	−1.323	−0.859
22	−3.505	−2.819	−2.508	−2.074	−1.717	−1.321	−0.858
23	−3.485	−2.807	−2.500	−2.069	−1.714	−1.319	−0.858
24	−3.467	−2.797	−2.492	−2.064	−1.711	−1.318	−0.857
25	−3.450	−2.787	−2.485	−2.060	−1.708	−1.316	−0.856
26	−3.435	−2.779	−2.479	−2.056	−1.706	−1.315	−0.856
27	−3.421	−2.771	−2.473	−2.052	−1.703	−1.314	−0.855
28	−3.408	−2.763	−2.467	−2.048	−1.701	−1.313	−0.855
29	−3.396	−2.756	−2.462	−2.045	−1.699	−1.311	−0.854
30	−3.385	−2.750	−2.457	−2.042	−1.697	−1.310	−0.854
35	−3.340	−2.724	−2.438	−2.030	−1.690	−1.306	−0.852
40	−3.307	−2.704	−2.423	−2.021	−1.684	−1.303	−0.851
45	−3.281	−2.690	−2.412	−2.014	−1.679	−1.301	−0.850
50	−3.261	−2.678	−2.403	−2.009	−1.676	−1.299	−0.849
60	−3.232	−2.660	−2.390	−2.000	−1.671	−1.296	−0.848
70	−3.211	−2.648	−2.381	−1.994	−1.667	−1.294	−0.847
80	−3.195	−2.639	−2.374	−1.990	−1.664	−1.292	−0.846
90	−3.183	−2.632	−2.369	−1.987	−1.662	−1.291	−0.846
100	−3.174	−2.626	−2.364	−1.984	−1.660	−1.290	−0.845
200	−3.131	−2.601	−2.345	−1.972	−1.652	−1.286	−0.843
500	−3.107	−2.586	−2.334	−1.965	−1.648	−1.283	−0.842
1000	−3.098	−2.581	−2.330	−1.962	−1.646	−1.282	−0.842

ν	$t_{0.800}$	$t_{0.900}$	$t_{0.950}$	$t_{0.975}$	$t_{0.990}$	$t_{0.995}$	$t_{0.999}$
1	1.376	3.078	6.314	12.706	31.820	63.656	318.294
2	1.061	1.886	2.920	4.303	6.965	9.925	22.327
3	0.978	1.638	2.353	3.182	4.541	5.841	10.214
4	0.941	1.533	2.132	2.776	3.747	4.604	7.173
5	0.920	1.476	2.015	2.571	3.365	4.032	5.893
6	0.906	1.440	1.943	2.447	3.143	3.707	5.208
7	0.896	1.415	1.895	2.365	2.998	3.499	4.785
8	0.889	1.397	1.860	2.306	2.896	3.355	4.501
9	0.883	1.383	1.833	2.262	2.821	3.250	4.297
10	0.879	1.372	1.812	2.228	2.764	3.169	4.144
11	0.876	1.363	1.796	2.201	2.718	3.106	4.025
12	0.873	1.356	1.782	2.179	2.681	3.055	3.930
13	0.870	1.350	1.771	2.160	2.650	3.012	3.852
14	0.868	1.345	1.761	2.145	2.624	2.977	3.787
15	0.866	1.341	1.753	2.131	2.602	2.947	3.733
16	0.865	1.337	1.746	2.120	2.583	2.921	3.686
17	0.863	1.333	1.740	2.110	2.567	2.898	3.646
18	0.862	1.330	1.734	2.101	2.552	2.878	3.o10
19	0.861	1.328	1.729	2.093	2.539	2.861	3.579
20	0.860	1.325	1.725	2.086	2.528	2.845	3.552
21	0.859	1.323	1.721	2.080	2.518	2.831	3.527
22	0.858	1.321	1.717	2.074	2.508	2.819	3.505
23	0.858	1.319	1.714	2.069	2.500	2.807	3.485
24	0.857	1.318	1.711	2.064	2.492	2.797	3.467
25	0.856	1.316	1.708	2.060	2.485	2.787	3.450
26	0.856	1.315	1.706	2.056	2.479	2.779	3.435
27	0.855	1.314	1.703	2.052	2.473	2.771	3.421
28	0.855	1.313	1.701	2.048	2.467	2.763	3.408
29	0.854	1.311	1.699	2.045	2.462	2.756	3.396
30	0.854	1.310	1.697	2.042	2.457	2.750	3.385
35	0.852	1.306	1.690	2.030	2.438	2.724	3.340
40	0.851	1.303	1.684	2.021	2.423	2.704	3.307
45	0.850	1.301	1.679	2.014	2.412	2.690	3.281
50	0.849	1.299	1.676	2.009	2.403	2.678	3.261
60	0.848	1.296	1.671	2.000	2.390	2.660	3.232
70	0.847	1.294	1.667	1.994	2.381	2.648	3.211
80	0.846	1.292	1.664	1.990	2.374	2.639	3.195
90	0.846	1.291	1.662	1.987	2.368	2.632	3.183
100	0.845	1.290	1.660	1.984	2.364	2.626	3.174
200	0.843	1.286	1.652	1.972	2.345	2.601	3.131
500	0.842	1.283	1.648	1.965	2.334	2.586	3.107
1000	0.842	1.282	1.646	1.962	2.330	2.581	3.098

Fuente: Canavos, G. C. (1998). *Probabilidad y estadística: aplicaciones y métodos.* Madrid: Editorial McGraw-Hill, pp. 621-622.

En una distribución t-student, por ejemplo, en una distribución muestral de medias con muestral de tamaño 30 (29 grados de libertad) hay una probabilidad igual a 0,05 de encontrar la media de una muestra a 2,0452 errores estándar de la media poblacional o más lejos aún en cualquiera de las direcciones, por arriba o por abajo (es una probabilidad a dos colas).El 97,5 % de las medias de las muestras de tamaño 30 estarán en el intervalo $-\infty$, + 2,0452 errores estándar de la media poblacional.

Tabla para los valores críticos (teóricos-esperados) de la distribución chi-cuadrado (χ^2), según los grados de libertad (en la primera columna con valores de *n* desde 1 hasta120) y para 1 – α (en la primera fila con valores desde 0,005 hasta 0,995)

n \ α	0.005	0.01	0.025	0.05	0.1	0.2	0.3	0.5	0.7	0.8	0.9	0.95	0.975	0.99	0.995
1	0.00004	0.00016	0.00098	0.00393	0.01579	0.06418	0.14847	0.45494	1.07419	1.64237	2.70554	3.84146	5.02389	6.63490	7.87944
2	0.01002	0.02010	0.05064	0.10259	0.21072	0.44629	0.71335	1.38629	2.40795	3.21888	4.60517	5.99146	7.37776	9.21034	10.59663
3	0.07172	0.11483	0.21580	0.35185	0.58437	1.00517	1.42365	2.36597	3.66487	4.64163	6.25139	7.81473	9.34840	11.34487	12.83816
4	0.20699	0.29711	0.48442	0.71072	1.06362	1.64878	2.19470	3.35669	4.87843	5.98862	7.77944	9.48773	11.14329	13.27670	14.86026
5	0.41174	0.55430	0.83121	1.14548	1.61031	2.34253	2.99991	4.35146	6.06443	7.28928	9.23636	11.07050	12.83250	15.08627	16.74960
6	0.67573	0.87209	1.23734	1.63538	2.20413	3.07009	3.82755	5.34812	7.23114	8.55806	10.64464	12.59159	14.44938	16.81189	18.54758
7	0.98926	1.23904	1.68987	2.16735	2.83311	3.82232	4.67133	6.34581	8.38343	9.80325	12.01704	14.06714	16.01276	18.47531	20.27771
8	1.34441	1.64650	2.17973	2.73264	3.48954	4.59357	5.52742	7.34412	9.52446	11.03009	13.36157	15.50731	17.53455	20.09024	21.95495
9	1.73493	2.08790	2.70039	3.32511	4.16816	5.38005	6.39331	8.34283	10.65637	12.24215	14.68366	16.91898	19.02277	21.66599	23.58935
10	2.15586	2.55821	3.24697	3.94030	4.86518	6.17908	7.26722	9.34182	11.78072	13.44196	15.98718	18.30704	20.48318	23.20925	25.18818
11	2.60322	3.05348	3.81575	4.57481	5.57778	6.98867	8.14787	10.34100	12.89867	14.63142	17.27501	19.67514	21.92005	24.72497	26.75685
12	3.07382	3.57057	4.40379	5.22603	6.30380	7.80733	9.03428	11.34032	14.01110	15.81199	18.54935	21.02607	23.33666	26.21697	28.29952
13	3.56503	4.10692	5.00875	5.89186	7.04150	8.63386	9.92568	12.33976	15.11872	16.98480	19.81193	22.36203	24.73560	27.68825	29.81947
14	4.07467	4.66042	5.62873	6.57063	7.78953	9.46733	10.82148	13.33927	16.22210	18.15077	21.06414	23.68479	26.11895	29.14124	31.31935
15	4.60092	5.22935	6.26214	7.26094	8.54676	10.30966	11.72117	14.33886	17.32169	19.31066	22.30713	24.99579	27.48839	30.57791	32.80132
16	5.14221	5.81221	6.90766	7.96165	9.31224	11.15212	12.62435	15.33850	18.41789	20.46508	23.54183	26.29623	28.84535	31.99993	34.26719
17	5.69722	6.40776	7.56419	8.67176	10.08519	12.00227	13.53068	16.33818	19.51102	21.61456	24.76904	27.58711	30.19101	33.40866	35.71847
18	6.26480	7.01491	8.23075	9.39046	10.86494	12.85905	14.43986	17.33790	20.60135	22.75955	25.98942	28.86930	31.52638	34.80531	37.15645
19	6.84397	7.63273	8.90652	10.11701	11.65091	13.71579	15.35166	18.33765	21.68913	23.90042	27.20357	30.14353	32.85233	36.19087	38.58226
20	7.43384	8.26040	9.59078	10.85081	12.44261	14.57844	16.26586	19.33743	22.77454	25.03751	28.41198	31.41043	34.16961	37.56623	39.99685
21	8.03365	8.89720	10.28290	11.59131	13.23960	15.44161	17.18227	20.33723	23.85779	26.17110	29.61509	32.67057	35.47888	38.93217	41.40106
22	8.64272	9.54249	10.98232	12.33801	14.04149	16.31001	18.10072	21.33704	24.93902	27.30145	30.81328	33.92444	36.78071	40.28936	42.79566
23	9.26042	10.19572	11.68855	13.09051	14.84796	17.18351	19.02109	22.33688	26.01857	28.42879	32.00690	35.17246	38.07563	41.63840	44.18128
24	9.88623	10.85636	12.40115	13.84842	15.65868	18.06180	19.94323	23.33673	27.09636	29.55332	33.19624	36.41503	39.36408	42.97982	45.55851
25	10.51965	11.52398	13.11972	14.61141	16.47341	18.93975	20.86703	24.33659	28.17192	30.67520	34.38159	37.65248	40.64647	44.31410	46.92789
26	11.16024	12.19815	13.84390	15.37916	17.29188	19.82019	21.79240	25.33646	29.24533	31.79401	35.56317	38.88514	41.92317	45.64168	48.28988
27	11.80759	12.87850	14.57338	16.15140	18.11390	20.70296	22.71924	26.33634	30.31929	32.91109	36.74122	40.11327	43.19451	46.96294	49.64892
28	12.46134	13.56471	15.30786	16.92788	18.93924	21.58797	23.64746	27.33623	31.39088	34.02657	37.91592	41.33714	44.46079	48.27824	50.99338
29	13.12115	14.25645	16.04707	17.70837	19.76774	22.47506	24.57699	28.33613	32.46117	35.13806	39.08747	42.55697	45.72229	49.58788	52.33562
30	13.78672	14.95346	16.79077	18.49266	20.59923	23.36411	25.50776	29.33603	33.53023	36.25019	40.25602	43.77297	46.97924	50.89218	53.67196
40	20.70654	22.16426	24.43304	26.50930	29.05052	32.34458	34.87594	39.33534	44.16487	47.26854	51.80506	55.75848	59.34171	63.69074	66.76596
50	27.99075	29.70668	32.35736	34.76425	37.68865	41.44921	44.31331	49.33494	54.72279	58.16480	63.16712	67.50481	71.42020	76.15389	79.48998
60	35.53449	37.48485	40.48175	43.18796	46.45889	50.64062	53.80013	59.33467	65.22651	68.97207	74.39701	79.08194	83.29767	88.37942	91.95170
120	83.85157	86.92328	91.57264	95.70464	100.62363	106.80561	111.41857	119.33400	127.61500	132.80628	140.23257	146.56736	152.21140	158.95017	163.64818

Fuente: Botella Rocamora, P., Alacreu García, M. y Martínez Beneito, M. A. (2013-2014). *Apuntes de Estadística en Ciencias de la Salud*. Memorias del Curso: Estadística en Ciencias de la Salud. España: Universidad Cardenal Herrera (Tablas anexas).

En una prueba de chi-cuadrado, con 20 grados de libertad, cuyo valor hallado sea de 34,17, la probabilidad de encontrar un resultado tan alejado o más de la hipótesis nula es de 0,025 (valor p = 0,025). Si se encontrase en la prueba un valor de chi-cuadrado superior a 34,17, el valor p sería inferior a 0,025.

Tabla para la probabilidad exacta de la distribución binomial $Pr(X = k) = \binom{n}{k} p^k q^{n-k}$, donde:

n es el número de intentos u observaciones, k el número de éxitos o eventos observados, p la probabilidad del evento o de éxitos (.05, .10, .15, ..., .50) y q = 1 - p, su complemento.

n	*k*	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
2	0	.9025	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025	.2500
	1	.0950	.1800	.2550	.3200	.3750	.4200	.4550	.4800	.4950	.5000
	2	.0025	.0100	.0225	.0400	.0625	.0900	.1225	.1600	.2025	.2500
3	0	.8574	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250
	1	.1354	.2430	.3251	.3840	.4219	.4410	.4436	.4320	.4084	.3750
	2	.0071	.0270	.0574	.0960	.1406	.1890	.2389	.2880	.3341	.3750
	3	.0001	.0010	.0034	.0080	.0156	.0270	.0429	.0640	.0911	.1250
4	0	.8145	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625
	1	.1715	.2916	.3685	.4096	.4219	.4116	.3845	.3456	.2995	.2500
	2	.0135	.0486	.0975	.1536	.2109	.2646	.3105	.3456	.3675	.3750
	3	.0005	.0036	.0115	.0256	.0469	.0756	.1115	.1536	.2005	.2500
	4	.0000	.0001	.0005	.0016	.0039	.0081	.0150	.0256	.0410	.0625
5	0	.7738	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0313
	1	.2036	.3280	.3915	.4096	.3955	.3602	.3124	.2592	.2059	.1563
	2	.0214	.0729	.1382	.2048	.2637	.3087	.3364	.3456	.3369	.3125
	3	.0011	.0081	.0244	.0512	.0879	.1323	.1811	.2304	.2757	.3125
	4	.0000	.0004	.0022	.0064	.0146	.0283	.0488	.0768	.1128	.1563
	5	.0000	.0000	.0001	.0003	.0010	.0024	.0053	.0102	.0185	.0313
6	0	.7351	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277	.0156
	1	.2321	.3543	.3993	.3932	.3560	.3025	.2437	.1866	.1359	.0938
	2	.0305	.0984	.1762	.2458	.2966	.3241	.3280	.3110	.2780	.2344
	3	.0021	.0146	.0415	.0819	.1318	.1852	.2355	.2765	.3032	.3125
	4	.0001	.0012	.0055	.0154	.0330	.0595	.0951	.1382	.1861	.2344
	5	.0000	.0001	.0004	.0015	.0044	.0102	.0205	.0369	.0609	.0938
	6	.0000	.0000	.0000	.0001	.0002	.0007	.0018	.0041	.0083	.0156
7	0	.6983	.4783	.3206	.2097	.1335	.0824	.0490	.0280	.0152	.0078
	1	.2573	.3720	.3960	.3670	.3115	.2471	.1848	.1306	.0872	.0547
	2	.0406	.1240	.2097	.2753	.3115	.3177	.2985	.2613	.2140	.1641
	3	.0036	.0230	.0617	.1147	.1730	.2269	.2679	.2903	.2918	.2734
	4	.0002	.0026	.0109	.0287	.0577	.0972	.1442	.1935	.2388	.2734
	5	.0000	.0002	.0012	.0043	.0115	.0250	.0466	.0774	.1172	.1641
	6	.0000	.0000	.0001	.0004	.0013	.0036	.0084	.0172	.0320	.0547
	7	.0000	.0000	.0000	.0000	.0001	.0002	.0006	.0016	.0037	.0078
8	0	.6634	.4305	.2725	.1678	.1001	.0576	.0319	.0168	.0084	.0039
	1	.2793	.3826	.3847	.3355	.2670	.1977	.1373	.0896	.0548	.0313
	2	.0515	.1488	.2376	.2936	.3115	.2965	.2587	.2090	.1569	.1094
	3	.0054	.0331	.0839	.1468	.2076	.2541	.2786	.2787	.2568	.2188

n	k	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
	4	.0004	.0046	.0185	.0459	.0865	.1361	.1875	.2322	.2627	.2734
	5	.0000	.0004	.0026	.0092	.0231	.0467	.0808	.1239	.1719	.2188
	6	.0000	.0000	.0002	.0011	.0038	.0100	.0217	.0413	.0703	.1094
	7	.0000	.0000	.0000	.0001	.0004	.0012	.0033	.0079	.0164	.0313
	8	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0007	.0017	.0039
9	0	.6302	.3874	.2316	.1342	.0751	.0404	.0207	.0101	.0046	.0020
	1	.2985	.3874	.3679	.3020	.2253	.1556	.1004	.0605	.0339	.0176
	2	.0629	.1722	.2597	.3020	.3003	.2668	.2162	.1612	.1110	.0703
	3	.0077	.0446	.1069	.1762	.2336	.2668	.2716	.2508	.2119	.1641
	4	.0006	.0074	.0283	.0661	.1168	.1715	.2194	.2508	.2600	.2461
	5	.0000	.0008	.0050	.0165	.0389	.0735	.1181	.1672	.2128	.2461
	6	.0000	.0001	.0006	.0028	.0087	.0210	.0424	.0743	.1160	.1641
	7	.0000	.0000	.0000	.0003	.0012	.0039	.0098	.0212	.0407	.0703
	8	.0000	.0000	.0000	.0000	.0001	.0004	.0013	.0035	.0083	.0176
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0008	.0020
10	0	.5987	.3487	.1969	.1074	.0563	.0282	.0135	.0060	.0025	.0010
	1	.3151	.3874	.3474	.2684	.1877	.1211	.0725	.0403	.0207	.0098
	2	.0746	.1937	.2759	.3020	.2816	.2335	.1757	.1209	.0763	.0439
	3	.0105	.0574	.1298	.2013	.2503	.2668	.2522	.2150	.1665	.1172
	4	.0010	.0112	.0401	.0881	.1460	.2001	.2377	.2508	.2384	.2051
	5	.0001	.0015	.0085	.0264	.0584	.1029	.1536	.2007	.2340	.2461
	6	.0000	.0001	.0012	.0055	.0162	.0368	.0689	.1115	.1596	.2051
	7	.0000	.0000	.0001	.0008	.0031	.0090	.0212	.0425	.0746	.1172
	8	.0000	.0000	.0000	.0001	.0004	.0014	.0043	.0106	.0229	.0439
	9	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0016	.0042	.0098
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010
11	0	.5688	.3138	.1673	.0859	.0422	.0198	.0088	.0036	.0014	.0005
	1	.3293	.3835	.3248	.2362	.1549	.0932	.0518	.0266	.0125	.0054
	2	.0867	.2131	.2866	.2953	.2581	.1998	.1395	.0887	.0513	.0269
	3	.0137	.0710	.1517	.2215	.2581	.2568	.2254	.1774	.1259	.0806
	4	.0014	.0158	.0536	.1107	.1721	.2201	.2428	.2365	.2060	.1611
	5	.0001	.0025	.0132	.0388	.0803	.1321	.1830	.2207	.2360	.2256
	6	.0000	.0003	.0023	.0097	.0268	.0566	.0985	.1471	.1931	.2256
	7	.0000	.0000	.0003	.0017	.0064	.0173	.0379	.0701	.1128	.1611
	8	.0000	.0000	.0000	.0002	.0011	.0037	.0102	.0234	.0462	.0806
	9	.0000	.0000	.0000	.0000	.0001	.0005	.0018	.0052	.0126	.0269
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0007	.0021	.0054
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0005
12	0	.5404	.2824	.1422	.0687	.0317	.0138	.0057	.0022	.0008	.0002
	1	.3413	.3766	.3012	.2062	.1267	.0712	.0368	.0174	.0075	.0029
	2	.0988	.2301	.2924	.2835	.2323	.1678	.1088	.0639	.0339	.0161
	3	.0173	.0852	.1720	.2362	.2581	.2397	.1954	.1419	.0923	.0537
	4	.0021	.0213	.0683	.1329	.1936	.2311	.2367	.2128	.1700	.1208
	5	.0002	.0038	.0193	.0532	.1032	.1585	.2039	.2270	.2225	.1934
	6	.0000	.0005	.0040	.0155	.0401	.0792	.1281	.1766	.2124	.2256
	7	.0000	.0000	.0006	.0033	.0115	.0291	.0591	.1009	.1489	.1934
	8	.0000	.0000	.0001	.0005	.0024	.0078	.0199	.0420	.0762	.1208
	9	.0000	.0000	.0000	.0001	.0004	.0015	.0048	.0125	.0277	.0537
	10	.0000	.0000	.0000	.0000	.0000	.0002	.0008	.0025	.0068	.0161
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0029
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
13	0	.5133	.2542	.1209	.0550	.0238	.0097	.0037	.0013	.0004	.0001
	1	.3512	.3672	.2774	.1787	.1029	.0540	.0259	.0113	.0045	.0016
	2	.1109	.2448	.2937	.2680	.2059	.1388	.0836	.0453	.0220	.0095
	3	.0214	.0997	.1900	.2457	.2517	.2181	.1651	.1107	.0660	.0349
	4	.0028	.0277	.0838	.1535	.2097	.2337	.2222	.1845	.1350	.0873
	5	.0003	.0055	.0266	.0691	.1258	.1803	.2154	.2214	.1989	.1571

n	k	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
	6	.0000	.0008	.0063	.0230	.0559	.1030	.1546	.1968	.2169	.2095
	7	.0000	.0001	.0011	.0058	.0186	.0442	.0833	.1312	.1775	.2095
	8	.0000	.0000	.0001	.0011	.0047	.0142	.0336	.0656	.1089	.1571
	9	.0000	.0000	.0000	.0001	.0009	.0034	.0101	.0243	.0495	.0873
	10	.0000	.0000	.0000	.0000	.0001	.0006	.0022	.0065	.0162	.0349
	11	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0012	.0036	.0095
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0016
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
14	0	.4877	.2288	.1028	.0440	.0178	.0068	.0024	.0008	.0002	.0001
	1	.3593	.3559	.2539	.1539	.0832	.0407	.0181	.0073	.0027	.0009
	2	.1229	.2570	.2912	.2501	.1802	.1134	.0634	.0317	.0141	.0056
	3	.0259	.1142	.2056	.2501	.2402	.1943	.1366	.0845	.0462	.0222
	4	.0037	.0349	.0998	.1720	.2202	.2290	.2022	.1549	.1040	.0611
	5	.0004	.0078	.0352	.0860	.1468	.1963	.2178	.2066	.1701	.1222
	6	.0000	.0013	.0093	.0322	.0734	.1262	.1759	.2066	.2088	.1833
	7	.0000	.0002	.0019	.0092	.0280	.0618	.1082	.1574	.1952	.2095
	8	.0000	.0000	.0003	.0020	.0082	.0232	.0510	.0918	.1398	.1833
	9	.0000	.0000	.0000	.0003	.0018	.0066	.0183	.0408	.0762	.1222
	10	.0000	.0000	.0000	.0000	.0003	.0014	.0049	.0136	.0312	.0611
	11	.0000	.0000	.0000	.0000	.0000	.0002	.0010	.0033	.0093	.0222
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0019	.0056
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0009
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
15	0	.4633	.2059	.0874	.0352	.0134	.0047	.0016	.0005	.0001	.0000
	1	.3658	.3432	.2312	.1319	.0668	.0305	.0126	.0047	.0016	.0005
	2	.1348	.2669	.2856	.2309	.1559	.0916	.0476	.0219	.0090	.0032
	3	.0307	.1285	.2184	.2501	.2252	.1700	.1110	.0634	.0318	.0139
	4	.0049	.0428	.1156	.1876	.2252	.2186	.1792	.1268	.0780	.0417
	5	.0006	.0105	.0449	.1032	.1651	.2061	.2123	.1859	.1404	.0916
	6	.0000	.0019	.0132	.0430	.0917	.1472	.1906	.2066	.1914	.1527
	7	.0000	.0003	.0030	.0138	.0393	.0811	.1319	.1771	.2013	.1964
	8	.0000	.0000	.0005	.0035	.0131	.0348	.0710	.1181	.1647	.1964
	9	.0000	.0000	.0001	.0007	.0034	.0116	.0298	.0612	.1048	.1527
	10	.0000	.0000	.0000	.0001	.0007	.0030	.0096	.0245	.0515	.0916
	11	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0074	.0191	.0417
	12	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0016	.0052	.0139
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0032
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
16	0	.4401	.1853	.0743	.0281	.0100	.0033	.0010	.0003	.0001	.0000
	1	.3706	.3294	.2097	.1126	.0535	.0228	.0087	.0030	.0009	.0002
	2	.1463	.2745	.2775	.2111	.1336	.0732	.0353	.0150	.0056	.0018
	3	.0359	.1423	.2285	.2463	.2079	.1465	.0888	.0468	.0215	.0085
	4	.0061	.0514	.1311	.2001	.2252	.2040	.1553	.1014	.0572	.0278
	5	.0008	.0137	.0555	.1201	.1802	.2099	.2008	.1623	.1123	.0667
	6	.0001	.0028	.0180	.0550	.1101	.1649	.1982	.1983	.1684	.1222
	7	.0000	.0004	.0045	.0197	.0524	.1010	.1524	.1889	.1969	.1746
	8	.0000	.0001	.0009	.0055	.0197	.0487	.0923	.1417	.1812	.1964
	9	.0000	.0000	.0001	.0012	.0058	.0185	.0442	.0840	.1318	.1746
	10	.0000	.0000	.0000	.0002	.0014	.0056	.0167	.0392	.0755	.1222
	11	.0000	.0000	.0000	.0000	.0002	.0013	.0049	.0142	.0337	.0667
	12	.0000	.0000	.0000	.0000	.0000	.0002	.0011	.0040	.0115	.0278
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0008	.0029	.0085
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0018
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

n	k	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
17	0	.4181	.1668	.0631	.0225	.0075	.0023	.0007	.0002	.0000	.0000
	1	.3741	.3150	.1893	.0957	.0426	.0169	.0060	.0019	.0005	.0001
	2	.1575	.2800	.2673	.1914	.1136	.0581	.0260	.0102	.0035	.0010
	3	.0415	.1556	.2359	.2393	.1893	.1245	.0701	.0341	.0144	.0052
	4	.0076	.0605	.1457	.2093	.2209	.1868	.1320	.0796	.0411	.0182
	5	.0010	.0175	.0668	.1361	.1914	.2081	.1849	.1379	.0875	.0472
	6	.0001	.0039	.0236	.0680	.1276	.1784	.1991	.1839	.1432	.0944
	7	.0000	.0007	.0065	.0267	.0668	.1201	.1685	.1927	.1841	.1484
	8	.0000	.0001	.0014	.0084	.0279	.0644	.1134	.1606	.1883	.1855
	9	.0000	.0000	.0003	.0021	.0093	.0276	.0611	.1070	.1540	.1855
	10	.0000	.0000	.0000	.0004	.0025	.0095	.0263	.0571	.1008	.1484
	11	.0000	.0000	.0000	.0001	.0005	.0026	.0090	.0242	.0525	.0944
	12	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0081	.0215	.0472
	13	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0021	.0068	.0182
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0016	.0052
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
18	0	.3972	.1501	.0536	.0180	.0056	.0016	.0004	.0001	.0000	.0000
	1	.3763	.3002	.1704	.0811	.0338	.0126	.0042	.0012	.0003	.0001
	2	.1683	.2835	.2556	.1723	.0958	.0458	.0190	.0069	.0022	.0006
	3	.0473	.1680	.2406	.2297	.1704	.1046	.0547	.0246	.0095	.0031
	4	.0093	.0700	.1592	.2153	.2130	.1681	.1104	.0614	.0291	.0117
	5	.0014	.0218	.0787	.1507	.1988	.2017	.1664	.1146	.0666	.0327
	6	.0002	.0052	.0301	.0816	.1436	.1873	.1941	.1655	.1181	.0708
	7	.0000	.0010	.0091	.0350	.0820	.1376	.1792	.1892	.1657	.1214
	8	.0000	.0002	.0022	.0120	.0376	.0811	.1327	.1734	.1864	.1669
	9	.0000	.0000	.0004	.0033	.0139	.0386	.0794	.1284	.1694	.1855
	10	.0000	.0000	.0001	.0008	.0042	.0149	.0385	.0771	.1248	.1669
	11	.0000	.0000	.0000	.0001	.0010	.0046	.0151	.0374	.0742	.1214
	12	.0000	.0000	.0000	.0000	.0002	.0012	.0047	.0145	.0354	.0708
	13	.0000	.0000	.0000	.0000	.0000	.0002	.0012	.0045	.0134	.0327
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0011	.0039	.0117
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0009	.0031
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0006
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
19	0	.3774	.1351	.0456	.0144	.0042	.0011	.0003	.0001	.0000	.0000
	1	.3774	.2852	.1529	.0685	.0268	.0093	.0029	.0008	.0002	.0000
	2	.1787	.2852	.2428	.1540	.0803	.0358	.0138	.0046	.0013	.0003
	3	.0533	.1796	.2428	.2182	.1517	.0869	.0422	.0175	.0062	.0018
	4	.0112	.0798	.1714	.2182	.2023	.1491	.0909	.0467	.0203	.0074
	5	.0018	.0266	.0907	.1636	.2023	.1916	.1468	.0933	.0497	.0222
	6	.0002	.0069	.0374	.0955	.1574	.1916	.1844	.1451	.0949	.0518
	7	.0000	.0014	.0122	.0443	.0974	.1525	.1844	.1797	.1443	.0961
	8	.0000	.0002	.0032	.0166	.0487	.0981	.1489	.1797	.1771	.1442
	9	.0000	.0000	.0007	.0051	.0198	.0514	.0980	.1464	.1771	.1762
	10	.0000	.0000	.0001	.0013	.0066	.0220	.0528	.0976	.1449	.1762
	11	.0000	.0000	.0000	.0003	.0018	.0077	.0233	.0532	.0970	.1442
	12	.0000	.0000	.0000	.0000	.0004	.0022	.0083	.0237	.0529	.0961
	13	.0000	.0000	.0000	.0000	.0001	.0005	.0024	.0085	.0233	.0518
	14	.0000	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0082	.0222
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0022	.0074
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0018
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003

n	k	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
20	0	.3585	.1216	.0388	.0115	.0032	.0008	.0002	.0000	.0000	.0000
	1	.3774	.2702	.1368	.0576	.0211	.0068	.0020	.0005	.0001	.0000
	2	.1887	.2852	.2293	.1369	.0669	.0278	.0100	.0031	.0008	.0002
	3	.0596	.1901	.2428	.2054	.1339	.0716	.0323	.0123	.0040	.0011
	4	.0133	.0898	.1821	.2182	.1897	.1304	.0738	.0350	.0139	.0046
	5	.0022	.0319	.1028	.1746	.2023	.1789	.1272	.0746	.0365	.0148
	6	.0003	.0089	.0454	.1091	.1686	.1916	.1712	.1244	.0746	.0370
	7	.0000	.0020	.0160	.0546	.1124	.1643	.1844	.1659	.1221	.0739
	8	.0000	.0004	.0046	.0222	.0609	.1144	.1614	.1797	.1623	.1201
	9	.0000	.0001	.0011	.0074	.0271	.0654	.1158	.1597	.1771	.1602
	10	.0000	.0000	.0002	.0020	.0099	.0308	.0686	.1171	.1593	.1762
	11	.0000	.0000	.0000	.0005	.0030	.0120	.0336	.0710	.1185	.1602
	12	.0000	.0000	.0000	.0001	.0008	.0039	.0136	.0355	.0727	.1201
	13	.0000	.0000	.0000	.0000	.0002	.0010	.0045	.0146	.0366	.0739
	14	.0000	.0000	.0000	.0000	.0000	.0002	.0012	.0049	.0150	.0370
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0013	.0049	.0148
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0013	.0046
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0011
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002
	19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	20	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

Fuente: Rosner, B. (2000). *Fundamentals of Biostatistics*. 5ta. ed. Belmont: Duxbury Press, pp. 811-815.

Tabla para la probabilidad exacta de la distribución de Poisson $P(X = x) = \frac{e^{-\mu}\mu^{k}}{k!}$ donde k es el número de éxitos o eventos observados, μ= λt los eventos o éxitos esperados promedio para un tiempo y ciertas condiciones y e = 2,71828, el valor de la base del logaritmo natural o neperiano

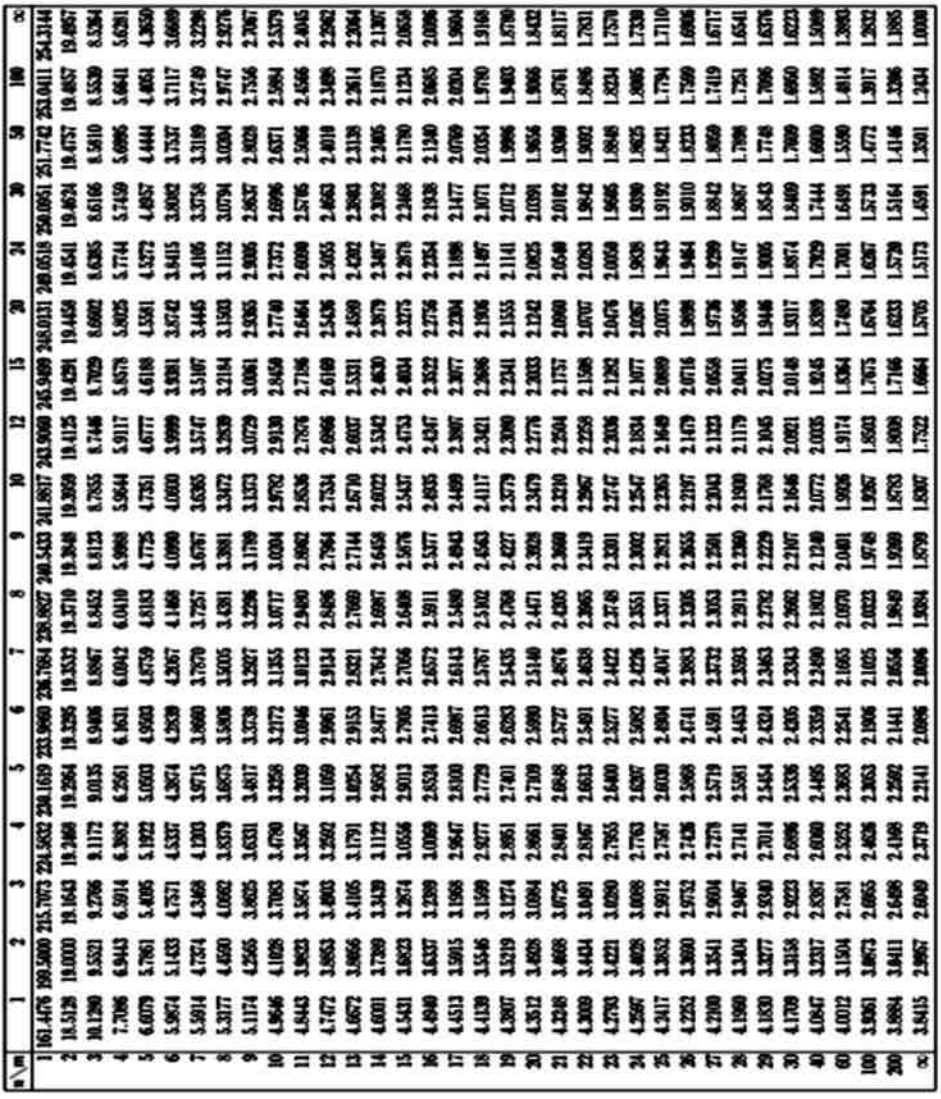
	μ									
k	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0
4	.1558	.1339	.1118	.0912	.0729	.0573	.0443	.0337	.0254	.0189
5	.1714	.1606	.1454	.1277	.1094	.0916	.0752	.0607	.0483	.0378
6	.1571	.1606	.1575	.1490	.1367	.1221	.1066	.0911	.0764	.0631
7	.1234	.1377	.1462	.1490	.1465	.1396	.1294	.1171	.1037	.0901
8	.0849	.1033	.1188	.1304	.1373	.1396	.1375	.1318	.1232	.1126
9	.0519	.0688	.0858	.1014	.1144	.1241	.1299	.1318	.1300	.1251
10	.0285	.0413	.0558	.0710	.0858	.0993	.1104	.1186	.1235	.1251
11	.0143	.0225	.0330	.0452	.0585	.0722	.0853	.0970	.1067	.1137
12	.0065	.0113	.0179	.0263	.0366	.0481	.0604	.0728	.0844	.0948
13	.0028	.0052	.0089	.0142	.0211	.0296	.0395	.0504	.0617	.0729
14	.0011	.0022	.0041	.0071	.0113	.0169	.0240	.0324	.0419	.0521
15	.0004	.0009	.0018	.0033	.0057	.0090	.0136	.0194	.0265	.0347
16	.0001	.0003	.0007	.0014	.0026	.0045	.0072	.0109	.0157	.0217
17	.0000	.0001	.0003	.0006	.0012	.0021	.0036	.0058	.0088	.0128
18	.0000	.0000	.0001	.0002	.0005	.0009	.0017	.0029	.0046	.0071
19	.0000	.0000	.0000	.0001	.0002	.0004	.0008	.0014	.0023	.0037
20	.0000	.0000	.0000	.0000	.0001	.0002	.0003	.0006	.0011	.0019
29	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0002	.0003	.0004
30	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0002
31	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001
32	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
33	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
21	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0003	.0005	.0009
22	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0002	.0004
23	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
24	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
25	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

	μ									
k	10.5	11.0	11.5	12.0	12.5	13.0	13.5	14.0	14.5	15.0
0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1	.0003	.0002	.0001	.0001	.0000	.0000	.0000	.0000	.0000	.0000
2	.0015	.0010	.0007	.0004	.0003	.0002	.0001	.0001	.0001	.0000
3	.0053	.0037	.0026	.0018	.0012	.0008	.0006	.0004	.0003	.0002
4	.0139	.0102	.0074	.0053	.0038	.0027	.0019	.0013	.0009	.0006
5	.0293	.0224	.0170	.0127	.0095	.0070	.0051	.0037	.0027	.0019
6	.0513	.0411	.0325	.0255	.0197	.0152	.0115	.0087	.0065	.0048
7	.0769	.0646	.0535	.0437	.0353	.0281	.0222	.0174	.0135	.0104
8	.1009	.0888	.0769	.0655	.0551	.0457	.0375	.0304	.0244	.0194
9	.1177	.1085	.0982	.0874	.0765	.0661	.0563	.0473	.0394	.0324
10	.1236	.1194	.1129	.1048	.0956	.0859	.0760	.0663	.0571	.0486
11	.1180	.1194	.1181	.1144	.1087	.1015	.0932	.0844	.0753	.0663
12	.1032	.1094	.1131	.1144	.1132	.1099	.1049	.0984	.0910	.0829
13	.0834	.0926	.1001	.1056	.1089	.1099	.1089	.1060	.1014	.0956
14	.0625	.0728	.0822	.0905	.0972	.1021	.1050	.1060	.1051	.1024
15	.0438	.0534	.0630	.0724	.0810	.0885	.0945	.0989	.1016	.1024
16	.0287	.0367	.0453	.0543	.0633	.0719	.0798	.0866	.0920	.0960
17	.0177	.0237	.0306	.0383	.0465	.0550	.0633	.0713	.0785	.0847
18	.0104	.0145	.0196	.0255	.0323	.0397	.0475	.0554	.0632	.0706
19	.0057	.0084	.0119	.0161	.0213	.0272	.0337	.0409	.0483	.0557
20	.0030	.0046	.0068	.0097	.0133	.0177	.0228	.0286	.0350	.0418
21	.0015	.0024	.0037	.0055	.0079	.0109	.0146	.0191	.0242	.0299
22	.0007	.0012	.0020	.0030	.0045	.0065	.0090	.0121	.0159	.0204
23	.0003	.0006	.0010	.0016	.0024	.0037	.0053	.0074	.0100	.0133
24	.0001	.0003	.0005	.0008	.0013	.0020	.0030	.0043	.0061	.0083
25	.0001	.0001	.0002	.0004	.0006	.0010	.0016	.0024	.0035	.0050
26	.0000	.0000	.0001	.0002	.0003	.0005	.0008	.0013	.0020	.0029
27	.0000	.0000	.0000	.0001	.0001	.0002	.0004	.0007	.0011	.0016
28	.0000	.0000	.0000	.0000	.0001	.0001	.0002	.0003	.0005	.0009
29	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0002	.0003	.0004
30	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0002
31	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001
32	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
33	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

	μ									
k	15.5	16.0	16.5	17.0	17.5	18.0	18.5	19.0	19.5	20.0
0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
2	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
3	.0001	.0001	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000
4	.0004	.0003	.0002	.0001	.0001	.0001	.0000	.0000	.0000	.0000
5	.0014	.0010	.0007	.0005	.0003	.0002	.0002	.0001	.0001	.0001
6	.0036	.0026	.0019	.0014	.0010	.0007	.0005	.0004	.0003	.0002
7	.0079	.0060	.0045	.0034	.0025	.0019	.0014	.0010	.0007	.0005
8	.0153	.0120	.0093	.0072	.0055	.0042	.0031	.0024	.0018	.0013
9	.0264	.0213	.0171	.0135	.0107	.0083	.0065	.0050	.0038	.0029
10	.0409	.0341	.0281	.0230	.0186	.0150	.0120	.0095	.0074	.0058
11	.0577	.0496	.0422	.0355	.0297	.0245	.0201	.0164	.0132	.0106
12	.0745	.0661	.0580	.0504	.0432	.0368	.0310	.0259	.0214	.0176
13	.0868	.0814	.0736	.0658	.0582	.0509	.0441	.0378	.0322	.0271
14	.0963	.0930	.0868	.0800	.0728	.0655	.0583	.0514	.0448	.0387
15	.1016	.0992	.0955	.0906	.0849	.0786	.0719	.0650	.0582	.0516
16	.0964	.0992	.0985	.0963	.0929	.0884	.0831	.0772	.0710	.0646
17	.0897	.0934	.0956	.0963	.0956	.0936	.0904	.0863	.0814	.0760
18	.0773	.0830	.0876	.0909	.0929	.0936	.0930	.0911	.0882	.0844
19	.0630	.0699	.0761	.0814	.0856	.0887	.0906	.0911	.0905	.0888
20	.0489	.0559	.0628	.0692	.0749	.0798	.0837	.0866	.0883	.0888
21	.0361	.0426	.0493	.0560	.0624	.0684	.0738	.0783	.0820	.0846
22	.0254	.0310	.0370	.0433	.0496	.0560	.0620	.0676	.0727	.0769
23	.0171	.0216	.0265	.0320	.0378	.0438	.0499	.0559	.0616	.0669
24	.0111	.0144	.0182	.0226	.0275	.0328	.0385	.0442	.0500	.0557
25	.0069	.0092	.0120	.0154	.0193	.0237	.0285	.0336	.0390	.0446
26	.0041	.0057	.0076	.0101	.0130	.0164	.0202	.0246	.0293	.0343
27	.0023	.0034	.0047	.0063	.0084	.0109	.0139	.0173	.0211	.0254
28	.0013	.0019	.0028	.0038	.0053	.0070	.0092	.0117	.0147	.0181
29	.0007	.0011	.0016	.0023	.0032	.0044	.0058	.0077	.0099	.0125
30	.0004	.0006	.0009	.0013	.0019	.0026	.0036	.0049	.0064	.0083
31	.0002	.0003	.0005	.0007	.0010	.0015	.0022	.0030	.0040	.0054
32	.0001	.0001	.0002	.0004	.0006	.0009	.0012	.0018	.0025	.0034
33	.0000	.0001	.0001	.0002	.0003	.0005	.0007	.0010	.0015	.0020
34	.0000	.0000	.0001	.0001	.0002	.0002	.0004	.0006	.0008	.0012
35	.0000	.0000	.0000	.0000	.0001	.0001	.0002	.0003	.0005	.0007
36	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0002	.0003	.0004
37	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002
38	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001
39	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
40	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

Fuente: Rosner, B. (2000). *Fundamentals of Biostatistics*. 5ta. ed. Belmont: Duxbury Press, pp. 815-817.

Tabla para los valores teóricos-esperados (cuantiles) de la distribución F-Snedecor Fisher según los grados de libertad en el numerador (toda la primera columna con valores de n desde 1 hasta ∞) y en el denominador (toda la primera fila con valores de m desde 1 hasta ∞) y para el valor 1 – α= 0,95



n\m	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	50	100	∞
1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817	243.9060	245.9499	248.0131	249.0518	250.0951	251.7742	253.0411	254.3144
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959	19.4125	19.4291	19.4458	19.4541	19.4624	19.4757	19.4857	19.4957
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855	8.7446	8.7029	8.6602	8.6385	8.6166	8.5810	8.5539	8.5264
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988	5.9644	5.9117	5.8578	5.8025	5.7744	5.7459	5.6995	5.6641	5.6281
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351	4.6777	4.6188	4.5581	4.5272	4.4957	4.4444	4.4051	4.3650
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	4.0600	3.9999	3.9381	3.8742	3.8415	3.8082	3.7557	3.7117	3.6689
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365	3.5747	3.5107	3.4445	3.4105	3.3758	3.3189	3.2749	3.2298
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.2839	3.2184	3.1503	3.1152	3.0794	3.0204	2.9747	2.9276
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	3.0729	3.0061	2.9365	2.9005	2.8637	2.8021	2.7556	2.7067
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.9130	2.8450	2.7740	2.7372	2.6996	2.6371	2.5884	2.5379
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536	2.7876	2.7186	2.6464	2.6090	2.5705	2.5066	2.4566	2.4045
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534	2.6866	2.6169	2.5436	2.5055	2.4663	2.4018	2.3498	2.2962
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.6710	2.6037	2.5331	2.4589	2.4202	2.3803	2.3128	2.2614	2.2064
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6022	2.5342	2.4630	2.3879	2.3487	2.3082	2.2405	2.1870	2.1307
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437	2.4753	2.4034	2.3275	2.2878	2.2468	2.1780	2.1234	2.0658
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377	2.4935	2.4247	2.3522	2.2756	2.2354	2.1938	2.1240	2.0685	2.0096
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943	2.4499	2.3807	2.3077	2.2304	2.1898	2.1477	2.0789	2.0234	2.0401
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563	2.4117	2.3421	2.2686	2.1906	2.1497	2.1071	2.0354	1.9790	1.9168
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227	2.3779	2.3080	2.2341	2.1555	2.1141	2.0712	2.0066	1.9493	1.8780
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928	2.3479	2.2776	2.2033	2.1242	2.0825	2.0391	1.9736	1.9168	1.8432
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660	2.3210	2.2504	2.1757	2.0960	2.0540	2.0102	1.9438	1.8861	1.8117
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419	2.2967	2.2259	2.1508	2.0707	2.0283	1.9842	1.9167	1.8586	1.7931
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201	2.2747	2.2036	2.1282	2.0476	2.0050	1.9605	1.8918	1.8327	1.7659
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002	2.2547	2.1834	2.1077	2.0267	1.9838	1.9390	1.8695	1.8095	1.7330
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821	2.2365	2.1649	2.0889	2.0075	1.9643	1.9192	1.8421	1.7794	1.7110
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655	2.2197	2.1479	2.0716	1.9898	1.9464	1.9010	1.8233	1.7589	1.6906
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501	2.2043	2.1323	2.0558	1.9728	1.9289	1.8842	1.8059	1.7419	1.6717
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360	2.1900	2.1179	2.0411	1.9586	1.9147	1.8697	1.7908	1.7251	1.6541
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2782	2.2229	2.1768	2.1045	2.0275	1.9446	1.9005	1.8543	1.7748	1.7086	1.6376
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107	2.1646	2.0921	2.0148	1.9317	1.8874	1.8409	1.7609	1.6950	1.6223
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240	2.0772	2.0035	1.9245	1.8389	1.7929	1.7444	1.6680	1.5982	1.5089
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.1070	2.0401	1.9926	1.9174	1.8364	1.7480	1.7001	1.6491	1.5550	1.4814	1.3893
100	3.9361	3.0873	2.6865	2.4626	2.3053	2.1906	2.1025	2.0323	1.9748	1.9267	1.8503	1.7675	1.6784	1.6307	1.5733	1.4772	1.3917	1.2932
200	3.8884	3.0411	2.6408	2.4168	2.2592	2.1441	2.0556	1.9849	1.9289	1.8789	1.8009	1.7166	1.6233	1.5729	1.5164	1.4146	1.3206	1.1885
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799	1.8307	1.7522	1.6664	1.5705	1.5173	1.4591	1.3501	1.2454	1.0000

Fuente: Botella Rocamora, P., Alacreu García, M. y Martínez Beneito, M. A. (2013-14). *Apuntes de Estadística en Ciencias de la Salud*. Memorias del Curso: Estadística en Ciencias de la Salud. España: Universidad Cardenal Herrera. (Tablas anexas).

En una prueba con el estadístico de contraste F-Snedecor-Fisher, con tres grados de libertad en el numerador y 10 en el denominador cuyo valor tabulado (teórico esperado)

de F sea de 3,708 o para el caso que tenga 15 grados de libertad en el numerador y 200 en el numerador con valor esperado de 1,717; la probabilidad de encontrar un resultado tan alejado o más de la hipótesis nula es de 0,05 (valor p = 0,05). Si se encontrase en la prueba de hipótesis un valor de F calculado-observado superior a 3,708 o de 1,717, respectivamente, para cada caso, el valor p sería inferior a 0,05.

Tabla para las probabilidades correspondientes a los puntos críticos del estadígrafo $D_{n,\alpha}$ (según n y α) mediante la prueba de Kolmogrov-Smirnov relacionada con la bondad de ajuste para la normalidad en la distribución de los valores de una variable:

n	Nivel de significación α							
	0,20	0,10	0,05	0,02	0,01	0,005	0,002	0,001
1	0,90000	0,95000	0,97500	0,99000	0,99500	0,99750	0,99900	0,99950
2	0,68337	0,77639	0,84189	0,90000	0,92929	0,95000	0,96838	0,97764
3	0,56481	0,63604	0,70760	0,78456	0,82900	0,86428	0,90000	0,92065
4	0,49265	0,56522	0,62934	0,68887	0,73424	0,77639	0,82217	0,85047
5	0,44698	0,50945	0,56328	0,62718	0,66853	0,70543	0,75000	0,78137
6	0,41037	0,46799	0,51926	0,57741	0,61661	0,65287	0,68571	0,72479
7	0,38148	0,43607	0,48342	0,53844	0,57581	0,60975	0,65071	0,67930
8	0,35831	0,40962	0,45427	0,50654	0,54179	0,57429	0,61368	0,64098
9	0,33910	0,38746	0,43001	0,47960	0,51332	0,54443	0,58210	0,60846
10	0,32260	0,36866	0,40925	0,45562	0,48893	0,51872	0,55500	0,58042
11	0,30829	0,35242	0,39122	0,43670	0,46770	0,49539	0,53135	0,55588
12	0,29577	0,33815	0,37543	0,41918	0,44905	0,47672	0,51047	0,53422
13	0,28470	0,32549	0,36143	0,40362	0,43247	0,45921	0,49189	0,51490
14	0,27481	0,31417	0,34890	0,38970	0,41762	0,44352	0,47520	0,49753
15	0,26589	0,30397	0,33750	0,37713	0,40420	0,42934	0,45611	0,48182
16	0,25778	0,29472	0,32733	0,36571	0,39201	0,41644	0,44637	0,46750
17	0,25039	0,28627	0,31796	0,35528	0,38086	0,40464	0,43380	0,45540
18	0,24360	0,27851	0,30936	0,34569	0,37062	0,39380	0,42224	0,44234
19	0,23735	0,27136	0,30143	0,33685	0,36117	0,38379	0,41156	0,43119
20	0,23156	0,26473	0,29408	0,32866	0,35241	0,37451	0,40165	0,42085
21	0,22517	0,25858	0,28724	0,32104	0,34426	0,36588	0,39243	0,41122
22	0,22115	0,25283	0,28087	0,31394	0,33666	0,35782	0,38382	0,40223
23	0,21646	0,24746	0,27491	0,30728	0,32954	0,35027	0,37575	0,39380
24	0,21205	0,24242	0,26931	0,30104	0,32286	0,34318	0,36787	0,38588
25	0,20790	0,23768	0,26404	0,29518	0,31657	0,33651	0,36104	0,37743
26	0,20399	0,23320	0,25908	0,28962	0,30963	0,33022	0,35431	0,37139
27	0,20030	0,22898	0,25438	0,28438	0,30502	0,32425	0,34794	0,36473
28	0,19680	0,22497	0,24993	0,27942	0,29971	0,31862	0,34190	0,35842
29	0,19348	0,22117	0,24571	0,27471	0,29466	0,31327	0,33617	0,35242
30	0,19032	0,21756	0,24170	0,27023	0,28986	0,30818	0,33072	0,34672
31	0,18732	0,21412	0,23788	0,26596	0,28529	0,30333	0,32553	0,34129
32	0,18445	0,21085	0,23424	0,26189	0,28094	0,29870	0,32058	0,33611
33	0,18171	0,20771	0,23076	0,25801	0,27577	0,29428	0,31584	0,33115
34	0,17909	0,21472	0,22743	0,25429	0,27271	0,29005	0,31131	0,32641
35	0,17659	0,20185	0,22425	0,25073	0,26897	0,28600	0,30597	0,32187
36	0,17418	0,19910	0,22119	0,24732	0,26532	0,28211	0,30281	0,31751
37	0,17188	0,19646	0,21826	0,24404	0,26180	0,27838	0,29882	0,31333
38	0,16966	0,19392	0,21544	0,24089	0,25843	0,27483	0,29498	0,30931
39	0,16753	0,19148	0,21273	0,23785	0,25518	0,27135	0,29125	0,39544

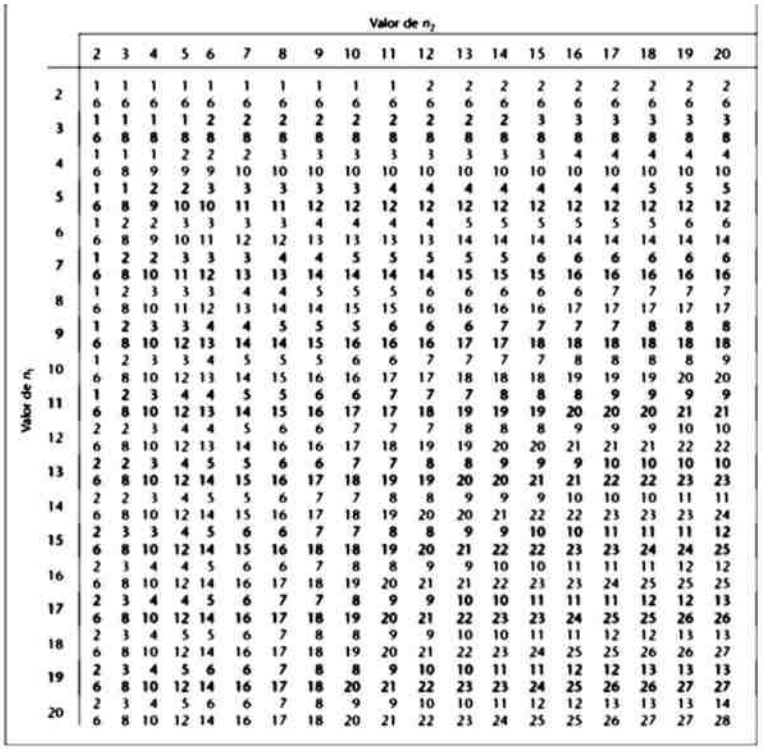
Fuente: Álvarez Cáceres, R. (2007). *Estadística aplicada a las ciencias de la salud*. España: Editorial Díaz de Santos, p. 982.

Tabla para los valores críticos del estadígrafo A^2 en la prueba de Anderson-Darling según 1 – α

Para el caso de:	Estadístico de prueba ajustado:	1-α = 0,900	1-α = 0,950	1-α = 0,975	1-α = 0,990
Se conocen todos los parámetros	A^2_n para n ≥ 5.	1,933	2,492	3,070	3,857
Distribución Normal: $\overline{X}(n)$, S^2 (n)	$(1 + 4/n - 25/n^2)$ A^2_n	0,632	0,751	0,870	1,029
Distrbución Exponencial: $\overline{X}(n)$,	$(1 + 0,6/n)$ A^2_n	1,070	1,326	1,587	1,943
Distribución Weibull: (α, β)	$(1 + 0,2/\sqrt{n})$ A^2_n	0,637	0,757	0,877	1,038
Distribución Log-logística: (α, β)	$(1+0,25/\sqrt{n})$ A^2_n	0,563	0,660	0,769	0,906

Fuente: Simulation modeling and analysis. Apéndice E, p. 137. Disponible en: http://simulacions-tefa.blogspot.com/feeds/posts/default

Tabla de los valores críticos inferior y superior (teóricos-esperados) del estadígrafo R (número de rachas) empleados en la prueba no paramétrica de rachas (Wald-Wolfowitz)



Valor de n_1	Valor de n_2 = 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
3	1	1	1	1	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
	6	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
4	1	1	1	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4
	6	8	9	9	9	10	10	10	10	10	10	10	10	10	10	10	10	10	10
5	1	1	2	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5
	6	8	9	10	10	11	11	12	12	12	12	12	12	12	12	12	12	12	12
6	1	2	2	3	3	3	3	4	4	4	4	5	5	5	5	5	5	6	6
	6	8	9	10	11	12	12	13	13	13	13	14	14	14	14	14	14	14	14
7	1	2	2	3	3	3	4	4	5	5	5	5	5	6	6	6	6	6	6
	6	8	10	11	12	13	13	14	14	14	14	15	15	15	16	16	16	16	16
8	1	2	3	3	3	4	4	5	5	5	6	6	6	6	6	7	7	7	7
	6	8	10	11	12	13	14	14	15	15	16	16	16	16	17	17	17	17	17
9	1	2	3	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8
	6	8	10	12	13	14	14	15	16	16	16	17	17	18	18	18	18	18	18
10	1	2	3	3	4	5	5	5	6	6	7	7	7	7	8	8	8	8	9
	6	8	10	12	13	14	15	16	16	17	17	18	18	18	19	19	19	20	20
11	1	2	3	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9
	6	8	10	12	13	14	15	16	17	17	18	19	19	19	20	20	20	21	21
12	2	2	3	4	4	5	6	6	7	7	7	8	8	8	9	9	9	10	10
	6	8	10	12	13	14	16	16	17	18	19	19	20	20	21	21	21	22	22
13	2	2	3	4	5	5	6	6	7	7	8	8	9	9	9	10	10	10	10
	6	8	10	12	14	15	16	17	18	19	19	20	20	21	21	22	22	23	23
14	2	2	3	4	5	5	6	7	7	8	8	9	9	9	10	10	10	11	11
	6	8	10	12	14	15	16	17	18	19	20	20	21	22	22	23	23	23	24
15	2	3	3	4	5	6	6	7	7	8	8	9	9	10	10	11	11	11	12
	6	8	10	12	14	15	16	18	18	19	20	21	22	22	23	23	24	24	25
16	2	3	4	4	5	6	6	7	8	8	9	9	10	10	11	11	11	12	12
	6	8	10	12	14	16	17	18	19	20	21	21	22	23	23	24	25	25	25
17	2	3	4	4	5	6	7	7	8	9	9	10	10	11	11	11	12	12	13
	6	8	10	12	14	16	17	18	19	20	21	22	23	23	24	25	25	26	26
18	2	3	4	5	5	6	7	8	8	9	9	10	10	11	11	12	12	13	13
	6	8	10	12	14	16	17	18	19	20	21	22	23	24	25	25	26	26	27
19	2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13	13
	6	8	10	12	14	16	17	18	20	21	22	23	23	24	25	26	26	27	27
20	2	3	4	5	6	6	7	8	9	9	10	10	11	12	12	13	13	13	14
	6	8	10	12	14	16	17	18	20	21	22	23	24	25	25	26	27	27	28

Fuente: Estadística (Apéndice A, pág.784).

- Los valores en esta tabla son los valores críticos (número de rachas = R), suponiendo una prueba de dos colas con un nivel de significación α = 0,05.
- La hipótesis nula (H_0) de aleatoriedad del orden de los datos en las muestras se rechaza si el número total de rachas R es menor que o igual al valor crítico inferior ó si es mayor que o igual al valor crítico superior en la tabla anterior.

Tabla para los valores críticos inferiores (teóricos-esperados) del estadígrafo S (número de veces que se repite el signo menos frecuente), en la prueba de los signos de Wilcoxon, según el número de observaciones por parejas (n) y el valor de alfa seleccionado para una o dos colas

Número de observaciones por parejas de los valores de la variable de libre distribución (n)	**Número de diferencias entre las parejas de los valores de la variable, con el signo menos frecuente (S o X)**			
	Para: α **= 0,005 (una cola) o** α **= 0,01 (2 colas)**	**Para:** α **= 0,01 (una cola) o** α **= 0,02 (2 colas)**	**Para:** α **= 0,025 (una cola) o** α **= 0,5 (2 colas)**	**Para:** α **= 0,5 (una cola) o** α **= 0,1 (2 colas)**
1	-	-	-	-
2	-	-	-	-
3	-	-	-	-
4	-	-	-	-
5	-	-	-	-
6	-	-	0	0
7	-	0	0	0
8	0	0	0	1
9	0	0	1	1
10	0	0	1	1
11	0	1	1	2
12	1	1	2	2
13	1	1	2	3
14	1	2	2	3
15	2	2	3	3
16	2	2	3	4
17	2	3	4	4
18	3	3	4	5
19	3	4	4	5
20	3	4	5	5
21	4	4	5	6
22	4	5	5	6
23	4	5	6	7
24	5	5	6	7
25	5	6	7	7

Fuente: Estadística (Apéndice A, p. 781).

– El símbolo en la casilla correspondiente indica que no es posible obtener un valor en la región crítica.
– La hipótesis nula (H_0) se rechaza si el número de veces que se repite el signo menos frecuente observado en la muestra de estudio (X o S observado) es menor o igual que el valor teórico-esperado que presenta la tabla anterior.

Tabla para los valores críticos inferiores (teóricos-esperados) del estadígrafo T+ (rango con signo) en la prueba de los rangos con signos de Wilcoxon, correspondientes a los diferentes niveles de significación (α) para un lado (*one sided* α) y dos lados (*two sided* α), para el número de observaciones correspondientes (n) por parejas, de valores de la variable cuantitativa de libre distribución en las dos muestras o grupos de datos

	One-Sided α												
	.05	.025	.01	.005	.05	.025	.01	.005	.05	.025	.01	.005
	Two-Sided α												
	.10	.05	.02	.01	.10	.05	.02	.01	.10	.05	.02	.01
n					*n*				*n*				
5	1	—	—	—	20	60	52	43	37	35	214	195	174	160
6	2	1	—	—	21	68	59	49	43	36	228	208	186	171
7	4	2	0	—	22	75	66	56	49	37	242	222	198	183
8	6	4	2	0	23	83	73	62	55	38	256	235	211	195
9	8	6	3	2	24	92	81	69	61	39	271	250	224	208
10	11	8	5	3	25	101	90	77	68	40	287	264	238	221
11	14	11	7	5	26	110	98	85	76	41	303	279	252	234
12	17	14	10	7	27	120	107	93	84	42	319	295	267	248
13	21	17	13	10	28	130	117	102	92	43	336	311	281	262
14	26	21	16	13	29	141	127	111	100	44	353	327	297	277
15	30	25	20	16	30	152	137	120	109	45	371	344	313	292
16	36	30	24	19	31	163	148	130	118	46	389	361	329	307
17	41	35	28	23	32	175	159	141	128	47	408	379	345	323
18	47	40	33	28	33	188	171	151	138	48	427	397	362	339
19	54	46	38	32	34	201	183	162	149	49	446	415	380	356
										50	466	434	398	373

Fuente: Van Belle, G., Lloyd, F., Heagerty, P. and Lumley, T. (2004). *Biostatistics: a methodology for the health sciences*. 2nd ed. United State of America: John Wiley and Sons, Inc., p. 834.

Se rechaza la hipótesis nula (H_0) si el estadístico de prueba T con signo observado (con el menor valor) es menor que el valor crítico inferior (esperado-tabulado) encontrado en esta tabla.

Tabla para los valores críticos superiores (teóricos-esperados) del estadígrafo U en la prueba de Mann-Whitney, correspondientes a los diferentes niveles de significación (α) para un lado o cola (*one sided* α) y dos lados o colas (*two sided* α) según el número de observaciones de cada muestra (n, m)

		One-Sided α												
		.10	.05	.025	.01	.005	.001		.10	.05	.025	.01	.005	.001
		Two-Sided α												
		.20	.10	.05	.02	.01	.002		.20	.10	.05	.02	.01	.002
n	m							n	m					
3	2	6	—	—	—	—	—	10	1	10	—	—	—	—	—
3	3	8	9	—	—	—	—	10	2	17	19	20	—	—	—
								10	3	24	26	27	29	30	—
4	2	8	—	—	—	—	—	10	4	30	33	35	37	38	40
4	3	11	12	—	—	—	—	10	5	37	39	42	44	46	49
4	4	13	15	16	—	—	—	10	6	43	46	49	52	54	57
								10	7	49	53	56	59	61	65
5	2	9	10	—	—	—	—	10	8	56	60	63	67	69	74
5	3	13	14	15	—	—	—	10	9	62	66	70	74	77	82
5	4	16	18	19	20	—	—	10	10	68	73	77	81	84	90
5	5	20	21	23	24	25	—								
								11	1	11	—	—	—	—	—
6	2	11	12	—	—	—	—	11	2	19	21	22	—	—	—
6	3	15	16	17	—	—	—	11	3	26	28	30	32	33	—
6	4	19	21	22	23	24	—	11	4	33	36	38	40	42	44
6	5	23	25	27	28	29	—	11	5	40	43	46	48	50	53
6	6	27	29	31	33	34	—	11	6	47	50	53	57	59	62
								11	7	54	58	61	65	67	71
7	2	13	14	—	—	—	—	11	8	61	65	69	73	75	80
7	3	17	19	20	21	—	—	11	9	68	72	76	81	83	89
7	4	22	24	25	27	28	—	11	10	74	79	84	88	92	98
7	5	27	29	30	32	34	—	11	11	81	87	91	96	100	106
7	6	31	34	36	38	39	42								
7	7	36	38	41	43	45	48	12	1	12	—	—	—	—	—
								12	2	20	22	23	—	—	—
8	2	14	15	16	—	—	—	12	3	28	31	32	34	35	—
8	3	19	21	22	24	—	—	12	4	36	39	41	43	45	48
8	4	25	27	28	30	31	—	12	5	43	47	49	52	54	58
8	5	30	32	34	36	38	40	12	6	51	55	58	61	63	68
8	6	35	38	40	42	44	47	12	7	58	63	66	70	72	77
8	7	40	43	46	49	50	54	12	8	66	70	74	79	81	87
8	8	45	49	51	55	57	60	12	9	73	78	82	87	90	96
								12	10	81	86	91	96	99	106
9	1	9	—	—	—	—	—	12	11	88	94	99	104	108	115
9	2	16	17	18	—	—	—	12	12	95	102	107	113	117	124
9	3	22	23	25	26	27	—								
9	4	27	30	32	33	35	—	13	1	13	—	—	—	—	—
9	5	33	36	38	40	42	44	13	2	22	24	25	26	—	—
9	6	39	42	44	47	49	52	13	3	30	33	35	37	38	—
9	7	45	48	51	54	56	60	13	4	39	42	44	47	49	51
9	8	50	54	57	61	63	67	13	5	47	50	53	56	58	62
9	9	56	60	64	67	70	74	13	6	55	59	62	66	68	73

		One-Sided α								One-Sided α					
		.10	.05	.025	.01	.005	.001			.10	.05	.025	.01	.005	.001
		Two-Sided α								Two-Sided α					
		.20	.10	.05	.02	.01	.002			.20	.10	.05	.02	.01	.002
n	m							n	m						
13	7	63	67	71	75	78	83	16	12	125	132	139	146	151	161
13	8	71	76	80	84	87	93	16	13	134	143	149	157	163	173
13	9	79	84	89	94	97	103	16	14	144	153	160	168	174	185
13	10	87	93	97	103	106	113	16	15	154	163	170	179	185	197
13	11	95	101	106	112	116	123	16	16	163	173	181	190	196	208
13	12	103	109	115	121	125	133								
13	13	111	118	124	130	135	143	17	1	17	—	—	—	—	—
								17	2	28	31	32	34	—	—
14	1	14	—	—	—	—	—	17	3	39	42	45	47	49	51
14	2	23	25	27	28	—	—	17	4	50	53	57	60	62	66
14	3	32	35	37	40	41	—	17	5	60	65	68	72	75	80
14	4	41	45	47	50	52	55	17	6	71	76	80	84	87	93
14	5	50	54	57	60	63	67	17	7	81	86	91	96	100	106
14	6	59	63	67	71	73	78	17	8	91	97	102	108	112	119
14	7	67	72	76	81	83	89	17	9	101	108	114	120	124	132
14	8	76	81	86	90	94	100	17	10	112	119	125	132	136	145
14	9	85	90	95	100	104	111	17	11	122	130	136	143	148	158
14	10	93	99	104	110	114	121	17	12	132	140	147	155	160	170
14	11	102	108	114	120	124	132	17	13	142	151	158	166	172	183
14	12	110	117	123	130	134	143	17	14	153	161	169	178	184	195
14	13	119	126	132	139	144	153	17	15	163	172	180	189	195	208
14	14	127	135	141	149	154	164	17	16	173	183	191	201	207	220
								17	17	183	193	202	212	219	232
15	1	15	—	—	—	—	—								
15	2	25	27	29	30	—	—	18	1	18	—	—	—	—	—
15	3	35	38	40	42	43	—	18	2	30	32	34	36	—	—
15	4	44	48	50	53	55	59	18	3	41	45	47	50	52	54
15	5	53	57	61	64	67	71	18	4	52	56	60	63	66	69
15	6	63	67	71	75	78	83	18	5	63	68	72	76	79	84
15	7	72	77	81	86	89	95								
15	8	81	87	91	96	100	106	18	6	74	80	84	89	92	98
15	9	90	96	101	107	111	118	18	7	85	91	96	102	105	112
15	10	99	106	111	117	121	129	18	8	96	103	108	114	118	126
15	11	108	115	121	128	132	141	18	9	107	114	120	126	131	139
15	12	117	125	131	138	143	152	18	10	118	125	132	139	143	153
15	13	127	134	141	148	153	163								
15	14	136	144	151	159	164	174	18	11	129	137	143	151	156	166
15	15	145	153	161	169	174	185	18	12	139	148	155	163	169	179
								18	13	150	159	167	175	181	192
16	1	16	—	—	—	—	—	18	14	161	170	178	187	194	206
16	2	27	29	31	32	—	—	18	15	172	182	190	200	206	219
16	3	37	40	42	45	46	—	18	16	182	193	202	212	218	232
16	4	47	50	53	57	59	62	18	17	193	204	213	224	231	245
16	5	57	61	65	68	71	75	18	18	204	215	225	236	243	258
16	6	67	71	75	80	83	88								
16	7	76	82	86	91	94	101	19	1	18	19	—	—	—	—
16	8	86	92	97	102	106	113	19	2	31	34	36	37	38	—
16	9	96	102	107	113	117	125	19	3	43	47	50	53	54	57
16	10	106	112	118	124	129	137	19	4	55	59	63	67	69	73
16	11	115	122	129	135	140	149	19	5	67	72	76	80	83	88

		One-Sided α												
		.10	.05	.025	.01	.005	.001		.10	.05	.025	.01	.005	.001
		Two-Sided α												
		.20	.10	.05	.02	.01	.002		.20	.10	.05	.02	.01	.002
n	m							n	m					
19	6	78	84	89	94	97	103	20	4	58	62	66	70	72	77
19	7	90	96	101	107	111	118	20	5	70	75	80	84	87	93
19	8	101	108	114	120	124	132	20	6	82	88	93	98	102	108
19	9	113	120	126	133	138	146	20	7	94	101	106	112	116	124
19	10	124	132	138	146	151	161	20	8	106	113	119	126	130	139
19	11	136	144	151	159	164	175	20	9	118	126	132	140	144	154
19	12	147	156	163	172	177	188	20	10	130	138	145	153	158	168
19	13	158	167	175	184	190	202	20	11	142	151	158	167	172	183
19	14	169	179	188	197	203	216	20	12	154	163	171	180	186	198
19	15	181	191	200	210	216	230	20	13	166	176	184	193	200	212
19	16	192	203	212	222	230	244	20	14	178	188	197	207	213	226
19	17	203	214	224	235	242	257	20	15	190	200	210	220	227	241
19	18	214	226	236	248	255	271	20	16	201	213	222	233	241	255
19	19	226	238	248	260	268	284	20	17	213	225	235	247	254	270
20	1	19	20	—	—	—	—	20	18	225	237	248	260	268	284
20	2	33	36	38	39	40	—	20	19	237	250	261	273	281	298
20	3	45	49	52	55	57	60	20	20	249	262	273	286	295	312

Fuente: Van Belle, G., Lloyd, F., Heagerty, P. and Lumley, T. (2004). *Biostatistics: a methodology for the health sciences.* 2nd ed. United State of America: John Wiley and Sons, Inc., pp. 835-837.

Se rechaza la hipótesis nula (H_0) si el estadístico de prueba U de Mann Whitney observado (con mayor valor de las dos muestra independientes) es mayor que el valor crítico superior (teórico-esperado) encontrado en esta tabla.

Tabla para los valores críticos inferior y superior (teóricos- esperados) del estadígrafo T en la prueba de la suma de rangos de Wilcoxon, según el tamaño de las muestras (n_1, n_2) y alfa (α) = 0,10; 0,05; 0,02; 0,01, para dos colas

	$\alpha = .10$ n_1*						$\alpha = .05$ n_1					
n_2*	4	5	6	7	8	9	4	5	6	7	8	9
	T_1 T_2	T_1 T_2	T_1 T_2	T_1 T_2	T_1 T_2	T_1 T_2	T_1 T_2	T_1 T_2	T_1 T_2	T_1 T_2	T_1 T_2	T_1 T_2
4	11–25	17–33	24–42	32–52	41–63	51–75	10–26	16–34	23–43	31–53	40–64	49–77
5	12–28	19–36	26–46	34–57	44–68	54–81	11–29	17–38	24–48	33–58	42–70	52–83
6	13–31	20–40	28–50	36–62	46–74	57–87	12–32	18–42	26–52	34–64	44–76	55–89
7	14–34	21–44	29–55	39–66	49–79	60–93	13–35	20–45	27–57	36–69	46–82	57–96
8	15–37	23–47	31–59	41–71	51–85	63–99	14–38	21–49	29–61	38–74	49–87	60–102
9	16–40	24–51	33–63	43–76	54–90	66–105	14–42	22–53	31–65	40–79	51–93	62–109
10	17–43	26–54	35–67	45–81	56–96	69–111	15–45	23–57	32–70	42–84	53–99	65–115
11	18–46	27–58	37–71	47–86	59–101	72–117	16–48	24–61	34–74	44–89	55–105	68–121
12	19–49	28–62	38–76	49–91	62–106	75–123	17–51	26–64	35–79	46–94	58–110	71–127
13	20–52	30–65	40–80	52–95	64–112	78–129	18–54	27–68	37–83	48–99	60–116	73–134
14	21–55	31–69	42–84	54–100	67–117	81–135	19–57	28–72	38–88	50–104	62–122	76–140
15	22–58	33–72	44–88	56–105	69–123	84–141	20–60	29–76	40–92	52–109	65–127	79–146
16	24–60	34–76	46–92	58–110	72–128	87–147	21–63	30–80	42–96	54–114	67–133	82–152
17	25–63	35–80	47–97	61–114	75–133	90–153	21–67	32–83	43–101	56–119	70–138	84–159
18	26–66	37–83	49–101	63–119	77–139	93–159	22–70	33–87	45–105	58–124	72–144	87–165
19	27–69	38–87	51–105	65–124	80–144	96–165	23–73	34–91	46–110	60–129	74–150	90–171
20	28–72	40–90	53–109	67–129	83–149	99–171	24–76	35–95	48–114	62–134	77–155	93–177
21	29–75	41–94	55–113	69–134	85–155	102–177	25–79	37–98	50–118	64–139	79–161	95–184
22	30–78	43–97	57–117	72–138	88–160	105–183	26–82	38–102	51–123	66–144	81–167	98–190
23	31–81	44–101	58–122	74–143	90–166	108–189	27–85	39–106	53–127	68–149	84–172	101–196
24	32–84	45–105	60–126	76–148	93–171	111–195	27–89	40–110	54–132	70–154	86–178	104–202
25	33–87	47–108	62–130	78–153	96–176	114–201	28–92	42–113	56–136	72–159	89–183	107–208
26	34–90	48–112	64–134	81–157	98–182	117–207	29–95	43–117	58–140	74–164	91–189	109–215
27	35–93	50–115	66–138	83–162	101–187	120–213	30–98	44–121	59–145	76–169	93–195	112–221
28	36–96	51–119	67–143	85–167	103–193	123–219	31–101	45–125	61–149	78–174	96–200	115–227
29	37–99	53–122	69–147	87–172	106–198	126–225	32–104	47–128	63–153	80–179	98–206	118–233
30	38–102	54–126	71–151	89–177	109–203	129–231	33–107	48–132	64–158	82–184	101–211	121–239
31	39–105	55–130	73–155	92–181	111–209	132–237	34–110	49–136	66–162	84–189	103–217	123–246
32	40–108	57–133	75–159	94–186	114–214	135–243	34–114	50–140	67–167	86–194	106–222	126–252
33	41–111	58–137	77–163	96–191	117–219	138–249	35–117	52–143	69–171	88–199	108–228	129–258
34	42–114	60–140	78–168	98–196	119–225	141–255	36–120	53–147	71–175	90–204	110–234	132–264
35	43–117	61–144	80–172	100–201	122–230	144–261	37–123	54–151	72–180	92–209	113–239	135–270
36	44–120	62–148	82–176	102–206	124–236	148–266	38–126	55–155	74–184	94–214	115–245	137–277
37	45–123	64–151	84–180	105–210	127–241	151–272	39–129	57–158	76–188	96–219	117–251	140–283
38	46–126	65–155	85–185	107–215	130–246	154–278	40–132	58–162	77–193	98–224	120–256	143–289
39	47–129	67–158	87–189	109–220	132–252	157–284	41–135	59–166	79–197	100–229	122–262	146–295
40	48–132	68–162	89–193	111–225	135–257	160–290	41–139	60–170	80–202	102–234	125–267	149–301
41	49–135	69–166	91–197	114–229	138–262	163–296	42–142	61–174	82–206	104–239	127–273	151–308
42	50–138	71–169	93–201	116–234	140–268	166–302	43–145	63–177	84–210	106–244	129–279	154–314
43	51–141	72–173	95–205	118–239	143–273	169–308	44–148	64–181	85–215	108–249	132–284	157–320
44	52–144	74–176	96–210	120–244	146–278	172–314	45–151	65–185	87–219	110–254	134–290	160–326
45	53–147	75–180	98–214	123–248	148–284	175–320	46–154	66–189	88–224	112–259	137–295	163–332
46	55–149	77–183	100–218	125–253	151–289	178–326	47–157	68–192	90–228	114–264	139–301	165–339
47	56–152	78–187	102–222	127–258	154–294	181–332	48–160	69–196	92–232	116–269	141–307	168–345
48	57–155	79–191	104–226	129–263	156–300	184–338	48–164	70–200	93–237	118–274	144–312	171–351
49	58–158	81–194	106–230	132–267	159–305	187–344	49–167	71–204	95–241	120–279	146–318	174–357
50	59–161	82–198	107–235	134–272	162–310	190–350	50–170	73–207	97–245	122–284	149–323	177–363

	$\alpha = .02$ n_1[a]						$\alpha = .01$ n_1					
n_2[b]	4	5	6	7	8	9	4	5	6	7	8	9
	T_1[c] T_2[d]	T_1 T_2	T_1 T_2	T_1 T_2	T_1 T_2	T_1 T_2	T_1 T_2	T_1 T_2	T_1 T_2	T_1 T_2	T_1 T_2	T_1 T_2
4	— —	15–35	22–44	29–55	36–66	48–78	— —	— —	21–45	28–56	37–67	46–80
5	10–30	16–39	23–49	31–60	40–72	50–85	— —	15–40	22–50	29–62	38–74	48–87
6	11–33	17–43	24–54	32–66	42–78	52–92	10–34	16–44	23–55	31–67	40–80	50–94
7	11–37	18–47	25–59	34–71	43–85	54–99	10–38	16–49	24–60	32–73	42–86	52–101
8	12–40	19–51	27–63	35–77	45–91	56–106	11–41	17–53	25–65	34–78	43–93	54–108
9	13–43	20–55	28–68	37–82	47–97	59–112	11–45	18–57	26–70	35–84	45–99	56–115
10	13–47	21–59	29–73	39–87	49–103	61–119	12–48	19–61	27–75	37–89	47–105	58–122
11	14–50	22–63	30–78	40–93	51–109	63–126	12–52	20–65	28–80	38–95	49–111	61–128
12	15–53	23–67	32–82	42–98	53–115	66–132	13–55	21–69	30–84	40–100	51–117	63–135
13	15–57	24–71	33–87	44–103	56–120	68–139	13–59	22–73	31–89	41–106	53–123	65–142
14	16–60	25–75	34–92	45–109	58–126	71–145	14–62	22–78	32–94	43–111	54–130	67–149
15	17–63	26–79	36–96	47–114	60–132	73–152	15–65	23–82	33–99	44–117	56–136	69–156
16	17–67	27–83	37–101	49–119	62–138	76–158	15–69	24–86	34–104	46–122	58–142	72–162
17	18–70	28–87	39–105	51–124	64–144	78–165	16–72	25–90	36–108	47–128	60–148	74–169
18	19–73	29–91	40–110	52–130	66–150	81–171	16–76	26–94	37–113	49–133	62–154	76–176
19	19–77	30–95	41–115	54–135	68–156	83–178	17–79	27–98	38–118	50–139	64–160	78–183
20	20–80	31–99	43–119	56–140	70–162	85–185	18–82	28–102	39–123	52–144	66–166	81–189
21	21–83	32–103	44–124	58–145	72–168	88–191	18–86	29–106	40–128	53–150	68–172	83 196
22	21–87	33–107	45–129	59–151	74–174	90–198	19–89	29–111	42–132	55–155	70–178	85–203
23	22–90	34 111	47–133	61–156	76–180	93–204	19–93	30–115	43–137	57–160	71–185	88–209
24	23–93	35 115	48 138	63–161	78–186	95–211	20–96	31–119	44–142	58–166	73–191	90–216
25	23–97	36 119	50 142	64–167	81–191	98–217	20–100	32–123	45–147	60–171	75–197	92–223
26	24–100	37–123	51–147	66–172	83–197	100–224	21–103	33–127	46–152	61–177	77–203	94–230
27	25–103	38–127	52–152	68–177	85–203	103–230	22–106	34–131	48–156	63–182	79–209	97–236
28	26–106	39–131	54–156	70–182	87–209	105–237	22–110	35–135	49–161	64–188	81–215	99–243
29	26–110	40–135	55–161	71–188	89–215	108–243	23–113	36–139	50–166	66–193	83–221	101–250
30	27–113	41–139	56–166	73–193	91–221	110–250	23–117	37–143	51–171	68–198	85–227	103–257
31	28–116	42–143	58–170	75–198	93–227	112–257	24–120	37–148	53–175	69–204	87–233	106–263
32	28–120	43–147	59–175	77–203	95–233	115–263	24–124	38–152	54–180	71–209	89–239	108–270
33	29–123	44–151	61–179	78–209	97–239	117–270	25–127	39–156	55–185	72–215	90–246	110–277
34	30–126	45–155	62–184	79–215	99–245	120–276	26–130	40–160	56–190	73–221	92–252	112–284
35	30–130	46–159	63–189	81–220	101–251	122–283	26–134	41–164	57–195	75–226	94–258	114–291
36	31–133	47–163	65–193	83–225	103–257	125–289	27–137	42–168	58–200	76–232	96–264	117–297
37	32–136	48–167	66–198	84–231	105–263	127–296	28–140	43–172	60–204	78–237	98–270	119–304
38	32–140	49–171	67–203	86–236	107–269	129–303	28–144	44–176	61–209	79–243	100–276	121–311
39	33–143	50–175	69–207	88–241	109–275	132–309	29–147	45–180	62–214	81–248	102–282	123–318
40	34–146	51–179	70–212	90–246	111–281	134–316	29–151	46–184	63–219	82–254	103–289	126–324
41	34–150	52–183	72–216	91–252	113–287	137–322	30–154	46–189	65–223	84–259	105–295	128–331
42	35–153	53–187	73–221	93–257	116–292	139–329	31–157	47–193	66–228	85–265	107–301	130–338
43	35–157	54–191	74–226	95–262	118–298	142–335	31–161	48–197	67–233	87–270	109–307	133–344
44	36–160	55–195	76–230	97–267	120–304	144–342	32–164	49–201	68–238	88–276	111–313	135–351
45	37–163	56–199	77–235	98–273	122–310	147–348	32–168	50–205	69–243	90–281	113–319	137–358
46	37–167	57–203	78–240	100–278	124–316	149–355	33–171	51–209	71–247	91–287	115–325	139–365
47	38–170	58–207	80–244	102–283	126–322	152–361	34–174	52–213	72–252	93–292	117–331	142–371
48	39–173	59–211	81–249	103–289	128–328	154–368	34–178	53–217	73–257	95–297	118–338	144–378
49	39–177	60–215	82–254	105–294	130–334	157–374	35–181	54–221	74–262	96–303	120–344	146–385
50	40–180	61–219	84–258	107–299	132–340	159–381	36–184	55–225	76–266	98–308	122–350	148–392

Fuente: Rosner, B. (2000). *Fundamentals of Biostatistics*. 5ta. ed. Belmont: Duxbury Press, pp. 831-832.

Se rechazala hipótesis nula (H_0) si el estadístico de prueba T observado (con el mayor valor para las dos muestras independientes) es mayor que el valor crítico superior (teó-

rico-esperado) o si el T observado (con el menor valor de las dos muestras) es menor que el encontrado en esta tabla.

donde:

$_1{}^a$: es el mínimo de los dos tamaños muestrales;

$n_2{}^b$: es el máximo de los dos tamanos muestrales;

$T_l{}^c$: valor crítico inferior para la suma de rangos en la primera muestra;

$T_r{}^d$: valor crítico superior para la suma de rangos en la primera muestra.

Tabla para los valores críticos (teóricos-esperados) del estadígrafo H en la prueba no paramétrica de Kruskal-Wallis, correspondientes con el número de observaciones de las muestras independientes n_1, n_2, n_3 y los diferentes niveles de significación alfa (α) = 0,10; 0,05; 0,02 y 0,01.

			α			
n_1	n_2	n_3	.10	.05	.02	.01
1	1	2	—	—	—	—
1	1	3	—	—	—	—
1	1	4	—	—	—	—
1	1	5	—	—	—	—
1	2	2	—	—	—	—
1	2	3	4.286	—	—	—
1	2	4	4.500	—	—	—
1	2	5	4.200	5.000	—	—
1	3	3	4.571	5.143	—	—
1	3	4	4.056	5.389	—	—
1	3	5	4.018	4.960	6.400	—
1	4	4	4.167	4.967	6.667	—
1	4	5	3.987	4.986	6.431	6.954
1	5	5	4.109	5.127	6.146	7.309
2	2	2	4.571	—	—	—
2	2	3	4.500	4.714	—	—
2	2	4	4.500	5.333	6.000	—
2	2	5	4.373	5.160	6.000	6.533
2	3	3	4.694	5.361	6.250	—
2	3	4	4.511	5.444	6.144	6.444
2	3	5	4.651	5.251	6.294	6.909
2	4	4	4.554	5.454	6.600	7.036
2	4	5	4.541	5.273	6.541	7.204
2	5	5	4.623	5.338	6.469	7.392
3	3	3	5.067	5.689	6.489	7.200
3	3	4	4.709	5.791	6.564	7.000
3	3	5	4.533	5.648	6.533	7.079
3	4	4	4.546	5.598	6.712	7.212
3	4	5	4.549	5.656	6.703	7.477
3	5	5	4.571	5.706	6.866	7.622
4	4	4	4.654	5.692	6.962	7.654
4	4	5	4.668	5.657	6.976	7.760
4	5	5	4.523	5.666	7.000	7.903
5	5	5	4.580	5.780	7.220	8.000

Fuente: Rosner, B. (2000). *Fundamentals of Biostatistics*. 5ta. ed. Belmont: Duxbury Press, p. 835.

Se rechaza la hipótesis nula (H_0) si el estadístico de prueba H observado es mayor que el valor crítico superior (teórico-esperado) encontrado la tabla anterior.

Tabla de probabilidades (p) asociadas con chi cuadrado téorico esperado (χ^2_r en la tabla) en la prueba de Friedman según el número de grupos de datos (k) y el tamaño de los grupos (N)

k = 3

N = 2		N = 3		N = 4		N = 5	
χ^2_r	p	χ^2_r	p	χ^2_r	p	χ^2_r	p
0	1.000	0.000	1.000	0.0	1.000	0.0	1.000
1	0.833	0.667	0.944	0.5	0.931	0.4	0.954
3	0.500	2.000	0.528	1.5	0.653	1.2	0.691
4	0.167	2.667	0.361	2.0	0.431	1.6	0.522
		4.667	0.194	3.5	0.273	2.8	0.367
		6.000	0.028	4.5	0.125	3.6	0.182
				6.0	0.042	4.8	0.124
				6.5	0.042	5.2	0.093
				8.0	0.0046	6.4	0.039
						7.6	0.024
						8.4	0.0085
						10.0	0.00077

N = 6		N = 7		N = 8		N = 9	
χ^2_r	p	χ^2_r	p	χ^2_r	p	χ^2_r	p
0.00	1.000	0.000	1.000	0.00	1.000	0.000	1.000
0.33	0.956	0.286	0.964	0.25	0.967	0.222	0.971
1.00	0.740	0.857	0.768	0.75	0.794	0.667	0.865
1.33	0.570	1.143	0.620	1.00	0.654	0.889	0.814
2.33	0.430	2.000	0.486	1.75	0.531	1.556	0.569
3.00	0.252	2.571	0.305	2.25	0.355	2.000	0.398
4.00	0.184	3.429	0.237	3.00	0.285	2.667	0.328
4.33	0.142	3.714	0.192	3.25	0.236	2.889	0.278
5.33	0.072	4.571	0.112	4.00	0.149	3.556	0.187
6.33	0.052	5.429	0.085	4.75	0.120	4.222	0.154
7.00	0.029	6.000	0.052	5.25	0.079	4.667	0.107
8.33	0.012	7.143	0.027	6.25	0.047	5.556	0.069
9.00	0.0081	7.714	0.021	6.75	0.038	6.000	0.057
9.33	0.0055	8.000	0.016	7.00	0.030	6.222	0.048
10.33	0.0017	8.857	0.0084	7.75	0.018	6.889	0.031
12.00	0.0001	10.286	0.0036	9.00	0.0099	8.000	0.019
	10.571	0.0027	9.25	0.0080	8.222	0.016	
	11.143	0.0012	9.75	0.0048	8.667	0.010	
	12.286	0.00032	10.75	0.0024	9.556	0.006	
	14.000	0.00002	12.00	0.0011	10.667	0.0035	
			12.25	0.0008	10.889	0.0029	
			13.00	0.0002	11.556	0.0013	
			14.25	0.0000	12.667	0.00066	
			16.00	0.0000	13.556	0.00035	

k = 4							
N = 2		N = 3		N = 4			
χ^2_r	P	χ^2_r	P	χ^2_r	P	χ^2_r	P
0,0	1,000	0,2	1,000	0,0	1,000	5,7	0,141
0,6	0,958	0,6	0,958	0,3	0,992	6,0	0,105
1,2	0,834	1,0	0,910	0,6	0,928	6,3	0,094
1,8	0,792	1,8	0,727	0,9	0,900	6,6	0,077
2,4	0,625	2,2	0,608	1,2	0,800	6,9	0,068
3,0	0,542	2,6	0,524	1,5	0,754	7,2	0,054
3,6	0,458	3,4	0,446	1,8	0,677	7,5	0,052
4,2	0,375	3,8	0,342	2,1	0,649	7,8	0,036
4,8	0,208	4,2	0,300	2,4	0,524	8,1	0,033
5,4	0,167	5,0	0,207	2,7	0,508	8,4	0,019
6,0	0,042	5,4	0,175	3,0	0,432	8,7	0,014
		5,8	0,148	3,3	0,389	9,3	0,012
		6,6	0,075	3,6	0,355	9,6	0,0069
		7,0	0,054	3,9	0,324	9,9	0,0062
		7,4	0,033	4,5	0,242	10,2	0,0027
		8,2	0,017	4,8	0,200	10,8	0,0016
		9,0	0,0017	5,1	0,190	11,1	0,00094
				5,4	0,158	12,0	0,00007

Fuente: Tablas de estadística.pdf.(pág.:38). Disponible en: http://materias.unq.edu.ar/pye/Trabajos%20Pr%C3%A1cticos/Tablas%20de%20Estadistica.pdf

Se rechaza la hipótesis nula (H_0) si el estadístico de prueba de Friedman chi-cuadrado (χ^2) observado-calculado es mayor que el valor crítico de χ^2_r (teórico-esperado) encontrado en esta tabla anterior y el mismo (observado) se corresponde con un valor de probabilidad (p) bajo H_0 menor que el nivel alfa preseleccionado por el investigador.

Tabla para los valores críticos superiores (teóricos-esperados) del coeficiente de correlación lineal de Pearson, en relación al número de observaciones (n) para las dos variables cuantitativas con distribución normal: x, y; según alfa (α) = 0,05 y 0,01, para una y dos colas.

	Una cola		Dos colas	
n	α=0.05	α=0.01	α=0.05	α=0.01
3	0.988	1.000	0.997	1.000
4	0.900	0.980	0.950	0.990
5	0.805	0.934	0.878	0.959
6	0.729	0.882	0.811	0.917
7	0.669	0.833	0.755	0.875
8	0.622	0.789	0.707	0.834
9	0.582	0.750	0.666	0.798
10	0.549	0.715	0.632	0.765
11	0.521	0.685	0.602	0.735
12	0.497	0.658	0.576	0.708
13	0.476	0.634	0.553	0.684
14	0.458	0.612	0.532	0.661
15	0.441	0.592	0.514	0.641
16	0.426	0.574	0.497	0.623
17	0.412	0.558	0.482	0.606
18	0.400	0.542	0.468	0.590
19	0.389	0.529	0.456	0.575
20	0.378	0.515	0.444	0.561
21	0.369	0.503	0.433	0.549
22	0.360	0.492	0.423	0.537
23	0.352	0.482	0.413	0.526
24	0.344	0.472	0.404	0.515
25	0.337	0.462	0.396	0.505
26	0.330	0.453	0.388	0.496
27	0.323	0.445	0.381	0.487
28	0.317	0.437	0.374	0.479
29	0.312	0.430	0.367	0.471
30	0.306	0.423	0.361	0.463
32	0.296	0.409	0.349	0.449
34	0.287	0.397	0.339	0.436
36	0.279	0.386	0.329	0.424
38	0.271	0.376	0.320	0.413
40	0.264	0.367	0.312	0.403
42	0.257	0.358	0.304	0.393
44	0.251	0.350	0.297	0.384
46	0.246	0.342	0.291	0.376
48	0.240	0.335	0.285	0.368
50	0.235	0.328	0.279	0.361
52	0.231	0.322	0.273	0.354
54	0.226	0.316	0.268	0.348
56	0.222	0.310	0.263	0.341
58	0.218	0.305	0.259	0.336
60	0.214	0.300	0.254	0.330
62	0.211	0.295	0.250	0.325
64	0.207	0.290	0.246	0.320
66	0.204	0.286	0.242	0.315
68	0.201	0.282	0.239	0.310
70	0.198	0.278	0.235	0.306
72	0.195	0.274	0.232	0.302
74	0.193	0.270	0.229	0.298
76	0.190	0.266	0.226	0.294
78	0.188	0.263	0.223	0.290
80	0.185	0.260	0.220	0.286
82	0.183	0.257	0.217	0.283
84	0.181	0.253	0.215	0.280
86	0.179	0.251	0.212	0.276
88	0.177	0.248	0.210	0.273
90	0.174	0.245	0.207	0.270
92	0.173	0.242	0.205	0.267
94	0.171	0.240	0.203	0.264
96	0.169	0.237	0.201	0.262
98	0.167	0.235	0.199	0.259
100	0.165	0.232	0.197	0.257

Fuente: Autores (2004). Métodos de análisis de datos en Ecología. España, Alcalá: Universidad de Alcalá, p. 30

Para probar H_0: $\rho = 0$ contra H_1: $\rho \neq 0$; rechace H_0 si el valor absoluto observado de r es mayor que el valor crítico en la tabla anterior.

Tabla para los valores críticos superiores (teóricos-esperados) del coeficiente de correlación por rangos de Spearman, en relación al número de observaciones (n) para los valores combinados de las dos variables cuantitativas de libre distribución: x, y; según los niveles de significación alfa (α) = .10, .05, .01 para dos lados o colas (*two sided* α).

	Two-Sided α		
	.10	.05	.01
n			
5	.900	—	—
6	.829	.886	—
7	.714	.786	.929
8	.643	.738	.881
9	.600	.700	.833
10	.564	.648	.794
11	.536	.618	.818
12	.497	.591	.780
13	.475	.566	.745
14	.457	.545	.716
15	.441	.525	.689
16	.425	.507	.666
17	.412	.490	.645
18	.399	.476	.625
19	.388	.462	.608
20	.377	.450	.591
21	.368	.438	.576
22	.359	.428	.562
23	.351	.418	.549
24	.343	.409	.537
25	.336	.400	.526
26	.329	.392	.515
27	.323	.385	.505
28	.317	.377	.496
29	.311	.370	.487
30	.305	.364	.478

Fuente: Rosner, B. (2000). *Fundamentals of Biostatistics*. 5ta. ed. Belmont: Duxbury Press, p. 834.

Para probar H_0: $\rho_s = 0$ contra H_1: $\rho_s \neq 0$; rechace H_0 si el valor absoluto observado de r_s es mayor que el valor crítico en la tabla anterior.

Tabla de los valores de P {S > s} según n ≤ 10, para el estadígrafo *S* presente en la fórmula matemática y representante del coeficiente de correlación de rangos de Kendall

S	Valores de *n*				*S*	Valores de *n*		
	4	5	8	9		6	7	10
0	0,625	0,592	0,548	0,540	1	0,500	0,500	0,500
2	0,375	0,408	0,452	0,460	3	0,360	0,386	0,431
4	0,167	0,242	0,360	0,381	5	0,235	0,281	0,364
6	0,042	0,117	0,274	0,306	7	0,136	0,191	0,300
8		0,042	0,199	0,238	9	0,068	0,119	0,242
10		0,0083	0,138	0,179	11	0,028	0,068	0,190
12			0,089	0,130	13	0,0083	0,035	0,146
14			0,054	0,090	15	0,0014	0,015	0,108
16			0,031	0,060	17		0,0054	0,078
18			0,016	0,038	19		0,0014	0,054
20			0,0071	0,022	21		0,00020	0,036
22			0,0028	0,012	23			0,023
24			0,00087	0,0063	25			0,014
26			0,00019	0,0029	27			0,0083
28			0,000025	0,0012	29			0,0046
30				0,00043	31			0,0023

Fuente: *Bioestadistica cualitativa*. (p. 63).

Tabla de los valores críticos de la varianza (S^2) presente en la fórmula del coeficiente de concordancia de rangos de Kendall para n < 8 y un valor de alfa (α) = 0,05 y 0,01

Valores de S^2						**Valores de S^2 (adicionales para n=3)**		
k	n=3	n=4	n=5	n=6	n=7	k	S^2	1-α
3			12,9	17,3	22,5	9	18,0	0,95
			15,1	20,5	26,5	9	25,3	0,99
4		12,4	17,7	23,9	31,0	12	24,0	0,95
		15,4	21,9	29,4	37,9	12	34,5	0,99
5		15,7	22,1	30,4	39,5	14	27,9	0,95
		20,1	28,5	38,2	49,1	14	40,6	0,99
6		18,9	27,2	36,9	47,9	16	31,9	0,95
		24,6	35,2	47,1	60,4	16	46,7	0,99
8	16,0	25,4	36,7	49,8	64,7	18	35,9	0,95
	22,3	34,3	48,5	64,7	82,8	18	52,9	0,99
10	20,0	32,0	46,2	62,8	81,6			0,95
	28,4	43,8	61,8	82,3	105,3			
15	29,9	48,2	70,0	95,1	123,6			
	43,7	67,5	95,4	126,4	161,4			
20	39,9	64,5	93,7	127,4	165,5			
	59,0	91,5	128,2	170,4	217,4			

Fuente: *Bioestadistica cualitativa*. (pp. 63-64).

Tabla de valores críticos para el coeficiente de correlación punto biserial

Grados de libertad	Valor de alfa		Grados de libertad	Valor de alfa	
n-2	0,05	0,01	n-2	0,05	0,01
1	0,097	1.00	24	0,388	0,496
2	0,950	0,990	25	0,381	0,487
3	0,878	0,959	26	0,374	0,478
4	0,811	0,917	27	0,367	0,470
5	0,754	0,874	28	0,361	0,463
6	0,707	0,834	29	0,355	0,456
7	0,666	0,798	30	0,349	0,449
8	0,632	0,756	35	0,325	0,418
9	0,576	0,708	40	0,304	0,393
10	0,576	0,708	45	0,288	0,372
11	0,553	0,684	50	0,273	0,354
12	0,532	0,661	60	0,250	0,325
13	0,514	0,641	70	0,232	0,302
14	0,497	0,623	80	0,217	0,283
15	0,482	0,606	90	0,205	0,267
16	0,468	0,590	100	0,195	0,254
17	0,456	0,575	125	0,174	0,228
18	0,444	0,561	150	0,159	0,208
19	0,433	0,549	200	0,138	0,181
20	0,423	0,537	300	0,113	0,148
21	0,413	0,526	400	0,098	0,128
22	0,404	0,515	500	0,088	0,115
23	0,396	0,505	1000	0,062	0,081

Fuente: *Bioestadistica cualitativa*. pp. 61-62.

Tabla de los valores críticos (teóricos esperados) o puntos de porcentage en la prueba de Bartholomew según el valor de (c) calculado y el nivel alfa especifico empleados

	Valores alfa				
c	.10	.05	.025	.01	.005
0.0	2,952	4,231	5,537	7,289	8,628
0.1	2,885	4,158	5,459	7,208	8,543
0.2	2,816	4,081	5,378	7,122	8,455
0.3	2,742	4,001	5,292	7,030	8,360
0.4	2,664	3,914	5,200	6,932	8,258
0.5	2,580	3,820	5,098	6,822	8,146
0.6	2,486	3,715	4,985	6,700	8,016
0.7	2,379	3,593	4,052	6,556	7,865
0.8	2,251	3,446	4,689	6,377	7,677
0.9	2,080	3,245	4,465	6,130	7,413
1.0	1,642	2,706	3,841	5,413	6,635

Tabla de los valores críticos-esperados o puntos de porcentaje para el valor de (c1) y (c2) calculado y el nivel alfa (alfa en la tabla) especifico mediante la prueba de Bartholomew

		C1							
C2	**ALFA**	**0,0**	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**
	.10	4,010							
	.05	5,435							
0,0	.025	6,061							
	.01	8,744							
	.005	10,171							
	.10	3,952	3,491						
	.05	5,112	5,105						
0,1	.025	6,194	6,724						
	.01	8,676	8,001						
	.005	10,090	10,020						
	.10	3,893	3,827	3, 756					
	.05	5,307	5,235	5,160					
0,2	.025	6,725	6,645	6,570					
	.01	8,602	8,522	8,437					
	.005	10,022	9,939	9,851					
	.10	3,891	3,760	3,685	3,606				
	.05	5,239	5,162	5,000	4,993				
0,3	.025	6,655	6,571	6,484	6,391				
	.01	8,525	8,438	8,346	8,266				
	.005	9,942	9,852	9,796	9,853				
	.10	3,765	3,688	3,607	3,519	3,423			
	.05	5,166	5,083	4,994	4,898	4,791			
0,4	.025	6,575	6,486	6,342	6,209	6,174			
	.01	8,447	8,348	8,247	8,137	8,014			
	.005	9,855	9,758	9,653	9,539	9,411			
	.10	3,695	3,610	3,521	3,423	3,313	3,187		
	.05	5,088	4,997	4,898	4,791	4,670	4,528		
0,5	.025	6,491	6,394	6,289	6,173	6,043	5,891		
	.01	8,352	8,246	8,136	8,013	7,873	7,789		
	.005	9,761	9,654	9,537	9,409	9,266	9,092		

		C1							
C2	ALFA	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7
	.10	3,617	3,523	3,422	3,310	3,183	3,031	2,837	
	.05	5,002	4,900	4,789	4,665	4,524	4,394	4,135	
0,6	.025	6,398	6,289	6,170	6,038	5,886	5,762	5,442	
	.01	8,251	8,135	8,000	7,867	7,703	7,504	7,244	
	.005	9,656	9,535	9,404	9,256	9,085	8,877	8,604	
	.10	3,530	3,422	3,305	3,172	3,017	2,822	2,550	1,987
	.05	4,904	4,787	4,657	4,510	4,337	4,118	3,665	3,137
0,7	.025	6,291	6,166	6,027	5,870	5,682	5,443	5,100	4,346
	.01	8,135	8,002	7,854	7,684	7,482	7,223	6,046	6,000
	.005	9,534	9,395	9,242	9,065	8,853	8,561	8,183	7,279
	.10	3,427	3,296	3,151	2,981	2,770	2,473	1,642	
	.05	4,707	4,644	4,483	4,294	4,056	3,715	2,706	
0,8	.025	6,163	6,011	5,838	5,634	5,375	4,999	3,041	
	.01	7,994	7,832	7,647	7,427	7,146	6,734	5,412	
	.005	9,385	9,217	9,025	8,795	8,500	8,064	6,635	
	.10	3,291	3,110	2,897	2,621	2,166			
	.05	4,631	4,432	4,195	3,883	3,353			
0,9	.025	5,990	5,770	5,523	5,102	4,591			
	.01	7,804	7,577	7,303	6,933	6,277			
	.005	9,103	8,948	8,661	8,273	7,576			
	.10	2,952							
	.05	4,231							
1,0	.025	5,537							
	.01	7,289							
	.005	8,628							

Tabla de los puntos de porcentaje o valores críticos según el número de muestras (K) y el nivel alfa especifico, mediante la prueba de Bartholomew, para evaluar el gradiente de variación de las proporciones ordenadas de hasta 12 muestras independientes con igual tamaño de muestra al ser comparadas

	Valores alfa				
K	.10	.05	.025	.01	.005
3	2,580	3,820	5,098	6,822	8,146
4	3,187	4,528	5,891	7,709	9,092
5	3,636	5,049	6,471	8,356	9,784
6	3,994	5,460	6,928	8,865	10,327
7	4,289	5,800	7,304	9,284	10,774
8	4,542	6,088	7,624	9,639	11,153
9	4,761	6,339	7,901	9,946	11,480
10	4,956	6,560	8,145	10,216	11,767
11	5,130	6,758	8,363	10,458	12,025
12	5,288	6,937	8,561	10,676	12,257

Fuente: Statistical Methods for Rates and Proportions(second edition).(pp. 291-294)

Tabla de valores teóricos-esperados en la distribución de probabilidades para el estadístico R de Cochran, en la prueba de homogeneidad de varianzas (comparación de más de dos varianzas), según el número de grupos (k) y el número de observaciones (n) en los mismos, para el nivel de significación alfa ($\alpha\alpha$) correspondiente.

$\alpha = 0.05$

	n													
k	2	3	4	5	6	7	8	9	10	11	17	37	145	1000
2	0.998	0.975	0.939	0.905	0.877	0.853	0.833	0.815	0.801	0.788	0.734	0.660	0.581	0.500
3	0.966	0.870	0.797	0.745	0.707	0.677	0.653	0.633	0.616	0.602	0.546	0.474	0.403	0.333
4	0.906	0.767	0.684	0.628	0.589	0.559	0.536	0.517	0.501	0.488	0.436	0.372	0.309	0.250
5	0.841	0.683	0.598	0.544	0.506	0.478	0.456	0.438	0.424	0.411	0.364	0.306	0.251	0.200
6	0.780	0.616	0.532	0.480	0.444	0.418	0.398	0.381	0.368	0.336	0.313	0.261	0.211	0.166
7	0.727	0.561	0.480	0.430	0.397	0.372	0.353	0.338	0.352	0.315	0.275	0.227	0.183	0.142
8	0.679	0.515	0.437	0.391	0.350	0.336	0.318	0.304	0.292	0.282	0.246	0.202	0.161	0.125
9	0.638	0.477	0.402	0.358	0.328	0.306	0.290	0.276	0.265	0.256	0.222	0.182	0.144	0.111
10	0.602	0.445	0.373	0.331	0.302	0.282	0.266	0.254	0.243	0.735	0.203	0.165	0.130	0.100

$\alpha = 0.01$

	n													
k	2	3	4	5	6	7	8	9	10	11	17	37	145	1000
2	0.999	0.995	0.979	0.958	0.937	0.917	0.898	0.882	0.867	0.853	0.794	0.706	0.606	0.500
3	0.993	0.942	0.883	0.831	0.793	0.760	0.733	0.710	0.691	0.674	0.605	0.515	0.423	0.333
4	0.967	0.864	0.781	0.711	0.676	0.641	0.612	0.589	0.570	0.553	0.488	0.405	0.325	0.250
5	0.927	0.788	0.693	0.632	0.587	0.553	0.525	0.503	0.485	0.469	0.409	0.335	0.264	0.200
6	0.882	0.721	0.625	0.563	0.519	0.486	0.460	0.440	0.422	0.408	0.352	0.285	0.222	0.166
7	0.837	0.664	0.568	0.508	0.465	0.434	0.410	0.391	0.375	0.361	0.310	0.249	0.192	0.142
8	0.794	0.615	0.520	0.462	0.422	0.393	0.370	0.352	0.337	0.324	0.277	0.221	0.170	0.125
9	0.754	0.572	0.481	0.425	0.387	0.359	0.337	0.320	0.306	0.295	0.251	0.199	0.152	0.111
10	0.717	0.535	0.446	0.393	0.357	0.330	0.310	0.294	0.251	0.270	0.281	0.181	0.137	0.100

Fuente: Tablas estadísticas (p. 3).Disponible en: http://www4.ujaen.es/~mpfrias/TablasInferencia.pdf

Tabla de valores críticos para el estadístico D en la prueba de normalidad de D'Agostino, según el número de observaciones para cada muestra y el nivel de significación alfa (α)

Nivel de significación α(expresados en por ciento)

n	0.5%	1%	2.5%	5%	10%
10	-4.66	-4.06	-3.25	-2.62	-1.99
12	-4.63	-4.02	-3.20	-2.58	-1.94
14	-4.57	-3.97	-3.16	-2.53	-1.90
16	-4.52	-3.92	-3.12	-2.50	-1.87
18	-4.47	-3.87	-3.08	-2.47	-1.85
20	-4.41	-3.83	-3.04	-2.44	-1.82
22	-4.36	-3.78	-3.01	-2.41	-1.81
24	-4.32	-3.75	-2.98	-2.39	-1.79
26	-4.27	-3.71	-2.96	-2.37	-1.77
28	-4.23	-3.68	-2.93	-2.35	-1.76
30	-4.19	-3.64	-2.91	-2.33	-1.75
32	-4.16	-3.61	-2.88	-2.32	-1.73
34	-4.12	-3.59	-2.86	-2.30	-1.72
36	-4.09	-3.56	-2.85	-2.29	-1.71
38	-4.06	-3.54	-2.83	-2.28	-1.70
40	-4.03	-3.51	-2.81	-2.26	-1.70
42	-4.00	-3.49	-2.80	-2.25	-1.69
44	-3.98	-3.47	-2.78	-2.24	-1.68
46	-3.95	-3.45	-2.77	-2.23	-1.67
48	-3.93	-3.43	-2.75	-2.22	-1.67
50	-3.91	-3.41	-2.74	-2.21	-1.66
60	-3.81	-3.34	-2.68	-2.17	-1.64
70	-3.73	-3.27	-2.64	-2.14	-1.61
80	-3.67	-3.22	-2.60	-2.11	-1.59
90	-3.61	-3.17	-2.57	-2.09	-1.58
100	-3.57	-3.14	-2.54	-2.07	-1.57

Fuente: *Testing for Normality* (p. 398).

Valores críticos de q (rango studentizado, empleado en la prueba de Duncan), según los grados de libertad (θ) (primera columna) y el número de paso de la tabla respectiva de las diferencias de medias entre grupos (primera fila) para un valor de alfa (α) = 0,05

	Paso								
θ	2	3	4	5	6	7	8	9	10
1	17.969	17.969	17.969	17.969	17.969	17.969	17.969	17.969	17.969
2	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085
3	4.501	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516
4	3.926	4.013	4.033	4.033	4.033	4.033	4.033	4.033	4.033
5	3.635	3.749	3.796	3.814	3.814	3.814	3.814	3.814	3.814
6	3.460	3.586	3.649	3.680	3.694	3.697	3.697	3.697	3.697
7	3.344	3.477	3.548	3.588	3.611	3.622	3.625	3.625	3.625
8	3.261	3.398	3.475	3.521	3.549	3.566	3.575	3.579	3.579
9	3.199	3.339	3.420	3.470	3.502	3.523	3.536	3.544	3.547
10	3.151	3.293	3.376	3.430	3.465	3.489	3.505	3.516	3.522
11	3.113	3.256	3.341	3.397	3.435	3.462	3.480	3.493	3.501
12	3.081	3.225	3.312	3.370	3.410	3.439	3.459	3.474	3.484
13	3.055	3.200	3.288	3.348	3.389	3.419	3.441	3.458	3.470
14	3.033	3.178	3.268	3.328	3.371	3.403	3.426	3.444	3.457
15	3.014	3.160	3.250	3.312	3.356	3.389	3.413	3.432	3.446
16	2.998	3.144	3.235	3.297	3.343	3.376	3.402	3.422	3.437
17	2.984	3.130	3.222	3.285	3.331	3.365	3.392	3.412	3.429
18	2.971	3.117	3.210	3.274	3.320	3.356	3.383	3.404	3.421
19	2.960	3.106	3.199	3.264	3.311	3.347	3.375	3.397	3.415
20	2.950	3.097	3.190	3.255	3.303	3.339	3.368	3.390	3.409
21	2.941	3.088	3.181	3.247	3.295	3.332	3.361	3.385	3.403
22	2.933	3.080	3.173	3.239	3.288	3.326	3.355	3.379	3.398
23	2.926	3.072	3.166	3.233	3.282	3.320	3.350	3.374	3.394
24	2.919	3.066	3.160	3.226	3.276	3.315	3.345	3.370	3.390
25	2.913	3.059	3.154	3.221	3.271	3.310	3.341	3.366	3.386

26	2.907	3.054	3.149	3.216	3.266	3.305	3.336	3.362	3.382
27	2.902	3.049	3.144	3.211	3.262	3.301	3.332	3.358	3.379
28	2.897	3.044	3.139	3.206	3.257	3.297	3.329	3.355	3.376
29	2.892	3.039	3.135	3.202	3.253	3.293	3.326	3.352	3.373
30	2.888	3.035	3.131	3.199	3.250	3.290	3.322	3.349	3.371
31	2.884	3.031	3.127	3.195	3.246	3.287	3.319	3.346	3.368
32	2.881	3.028	3.123	3.192	3.243	3.284	3.317	3.344	3.366
33	2.877	3.024	3.120	3.188	3.240	3.281	3.314	3.341	3.364
34	2.874	3.021	3.117	3.185	3.238	3.279	3.312	3.339	3.362
35	2.871	3.018	3.114	3.183	3.235	3.276	3.309	3.337	3.360
36	2.868	3.015	3.111	3.180	3.232	3.274	3.307	3.335	3.358
37	2.865	3.013	3.109	3.178	3.230	3.272	3.305	3.333	3.356
38	2.863	3.010	3.106	3.175	3.228	3.270	3.303	3.331	3.355
39	2.861	3.008	3.104	3.173	3.226	3.268	3.301	3.330	3.353
40	2.858	3.005	3.102	3.171	3.224	3.266	3.300	3.328	3.352
48	2.843	2.991	3.087	3.157	3.211	3.253	3.288	3.318	3.342
60	2.829	2.976	3.073	3.143	3.198	3.241	3.277	3.307	3.333
80	2.814	2.961	3.059	3.130	3.185	3.229	3.266	3.297	3.323
120	2.800	2.947	3.045	3.116	3.172	3.217	3.254	3.286	3.313
240	2.786	2.933	3.031	3.103	3.159	3.205	3.243	3.276	3.304
∞	2.772	2.918	3.017	3.089	3.146	3.193	3.232	3.265	3.294

Fuente: Polo, J. C. (2011). *Apuntes sobre bioestadística aplicada a las ciencias farmacéuticas*. La Habana: Editorial Félix Varela. (Apéndice 13).

Tabla de números aleatorios (ejemplo reducido).

25 19 64 82 84	62 74 29 92 24	61 03 91 22 48	23 02 41 56 04
44 31 52 43 07	44 06 03 09 34	55 85 66 96 28	28 30 62 58 83
65 68 62 42 45	68 45 19 69 59	35 14 82 56 80	22 06 52 26 39
69 31 46 29 85	18 88 26 95 54	01 02 14 03 05	37 31 61 28 98
94 61 47 02 10	67 80 84 41 26	66 42 19 24 94	13 13 28 69 96
76 69 76 24 13	33 65 78 12 35	91 59 11 38 44	23 31 48 75 74
76 32 12 19 35	22 95 30 19 29	57 74 43 20 90	43 33 42 01 59
20 39 84 95 61	58 22 04 02 99	28 31 93 43 94	87 73 19 38 47
54 36 90 98 10	97 19 21 63 34	69 33 17 03 02	11 15 50 46 08
82 80 37 14 20	56 39 59 89 63	33 90 38 44 50	03 68 03 13 60
64 13 09 37 11	86 02 57 41 99	65 16 58 11 01	98 78 80 63 23
07 37 66 20 56	24 65 58 57 04	18 62 85 28 24	26 45 17 82 76
02 72 61 07 75	85 66 48 38 73	75 10 96 59 31	79 16 78 63 99
43 61 00 66 42	76 26 71 24 33	04 75 14 93 39	68 52 16 83 34
64 09 44 62 58	40 64 64 57 60	97 00 12 91 33	22 14 73 01 11
06 21 07 34 26	01 51 48 69 57	19 17 53 55 96	62 40 03 87 10
96 88 22 46 94	35 56 60 94 20	00 98 48 18 97	91 51 63 27 95
74 25 84 03 07	50 64 19 18 91	98 55 83 46 09	49 66 41 12 45
38 54 52 25 78	01 98 00 89 85	86 12 22 89 25	46 86 80 97 78
65 12 64 64 70	58 41 05 49 08	90 72 92 33 10	09 12 81 93 63
60 30 02 04 26	66 21 41 77 60	99 25 72 61 22	52 40 74 67 29
87 05 46 52 76	89 96 34 22 37	27 11 57 04 19	46 90 61 03 06
89 85 32 22 80	34 89 12 29 37	18 88 53 06 09	81 83 33 98 29
91 27 59 43 09	11 05 92 06 97	68 82 34 08 83	25 40 58 40 64
33 94 24 20 28	62 44 07 12 63	34 39 02 92 31	24 89 74 75 61
61 02 73 36 85	67 28 50 49 85	15 19 74 67 23	61 38 93 73 68
76 23 15 58 20	05 64 12 70 88	80 58 35 06 88	73 48 27 39 43
57 49 36 44 06	74 93 55 39 26	27 70 98 76 68	77 82 96 96 97
60 42 17 18 48	16 34 92 19 52	24 10 70 06 51	59 62 37 95 42
53 67 14 95 29	50 00 07 78 23	49 54 36 85 14	18 50 54 18 82

Fuente: Colectivo de autores (2004). *Informática médica: Bioestadística*. Tomo 2. La Habana: Editorial Ciencias Médicas, p. 630.

Anexo 2

Formularios para el manejo de muestras y variables relacionados con los contrastes de hipótesis de la inferencia estadística y algunos aspectos afines

Estadísticos de prueba (una población con una variable)

Métodos paramétricos

Inferencia acerca de μ, elección entre la distribución normal y la t-student:

- Distribución normal: se conoce σ y la población se distribuye normalmente (n > 30).
- Distribución t-student: se desconoce σ y la población se distribuye normalmente (n > 30).
- Una media con σ conocida:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

- Una media con σ desconocida:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

- Una proporción:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Métodos no paramétricos

- Una varianza o desviación estándar:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Nota: cuando la variable no sigue ni se aproxima a una distribución normal y se desconoce σ y $n \leq 30$ se puede emplear en los programas estadísticos métodos no paramétricos transformando los valores de los datos a rangos.

Intervalos de confianza (una población, una variable)

– Una media con σ conocida:

$\bar{x} - E < \bar{x} + E$

donde:

$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$, para poblaciones infinitas, grandes;

$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$, para poblaciones finitas, pequeñas.

– Una media con σ desconocida:

$\bar{x} - E < \mu < \bar{x} + E$

donde:

$E = z_{\alpha/2} \frac{s}{\sqrt{n}}$

– Una proporción:

$\hat{p} - E < \hat{p} + E$

donde:

$E = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \sqrt{\frac{N-n}{N-1}}$, para población infinita o muy grande y se realiza muestreo con reemplazo;

$E = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \sqrt{\frac{N-n}{N-1}}$, para población finita o cuando se realiza muestreo sin reemplazo.

– Una varianza cuando n ≤ 101, es decir, para grados de libertad n − 1 ≤ 100:

$\frac{(n-1)s^2}{\chi^2_R} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L}$

donde:

χ^2_L y χ^2_R son los valores críticos tabulados de chi-cuadrado, que son calculados a partir del número de grados de libertad y el nivel alfa seleccionado; o sea, para la cola izquierda es χ^2_L o $\chi^2_{1-\alpha/2}$ y la cola derecha χ^2_R o $\chi^2_{\alpha/2}$, respectivamente.

En caso contrario n > 101, o sea, para grados de libertad n − 1 > 100, se puede aproximar χ^2_I y χ^2_D utilizando la expresión:

$\chi^2 = \frac{1}{2}\left[\pm z_{\alpha/2} + \sqrt{2k-1}\right]^2$ donde k = n − 1 (grados de libertad).

Estadísticos de prueba (una población con dos variables)

Métodos paramétricos

– Correlación lineal simple entre dos variables cuantitativas con normalidad (contraste de independencia para el coeficiente de correlación lineal simple):

$$t = \sqrt{\frac{(n-2)r^2}{1-r^2}} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

donde: $r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$

– Regresión lineal simple entre dos variables cuantitativas con normalidad (contraste de independencia entre las variables por medio del coeficiente de regresión lineal simple):

$$t = \frac{b - B}{\sqrt{\frac{s^2}{(xx)}}}$$

– Ecuación estimada de la recta de regresión lineal simple entre dos variables cuantitativas:

$$\hat{y} = b_0 + b_1 x$$

donde:

$$b_0 = \overline{y} - b_1 \overline{x}$$

$$b_0 = \frac{(\Sigma y)(\Sigma x)^2 - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b_0 = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

– ANOVA simple de Fisher para evaluar el ajuste del modelo de regresión lineal simple (ecuación de regresión) a los datos muestrales (experimentales):

$F = \frac{CM_e}{CM_{ne}} = \frac{\sum[\hat{y}_i - \overline{y}]^2}{\frac{\sum[y_i - y_i]^2}{n-2}}$ que sigue una distribución de probabilidades F de Fisher-Snedecor con 1 grado de libertad en el numerador y de n – 2 en el denominador.

Métodos no paramétricos

– Correlación de rangos ordenados (contraste de independencia entre las dos variables cuantitativas de libre distribución por medio del coeficiente por rangos de Spearman, se emplea para n > 30:

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

Nota: se puede emplear: $\frac{\pm z}{\sqrt{n-1}}$ para estimar el valor crítico de r_S cuando n > 30.

– Correlación de rangos ordenados (contraste de independencia entre las dos variables cuantitativas de libre distribución por medio del coeficiente por rangos de Kendall, se emplea para n > 10):

$$Z = \frac{r_K}{\sqrt{\frac{4n+10}{9n(n-1)}}} = r_K\sqrt{\frac{9n(n-1)}{4n+10}}$$

– Concordancia de rangos ordenados: contraste de concordancia de Kendall empleando la distribución chi-cuadrado mediante el procedimiento de la prueba de rangos de Friedman para rangos adjudicados a los datos de una variable cuantitativa por calificadores o evaluadores:
$\chi^2 = k(n-1)r_{CK}$ con k – 1 grados de libertad.

– Independencia o asociación entre dos variables cualitativas, empleando tablas de contingencia (fxc):

$\chi^2 = \sum\frac{(O-E)^2}{E}$ con (f – 1) (c – 1) grados de libertad;

donde:

$$E = \frac{(\text{Total por renglón})(\text{Total por columna})}{(\text{Gran total})}$$

– Intervalo de confianza (una población, dos variables cuantitativas):

- Coeficiente de correlación lineal simple poblacional estimado (ρ):

$$\frac{e^{2w_l}-1}{e^{2w_l}+1} < \frac{e^{2w_b}-1}{e^{2w_b}+1}$$

donde:

$$w_l = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - z_{\alpha/2}\cdot\frac{1}{\sqrt{n-3}}\ ;$$

$$w_l = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) + z_{\alpha/2}\cdot\frac{1}{\sqrt{n-3}}\ ;$$

- Coeficiente de regresión lineal simple poblacional estimado (β):

$$\beta = \left[b \pm (t_{1-\alpha/2,n-2}) \sqrt{\frac{s^2}{(xx)}} \right]$$

- Ecuación de la recta de regresión lineal simple poblacional estimada (y):

  $$\hat{y} - E < y < \hat{y} + E$$

  donde:

  $$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}}$$

  $$s_e = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n-2}} = \sqrt{\frac{\Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy}{n-2}}$$

Estadísticos de prueba (dos poblaciones con una variable)

Métodos paramétricos

– Dos medias no pareadas con σ_1 y σ_2 conocidas:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

– Dos medias no pareadas con σ_1 y σ_2 desconocidas, pero se suponen iguales, con grados de libertad $n_1 + n_2 - 2$:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

donde:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

– Dos medias no pareadas con σ_1 y σ_2 desconocidas, pero no se suponen iguales, y con grados de libertad el menor de $n_1 - 1$, $n_2 - 1$:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

– Dos medias pareadas con grados de libertad $n - 1$:

$$t = \frac{\bar{d} - \mu_d}{s_d \sqrt{n}}$$

– Dos proporciones:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\overline{pq}}{n_1} + \frac{\overline{pq}}{n_2}}}$$

– Dos varianzas o desviaciones estándar:

$$F = \frac{s_1^2}{s_2^2}$$

donde:

$s_1^2 \geq s_2^2$;

$n_1 - 1$: grados de libertad para el numerador;
$n_2 - 1$: grados de libertad para el denominador.

Métodos no paramétricos

– Prueba de homogeneidad en las dos muestras independientes:

$\chi^2 = \sum \frac{(O-E)^2}{E}$ por tablas de contingencia fxc con (f – 1) (c – 1) grados de libertad;

donde:

$$E = \frac{(\text{Total por renglón})(\text{Total por columna})}{(\text{Gran total})}$$

– Prueba de McNemar para homogeneidad en las dos muestras pareadas para una variable cualitativa dicotómica:

$\chi^2 = \frac{(|b-c|-1)^2}{b+c}$ por tablas de contingencia fxc con (f – 1) (c – 1) = 1 grados de libertad;

– Prueba del signo de Wilcoxon (dos muestras pareadas para n > 25):

$$z = \frac{(x+0{,}5) - (n/2)}{\sqrt{n}/2} \ .$$

– Prueba de rangos y signos de Wilcoxon (dos muestras pareadas para n > 30):

$$z \frac{T - n(n+1)/4}{\sqrt{\frac{n(1-1)(2n+1)}{24}}}$$

– Prueba de Mann Whitney (dos muestras independientes para $n_1 + n_2 > 20$):

$$z = \frac{U - (n_1 n_2)/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}}$$

– Prueba de la suma de rangos de Wilcoxon (dos muestras independientes para $n_1 + n_2 > 20$):

$$z = \frac{R - \mu_R}{\sigma_R} = \frac{R - \frac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

– Prueba de rachas o de tendencias (una muestra dividida en dos grupos de datos o submuestras, según una determinada característica para n_1 y $n_2 > 20$):

$$z = \frac{G - \mu_G}{\sigma_G} = \frac{G - \left(\frac{2n_1 n_2}{n_1 + n_2} + 1\right)}{\sqrt{\frac{(2n_1 n_2)(2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}}$$

– Intervalos de confianza (dos poblaciones, una variable):

- $(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E$
  donde:

  $E = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ para la diferencia de dos medias no pareadas con σ_1 y σ_2 conocidas;

  $E = t_{\alpha/2}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$ y $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$ para la diferencia de dos medias no pareadas con σ_1 y σ_2 desconocidas ; pero se suponen iguales;

  $E = t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ para la diferencia de dos medias no pareadas con σ_1 y σ_2 desconocidas, pero no se suponen iguales.

- $\bar{d} - E < \mu_d < \bar{d} + E$
  donde:

  $E = z_{\alpha/2}\frac{s_d}{\sqrt{n}}$ para la diferencia de dos medias en muestras pareadas;

- $(\hat{p}_1 - \hat{p}_2) - E < (p_1 - p_2) < (\hat{p}_1 - \hat{p}_2) + E$
  donde:

  $E = z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$ para la diferencia de dos proporciones.

Estadísticos de prueba (más de dos poblaciones y una variable)

Métodos paramétricos

– Para tamaños muestrales (n) iguales:
  - Primer método: ANOVA de un factor, de una vía, de una entrada:

$$F = \frac{\text{Varianza entre las muestras}}{\text{Varianza de las muestras}} = \frac{\text{Varianza para las medias muestrales}}{\text{Media de las varianzas muestrales}}$$

$$F = \frac{ns_{\bar{x}}^2}{s_p^2}$$

Donde:

$s_{\bar{x}}^2$: varianza de las medias muestrales o de los grupos de datos;

n: tamaño de cada una de las muestras;

s_p^2 : varianza agrupada que se obtiene al calcular la media de las varianzas muestrales.

k – 1: grados de libertad en el numerador;

k(n – 1) o N – k, donde N = ∑n, para el denominador.

- Segundo método: ANOVA de una vía por el método de la suma de cuadrados (SC) y cuadrados medios (CM):

$$F = \frac{CM(\text{del tratamiento})}{CM(\text{del error})}$$

donde:

k – 1: grados de libertad en el numerador;

N – k: grados de libertad en el denominador;

$CM(\text{del tratamiento}) = \frac{SC(\text{del tratamiento})}{k-1}$: intervarianza entre grupos, entre muestras y entre tratamientos);

$CM(\text{del error}) = \frac{SC(\text{del error})}{N-k}$: intravarianza dentro de los grupos, dentro de las muestras y del error;

$CM(\text{total}) = \frac{SC(\text{total})}{N-1}$: cuasivarianza total;

$SC(\text{del tratamiento}) = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \ldots + n_k(\bar{x}_k - \bar{\bar{x}})^2 = \Sigma n_i(\bar{x}_i - \bar{\bar{x}})^2$: entre los grupos, entre las muestras y de tratamientos;

$SC(\text{del error}) = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_k - 1)s_k^2 = \Sigma(n_i - 1)s_i^2$: dentro de los grupos, dentro de las muestras y del error;

$SC(\text{total}) = \Sigma(x - \bar{\bar{x}})^2 = SC(\text{del tratamiento}) + SC(\text{del error})$

– Para tamaños muestrales (n) desiguales:

$$F = \frac{\text{Varianza entre las muestras}}{\text{Varianza dentro de las muestras}} = \frac{\left[\frac{\Sigma n_i(\bar{x}_i - \bar{\bar{x}})^2}{k-1}\right]}{\left[\frac{\Sigma(n_i - 1)s_i^2}{\Sigma(n_i - 1)}\right]}$$

donde:

$\bar{\bar{x}}$: media de todos los valores muestrales combinados;

k: número de medias poblacionales estimadas (grupos de datos) que se están comparando;

n_i: número de valores en la i-ésima muestra;

$\bar{x}_i$: media de los valores en la i-ésima muestra;

s_i^2 : varianza de los valores en la i-ésima muestra.

k – 1: grados de libertad en el numerador;

N – k: grados de libertad en el denominador.

Métodos no paramétricos

- Prueba de Kruskal-Wallis con una variable cuantitativa, más de dos grupos de datos independientes o no relacionados a comparar:

$$\text{K de Kruskal-Wallis} = \frac{12}{N(N+1)}\left(\frac{R_1^2}{n_1}+\frac{R_2^2}{n_2}+\ldots+\frac{R_k^2}{n_k}\right)-3(N+1)$$

que sigue una distribución de chi-cuadrado con k – 1 grados de libertad (k es el número de grupos de datos independientes) para N > 15.

- Prueba de Friedman con una variable cuantitativa, más de dos grupos de datos relacionados a comparar:

$$\text{Q de Friedman} = \frac{12}{nk(k+1)}\left[\sum_{j=1}^{k} R_j^2\right]-3n(k+1)$$

que sigue una distribución de chi-cuadrado con k – 1 grados de libertad (k es el número de grupos de datos relacionados).

- Prueba de Cochran empleando una variable cualitativa dicotómica, más de dos grupos de datos relacionados a comparar:

$$\text{Q de Cochran} = \frac{J(J-1)\Sigma T_{+j}^2-(J-1)T^2}{JT-\Sigma T_{i+}^2} = (m-1)\cdot\frac{m\sum_{j=1}^{m} T_j^2-T^2}{mT-\sum_{n=1}^{N} S_n^2}$$

que sigue una distribución de chi-cuadrado con k – 1 o m – 1 grados de libertad (k o m es el número de grupos de datos).

- Prueba de Levene con una variable cuantitativa, para dos o más grupos de datos independientes a comparar sus varianzas:

$F = \dfrac{\text{CM intergrupos}}{\text{CM intragrupos}}$, que sigue una distribución F de Fisher-Snedecor con k – 1 grados de libertad (intergrupos) y $\sum_{j=1}^{k}(n_j-1)^2$ grados de libertad (intragrupos)

donde:

$$\text{CM intergrupos} = \frac{\sum_{i=1}^{k} n_j (\bar{D}_j - \bar{D}_t)^2}{k-1}$$

$$\text{CM intragrupos} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} (\bar{D}_{ij} - \bar{D}_j)^2}{\sum_{j=1}^{k} (n_j - 1)^2}$$ (k es el número de grupos de datos).

- Prueba de Barlett con una variable cuantitativa, más de dos grupos de datos independientes a comparar sus varianzas:

$$\chi^2 \text{de Barlett} = \frac{\left[\ln \frac{\Sigma\sigma^2(n-1)}{(n-1)} \Sigma(n-1)\right] - \ln\sigma^2(n-1)}{1 + \frac{K+1}{3(K-1)(N-K)}}$$

que sigue una distribución chi-cuadrado con K – 1 grados de libertad (K es el número de grupos de datos).

- Prueba de Cochran con una variable cuantitativa, más de dos grupos de datos independientes de iguales tamaños, a comparar sus varianzas:

$$\text{R de Cochran} = \frac{\text{máxima}\left(S_i^2\right)_{i=1}^{f}}{\sum_{i=1}^{t} S_i^2}$$

donde:

$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$ que sigue una distribución de Cochran n – 1, t o k, 1 – α.

Determinación del tamaño de muestra

Para diseños de estudio que emplean estimadores como:

- Una proporción $\hat{p}$ y $\hat{q}$ desconocidas y estimadas para un valor de 0,5 para cada una que aporta su multiplicación 0,25, lo cual está en la fórmula:

$n = \frac{(Z_{\alpha/2})^2 \cdot 0{,}25}{E^2}$ donde E es el margen de error para la proporción muestral que el investigador asume. Se emplea para población infinita o muy grande o cuando el muestreo es con reeemplazo.

- Una proporción con $\hat{p}$ y $\hat{q}$ conocidas y estimadas a partir de un estudio anterior, estudio piloto o criterio de experto:

$n = \frac{(Z_{\alpha/2})^2 \hat{p}\hat{q}}{E^2}$ donde E es el margen de error para la proporción muestral que el investigador asume. Se emplea para población infinita o muy grande o cuando el muestreo es con reemplazo.

– Una proporción con $\hat{p}$ y $\hat{q}$ conocidas y estimadas a partir de un estudio anterior, estudio piloto o criterio de experto:

$n = \frac{N\hat{p}\hat{q}(Z_{\alpha/2})^2}{\hat{p}\hat{q}(Z_{\alpha/2})^2 + (N-1)E^2}$ donde E es el margen de error para la proporción muestral que el investigador asume. Se emplea para población finita o pequeña o cuando el muestreo es sin reemplazo.

Cálculo del estimado de P (proporción poblacional) y de E (margen de error) desde un intervalo de confianza:

- Estimado puntual de P, a partir de la proporción muestral ($\hat{p}$):

$$\hat{p} = \frac{(\text{Límite de confianza superior}) + (\text{Límite de confianza inferior})}{2}$$

- Estimado de E (margen de error para la proporción poblacional estimada):

$$E = \frac{(\text{Límite de confianza superior}) + (\text{Límite de confianza inferior})}{2}$$

– Una media con σ conocida:

$n = \left(\frac{Z_{\alpha} / 2\sigma}{E}\right)^2$ donde E es el margen de error para la media poblacional estimada de la muestral que el investigador asume. Se emplea para población infinita o muy grande o cuando el muestreo es con reemplazo.

– Una media con σ conocida:

$n = \frac{N\sigma^2(Z_{\alpha/2})^2}{(N-1)E^2 + \sigma^2(Z_{\alpha/2})^2}$ donde E es el margen de error para la media poblacional estimada de la muestral que el investigador asume. Se emplea para población finita o pequeña o cuando el muestreo es sin reemplazo.

– Para dos medias:

$n = \frac{2s^2(Z_{\alpha/2} + Z_{\beta})^2}{d^2}$ para la diferencia de dos medias poblacionales estimadas de dos muestrales, el valor de n corresponde a una de las muestras por lo que se debe multiplicar por 2 para calcular el n de ambas muestras de igual tamaño.

– Para dos proporciones:

$n = \frac{2p_M q_M(Z_{\alpha/2} + Z_{\beta})^2}{(p_A - p_B)^2}$ donde $p_M = \frac{p_A + p_B}{2}$ para la diferencia de dos proporciones poblacionales estimadas de dos muestrales, el valor de n corresponde a una de las muestras por lo que se debe multiplicar por 2 para calcular el n de ambas muestras de igual tamaño.

Los procedimientos para calcular el tamaño muestral necesario para poder estimar σ^2 son mucho más complejos que los procedimientos resumidos anteriormente. A continuación se muestra una tabla de referencia a emplear para este objetivo:

Tamaño muestral para σ^2	
Para tener una confianza del 95 % de que s^2 está dentro del	**Del valor de σ^2m el tamaño muestral n debe ser al menos**
1 %	77,207
5 %	3,148
10 %	805
20 %	210
30 %	97
40 %	56
50 %	37

Nota: Para tener un nivel de confianza del 95 % de que s^2 está dentro del 10 % del valor de σ^2, es necesario la selección al azar de una muestra de 805 individuos para permitirnos realizar la inferencia relacionada con la varianza poblacional estimada a partir de la muestral empleada.

Propiedades relacionadas con las probabilidades

- Si A y B son mutuamente excluyentes:
  P(A o B) = P(A) + P(B)
- Si A y B no son mutuamente excluyentes:
  P(A o B) = P(A) + P(B) – P(A y B)
- Si A y B son independientes:
  P(A y B) = P(A) · P(B)
- Si A y B son dependientes:
  P(A y B) = P(A) · P(B|A)
- Regla de los sucesos complementarios:

$$P(\bar{A}) = 1 - P(A)$$

- Permutaciones sin elementos iguales:

$$nP_r = \frac{n!}{(n-r)!}$$

- Permutaciones (n_1 iguales,…):

$$\frac{n!}{n_1!n_2!...n_k!}$$

- Combinaciones:

$$nC_r = \frac{n!}{(n-r)!r!}$$

Distribuciones de probabilidad

- Distribución de probabilidad para la media:
  $\mu = \Sigma x \cdot P(x)$
- Distribución de probabilidad para la desviación estándar:

$$\sigma = \sqrt{[\Sigma x^2 \cdot P(x)] - \mu^2}$$

- Probabilidad binomial:

$$P(x) = \frac{n!}{(n-x)!x!} \cdot p^x \cdot q^{n-x}$$

- Media binomial:
  $\mu = n \cdot p$
- Varianza binomial:
  $\sigma^2 = npq$
- Desviación estándar binomial:

$$\sigma = \sqrt{npq}$$

- Probabilidad para la distribución de Poisson:

$$P(x) = \frac{\mu^x \cdot e^{-\mu}}{x!}$$ donde e = 2,71828.

- Para puntuación estándar o tipificación del valor x de una variable:

$$z = \frac{x - \overline{x}}{s} = \frac{x - \mu}{\sigma}$$

- Para aplicar el teorema del límite central:

$$\mu_{\overline{x}} = \mu \text{ y } \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

Anexo 3

Glosario de términos

Alfa (α): símbolo empleado para representar la probabilidad de un error tipo I, también denominado *nivel de significación*.
Análisis de varianza: método para analizar las varianzas poblacionales, que permite hacer pruebas de hipótesis acerca de medias de poblaciones.
Análisis de varianza de un factor: análisis de varianza que implica datos clasificados en grupos de acuerdo con un solo criterio.
Anchura de clase: la diferencia entre dos límites de clase inferiores consecutivos en una distribución de frecuencias.
Aproximación de la probabilidad por frecuencia relativa: valor de probabilidad estimado con base en observaciones reales.
Beta (β): símbolo empleado para representar la probabilidad de un error tipo II.
Bimodal: se dice que un conjunto de datos es bimodal cuando tiene dos modas.
Bloque: grupo de individuos similares con respecto a las formas en que pueden afectar el resultado de un experimento.
Cambio marginal: para variables relacionadas por una ecuación de regresión, la magnitud del cambio en la variable dependiente, cuando una de las variables independientes cambia en una unidad y las demás variables independientes se mantienen constantes.
Celda: categoría empleada para separar datos cualitativos (o de atributo).
Censo: recolección de datos de cada elemento de una población.
Centroide: el punto determinado a partir de un conjunto de datos bivariados.
CM (del error): cuadrado medio del error. Se usa en el análisis de varianza.
CM (del tratamiento): cuadrado medio de tratamientos. Se usa en el análisis de varianza.
CM (total): cuadrado medio de la variación total. Se usa en el análisis de varianza.
Coeficiente de confianza: probabilidad de que un parámetro de población esté contenido dentro de un intervalo de confianza particular, también se denomina *nivel de confianza* o *grado de confianza*.
Coeficiente de correlación: medida de la fuerza de la relación entre dos variables x, y.

Coeficiente de correlación de rangos ordenados (de Spearman): medida de la fuerza de la relación entre dos variables. Se basa en los rangos ordenados de los valores.
Coeficiente de correlación lineal o producto-momento de Pearson: medida de la fuerza de la relación entre dos variables.
Coeficiente de determinación: cantidad de la variación de y que se explica con la línea de regresión.
Coeficiente de determinación ajustado: coeficiente de determinación múltiple R^2, modificado para justificar el número de variables y del tamaño de la muestra.
Coeficiente de determinación múltiple: medida de qué tan bien una ecuación de regresión múltiple se ajusta a los datos muestrales.
Coeficiente de variación (CV): razón entre la desviación estándar y la media, que se expresa como un porcentaje.
Complemento de un suceso: todos los resultados en los que el suceso original no ocurre.
Confusión: situación que se presenta cuando no es posible distinguir entre los efectos de dos o más variables.
Corrección por continuidad: ajuste que se hace cuando una variable aleatoria discreta se aproxima con una variable aleatoria continua.
Correlación: asociación estadística entre dos variables.
Cuartil medio: la mitad de la suma de los cuartiles primero y tercero.
Cuartiles: los tres valores que dividen datos ordenados en cuatro grupos, con aproximadamente el 25 % de los valores en cada grupo.
Curva de densidad: gráfica de una distribución de probabilidad continua.
Datos: información o números que describen alguna característica.
Datos apareados: se dice que en dos muestras hay datos apareados cuando cada valor de una de ellas está apareado con un valor correspondiente de la otra muestra.
Datos bivariados: datos ordenados en pares.
Datos categóricos: datos que pueden dividirse en diferentes categorías y que se distinguen por alguna característica no numérica.
Datos continuos: datos que se obtienen de un número infinito de valores posibles, que corresponden a puntos de una escala continua que abarca un rango de valores sin huecos, saltos ni interrupciones.
Datos cualitativos: datos que pueden dividirse en diferentes categorías que se distinguen por alguna característica no numérica.
Datos cuantitativos: datos que consisten en números que representan conteos o mediciones.
Datos de atributo: datos que pueden dividirse en diferentes categorías que se distinguen por alguna característica no numérica.
Datos de proceso: datos ordenados según alguna secuencia de tiempo, que miden una característica de bienes o servicios que resultan de alguna combinación de equipos, personas, materiales, métodos y condiciones.
Datos de series de tiempo: datos reunidos en diferentes puntos del tiempo.

Datos discretos: datos con la propiedad de que el número de valores posibles es un valor finito o que puede contarse, que resulta en 0 posibilidades, 1 posibilidad o 2 posibilidades, etcétera.

Datos numéricos: datos que consisten en números que representan conteos o mediciones.

Datos ordenados: datos acomodados en orden.

Desviación: magnitud de la diferencia entre un puntaje y la media. Se expresa como:

- Desviación absoluta: la medida de variación que es igual a la suma de las desviaciones de cada valor respecto a la media, dividida entre el número de valores.
- Desviación estándar: medida de variación igual a la raíz cuadrada de la varianza.
- Desviación explicada: para un par de valores de un conjunto de datos bivariados, la diferencia entre el valor de y predicho y la media de todos los valores de y.
- Desviación media absoluta: medida de variación que es igual a la suma de las desviaciones de cada valor respecto a la media, dividida entre el número de valores.
- Desviación no explicada: para un par de valores en un conjunto de datos bivariados, la diferencia entre la coordenada y y el valor predicho.
- Desviación total: suma de la desviación explicada y la desviación no explicada para un par dado de valores en un conjunto de datos bivariados.

Diagrama de árbol: representación gráfica de los diferentes resultados posibles en un suceso compuesto.

Diagrama de dispersión: representación gráfica de datos (x, y) apareados.

Diseño de bloques aleatorizado: diseño en el que se obtiene una medición para cada tratamiento aplicado a cada uno de varios individuos apareados según características similares.

Diseño rigurosamente controlado: diseño experimental en el que se obliga a que todos los factores sean constantes para eliminar los efectos de factores ajenos.

Diseño totalmente aleatorizado: procedimiento de un experimento en el que cada elemento tiene la misma probabilidad de pertenecer a las diferentes categorías o tratamientos.

Distribución de frecuencias: lista de valores de datos (individualmente o por grupos de intervalos) junto con sus frecuencias (o conteos) correspondientes.

Distribución de frecuencias acumulativas: distribución de frecuencias en la que cada clase y frecuencia representa los datos acumulativos hasta esa clase inclusive.

Distribución de frecuencias relativas: variación de la distribución básica de frecuencias en la que la frecuencia de cada clase se divide entre el total de todas las frecuencias.

Distribución de Poisson: distribución de probabilidad discreta que se aplica a ocurrencias de algún suceso durante un intervalo de tiempo, distancia, área, volumen u otra unidad similar que se especifique.

Distribución de probabilidad: conjunto de valores de una variable aleatoria junto con sus probabilidades correspondientes.

Distribución muestral de medias muestrales: distribución de las medias muestrales que se obtiene al seleccionar repetidamente muestras de igual tamaño de la misma población.
Distribución muestral de proporciones: distribución de probabilidad de las proporciones muestrales, donde todas las muestras tienen el mismo tamaño muestral n.
Distribución normal: Distribución de probabilidad con forma de campana.
Distribución normal bivariada: Distribución de datos apareados en la que, para cualquier valor fijo de una variable, los valores de la otra variable están distribuidos normalmente.
Distribución normal estándar: distribución normal con una media igual a 0 y una desviación estándar igual a 1.
Distribución t o de t-student: distribución normal que suele estar asociada con datos muestrales de una población con una desviación estándar desconocida.
Distribución uniforme: distribución de probabilidad en la que todos los valores de la variable aleatoria son igualmente probables.
Ecuación de regresión: ecuación algebraica que describe la relación entre variables.
Ecuación de regresión múltiple: ecuación que expresa una relación lineal entre una variable dependiente y dos o más variables independientes (x_1, x_2,..., x_k).
Efecto placebo: efecto que ocurre cuando un sujeto que no recibe tratamiento cree incorrectamente que sí lo está recibiendo y reporta una mejoría en sus síntomas.
Eficiencia: medida de la sensibilidad de una prueba no paramétrica en comparación con una prueba paramétrica correspondiente.
Enfoque clásico a la probabilidad: enfoque en el que la probabilidad de un evento se determina dividiendo el número de maneras en que este puede suceder entre el número total de resultados posibles.
Error de muestreo: diferencia entre el resultado de una muestra y el resultado real de la población. Se debe a fluctuaciones aleatorias en las muestras.
Error estándar de estimación: medida de la dispersión de puntos muestrales alrededor de la línea de regresión.
Error estándar de la media: desviación estándar de todas las posibles medias muestrales.
Error tipo I: error que se comete al rechazar la hipótesis nula cuando esta es verdadera.
Error tipo II: error que se comete al no rechazar la hipótesis nula cuando esta es falsa.
Errores que no son de muestreo: errores debidos a factores externos no relacionados con el muestreo.
Espacio muestral: conjunto de todos los posibles resultados o sucesos de un experimento que no se pueden descomponer más.
Estadística descriptiva: métodos empleados para resumir las características clave de datos conocidos.
Estadística inferencial: métodos que implican el uso de datos muestrales para hacer generalizaciones o inferencias acerca de una población.

Estadístico o estadígrafo: característica medida de una muestra.
Estadístico de prueba: estadístico muestral que se basa en los datos muestrales; sirve para tomar la decisión respecto a rechazar o no la hipótesis nula.
Estimado: valor específico o intervalo de valores que se usa para aproximar algún parámetro de población.
Estimado conjunto de p_1 y p_2: probabilidad que se obtiene combinando los datos de dos proporciones muestrales y dividiendo el número total de éxitos entre el número total de observaciones.
Estimado conjunto de σ^2: estimado de la varianza que es común a dos poblaciones. Se obtiene calculando un promedio ponderado de las dos varianzas muestrales.
Estimado de intervalo: rango de valores usado para estimar algún parámetro de población con un nivel de confianza específico; también se conoce como *intervalo de confianza*.
Estimado puntual: valor individual que sirve como estimado de un parámetro de población.
Estimador: estadístico de muestra (como la media de muestra), que sirve para aproximar un parámetro de población.
Estimador sin sesgo: estadístico muestral que tiende a acercarse al parámetro de población para cuya estimación se usa.
Estudio ciego: procedimiento utilizado en experimentos, en el que el sujeto no sabe si está recibiendo un tratamiento o un placebo.
Estudio de casos y controles: estudio en el que se reúnen datos del pasado (a través del examen de registros, entrevistas y otros recursos).
Estudio de cohorte: estudio de sujetos en grupos identificados que comparten factores comunes (denominados cohortes), en el que los datos se reunirán en el futuro.
Estudio doble ciego: procedimiento utilizado en un experimento, en el que el sujeto no sabe si está recibiendo un tratamiento o un placebo y la persona que aplica el tratamiento tampoco lo sabe.
Estudio longitudinal: estudio de sujetos en grupos identificados que comparten factores comunes (llamados cohortes) y en el que los datos se reunirán en el futuro.
Estudio observacional: estudio en el que se observan y miden características específicas, pero no se intenta manipular o modificar a los sujetos en estudio.
Estudio prospectivo: estudio de sujetos en grupos identificados que comparten factores comunes (denominados cohortes), en el que los datos se reunirán en el futuro.
Estudio retrospectivo: estudio en el que se reúnen datos del pasado (a través del examen de registros, entrevistas y otros recursos).
Estudio transversal: estudio en el que los datos se observan, miden y reúnen en un punto del tiempo.
Experimento: la aplicación de un tratamiento y la posterior observación de sus efectos sobre los sujetos.

Experimento binomial: experimento que tiene un número fijo de ensayos independientes y en el que cada resultado pertenece exactamente a una de dos categorías.
Experimento multinomial: experimento que tiene un número fijo de ensayos independientes y en el que cada resultado pertenece exactamente a una de varias categorías.
Factor: en análisis de varianza, propiedad o característica que permite distinguir unas poblaciones de otras.
Factor de corrección por población finita: factor para corregir el error estándar de la media cuando el tamaño de una muestra excede el 5 % del tamaño de una población finita.
Fórmula de probabilidad binomial: expresión utilizada para calcular probabilidades en un experimento binomial.
Frecuencia acumulativa: suma de las frecuencias de una clase y de todas las clases precedentes.
Frecuencia esperada: frecuencia teórica para una celda de una tabla de contingencias o tabla multinomial.
Frecuencia observada: conteo de frecuencia real registrado en una celda de una tabla de contingencias o tabla multinomial.
Frecuencia relativa: frecuencia de una clase, dividida entre el total de todas las frecuencias.
Fronteras de clase: valores que se obtienen de una distribución de frecuencias aumentando los límites de clase superior y reduciendo los límites de clase inferior en la misma cantidad, de forma que no haya huecos entre clases consecutivas.
Grado de confianza: probabilidad de que un parámetro de población esté contenido dentro de un intervalo de confianza particular. También se denomina *nivel de confianza*.
Grados de libertad: número de valores que pueden variar después de haberse impuesto ciertas restricciones a todos los valores.
Grados de libertad del denominador: grados de libertad que corresponden al denominador del estadístico de prueba F.
Grados de libertad del numerador: grados de libertad que corresponden al numerador del estadístico de prueba F.
Gráfico de puntos: gráfica en la que cada valor de los datos se grafica como un punto sobre una escala de valores.
Gráfico de rachas: gráfica secuencial de valores de datos individuales a lo largo del tiempo, donde se usa un eje (casi siempre el vertical) para los valores de datos y el otro eje (casi siempre el horizontal) para la secuencia de tiempo.
Gráfico de rango: gráfica de control basada en rangos muestrales. Sirve para vigilar la variación en un proceso.
Gráfico de tallo y hojas: método para clasificar y acomodar datos para revelar su distribución.

Gráfico normal cuantilar: gráfica de puntos (x, y), donde cada valor de x pertenece al conjunto original de datos muestrales y cada valor y es una puntuación z correspondiente a un valor cuantilar de la distribución normal estándar.
Grupo de control: grupo de sujetos en un experimento a quienes no se les administra un tratamiento específico.
Grupo de tratamiento: grupo de sujetos que reciben algún tratamiento en un experimento.
Hipótesis: declaración o afirmación acerca de alguna propiedad de una población.
Hipótesis alternativa: afirmación que equivale a la negación de la hipótesis nula. Se denota con H_1.
Hipótesis nula: aseveración acerca de alguna característica de población, que por lo regular implica la ausencia de una diferencia; se denota con H_0.
Histograma: gráfica de barras verticales que representa la distribución de frecuencia de un conjunto de datos.
Histograma de frecuencias relativas: variación del histograma básico en la que las frecuencias se sustituyen por frecuencias relativas.
Histograma de probabilidad: histograma en el que los resultados se listan a lo largo del eje horizontal y las probabilidades se listan a lo largo del eje vertical.
Interacción: en el análisis de varianza de dos factores, el efecto que se observa cuando uno de ellos varía para diferentes categorías del otro factor.
Intercepto y: punto en el que una línea recta cruza el eje y.
Intervalo: nivel de medición de datos; caracteriza datos que pueden acomodarse en orden y para los que las diferencias entre los valores de los datos son significativas.
Intervalo de confianza: rango de valores empleado para estimar algún parámetro de población con un nivel de confianza específico. También se denomina *estimado de intervalo*.
Intervalo de predicción: estimado del intervalo de confianza de un valor predicho de y.
Límites de clase inferiores: los números más pequeños que pueden pertenecer a las diferentes clases de una distribución de frecuencias.
Límites de clase superiores: los números más grandes que pueden pertenecer a las diferentes clases de una distribución de frecuencias.
Límites del intervalo de confianza: dos números que se usan como fronteras superior e inferior de un intervalo de confianza.
Línea central: línea de una gráfica de control que representaun valor central de las mediciones características.
Línea de regresión: línea recta que mejor se ajusta a un conjunto de puntos que representan datos muestrales apareados.
Margen de error: máxima diferencia probable (con probabilidad $1 - \alpha$) entre el estadístico muestral observado y el verdadero valor del parámetro poblacional.

MANCOVA: análisis múltiple de covarianzas; es decir, análisis múltiple de varianzas con presencia de covariables.
MANOVA: análisis múltiple de varianzas.
Media: la suma de un conjunto de valores, dividida entre el número de valores.
Media aritmética: suma de un conjunto de valores dividida entre el número de valores. Normalmente se denomina *media*.
Media ponderada: media de un conjunto de valores a los que se han asignado diferentes grados de importancia.
Mediana: valor que está a la mitad de un conjunto de valores acomodados en orden por magnitud.
Medida de tendencia central: valor que pretende indicar el centro de los valores de una colección de datos.
Medida de variación: cualquiera de varias medidas diseñadas para reflejar la magnitud de la variación o dispersión de un conjunto de valores.
Metanálisis: es una metodología para la revisión sistemática y cuantitativa de la investigación, ampliamente consolidada y aplicada en las ciencias de la salud.
Método clásico de prueba de hipótesis: método para probar hipótesis que se basa en una comparación del estadístico de prueba con los valores críticos.
Método tradicional de prueba de hipótesis: método para probar hipótesis que se basa en una comparación del estadístico de prueba y los valores críticos.
Mitad del rango: mitad de la suma de los valores máximo y mínimo.
Moda: valor que se presenta con mayor frecuencia.
Modelo matemático: función matemática que se "ajusta" o que describe datos de la vida real.
Muestra: subconjunto de una población.
Muestra aleatoria: muestra seleccionada de tal manera que permite a cada miembro de la población tener la misma probabilidad de ser elegido.
Muestra aleatoria simple: muestra de cierto tamaño seleccionada de manera que toda posible muestra del mismo tamaño tenga la misma probabilidad de ser elegida.
Muestra autoseleccionada: muestra en la que los sujetos deciden por sí mismos ser incluidos. También se llama *muestra de respuesta voluntaria*.
Muestra de respuesta voluntaria: muestra en la que los sujetos deciden por sí mismos ser incluidos.
Muestra dependiente: muestra cuyos valores están relacionados con los valores de otra muestra.
Muestra independiente: muestra cuyos valores no están relacionados con los valores de otra muestra.
Muestras apareadas: dos muestras que son dependientes en el sentido de que los valores de los datos forman pares.

Muestreo de aceptación: elementos muestrales sin reemplazo y que permiten rechazar todo el lote con base en el número de defectos obtenidos.
Muestreo de conveniencia: muestreo en el que se seleccionan datos porque se encuentran disponibles.
Muestreo estratificado: muestreo en el que se obtienen muestras de cada estrato (clase).
Muestreo por conglomerados: tipo de muestreo en el que se divide el área de población en secciones (o conglomerados) y luego se seleccionan en forma aleatoria algunas de esas secciones, después se eligen todos los miembros de las secciones elegidas.
Muestreo sistemático: muestreo en el que se selecciona cada k-ésimo elemento.
Multimodal: se dice que un conjunto de datos es multimodal cuando tiene más de dos modas.
Nivel de confianza: probabilidad de que un parámetro de población esté contenido en un intervalo de confianza particular. También se llama *grado de confianza*.
Nivel de significación: probabilidad de cometer un error tipo I al realizar una prueba de hipótesis.
Nominal: nivel de medición de datos. Caracteriza datos que consisten únicamente en nombres, etiquetas o categorías.
Ordinal: nivel de medición de datos. Caracteriza datos que podrían estar acomodados en orden, pero las diferencias entre los valores de los datos no pueden determinarse, o bien, carecen de sentido.
Parámetro: característica medida de una población.
Pares discordantes: pares de categorías en los que las dos categorías son diferentes. Se usan en la prueba de McNemar.
Pendiente: medida de la inclinación de una línea recta.
Percentil: los 99 valores que dividen datos clasificados en 100 grupos, con aproximadamente el 1 % de los valores en cada grupo.
Población: colección entera y completa de elementos por estudiar.
Polígono de frecuencias: representación gráfica de la distribución de los datos, que utiliza segmentos de línea recta conectados.
Posibilidades a favor: razón entre la probabilidad de que un suceso ocurra y la probabilidad de que no ocurra. Suele expresarse como la proporción de dos enteros sin factores comunes.
Posibilidades en contra: razón entre la probabilidad de que un suceso no ocurra y la probabilidad de que ocurra. Suele expresarse en la forma a:b, donde a y b son enteros sin factores comunes.
Posibilidades reales a favor: el recíproco de las posibilidades reales en contra de un suceso.
Posibilidades reales en contra: el cociente generalmente expresado en la forma a:b (o "a es a b").
Potencia de una prueba: la probabilidad (1 – β) de rechazar una hipótesis nula falsa.

Probabilidad: medida de la posibilidad de que ocurra un suceso dado. Se expresa como un número entre 0 y 1.
Probabilidad condicional: la probabilidad de un suceso, dado que algún otro suceso ya ocurrió.
Probabilidad subjetiva: conjetura o estimado de una probabilidad con base en un conocimiento de las circunstancias relevantes.
Procedimientos de comparación múltiple: procedimientos para identificar cuáles medias específicas son diferentes, después de concluir que tres o más medias no son iguales.
Proceso estadísticamente estable: proceso que solo tiene variación natural, sin patrones, ciclos ni puntos inusitados.
Promedio: cualquiera de varias medidas diseñadas para revelar la tendencia central de un conjunto de datos.
Propiedad de mínimos cuadrados: propiedad que afirma que, para una línea de regresión, la suma de los cuadrados de las desviaciones verticales de los puntos muestrales, respecto a la línea de regresión, es la más pequeña posible.
Prueba de bondad de ajuste: prueba para determinar qué tan bien alguna distribución de frecuencias observada se ajusta a una distribución teórica.
Prueba de hipótesis unilateral de cola derecha: prueba de hipótesis en la que la región crítica se ubica en el área extrema derecha de la distribución de probabilidad.
Prueba de hipótesis unilateral de cola izquierda: prueba de hipótesis en la que la región crítica está situada en el área extrema izquierda de la distribución de probabilidad.
Prueba de hipótesis bilateral de dos colas o dos lados: prueba de hipótesis en la que la región crítica se divide entre las áreas extremas izquierda y derecha de la distribución de probabilidad.
Prueba de hipótesis: método para probar afirmaciones acerca de poblaciones; también se llama prueba de significación.
Prueba de homogeneidad: prueba de la aseveración de que diferentes poblaciones tienen la misma proporción de alguna característica.
Prueba de independencia: prueba de la hipótesis nula que afirma que, en una tabla de contingencias, la variable de renglón y la variable de columna no están relacionadas.
Prueba de Kruskal-Wallis: prueba de hipótesis no paramétrica que sirve para comparar tres o más muestras independientes. También se llama *prueba H*.
Prueba de McNemar: usa conteos de frecuencia de datos nominales pareados de dos categorías para poner a prueba la hipótesis nula de que las frecuencias de pares discordantes ocurren en la misma proporción.
Prueba de rachas: método no paramétrico que sirve para detectar aleatoriedad en la extracción de la muestra.
Prueba de rangos con signo (o de rangos signados) de Wilcoxon: prueba de hipótesis no paramétrica que se utiliza para comparar dos muestras dependientes.
Prueba de significación: prueba de hipótesis.

Prueba de suma de rangos de Wilcoxon: prueba de hipótesis no paramétrica que se utiliza para comparar dos muestras independientes.
Prueba del signo: prueba de hipótesis no paramétrica que sirve para comparar muestras de dos poblaciones.
Prueba H: prueba no paramétrica de Kruskal-Wallis.
Prueba U de Mann-Whitney: prueba de hipótesis que equivale a la prueba de suma de rangos de Wilcoxon para dos muestras independientes.
Pruebas no paramétricas: procedimientos estadísticos para hacer pruebas de hipótesis o estimar parámetros, en los que no es preciso hacer suposiciones acerca de la naturaleza o forma de las distribuciones de las poblaciones. También se denominan *pruebas de distribución libre*.
Pruebas paramétricas: procedimientos estadísticos basados en parámetros de población para probar hipótesis o estimar parámetros.
Punto influyente: punto que afecta fuertemente la gráfica de una línea de regresión.
Punto medio de clase: en una clase de una distribución de frecuencias, el valor que está a la mitad entre el límite de clase superior y el límite de clase inferior.
Puntuación estándar: número de desviaciones estándar que un valor dado está por arriba o por debajo de la media. También se llama *puntuación Z*.
Puntuación Z: número de desviaciones estándar que un valor dado está por arriba o por debajo de la media.
Racha o tendencia: secuencia de datos que presentan la misma característica. Se usa en la prueba de rachas para detectar aleatoriedad.
Rango: medida de variación que es la diferencia entre los valores máximo y mínimo.
Rango de percentiles 10-90: diferencia entre los percentiles décimo y nonagésimo.
Rango intercuartilar: la diferencia entre los cuartiles primero y tercero.
Rango ordenado: posición numérica de un elemento en un conjunto muestral acomodado en orden.
Rango semi-intercuartilar: la mitad de la diferencia entre los cuartiles primero y tercero.
Razón: nivel de medición de los datos. Caracteriza datos que pueden ser acomodados en orden, para los que las diferencias entre los valores son significativas y existe un punto de partida cero inherente.
Región crítica: el conjunto de todos los valores del estadístico de prueba que harían que se rechazara la hipótesis nula.
Regla de combinaciones: regla para determinar el número de combinaciones diferentes de elementos seleccionados.
Regla de conteo fundamental: regla que dice que, para una secuencia de dos sucesos en la que el primer suceso puede ocurrir de m maneras y el segundo de n maneras, los sucesos juntos pueden ocurrir en un total de $m \cdot n$ maneras.
Regla de la multiplicación: regla para determinar la probabilidad
de que ocurra el suceso A en un ensayo y de que ocurra el suceso B en un segundo ensayo.

Regla de la suma: regla para determinar la probabilidad de que, en un solo ensayo, ocurra el suceso A o el suceso B, o bien, de que ocurran ambos.
Regla del suceso infrecuente: si, bajo un supuesto dado, la probabilidad de un resultado específico observado es muy pequeña, se concluye que posiblemente el supuesto no sea correcto.
Regresión logística: método usado en la regresión múltiple cuando la variable ficticia es la variable de respuesta (y).
Regresión múltiple: estudio de relaciones lineales entre tres o más variables.
Regresión por pasos: proceso que implica el uso de diferentes combinaciones de variables hasta obtener el mejor modelo; se usa en regresión múltiple.
Residual: diferencia entre un valor muestral y observado, y el valor de Y que se predice con una ecuación de regresión.
Robustez: propiedad que se describe para un método o estadístico cuando no ocurre la alteración del mismo de una forma significativa, pese a no cumplirse los supuestos de partida en los que están basados.
SC (del error): suma de cuadrados que representa la variabilidad que se supone es común a todas las poblaciones consideradas. Se usa en el análisis de varianza.
SC (del tratamiento): medida de la variación entre las medias muestrales. Se usa en el análisis de varianza.
SC (total): medida de la variación total (alrededor de) en todos los datos muestrales combinados. Se usa en el análisis de varianza.
Selección aleatoria: selección de elementos muestrales de forma que todos los elementos disponibles para ser seleccionados tengan la misma probabilidad de ser elegidos.
Simetría: propiedad de datos cuya distribución puede dividirse en dos mitades que son aproximadamente imágenes especulares al trazar una línea vertical por la mitad.
Simulación: proceso que se comporta de forma similar a algún experimento, por lo que se obtienen resultados similares.
Suceso: resultado de un experimento.
Suceso compuesto: combinación de sucesos simples.
Suceso simple: resultado experimental que no puede descomponerse más.
Sucesos dependientes: sucesos para los cuales la ocurrencia de cualquier suceso individual afecta las probabilidades de ocurrencia de los demás sucesos.
Sucesos disjuntos o mutuamente excluyentes: sucesos que no pueden ocurrir de manera simultánea.
Sucesos independientes: sucesos para los cuales la ocurrencia de cualquiera de ellos no afecta las probabilidades de ocurrencia de los demás.
Sucesos mutuamente excluyentes: sucesos que no pueden ocurrir de manera simultánea.
Tabla de contingencias o cruzadas: tabla de frecuencias observadas en la que los renglones corresponden a una variable de clasificación y las columnas corresponden a otra variable de clasificación. También se denomina *tabla de dos factores*.

Tabla de frecuencias: lista de categorías de valores junto con sus frecuencias correspondientes.
Tamaño de muestra: número de elementos de una muestra.
Teorema central del límite: teorema que afirma que las medias muestrales tienden a distribuirse normalmente, con una media y una desviación estándar.
Unidades experimentales: sujetos de un experimento.
Valor crítico: valor que separa la región crítica de los valores del estadístico de prueba que no conducirían al rechazo de la hipótesis nula.
Valor esperado para una variable aleatoria discreta, la media de los resultados.
Valor P: probabilidad de que un estadístico de prueba (teórico-esperado) en una prueba de hipótesis sea al menos tan extremo como el que en realidad se obtuvo (observado-calculado).
Valores extremos: valores poco comunes, en el sentido de que están muy lejos de la mayoría de los datos.
Valores predichos: valores de una variable dependiente que se obtienen usando valores de variables independientes en una ecuación de regresión.
Variabilidad de muestreo: variación de un estadístico en distintas muestras.
Variable aleatoria: variable (casi siempre representada con x) que tiene un solo valor numérico (determinado por el azar) para cada resultado de un experimento.
Variable aleatoria continua: variable aleatoria con valores infinitos, que se puede asociar con los puntos en un intervalo lineal continuo.
Variable aleatoria discreta: variable aleatoria que tiene un número finito de valores o bien un número de valores que pueden contarse.
Variable de respuesta: variable y en una ecuación de regresión o en una ecuación de regresión múltiple.
Variable dependiente: variable y de una ecuación de regresión o de regresión múltiple.
Variable dicotómica: variable que tiene dos valores discretos posibles.
Variable ficticia o indicadora: variable dicotómica con valores posibles de 0 y 1. Se emplea en el análisis de regresión.
Variable independiente: la variable x de una ecuación de regresión o una de las variables x de una ecuación de regresión múltiple.
Variable interventora: variable que afecta las variables que se están estudiando, pero que no está incluida en el estudio.
Variables de predicción: variables independientes en una ecuación de regresión.
Variación aleatoria: tipo de variación en un proceso que se debe al azar. El tipo de variación inherente a cualquier proceso que no puede generar todos los bienes o servicios exactamente de la misma forma todo el tiempo.
Variación asignable: tipo de variación en un proceso, que es el resultado de causas identificables.
Variación dentro de muestras: en el análisis de varianza, la variación que se debe al azar.

Variación explicada: suma de los cuadrados de las desviaciones explicadas para todos los pares de datos bivariados de una muestra.
Variación no explicada: suma de los cuadrados de las desviaciones no explicadas para todos los pares de datos bivariados en una muestra.
Variación total: suma de los cuadrados de la desviación total para todos los pares de datos bivariados de una muestra.
Varianza: medida de variación que es igual al cuadrado de la desviación estándar.

Bibliografía

Abraira, V. (1996). *Métodos multivariantes en bioestadística*. Centro de Estudios Ramón Areces.

Ahnn, S. y Anderson, S. J. (1995). Sample size determination for comparing more than two survival distributions. *Stat Med*, 14, 227322-82.

Albert, J. (1996). *Bayesian computation using Minitab*. Belmont, California: Duxbury Press, Wadsworth Publishing Company.

Altman, D. G. (1991). *Practical statistics for medical research*. Boca Ratón: Chapman & Hall/CRC.

Álvarez Cáceres, R. (2007). *Estadística aplicada a las ciencias de la salud*. España: Díaz de Santos.

Álvarez Suárez, M. M., Caballero, A. y Pérez Lechuga, G. (2006). Análisis multivariante: clasificación, organización y validación de resultados. Procede del Acta del Fourth LACCEI International Latin Anerican and Caribbean Conference for Engineering and Technology: Breaking Frontiers and Barriers in Engineering : Educational, Research and Practice. Puerto Rico.

Armitage, P. (1998). McNemar test. En: Armitage, P. y Colton, T. (Eds). *Encyclopedia of Biostatistics*. Chichester: John Wiley & Sons, pp. 2486-2487.

Armitage, P. y Berry, G. (1992). *Estadística para la investigación biomédica*. Barcelona: Ediciones Doyma.

Armitage, P. y Berry, G. (1994). *Statistical Methods in Medical Research*. 3rd ed. Oxford: Blackwell Scientific Publications.

Armitage, P. y Berry, G. (1997). *Estadística para la investigación biomédica*. 3ra edición. España: Harcourt Brace.

Arpajón Peña, Y. y Sosa Pérez, A. L. (2014). Procedimientos para la estimación por intervalos de confianza en las investigaciones biomédicas. *Rev Cubana Angiol Cir Vasc*, 2(1), 43-49.

Bacallao Gallestey, J. (1999). *Manual de investigación educacional*. Maestría en Educación Médica. La Habana: Centro Nacional de Perfeccionamiento Médico.

Bacallao Gallestey, J. (2012). Enfoques modernos del sesgo y la causalidad en la investigación epidemiológica. *Revista Cubana de Salud Pública*, 38(5), 686-701.

Bayarre Vea, H. y Hosford Saing, R. (2007). Métodos y técnicas aplicadas a la investigación en Atención Primaria de Salud. La Habana.

Behar Gutiérrez, R. y Grima Cintas, P. (2004). *55 Respuestas a dudas típicas de estadística*. España: Díaz de Santos.

Benavides, A. C. y Silva, L. C. (2002). Insuficiencias del paradigma frecuentista y el enfoque bayesiano como alternativa. Trabajo para optar por el Grado Científico de Doctor en Ciencias de la Salud. Escuela Nacional de Salud, La Habana.

Berry, D. A. (1996). *Statistics: A bayesian perspective*. Belmont, California: Duxbury Press.
Bland, J. M. y Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 1, 307-310.
Botella Rocamora, P., Alacreu García, M. y Martínez Beneito, M. A. (2013-14). *Apuntes de Estadística en Ciencias de la Salud*. Memorias del Curso: Estadística en Ciencias de la Salud. España: Universidad Cardenal Herrera.
Brown, L. D., Cai, T. T., DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16(2), 101-133.
Canavos, G. C. (1998). *Probabilidad y estadística: aplicaciones y métodos*. Madrid: Editorial McGraw-Hill.
Cantor, A.B. (1996). Sample size calculations for Cohen's kappa. *Psychological Methods*, 1(2), 150-153.
Checkoway, H., Pearce, N. y Kriebel, D. (2004). *Research Methods in Occupational Epidemiology*. 2nd ed. Oxford: University Press.
Chernick, M. R. y Friis, R. H. (2003). *Introductory Biostatistics for the Health Sciences: Modern Applications Including Bootstrap*. United State of America: Wiley Interscience.
Clark, M. (2014). Los valores P y los intervalos de confianza: ¿en qué confiar? *Rev Panam Salud Pública*, 15(5), 300-306. Disponible en: http://www.dx.doi.org/10.1590/S1020-49892004000500001
Cleves, M. A. (1995). Robust test for the equality of variances. *Stata Technical Bulletin*, 13-5.
Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, 37, 256-266.
Colectivo de autores (2004). *Informática médica: Bioestadística*. Tomo 2. La Habana: Editorial Ciencias Médicas.
Colectivo de autores (2004). *Métodos de análisis de datos en ecología*. España, Alcalá: Universidad de Alcalá.
Colectivo de autores (2006). Programa computarizado para el análisis epidemiológico de datos: EPIDAT 3.1. Ayuda de Tablas de contingencia. Disponible en: http://www.sergas.es/Saude-publica/EPIDAT
Colectivo de autores (2014). Programa computarizado para el análisis epidemiológico de datos: EPIDAT 4.1. Ayuda de análisis descriptivo. Disponible en: http://www.sergas.es/Saude-publica/EPIDAT
Colectivo de autores (2014). Programa computarizado para el análisis epidemiológico de datos: EPIDAT 4.1. Ayuda de distribuciones de probabilidad. Disponible en: http://www.sergas.es/Saude-publica/EPIDAT
Colectivo de autores (2014). Programa computarizado para el análisis epidemiológico de datos: EPIDAT 4.1. Ayuda de inferencia sobre parámetros. Disponible en: http://www.sergas.es/Saude-publica/EPIDAT
Colectivo de autores (2014). Programa computarizado para el análisis epidemiológico de datos. EPIDAT 4.1. Ayuda de Regresión logística. Disponible en: http://www.sergas.es/Saude-publica/EPIDAT
Colectivo de autores (2016). Programa computarizado para el análisis epidemiológico de datos: EPIDAT 4.2. Ayudas de análisis bayesiano. Disponible en: http://www.sergas.es/Saude-publica/EPIDAT

Connor, R. J. (1987). Sample size for testing differences in proportions for the paired-sample design. *Biometrics*, 43, 207-211.

Conover, W. J. (1980). *Practical non parametric statistics*. 2nd ed. New York: John Wiley & Sons.

Cornfield, J., Gordon, T. and Smith, W. N. (1961). Quantal response curves for experimentally uncontrolled variables. *Bulletin of the International Statistical Institute*, 38, 97-115.

Coughlin, S. S., Benichou, J. y Weed, D. L. (1996). Estimación del riesgo atribuible en los estudios de casos y controles. *Bol Oficina Sanit Panam*, 121(2).

Cox, D. R. y Snell, E. J. (1989). *The analysis of binary data*. 2nd ed. London: Chapman and Hall.

D'Agostino, R. B. (1972). Small sample probability points for the D test of normality. *Biometrika*, 59, 219-221.

D'Agostino, R. B., Belanger, A., D'Agostino, R. B. Jr. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician*, 44(4), 316-21.

Daniel, W. W. (1991). *Biostatistics: a foundation for analysis in the health of sciences*. 5th ed. United State of America: John Wiley & Sons, Inc.

Daniel, W. W. (1997). *Bioestadística. Base para el análisis de las ciencias de la salud*. 3ra ed. México DF: Limusa.

De Irala, J., Martínez, M. A. y Guillén, F. (2001). ¿Qué es una variable de confusión? *Med Clin Barc*, 117, 377-385. Disponible en: http://simulacionstefa.blogspot.com/feeds/posts/default

Efron, B. y Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Egaña Morales, E. (2010). *Bioestadística cualitativa*. La Habana: Editorial Ciencias Médicas.

Escrig, J., Miralles, J., Martínez, D. y Rivadulla, I. (2007). Intervalos de confianza: ¿por qué usarlos? *Cir Esp*, 81(3), 121-125.

Esper, R. y Machado, R. (2008). *La investigación en medicina: bases teóricas y prácticas. Elementos de bioestadística*. Buenos Aires: La Prensa Médica Argentina.

Fay, M. (2010) Confidence intervals that match Fisher's exact or Blaker's exact tests. *Biostat*, 11(2), 373-374.

Feinsinger, P. (2012). Lo que es, lo que podría ser y el análisis e interpretación de los datos de un estudio de campo. *Ecol Bolivia*, 47(1), 1-6.

Fernández-Abascal, H., Guijarro, M. M., Rojo, J. L. y Sanz, J. A. (1994). *Cálculo de probabilidades y estadística*. Barcelona: Editorial Ariel.

Fleiss, J. L. (1981) *Statistical Methods for Rates and Proportions*. 2nd ed. New York: John Wiley & Sons.

Freedman, L. S. (1982). Tables of the number of patients required in clinical trials using the logrank test. *Stat Med*, 1, 121-129.

Freixa, M., Salafranca, L. L., Guaridia, J., Ferrer, R. y Turbany, J. (1992). *Análisis exploratorio de datos: nuevas técnicas estadísticas*. Barcelona: Colección LCT-36 (Letras, Ciencia y Técnica).

Garcés, D. y Barragán, F. J. (2014). Ronda clínica y epidemiológica. Introducción al análisis multivariante (parte 1). *IATREIA*, 327(3), 355-363.

García, J. A. (2005). *Apuntes de Bioestadística*. México: Laboratorio de Biología Teórica, Postgrado e Investigación de la Universidad de Salle.

Gardner, M. J. y Altman, D. G. (1989). *Statistics with confidence: confidence intervals and statistical guidelines*. London: British Medical Journal.

Greenland, S. (2006). Bayesian perspectives for epidemiological research: foundations and basic methods. *Int J Epidemiol*, 35(3), 765-775.

Hair, J. F. Jr, Anderson, R. E., Tathan, R. L. y Black, W. C. (1999*). Análisis multivariante*. 5ta ed. Madrid: Prentice Hall Iberia.

Haynam, G. E., Govindarajulu, Z. y Leone, F. C. (1970). Tables of the cumulative non-central chi-square distribution. En: *Selected tables in mathematical statistics*. New York: American Mathematical Society, pp. 1-78.

Hernández Sampieri, R., Fernández Collado, C. y Baptista Lucio, P. (2006). *Metodología de la Investigación*. 4ta ed. México: Mc Graw-Hill Interamericana.

Hernández, B. y Velasco, M. H. (2002). Encuestas transversales. *Salud Pub Mex*, 42(5), 447-455.

Herrerías Pleguezuelo, R. y Palacios González, F. (2007). *Curso de inferencia estadística y del modelo lineal simple*. Madrid: Delta Publicaciones Universitarias.

Hervada Vidal, X., Santiago Pérez, M. I., Vázquez Fernández, E., Castillo Salgado, C., Loyola Elizondo, E. y Silva Ayçaguer, L. C. (2004). Epidat 3.0: programa para análisis epidemiológico de datos tabulados. *Rev Esp Salud Pública*, 78, 277-280.

Higgins, J. P. and Green, S (editors) (2011). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. *The Cochrane Collaboration*. Disponible en: https://www.cochrane-handbook.org.

Hoel, P. G. (1971). *Introduction to mathematical statistics*. 4ta ed. New York: John Wiley & Sons, Inc.

Hosmer, D. W. Jr and Lemeshow, S. (1989). *Applied Logistic Regression*. New York: John Wiley & Sons.

Jones, R. H. (1975). Probability estimation using a multinomial logistic function. *Journal of Statistical and Computer Simulation*, 3, 315-329.

Kaul, S. and Diamond, G. (2010). Trial and error how to avoid commonly encountered limitations of published clinical trials. *J Am Coll Cardiol*, 55, 415-427.

Kirkwood, B. R. (1988). *Essentials of medical stastistics*. Oxford: Blackwell Science.

Laake, P., Breien Benestad, H. and Reino Olsen, B. (2007). *Research Methodology in the Medical and Biological Sciences*. United State of America: ELSEVIER.

Lantz, C. A. and Nebenzahl, E. (1996). Behavior and interpretation of the k statistic: Resolution of the two paradoxes. *J Clin Epidemiol*, 49(4), 431-434.

Last, J. M. (2001). *A dictionary of epidemiology*. 4ta ed. New York: Oxford University Press.

Lautenbach, E., Woeltje, K. F. and Malani, P. N. (2010). *Practical Healthcare Epidemiology*. 3rd ed. United State of America: The University of Chicago.

Le, C. T. (2003). *Introductory Biostatistics*. United State of America: Wiley Interscience.

Lemeshow, S. and Hosmer, D. W. Jr. (1982). A review of goodness of fit statistics for use in the development of logistic regression models. *Am J Epidemiol*, 115, 92-106.

Levy, P. S. and Stolte, K. (2000). Statistical methods in public health and epidemiology: a look at the recent past and projections for the next decade. *Stat Methods Med Res*, 9, 41-55.

Limón, J., Rodríguez, M., Báez, Y. y Tlapa, D. (2011). Evaluación de la robustez del sistema Mahalanobis-Taguchi a diferentes arreglos factoriales. Información *Tecnolog*, 22(4), 85-96.

Lwanga, S. K. y Lemeshow, S. (1991). *Determinación del tamaño de las muestras en los estudios sanitarios. Manual práctico*. Ginebra: OMS.

Machin, D., Campbell, M. J., Fayers, P. M. and Pinol, A. P. (1997). *Sample size tables for clinical studies*. 2da ed. United State of America: Blackwell Science Ltd.

Manchikanti, L., Benyamin, R., Helm, S. and Hirsch, J. A. (2009). Evidence-Based Medicine, Systematic Reviews and Guidelines in Interventational Pain Management: Part 3: Systematic Reviews and Meta-Analysis of Randomized Trials. *Pain Physician*, 12, 35-72.

Martín Matew, M., Horna Campos, O., Nedel Borges, F. y Navarro Giné, A. (2010). *Fundamentos de Estadística en Ciencias de la Salud*. Barcelona: Universitat Autonóma de Barcelona Servei de Publicacions Bellaterra.

Martín Pliego, J. y Ruiz Maya, L. (2004). *Estadística I: Probabilidad*. 2da ed. Madrid: Thomson.

Martínez González, M. A., Sánchez Villegas, A. y Faulín Fajardo, F. J. (Eds.) (2006). *Bioestadística amigable*. 2da ed. España: Díaz de Santos.

Matthews, R. A. (1999). Significance levels for the assessment of anomalous phenomena. *Journal of Scientific Exploration*, 13(1), 1-7.

Meyer, P. L. (1973). *Probabilidad y aplicaciones estadísticas*. 2da ed. Bogotá: Fondo Educativo Interamericano.

Mittlböck, M. and Schemper, M. (1987). Explained variation for logistic regression. *Stat Med*, 15, 1987-1997.

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D. and Stroup. D. F. (1999). Improving the quality ofreports of metanalyses of randomised controlled trials: the QUOROM statement. *Lancet*, 54, 1896-900.

Moher, D., Liberati, A., Tetzlaff, J. and Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analysis: the PRISMA statement. *BMJ*, 339, b2535.

Montgomery, D. C. (1991). *Design and analysis of experiments*. United State of America: John Wiley and Sons, Inc.

Mood, A. M. and Graybill, F. A. (1963). *Introduction to the theory of statistics*. New York: McGraw-Hill.

Nagelkerke, N. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78, 691-692.

Narula, S. C. and Desu, M. M. (1981). Algorithm AS 170: Computation of probability and non-centrality parameter of a non-central chi-squared distribution. *Journal of the Royal Statistical Society*, 30(3), 349-352.

Newcombe, R. y Merino, C. (2006). Intervalos de confianza para las estimaciones de proporciones y las diferencias entre ellas. *Interdisciplinaria*, 23(2), 141-154.

Newman, S. C. (2001). *Biostatistical Methods in Epidemiology*. United State of America: John Wiley & Sons, Inc.

Nunnally, J. C. and Bernstein, I. H. (1994). *Psychometric theory*. 3era ed. New York: McGraw-Hill.

Obuchowski, N. A. (1998). Sample size calculations in studies of test accuracy. *Statistical Methods in Medical Research*, 7, 371-392.

Palta, M. (2003). *Quantitative Methods in Population Health: extensions of ordinary regression*. United State of America: John Wiley & Sons. Inc.

Pita Fernández, S., Pértegas Díaz, S. (2003). Pruebas diagnósticas: Sensibilidad y especificidad. *Cad Aten Primaria*, 10, 120-124.

Polo, J. C. (2011). *Apuntes sobre bioestadística aplicada a las ciencias farmacéuticas*. La Habana: Editorial Félix Varela.

Rosner, B. (2000). *Fundamentals of Biostatistics*. 5ta. ed. Belmont: Duxbury Press.

Rosner, B. (2010). *Fundamentals of Biostatistics*. 7ma ed. Belmont: Duxbury Press.

Rothman, K. J. (1998). *Epidemiología moderna*. Madrid: Díaz de Santos.
Rothman, K. J. and Greenland, S. (1998). *Modern epidemiology*. 2nd ed. United State of America: Lippincott- Raven.
Royston, J. P. (1991). Comment on sg3.4 and an improved D'Agostino test. *Stata Technical Bulletin*, 3, 23-24.
Royston, J. P. (1992). sg7: Centile estimation command. *Stata Technical Bulletin*, 8, 12-15.
Royston, J. P. (1993). A pocket-calculator algorithm for the Shapiro-Francia test for non-normality: an application to medicine. *Statistics in Medicine*, 12, 181-184.
Sánchez, L., Pérez, D., Cruz, G., Silva, L. C., Boelaert, M. y Van der Stuyft, P. (2004). Participación comunitaria en el control de Aedes aegypti: opiniones de la población en un municipio de La Habana, Cuba. *Rev Panam Salud Pública*, 15(1), 19-25.
Sarría Castro, M., Silva Ayçaguer, L. C. (2004). Las pruebas de significación estadística en tres revistas biomédicas en lengua española: una revisión crítica. *Rev Panam Salud Pública*, 5(5), 300-306.
Schoenfeld, D. A. (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics*, 39, 499-503.
Schouten, H. J. (1999). Sample size formula with a continuous outcome for unequal group sizes and unequal variances. *Stat Med*, 18, 87-91.
Silva Ayçaguer, L. C y Suárez, P. (2000). ¿Qué es la inferencia bayesiana? *JANO, Medicina y Humanidades*, 58(1338), 65-66.
Silva Ayçaguer, L. C. (1995). *Excursión a la regresión logística en Ciencias de la Salud*. Madrid: Díaz de Santos.
Silva Ayçaguer, L. C. (1997). *Cultura estadística e investigación científica en el campo de la salud: una mirada crítica*. Madrid: Díaz de Santos.
Silva Ayçaguer, L. C. (1999). La crisis de las pruebas de significación y la alternativa bayesiana. Memorias del XI Congreso de la Sociedad Gallega de Estadística e Investigación Operativa, Santiago de Compostela.
Silva Ayçaguer, L. C. (2000). *Diseño razonado de muestras y captación de datos para la investigación sanitaria*. Madrid: Díaz de Santos.
Silva Ayçaguer, L. C. (2001). Errores garrafales en el uso del muestreo. Apuntes de una experiencia. *Brotes*, 2 (1), 50-54.
Silva Ayçaguer, L. C. (2005). Una Ceremonia Estadística para identificar Factores de Riesgo. *Buenos Aires Salud Colectiva*, 1(3), 309-322.
Silva Ayçaguer, L. C. (2008). Los límites de las pruebas de significación estadística y los valores de p. *AMF*, 4(2), 62-63.
Silva Ayçaguer, L. C. (2009). *La investigación biomédica y sus laberintos: en defensa de la racionalidad para la ciencia del siglo XXI.* Madrid: Díaz de Santos.
Silva Ayçaguer, L. C. y Barroso J. (2004). *Regresión Logística*. Cuaderno 27. Madrid: La Muralla.
Silva Ayçaguer, L. C. y Barroso, I. M. (2001). Selección algorítmica de modelos en las aplicaciones biomédicas de la regresión múltiple. *Med Clin Barc*, 116, 741-745.
Silva Ayçaguer, L. C. y Benavides A. (2000). Contra la sumisión estadística: un apunte sobre las pruebas de significación. *Metas*, 27, 35-40.
Silva Ayçaguer, L. C., Pérez, C. y Cuellar, I. (1994). Uso de la estadística en la investigación de salud contemporánea. *Gac Sanit*, 9(48), 189-195.

Snedecor, G. W. and Cochran, W. G. (1989). *Statistical methods*. 8va ed. United State of America: Iowa State University Press.

Streiner, G. R. y Norman, D. L. (1998). *Bioestadística*. Madrid: Harcourt Brace.

Stroup, D. F., Berlin, J. A. and Morton, S. C. (2000). Meta-analysis of Observational Studies in Epidemiology: A Proposal for Reporting. *JAMA*, 283, 2008-2012.

Talavera, J. O., Wacher Rodarte, N. H. y Rivas Ruiz, R. (2011). Estudios de proceso (prueba diagnóstica). *Rev Med Inst Mex Seguro Soc*, 49 (2), 163-170.

Taucher, E. (1999*). Bioestadística.* 2da ed. España: Editorial Universitaria.

Thode, H. C. Jr. (2002). *Testing for Normality*. New York: Marcel Dekker, Inc.

Triola, M. F. (2004). *Probabilidad y estadística*. 9na ed. México: Pearson Educación Inc.

Triola, M. F. (2009). *Estadística*. 10ma ed. México: Pearson Education Inc.

Van Belle, G., Lloyd, F., Heagerty, P. and Lumley, T. (2004). *Biostatistics: a methodology for the health sciences*. 2nd ed. United State of America: John Wiley and Sons, Inc., pp. 834-838.

Van Belle, G., Lloyd, F., Heagerty, P. and Lumley, T. (2004). Point and Interval Estimates. En: Balding D, Cressie N, Fisher N. (Eds). *Biostatistics: a methodology for the health sciences*. United State of America: John Wiley and Sons, Inc., pp. 85-87.

Varela Mollou, J., Rial Boubeta, A. y Picón Prado, E. (2016). Análisis multivariante. Universidad de Santiago de Compostelas. Facultad de Psicología, Curso 2015-2016. Disponible en: www.usc.es/psicom, www.miavizor.es, www.sellodetransparencia.es

Vélez Ibarrola, R. y García Pérez, A. (2012). *Principios de inferencia estadística*. Madrid: Universidad Nacional de Educación a Distancia.

Walker, S. H. and Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, S4, 167-179.

Wassertheil Smoller, S. (2003). *Biostatistics and Epidemiology*. 3rd ed. United State of America: Springer.

Elementos básicos para el manejo de muestras y variables en bioestadística

Dr. Jesús Romero Madero

El eficaz procesamiento estadístico de la información en los diferentes diseños de estudios investigativos requiere cada vez más del dominio actualizado de los elementos conceptuales básicos de la bioestadística y de novedosos algoritmos que faciliten el manejo de muestras y variables para una mejor selección y ejecución de las distintas opciones de técnicas de la inferencia estadística. Esta obra pretende proveer al estudiante de pregrado y de posgrado, y al investigador joven, de una herramienta de consulta bioestadística o tutorial que le ayude a orientarse en el correcto manejo de muestras y variables en relación con las opciones y procedimientos que se pueden seleccionar en la inferencia estadística empleando el método frecuentista y el bayesiano, así como con las posibles interpretaciones de los resultados que se obtengan. Ello se complementa con la exposición de los elementos conceptuales básicos involucrados en cada caso.

Jesús Romero Madero es Doctor en Medicina (1993), Especialista de II Grado en Gastroenterología (2007), Máster en Salud de los Trabajadores (2006) y Profesor Auxiliar (2019). Se vinculó con la docencia médica en la etapa de pregrado, como alumno ayudante de Medicina Interna, y en el posgrado por más de 15 años en las Facultades de Ciencias Médicas de las provincias La Habana y Artemisa. Ha recibido reconocimientos y premios en jornadas científicas provinciales y nacionales durante la etapa de pregrado, tiene realizadas numerosas publicaciones en revistas nacionales y, por más de 20 años, ha participado en congresos, simposios y eventos nacionales e internacionales de la especialidad de gastroenterología. Es miembro de la Sociedad Cubana de Gastroenterología. Actualmente labora como médico especialista y jefe del Servicio de Gastroenterología en el Hospital General Docente Iván Portuondo Echevarría, de Artemisa, y pertenece al claustro de profesores de la Facultad de Ciencias Médicas de esa provincia.

ecimed
EDITORIAL CIENCIAS MÉDICAS

www.ecimed.sld.cu