A Jay
Sidney Siegel
(Primera edición)

A Caryn, Norman y Tanya
N. John Castellan
(Segunda edición)

Traducción: **Mtra. Laura Edna Aragón Borja**
Profesor asociado B, tiempo completo, división de estudios de postgrado, UNAM campus Iztacala.

Mtro. Luis Enrique Fierros Dávila
Profesor asociado D, tiempo completo departamento de psicología y ciencias de la comunicación de la Universidad de Sonora.

Revisión Técnica: **Mtro. Arturo Silva Rodríguez**
Profesor titular A, tiempo completo, UNAM campus Iztacala.

ESTADÍSTICA NO PARAMÉTRICA

Aplicada a las ciencias de la conducta

Sidney Siegel
N. John Castellan

EDITORIAL TRILLAS
México, Argentina, España,
Colombia, Puerto Rico, Venezuela

Catalogación en la fuente

Siegel, Sidney
Estadística no paramétrica : aplicada a las ciencias de la conducta. -- 4a ed. -- México : Trillas, 1995 (reimp. 1998).
437 p. ; 23 cm.
Traducción de: Nonparametric statistics for the behavioral sciences
Bibliografía: p. 429-432
Incluye índices
ISBN 968-24-5101-9

1. Psicometría. 2. Ciencias sociales - Modelos matemáticos. I. Castellan, N. John. II. t.

D- 519.5'5789e LC- BF39'55.4 223

Título de esta obra en inglés:
Non Parametric Statistics
for the Behavioral Sciences.

Versión autorizada en español de la
segunda edición publicada en inglés por
© McGraw-Book Company
Nueva York, E. U. A.

División Administrativa, Av. Río Churubusco 385,
Col. Pedro María Anaya, C. P. 03340, México, D. F.
Tel. 6884233, FAX 6041364

División Comercial, Calz. de la Viga 1132, C. P. 09439
México, D. F. Tel. 6330995, FAX 6330870

Miembro de la Cámara Nacional de la
Industria Editorial. Reg. núm. 158

Primera edición en español, 1970
Segunda edición revisada en español, 1972 (ISBN 968-24-0146-1)
Reimpresiones, 1974, 1975, 1976, 1978, 1979, 1980, 1982,
1983, 1985, 1986 y 1988
Tercera edición en español, 1990 (ISBN 968-24-3896-9)
Reimpresiones, 1991 y 1994
Cuarta edición en español, 1995 (ISBN 968-24-5101-9)

Primera reimpresión, mayo 1998

Impreso en México
Printed in Mexico

Acerca de los autores

Sidney Siegel era profesor e investigador de Psicología en la Universidad del Estado de Pensilvania cuando ocurrió su muerte en 1961, a la edad de 45 años. Nativo de la ciudad de Nueva York, fue educado en California y obtuvo su doctorado en Psicología en la Universidad de Stanford en 1953. Desde entonces hasta su muerte, trabajó en la Facultad de la Universidad de Pensilvania, excepto por un año en que fue miembro del Centro de Estudios Avanzados en las Ciencias de la Conducta de la Universidad de Stanford.

Fue autor o coautor de cuatro libros publicados por McGraw-Hill: *Bargaining and Group Decision Making* (1960), con Lawrence E. Fouraker, y obtuvo el Premio de la Academia de Artes y Ciencias en 1959. Le siguió el libro *Bargaining Behavior* (1963), también en coautoría con Fouraker. En 1964 McGraw-Hill publicó *Choice, Strategy, and Utility*, después de que fue completado de manera póstuma por Alberta, E. Siegel y Julia McMichael Andrews. McGraw-Hill también publicó su colección de escritos en 1964, con el título *Decision and Choice*, editado por Samuel Messick y Arthur H. Brayfield. Se incluye también una memoria escrita por la señora Siegel. El antecedente de este libro fue *Nonparametric Statistics for the Behavioral Sciences* (1956), que, además de en inglés, apareció en japonés, italiano, alemán y español.

N. John Castellan, Jr. es profesor de Psicología en la Universidad de Indiana, en Bloomington. Recibió su A. B. de la Universidad de Stanford y su doctorado en la Universidad de Colorado. Ha trabajado como investigador asociado visitante en el Instituto de Investigación de Oregón y como profesor visitante de las Ciencias de la Computación en la Universidad de Colorado.

El profesor Castellan se ha desempeñado como asesor sobre estadística y computación en empresas y la industria; así mismo, ha trabajado como decano en la investigación y licenciatura en la Universidad de Indiana, y, en comités editoriales de varias revistas profesionales. Fue presidente de la Sociedad para el Cómputo en Psicología. Es editor de *Judgment/Decision Making Newsletter* y miembro de la Asociación Psicológica Estadounidense y de la Asociación Estadounidense para el Desarrollo de la Ciencia.

Es coautor de *Introduction to the Statistical Method* (2a. ed.) y fue coeditor de tres volúmenes de la serie monográfica *Cognitive Theory*. Ha publicado cerca de 50 artículos sobre estadística, toma de decisiones y la aplicación de las computadoras a la investigación y la instrucción.

Agradecimientos

Expreso mi agradecimiento a los siguientes editores y autores, quienes amablemente han otorgado su permiso para la reproducción de una o más tablas de la sección de Apéndices.

Agradezco profundamente al revisor literario de sir Ronald A. Fisher, F. S. R. a Frank Yates, F. S. R. y al Longman Group Ltd., Londres, por su autorización para reproducir las tablas III y IV de su libro *Statistical Tables for Biological, Agricultural and Medical Research* (6a. ed., 1974).

A la Administración de Biometrika, editores de *Biometrika y Biometrika Tables for Staticians, Volume I*, (3a. ed., 1966).

A Charles Griffin & Co. Ltd., por los materiales para los *Métodos de correlación de rangos* de Kendall (4a. ed., 1970).

A la American Statistical Association, editora de *Journal of the American Statistical Association and Technometrics*; la Biometric Society, editora de *Biometrics*; el Institute of Mathematical Statistics, editor de *Annals of Mathematical Statistics*; Gordon and Breach Science Publishers, Inc., editores de *Journal of Statistical Computation and Simulation*; Alfred A. Knopf; John Wiley; Macmillan, y McGraw-Hill.

Estoy, asimismo, en deuda con W. J. Dison, C. W. Dunnett, M. A. Fligner, M. H. Gail, S. S. Gupta, K. R. Hammond, M. Hollander, J. E. Householder, F. J. Massey Jr., C. Eisenhart, S. Maghsoodloo, M. R. Mickey Jr., R. E. Odeh, E. B., Page, D. W. Stilson y J. H. Zar, por otorgar la autorización para reproducir tablas estadísticas de sus trabajos publicados.

Prefacio a la segunda edición

Al revisar *Estadística no paramétrica para las ciencias de la conducta* he incluido técnicas que, según creo, son de especial valor para los científicos de la conducta. Debido al desarrollo de estadísticas no paramétricas y de distribución libre aparecido posteriormente a la primera edición, varios procedimientos han sido reemplazados por nuevas técnicas y algunos temas se han desarrollado considerablemente. En particular, las técnicas para *k* muestras (capítulos 6 y 7) se han ampliado y se comparan los procedimientos. Las medidas de asociación han sido desarrolladas significativamente.

Un rasgo distintivo de la primera edición fue la descripción paso a paso de la aplicación de cada procedimiento a datos reales. He tratado de mantener esta característica en la presente edición. Aunque algunos ejemplos de la primera edición han sido reemplazados, otros se mantienen. El objetivo es proporcionar una clara ilustración de exposición razonada, uso, cálculo e interpretación de cada estadístico.

Debido a la gran variedad de procedimientos no paramétricos y a la limitación de espacio, la elección de los métodos ha resultado difícil. Mi elección se ha basado en parte en la utilidad de cada procedimiento y en el esfuerzo por reducir la inclusión de pruebas similares.

Algunas elecciones merecen mención especial: he optado por incluir la prueba Ji cuadrada de Pearson para tablas de contingencia, más que modelos log-lineales. La razón es doble: he comprobado que los estudiantes dominan más fácilmente los conceptos de la prueba de Pearson y que la evidencia sugiere que dicha prueba es más apta para muestras pequeñas.

He omitido las pruebas de multivarianza, excepto por un par de ejemplos, en pruebas de secuencias de conductas. Aunque existen temas importantes para los científicos de la conducta, cada uno de estos requiere explicaciones extensas para tener una presentación adecuada.

Los lectores con un mínimo de conocimientos de estadística pueden utilizar este libro; sin embargo, estos lectores encontrarán los capítulos 1 y 2 más concisos, pero completos. Aquellos que han llevado uno o más cursos de estadística, pueden revisar superficialmente dichos capítulos.

Un aspecto importante del libro es que obliga a la controversia. En la primera edición, las escalas de medición se destacaron a lo largo del texto. En esta revisión he incluido una extensa exposición de las escalas de medición (capítulo 2), he "suavizado" la mayor parte del lenguaje relacionado con la importancia de las escalas en las explicaciones de las técnicas en particular. El papel de las escalas de medición en *investigación* es complicado y éste a menudo se considera independiente de la *estadística*. Mi experiencia en la enseñanza y la asesoría me ha inducido a creer que con demasiada frecuencia se le otorga poca importancia, con resultados desafortunados. Las mediciones afectan la *interpretación* de los datos que se obtienen en las investigaciones, y he comprobado que el énfasis en las escalas ayuda a los investigadores a hacer interpretaciones adecuadas de sus datos. Aunque algunos defensores de las distintas perspectivas relacionadas con el papel de las escalas de medición en la estadística pueden no estar satisfechos con el énfasis que pongo, creo que un mejor balance ayudará a los investigadores a realizar su trabajo más correctamente.

Un rasgo adicional de esta edición es la inclusión de listas de programas para computadora, que resultarán útiles para algunos de los procedimientos. El cálculo de muchas de las técnicas presentadas en el texto puede ser realizado manualmente o por medio de una calculadora electrónica de bolsillo. Sin embargo, otras técnicas implican cálculos difíciles o tediosos. Para éstas, se incluyen listas de programas en el Apéndice II. Estas listas se encuentran en BASIC porque su lenguaje es accesible, virtualmente, a todos los usuarios de microcomputadoras (y sistemas mayores). Se ha realizado un esfuerzo por hacer que las listas sean fáciles de interpretar, de tal suerte que pueda entenderse la lógica de cada programa, sin requerir ningún programa adicional. Como resultado, tenemos una serie de programas que no son tan eficaces o elegantes como pudieran serlo. De nuevo, la meta fue la claridad y facilidad de uso. Como se advierte, hasta el momento no existe un solo paquete de programas para computadora que pueda realizar todos los análisis descritos en el libro.*

En el momento de preparar esta edición, quiero expresar mi reconocimiento al ánimo y apoyo recibido de Alberta Siegel, al inicio y durante la elaboración del trabajo. Quiero expresar mi gratitud a todos los estudiantes que colaboraron en los primeros borradores de esta revisión, quienes ofrecieron un sinnúmero de críticas que enriquecieron el trabajo. Estoy particularmente agradecido a los colegas que leyeron y comentaron uno o más borradores del manuscrito: Helena Chmura Kraemer, Richard Lehman, Thomas Nygren, James L. Phillips, J. B. Spalding y B. James Starr. Finalmente, el gran apoyo de mi esposa e hijos, quienes, si no siempre entendieron lo que me encontraba haciendo, me dieron ánimo y estímulo para terminar la tarea.

N. JOHN CASTELLAN, JR.

* El conjunto completo de los procedimientos presentados en este libro está disponible en un paquete de programas para microcomputadora de técnicas estadísticas no paramétricas. Para información adicional relacionada con dicho paquete el lector puede dirigirse a N. John Castellan Jr., Departamento de Psicología, Indiana University, Bloomington, In., 47405.

Prefacio a la primera edición

Considero que entre las técnicas estadísticas utilizadas para evaluar hipótesis, las que mejor se adecuan a los datos de las ciencias conductuales son las no paramétricas. Los dos nombres alternativos que frecuentemente se les da a estas pruebas sugieren dos razones para esa adecuación. A menudo se les denomina de *distribución libre* y uno de sus rasgos principales es que no suponen que las puntuaciones que se analizan fueron extraídas de una población distribuida de una cierta manera, por ejemplo, de una población que presenta una distribución normal. Alternativamente, muchas de estas pruebas se identifican como *pruebas de rangos* y ese título sugiere otro rasgo fundamental: las técnicas no paramétricas pueden utilizarse con puntuaciones que, en sentido estricto, no son numéricas, pero que son simplemente rangos. Una tercera ventaja de dichas técnicas es, por supuesto, la sencillez de sus cálculos. Muchos creen que los investigadores y estudiantes de las ciencias de la conducta requieren un mayor tiempo y reflexión para la formulación cuidadosa de sus problemas de investigación, así como para la recolección de los datos precisos y relevantes. Tal vez prestarían más atención a esta búsqueda si se les auxiliara en la necesidad de realizar los cálculos estadísticos que son complicados y que consumen tiempo. Una ventaja final de las pruebas no paramétricas es su aplicación a muestras pequeñas, un rasgo que podría ser útil al investigador que recaba datos de estudios pilotos y a aquel cuyas muestras son pequeñas dada su naturaleza (por ejemplo, muestras de personas con una rara forma de enfermedad mental, o muestras de culturas).

Hasta la fecha no existe fuente disponible que presente las técnicas no paramétricas en una forma útil y en términos que sean familiares a los científicos de la conducta. Las técnicas se describen en distintas publicaciones de matemáticas y estadística. La mayor parte de los científicos de la conducta no tienen la completa preparación matemática requerida para consultar estas fuentes. Adicionalmente, ciertos escritores han presentado resúmenes de las técnicas en artículos dirigidos a científicos sociales. Notables entre éstos son Blum y Fattu (1954), Moses (1952*a*), Mosteller y Bush (1954) y Smith (1953). Aún más, algunos de los nuevos textos de estadística para científicos sociales contienen capítulos en los que se desarrollan métodos no paramétricos. Entre ellos podemos citar a Edwards (1954), McNemar (1955) y Walker y Lev (1953). A pesar de lo valioso de esas fuentes, los autores fueron sumamente selectivos en las técnicas presentadas y no incluyeron las tablas de los valores de significancia que se utilizan en varias de las pruebas. Por tanto, creí que sería deseable un texto de métodos no paramétricos que incrementara la bibliografía integrada por los textos mencionados.

En este libro he presentado las pruebas de acuerdo con los diseños de investigación a los que son aplicables. En el análisis de cada prueba he intentado señalar su "función", por ejemplo, indicar el tipo de datos a los que es aplicable, por convenir a alguna noción del razonamiento o a la prueba que subyace a la técnica, para explicar su cálculo, para proporcionar ejemplos de su aplicación en la investigación de las ciencias de la conducta y para comparar la prueba con su equivalente paramétrica, si es posible, y con cualquier prueba no paramétrica de función similar.

El lector puede sorprenderse por la cantidad de espacio dedicado a los ejemplos del uso de estas pruebas, y aun, asombrarse con la reiteración con la que se presentan las mismas. Puedo justificar de la siguiente manera el espacio dedicado: *a*) los ejemplos ayudan a enseñar los cálculos de las pruebas; *b*) ilustran la aplicación de la prueba a problemas de investigación en las ciencias conductuales, y *c*) el uso de los mismos seis pasos en cada prueba de hipótesis demuestra que una lógica idéntica subyace a cada una de las numerosas técnicas estadísticas, un hecho que no es muy bien entendido por muchos investigadores.

Puesto que he intentado presentar todos los datos en "bruto" para cada uno de los ejemplos, no logré obtener dichos datos de fuentes de un grupo católico. Al publicar investigaciones típicamente se presentan datos brutos y, por tanto, me sentí obligado a obtenerlos de fuentes de un grupo parroquial para la mayor parte de los ejemplos, fuentes de las cuales los datos eran fácilmente asequibles. El lector entenderá que esto es una disculpa por la frecuencia con que he presentado en los ejemplos mi propia investigación y la de mis colegas cercanos. En ocasiones no encontré los datos apropiados para ilustrar el uso de una prueba y, por lo mismo, los "inventé" para ese propósito.

Al escribir este libro he sido muy cuidadoso respecto a la influencia que varios profesores y colegas han ejercido en mi pensamiento. El profesor Quinn McNemar fue quien fundamentalmente me capacitó en estadística inferencial y me introdujo en la importancia de los supuestos que subyacen a varias pruebas estadísticas. El profesor Lincoln Moses enriqueció mi comprensión de la estadística y fue el primero que hizo que me interesara en la bibliografía de la estadística no paramétrica. Mis estudios con el profesor George Polya redituaron interesantes pensamientos en teoría de la probabilidad. Los profesores Kenneth J. Arrow, Albert H. Bowker, Douglas H. Lawrence y J. C. C. McKinsey contribuyeron significativamente a mi entendimiento de la estadística y del diseño experimental. Mi comprensión de la teoría de la medida fue profundamente influida por mi colaboración en una investigación con los profesores Donald Davidson y Patrick Suppes.

Este libro se benefició enormemente con las estimulantes y detalladas sugerencias y críticas que me proporcionaron los profesores James B. Bartoo, Quinn. McNemar y Lincoln Moses, después de haber leído el manuscrito. Estoy profundamente agradecido con cada uno de ellos por el valioso tiempo que me dedicaron, así como a su conocimiento. Estoy en deuda además, con los profesores John F. Hall y Robert E. Stover, quienes me animaron a escribir este libro y contribuyeron con sus comentarios críticos en algunos de los capítulos. Por supuesto, ninguna de estas personas es responsable, de manera alguna, por las fallas que puedan encontrarse: éstas son enteramente de mi responsabilidad, y estaré sumamente agradecido si alguno de los lectores que detecten errores me los hacen saber.

Gran parte de la utilidad de esta obra se debe a la generosidad de muchos autores y editores, quienes amablemente me permitieron adaptar o reproducir tablas y otros materiales publicados originalmente por ellos. He mencionado cada una de las fuentes en donde aparecen los materiales, y además, deseo hacer mención de mi gratitud a Donovan Auble, Irvin L. Child, Frieda Swed Cohn, Churchill Eisenhart, D. J. Finney, Milton Friedman, Leo A. Goodman, M. G. Kendall, William Kruskal, Joseph Lev, Henry B. Mann, Frank J. Massey Jr., Edwin G. Olds, George W. Snedecor, Helen M. Walker, W. Allen Wallis, John E. Walsh, John W. M. Whiting, D. R. Whitney, Frank Wilcoxon y al Institute of Mathematical Statistics, la American Statistical Association, Biometrika, la American Psychological Association, a la Iowa State College Press, a la Yale University Press, al Institute of Educational Research en Indiana University, la American Cyanamid Company, Charles Griffin & Co. Ltd., John Wiley & Sons Inc. y Henry Holt and Company Inc. Estoy sumamente agradecido al profesor sir Ronald A. Fisher, Cambridge, al doctor Frank Yates, Rothamsted, y a los señores Oliver and Boyd Ltd., Edimburg, por su autorización para reproducir las tablas III y IV de su libro *Statistical Tables for Biological, Agricultural, and Medical Research.*

Mi gran deuda personal es con mi esposa, la doctora Alberta Engvall Siegel, sin cuya ayuda este libro no se hubiera escrito. Ella trabajó estrechamente conmigo en cada fase de su planeación y redacción. Sé que el libro no solamente se benefició de su conocimiento de las ciencias de la conducta, sino de una cuidadosa edición, cuestiones que deben ser destacadas en cualquier exposición de méritos que el libro pudiera tener.

Sidney Siegel

Índice de contenido

Glosario de símbolos

Nota: El número entre paréntesis indica el(los) capítulo(s) del libro en donde se definieron los símbolos o donde se utilizaron inicialmente.

a_{ij} Notación preferente utilizada en el cálculo del coeficiente de acuerdos de Kendall (8).

$A(X_i)$ Indica el atributo de un objeto x_i (2).

α Alfa. Probabilidad de cometer un error de tipo I: la probabilidad de rechazar H_0 cuando ésta es verdadera.

β Beta. Probabilidad de cometer un error de tipo II: la probabilidad de rechazar H_1 cuando ésta es verdadera.

C Coeficiente de Cramér (8).

C_j Indica la sumatoria de las frecuencias en la j-ésima columna en una tabla de contingencia (7 y 8).

γ Gamma. Índice poblacional gamma de la asociación entre variables ordenadas (8).

d_{BA} *d* de Somers, un índice de la asociación asimétrica para variables ordenadas (8).

d_i Diferencia entre puntuaciones igualadas: X_i-Y_i. Se utiliza en la prueba de Wilcoxon (4), en la prueba de las permutaciones para pares replicados (4) y la correlación rango-orden de Spearman (8).

d_{ij} Residuos ajustados o estandarizados utilizados al evaluar las desviaciones individuales de cada celdilla en la prueba ji cuadrada (7).

$D_{m,n}$ Estadístico asociado a las pruebas de Kolmogorov-Smirnov (3 y 5).

$D(X_j)$ Índice de dispersión en la prueba de rangos de Moses para escalas de diferencias (5).

gl Grados de libertad asociados a varias pruebas estadísticas, generalmente pruebas ji cuadrada y pruebas t.

Δ_{BA} Delta. La población paramétrica correspondiente a la d de Somers, un índice de la asociación asimétrica para variables ordenadas (8).

E_i Valor esperado utilizado en las pruebas ji cuadrada (3 y 4).

E_{ij} Valor esperado utilizado en las pruebas ji cuadrada (5 y 7).

$F_0(X)$ Distribución de la frecuencia acumulada especificada por la hipótesis nula en la prueba Kolmogorov-Smirnov (3).

F_r Estadístico del análisis de varianza bifactorial por rangos, de Friedman (6).

G Estadístico gamma para medir la asociación entre variables ordenadas (8).

H_0 Indica la hipótesis nula.

H_1 Indica la hipótesis alterna.

θ_x Theta. La mediana poblacional de la variable X.

J Prueba de Jonckheere para alternativas estadísticas ordenadas (7).

J^* Aproximación de la prueba estadística de Jonckheere para muestras grandes (7).

K Estadístico Kappa, un índice para los acuerdos entre datos en escala nominal (8).

$K_{m,n}$ Estadístico asociado con la forma para muestras grandes de la prueba del momento del cambio (3).

KW Estadístico del análisis de varianza unifactorial por rangos de Kruskal-Wallis (7).

κ Kappa. Índice poblacional kappa de acuerdos para datos en escala nominal (8).

L Estadístico de la prueba de Page para alternativas ordenadas (6).

$L(X_i)$ Indica la función de etiquetación para un objeto x_i (2).

L_B, L_A Estadístico lambda para medir la asociación asimétrica entre variables en escala nominal (8).

λ_B, λ_A Lambda. Índice poblacional lambda de la asociación asimétrica entre variables en escala nominal (8).

$\mathbf{M}_{ij}^{+}$, $\mathbf{N}_{ij}^{+}$ Acción de contar para tablas de contingencia. Se utiliza en el cálculo del estadístico gamma (8).

m Muestra de mayor tamaño en pruebas de dos muestras.

m', n' Tamaños de las muestras ajustados en la prueba de rangos de Moses para escalas de diferencias (5).

$\max(X)$ Valor máximo de la variable X.

$\text{med}(X)$ Mediana de la variable X.

$\text{med}(X_i, X_j, X_k)$ Mediana de las variables X_i, X_j, X_k.

$\min(X)$ Valor mínimo de la variable X.

μ Mu. Media poblacional.

μ_x Media poblacional de la variable X.

n Muestra de menor tamaño en pruebas de dos muestras.

n_{ij} Valor observado, utilizado en pruebas ji cuadrada (5 y 7).

N Tamaño de la muestra.

$\binom{N}{k} = \dfrac{N!}{k!(N-k)!}$ Coeficiente binomial. Expresa el número de combinaciones de N objetos tomados en k en cada ocasión (3).

$N!$ Factorial. $N! = N\ (N-1)\ (N-2)\ (N-3)\ldots(2)\ (1)$, por ejemplo, $5! = (5)\ (4)\ (3)\ (2)\ (1) = 120$. *Nota*: Por definición $0! = 1$ (3)

O_i Valor observado, utilizado en pruebas ji cuadrada (3 y 4).

p Probabilidad. Se utiliza en lugar de $P\,[X]$ cuando el contexto es claro.

$P[H]$ Probabilidad de la variable aleatoria H.

q Probabilidad. Generalmente se utiliza para indicar la probabilidad asociada con un resultado binario, $q = 1 - p$ (3).

$q(\alpha, \#c)$ Estadístico utilizado en la comparación de un grupo control con grupos o condiciones relacionados (6).

Q Estadístico de la prueba Q de Cochran para comparar proporciones correlacionadas (6).

r Número de series en la prueba para una muestra de series (3).

r_ϕ Coeficiente phi para tablas de contingencia de 2×2 (8).

r_s Coeficiente de correlación de rangos ordenados de Spearman (8).

R_i Indica la sumatoria de las frecuencias en la i-ésima columna en una tabla de contingencia (7 y 8).

R_j Sumatoria de rangos en el j-ésimo grupo (6, 7 y 8).

$\overline{R}_j$ Promedio de los rangos en el j-ésimo grupo (6, 7 y 8).

ρ_s Rho. Coeficiente poblacional de correlación por orden de rangos de Spearman (8).

S Número de acuerdos menos el número de desacuerdos en el ordenamiento por rangos de dos conjuntos de datos. Se utiliza para calcular el coeficiente de correlación por orden de rangos de Kendall (8).

$S_N(X)$ Distribución de la frecuencia acumulada para muestras de tamaño N. Se utiliza en la prueba de Kolmogorov-Smirnov (3 y 5).

σ Sigma. Desviación estándar poblacional.

σ_x Desviación estándar poblacional de la variable X.

$\sigma_{\bar{x}}$ Error estándar poblacional de la media.

σ^2 Varianza poblacional.

t Estadístico de la prueba t de Student.

t_j Número de rangos empatados en el j-ésimo grupo de empates. Se utiliza en pruebas donde los datos son rangos (5).

T, T_{xy} Coeficiente de correlación por orden de rangos de Kendall (8).

$T_{xy \cdot z}$ Coeficiente de correlación parcial rangos de Kendall (8).

T_c Correlación entre varios jueces y un criterio de ordenamiento por rangos (8).

T^+ Sumatoria de las diferencias positivas en la prueba de signos por rangos de Wilcoxon (4).

T^- Sumatoria de las diferencias negativas en la prueba de signos por rangos de Wilcoxon (4).

T_x, T_y Factor de corrección para rangos empatados en el coeficiente de correlación por orden de rangos de Spearman (8).

T_x, T_y Factor de corrección para rangos empatados en el coeficiente de correlación por orden de rangos de Kendall (8). (Los valores T_x y T_y serán diferentes dependiendo de si se aplica el coeficiente de Kendall o el de Spearman.)

τ Tau. Coeficiente de correlación poblacional por orden de rangos de Kendall (8).

$\tau_{xy \cdot z}$ Coeficiente de correlación parcial poblacional por orden de rangos de Kendall (8).

$\bar{\tau}$ Tau poblacional promedio para evaluar la significación de el coeficiente de Kendall para los acuerdos cuando los datos son rangos (8).

u Coeficiente de Kendall para los acuerdos (8).

U_{ij} Notación del estadístico U de Mann-Whitney. Utilizado en el cálculo del estadístico de Jorckheere (7).

$U(YX)$ Ubicación promedio de un conjunto X de puntuaciones respecto a un conjunto Y de puntuaciones. Se utiliza en la prueba poderosa por orden de rangos (5).

$U(YX_i)$ Ubicación de un conjunto X de puntuaciones respecto a las puntuaciones Y. Se utiliza en la prueba poderosa de rangos ordenados (5).

$\grave{U}$ Prueba estadística para la prueba poderosa de rangos ordenados (5).

υ Ípsilon. Parámetro poblacional para el coeficiente de Kendall para los acuerdos cuando los datos son comparaciones apareadas (8).

$\varnothing$ Phi. Subíndice utilizado para $r_\varnothing$, el coeficiente Phi (8).

V_x, V_y Estadístico similar a la varianza para la prueba poderosa de rangos ordenados (5).

W Coeficiente de Kendall de acuerdos entre ordenamientos múltiples por rangos (8).

W_T Índice de acuerdo entre juicios. Similar al coeficiente de Kendall para acuerdos (8).

W_x Sumatoria de rangos para el grupo X en la prueba Wilcoxon-Mann-Whitney (5). Además se utiliza en la prueba Siegel-Tukey para escalas de diferencias (5).

X, X_i Dato o puntuación observado.

$\overline{X}$ Media muestral de la variable X.

X^2 Estadístico de la prueba ji cuadrada (3, 4, 5 y 7).

X_i^2 Estadístico de la prueba ji cuadrada para particiones de una tabla de contingencia (5 y 7).

χ^2 Ji cuadrada. Distribución ji cuadrada (3, 4, 5 y 7).

z Puntuación z. Generalmente se utiliza para indicar una variable transformada a una *forma estándar*; por ejemplo, con media igual a cero y desviación estándar igual a uno.

\# Procedimiento de contar. Por ejemplo:

\#H Número de cabezas (1).

\#(+) Número de acuerdos en la ordenación de los objetos de dos grupos (8).

\#(−) Número de desacuerdos en la ordenación de los objetos de dos grupos (8).

Introducción

Los estudiantes de las ciencias de la conducta y sociales están acostumbrados a utilizar palabras de uso común en formas que, en un principio, no les resultaban familiares. Durante el transcurso de sus estudios, aprenden que el científico conductual que habla de *sociedad* no se está refiriendo a ese privilegiado grupo de personas cuyos nombres aparecen en las páginas de sociales de los periódicos. Saben también que, aunque un estudiante de secundaria pueda desdeñar o despreciar a alguno de sus compañeros por "no tener personalidad", la denotación científica del término *personalidad* tiene poco o nada en común con el significado que le pueda dar un adolescente. Los estudiantes aprenden así mismo que el término *cultura*, cuando es usado técnicamente, abarca mucho más que un refinamiento estético. Por otro lado, ellos no caerán en el error de decir el disparate de que un vendedor "usa" la *psicología* para persuadir a un cliente de comprar un producto en particular.

De manera similar, los estudiantes descubren que el campo de la *estadística* es completamente diferente de la concepción común que se tiene de él. Tanto en los periódicos como en la radio y la televisión se presenta al estadístico como aquella persona que recaba una gran cantidad de información cuantitativa, la resume, la procesa y la difunde. Así, estamos familiarizados con la noción de que el trabajo del estadístico consiste en la determinación del salario por hora promedio en una industria o el número promedio de niños en la familia urbana estadounidense: a algunos les resulta más familiar el papel del estadístico en los acontecimientos deportivos. Pero los estudiantes que han tomado un curso de estadística, aunque sea introductorio, saben que la descripción es sólo una de las funciones del estadístico.

Una función central de la estadística moderna es la *inferencia estadística*. La estadística inferencial está interesada en dos tipos de problemas: la estimación de los parámetros de la población y las pruebas de hipótesis. Estas últimas, serán el tema principal de este libro.

El verbo *inferir* significa "obtener conclusiones como una consecuencia o como una probabilidad". Cuando vemos que una mujer no usa anillo alguno en los dedos de su mano izquierda, podemos *inferir* que no está casada. Sin embargo, esta inferencia pudiera ser incorrecta. Por ejemplo, esa mujer podría ser originaria de Europa, en donde el anillo de bodas se usa con frecuencia en la mano derecha, o simplemente, que ella haya decidido no usar ese anillo.

En la inferencia estadística estamos interesados en cómo obtener conclusiones acerca de grandes grupos de sujetos o de eventos, sobre la base de observaciones de pocos sujetos o de lo que ha ocurrido en el pasado. La estadística proporciona instrumentos que formalizan y estandarizan nuestros procedimientos para obtener tales conclusiones. Por ejemplo, si quisiéramos determinar cuál de tres variedades de salsa de tomate es la más popular en las cocinas estadounidenses, podríamos recabar información sobre este tema parándonos cerca de la sección de salsas de tomate de una tienda y contando el número de envases de cada tipo que la gente adquiere en el curso de un día; con seguridad, el número de elecciones de las tres variedades de salsa será diferente. Pero, ¿podemos *inferir* que la variedad preferida *ese* día en esa tienda por los compradores de *ese* día sea realmente la más popular en las cocinas estadounidenses? El poder hacer tal inferencia debe depender del margen de popularidad sostenido por la marca más frecuentemente elegida, por la representatividad de la tienda y también por la representatividad del grupo de compradores que hemos observado.

Los procedimientos de la inferencia estadística introducen orden en cualquier intento de obtener conclusiones de las evidencias proporcionadas por las muestras. La lógica de los procedimientos dicta algunas de las condiciones en las cuales la evidencia debe reunirse, y las pruebas estadísticas determinan si, de la evidencia que hemos reunido, podemos tener confianza en lo que hemos concluido acerca de un gran grupo, derivado de sólo los pocos sujetos que hemos muestreado.

Un problema común para la inferencia estadística es determinar, en términos de probabilidad, si las diferencias observadas entre dos muestras significa que las poblaciones muestreadas son realmente diferentes. Aun si reuniéramos dos grupos de puntuaciones tomando al azar muestras de la misma población, probablemente encontraríamos que las puntuaciones difieren en algún grado. ¿Ocurren estas diferencias simplemente por factores aleatorios? ¿Cómo podemos determinar en cualquier caso dado si las diferencias observadas entre dos muestras se deben meramente al azar o son causadas por otros factores? Los procedimientos de la inferencia estadística nos permiten determinar si las diferencias observadas están o no dentro del grado en que podrían haber ocurrido simplemente por azar. Otro problema común es determinar si una muestra de puntuaciones pertenece a alguna población específica. Un problema adicional consiste en decidir si podemos inferir legítimamente que varios grupos difieren entre ellos. En este libro trataremos con cada uno de estos problemas de la inferencia estadística.

En el desarrollo de los métodos estadísticos modernos, las primeras técnicas de inferencia que aparecieron fueron aquellas que hicieron suposiciones acerca de la naturaleza de las poblaciones de las cuales se derivaron las observaciones y los datos. Estas técnicas estadísticas se llaman *paramétricas*. Por ejemplo, una técnica de inferencia puede estar basada en la suposición de que los datos se derivan de una *población normalmente distribuida*. Otra técnica de inferencia puede estar basada en la suposición de que dos conjuntos de datos se tomaron de poblaciones que tienen la misma varianza (σ^2) o dispersión de puntuaciones. Tales técnicas proporcionan conclusiones de la forma siguiente: "Si las suposiciones acerca de la forma de la distribución de la población son válidas, entonces podemos concluir que. . .". Debido a las suposiciones comunes, tales pruebas se sistematizan fácilmente y son también muy fáciles de enseñar y aplicar.

Un poco más recientemente hemos presenciado el desarrollo de un gran número de técnicas de inferencia que no hacen suposiciones numerosas o rigurosas acerca de la población de la cual se han muestreado los datos. Estas técnicas de *distribución libre o no paramétricas* dan como resultado conclusiones que requieren menos calificaciones. Si hemos usado una de estas técnicas, seremos capaces de decir que: "Sin considerar la(s) forma(s) de la(s) población(es), podemos concluir que. . .". En este libro expondremos este tipo de técnicas.

Algunas técnicas no paramétricas son *pruebas de rangos* o *pruebas de orden*, y estos términos sugieren otro modo en el que las pruebas no paramétricas difieren de las pruebas paramétricas. Cuando usamos cualquier prueba estadística, implícitamente hacemos ciertas suposiciones acerca de las asignaciones numéricas de los objetos observados. Como veremos en el capítulo 2, las reglas para la asignación numérica constituyen una escala de medición. La regla de asignación que usamos (es decir, la escala) impone restricciones al tipo de interpretaciones y operaciones que son apropiadas a esas asignaciones. Cuando la aplicación de la prueba estadística transforma los valores de la escala de manera inapropiada, se dificulta interpretar el resultado. Aunque podemos computar una prueba estadística paramétrica para datos de cualquier tipo, la facilidad en la interpretación de la prueba depende de la manera en que las observaciones se tranforman en números para su análisis. Por otra parte, muchas pruebas no paramétricas se centran, más que en sus valores "numéricos", en el orden o el rango de sus puntuaciones; e incluso otras técnicas no paramétricas son útiles con datos para los que el ordenamiento es imposible (esto es, con datos clasificatorios). Mientras que una prueba paramétrica puede centrarse en las diferencias entre las medias de dos poblaciones, la prueba no paramétrica análoga se enfoca en las diferencias entre las medianas. En las ciencias de la conducta, las ventajas de los estadísticos basados en el ordenamiento de los datos (¡en las cuales las puntuaciones "numéricas" pueden ser numéricas sólo en apariencia!) son aparentes. Examinaremos este tema con más detalle en el capítulo 2, en el que se contrastan las pruebas paramétricas y las no paramétricas.

De los ocho capítulos de este libro, cinco se dedican a la presentación de una gran variedad de pruebas estadísticas no paramétricas. Las pruebas se analizan en los capítulos de acuerdo con el diseño de investigación para el cual resultan adecuadas. El capítulo 3 contiene pruebas que pueden usarse cuando se desea determinar si una muestra simple pertenece a alguna población específica. En los dos capítulos siguientes se presentan pruebas que pueden emplearse cuando el propósito es comparar las puntuaciones obtenidas por dos muestras; en el capítulo 4 se consideran las pruebas para dos muestras relacionadas, mientras que en el capítulo 5 se examinan las pruebas para dos muestras independientes. De manera similar, los dos siguientes capítulos se dedican a pruebas de significancia para tres o más muestras; el capítulo 6 presenta pruebas para tres o más muestras relacionadas y el capítulo 7, para tres o más muestras independientes. En el capítulo 8 se estudian las medidas de correlación no paramétricas y sus pruebas de significancia.

Además, hemos tratado de hacer el libro inteligible al lector cuyo conocimiento matemático se limite al álgebra elemental. Esta orientación implica excluir la presentación de muchas derivaciones. Siempre que ha sido posible hemos tratado de comunicar un entendimiento "intuitivo" de la racionalización que subyace a una prueba, ya que creemos que este entendimiento será más útil que un intento de seguir la derivación. Al lector con conocimientos de matemáticas más complejos que quiera dedicarse a los temas cubiertos en este libro, se le remite a las fuentes bibliográficas a las que hacemos referencia.

Los lectores cuyo conocimiento matemático sea limitado y especialmente aquellos cuya experiencia educativa sea tal que le haya desarrollado respuestas emocionales negativas a los símbolos, a menudo encuentran los libros de estadística difíciles debido al extenso uso que se hace de ellos. Tales lectores pueden descubrir que gran parte de esta dificultad desaparece si le prestan más atención de la acostumbrada y relacionan la presentación textual con las presentaciones tabulares de los datos. Además, se invita al lector a aprender a leer ecuaciones y fórmulas como si fueran oraciones, sustituyendo los nombres de las variables por los nombres de los símbolos. Desde luego, no se espera que un estudiante de ciencias de la conducta o sociales pueda mantener la misma rapidez de lectura en un libro de estadística que, por ejemplo, en un libro de personalidad, de hostilidad intergrupal o del papel que la geografía desempeña en las diferencias culturales. Los textos estadísticos son más densos que la mayoría de los de las ciencias sociales —nosotros usamos símbolos para mayor brevedad, así como para mayor precisión— y, por tanto, se requiere una lectura más lenta. El lector que encuentre dificultad en los símbolos, puede apoyarse en el glosario que se ha incluido. En él se resumen los significados de varios símbolos empleados en el libro. Una razón de que el uso extensivo de símbolos haga el material más difícil, puede ser que los símbolos son términos generales o abstractos, que adquieren una variedad de significados específicos en una variedad de casos específicos. Así, por ejemplo, cuando hablamos de k muestras, queremos decir cualquier número de muestras; 3, 4, 8, etc. Naturalmente, en estos ejemplos cada símbolo adquiere un valor numérico específico, y los ejemplos pueden servir para "concretar" la exposición al lector.

Muchos de los lectores cuentan con calculadoras electrónicas en las que pueden computar la mayoría de los estadísticos descritos en esta obra. Otros lectores tienen acceso a "paquetes" estadísticos para usar en computadoras. Aunque las computadoras pueden hacer de cualquier trabajo de análisis de datos pesado un trabajo mínimo, es importante que el usuario entienda el estadístico, sus suposiciones y lo que hace con los datos. Un modo mejor de comprender las técnicas estadísticas es computarlas con nuestros propios datos. Al presentar las técnicas hemos escogido nuestros datos de escritorio para describir los procedimientos de análisis de una manera amena. Aunque ciertamente se pueden usar los paquetes de computadora (y en muchos casos deben utilizarse), con frecuencia es más fácil analizar pequeños conjuntos de datos "a mano", auxiliándonos con una calculadora. Para algunas de las estadísticas más complicadas, hemos incluido una lista de programas simples de computación que ayudarán a analizar los datos, si el procedimiento no está fácilmente disponible en otros paquetes.

Por último, el lector con conocimientos matemáticos limitados también encontrará los ejemplos especialmente útiles. Para cada prueba estadística presentada en este libro, se da un ejemplo de su uso en investigación. Los ejemplos sirven así mismo para ilustrar la importancia de los estadísticos en la investigación del científico conductual. Ésta quizá sea su función más útil, debido a que esta obra se dirige al investigador cuyo interés principal está en los campos de conocimiento de las ciencias de la conducta y sociales, más que en su metodología. Los ejemplos demuestran la íntima interrelación de la materia y el método en estas ciencias.

El uso de pruebas estadísticas en la investigación

En las ciencias de la conducta llevamos a cabo investigaciones con el propósito de probar hipótesis que derivamos de las teorías de la conducta. Una vez establecida una hipótesis estadística que nos parece importante para cierta teoría, recabamos datos que nos permitan decidir acerca de esa hipótesis. Nuestra decisión puede conducirnos a sostener, revisar o rechazar la hipótesis y la teoría de la cual se originó.

Para lograr una decisión objetiva acerca de si la hipótesis particular es confirmada por un conjunto de datos, debemos tener un procedimiento objetivo para rechazar o bien aceptar tal hipótesis. Se destaca la objetividad debido a que un aspecto importante del método científico es que se debe llegar a conclusiones por medio de métodos que sean del dominio público y que puedan ser repetidos por otros investigadores competentes.

Este procedimiento objetivo debe estar basado en la información o los datos que obtenemos de nuestra investigación y en el riesgo que estamos dispuestos a correr de que nuestra decisión acerca de la hipótesis sea incorrecta.

El procedimiento que generalmente se sigue incluye varios pasos. A continuación exponemos estos pasos en orden de ejecución: éste y el siguiente capítulo están dedicados a examinarlos con algún detalle. Los enumeramos aquí con el propósito de que el lector tenga una visión general del procedimiento total.

i. Establecer la hipótesis nula (H_0) y la alterna (H_1). Decidir qué datos se van a recabar y en qué condiciones.
Seleccionar una prueba estadística (con su modelo estadístico asociado) para probar H_0.

ii. De entre varias pruebas que pueden usarse con un diseño de investigación determinado, elegir el modelo de prueba que se aproxime lo más cercanamente posible a las condiciones de la investigación en términos de las suposiciones en las cuales está basada la prueba.

iii. Especificar un nivel de significancia (α) y un tamaño de muestra (N).

iv. Encontrar la distribución muestral de la prueba estadística bajo la suposición de que H_o es verdadera.

v. Con base en los puntos *ii*, *iii* y *iv*, definir la región de rechazo para la prueba estadística.

vi. Recabar los datos. Usando los datos obtenidos de la(s) muestra(s), computar el valor de la prueba estadística. Si ese valor está en la región de rechazo, la decisión es rechazar H_o; si ese valor está fuera de esta región, la decisión es que H_o no puede ser rechazada en el nivel de significación elegido.

En este libro se presentan varias pruebas estadísticas. En la mayoría de las presentaciones, se dan uno o más ejemplos para ilustrar el uso de la prueba. En cada ejemplo se siguen los seis pasos mencionados. Para entender la importancia de la estadística en la confirmación de hipótesis se requiere una comprensión básica de la razón de cada uno de estos pasos.

LA HIPÓTESIS NULA

El primer paso en el procedimiento de toma de decisiones es establecer la hipótesis nula (H_0). La *hipótesis nula* es una hipótesis de "no efecto" y por lo general se formula con el propósito expreso de ser rechazada; vale decir, es la negación del punto que se está tratando de probar. Si es rechazada, se apoya la hipótesis alterna (H_1). La *hipótesis alterna* es la declaración operacional de la hipótesis de investigación del experimentador. La *hipótesis de investigación* es la predicción derivada de la teoría sometida a prueba.

Cuando queremos tomar decisiones acerca de diferencias, probamos H_0 contra H_1. H_1 constituye la aseveración o hipótesis que se acepta si se rechaza H_0.

Supongamos que cierta teoría psicosocial nos conduce a predecir que dos grupos específicos de personas difieren en la cantidad de tiempo que dedican a leer periódicos. Esta predicción podría ser nuestra hipótesis de investigación; en otras palabras, nuestra hipótesis de investigación es que los grupos difieren. La confirmación de esa predicción apoya la teoría de la cual fue derivada. Para probar esta hipótesis de investigación, la formulamos en forma operacional como la hipótesis alterna H_1. Pero ¿cómo? Se podría usar la cantidad media de tiempo que cada grupo dedica a la lectura de periódicos. Entonces H_1 podría ser que $\mu_1 \neq \mu_2$, esto es, la cantidad media de tiempo dedicada a leer periódicos por los miembros de las dos poblaciones es diferente. H_0 podría ser que $\mu_1 = \mu_2$, esto es, la cantidad media de tiempo dedicada a leer periódicos por los miembros de las dos poblaciones es la misma. Si los datos nos permiten rechazar H_0, entonces podríamos aceptar H_1, ya que los datos apoyan la hipótesis de investigación y su teoría subyacente.

La naturaleza de la hipótesis de investigación determina cómo debe establecerse H_1. Si la hipótesis de investigación simplemente establece que dos grupos diferirán respecto a sus medias, entonces H_1 es que $\mu_1 \neq \mu_2$. Pero si la teoría predice la *dirección* de la diferencia, es decir, que un grupo especificado tendrá una media mayor que el otro, entonces H_1 pudiera ser que $\mu_1 > \mu_2$ o que $\mu_1 < \mu_2$, esto es, la media del grupo 1 es mayor que o menor que la media del grupo 2, respectivamente.

Se puede notar que, aunque podemos decir que los datos apoyan H_1 y nos gustaría aceptar esa hipótesis, no podemos afirmar que H_1 sea verdadera. Como veremos en la sección dedicada al nivel de significación y el tamaño de la muestra, nuestros datos sólo nos permiten hacer juicios probabilísticos acerca de las hipótesis. Aunque podemos *decir* que estamos rechazando una hipótesis y aceptando su alterna, no podemos decir que la hipótesis alterna sea cierta.

LA ELECCIÓN DE LA PRUEBA ESTADÍSTICA

El campo de la estadística ha tenido un desarrollo tan grande, que ahora tenemos, para casi cualquier diseño de investigación, pruebas estadísticas alternativas válidas que podemos utilizar para decidir acerca de una hipótesis. Teniendo pruebas alternativas válidas, necesitamos algunas bases racionales para elegir entre ellas. Ya que este libro se centra en la estadística no paramétrica, la decisión entre procedimientos estadísticos paramétricos y no paramétricos es uno de sus temas centrales. El examen de este aspecto se reserva para un capítulo separado. En el capítulo 2 se presenta un análisis de las bases para elegir entre varias pruebas aplicables a un diseño de investigación determinado. Aunque aquí no tengamos una exposición detallada, es importante recordar que la elección de pruebas estadísticas es el segundo paso del procedimiento.

EL NIVEL DE SIGNIFICACIÓN Y EL TAMAÑO DE LA MUESTRA

Cuando se han establecido la hipótesis nula y la hipótesis alterna, y cuando se ha seleccionado la prueba estadística adecuada, el siguiente paso consiste en especificar un nivel de significación (α) y seleccionar un tamaño de muestra (N).

Brevemente, éste es nuestro procedimiento de toma de decisiones: antes de recabar los datos, especificamos un conjunto de todas las muestras posibles que pudieran ocurrir si H_0 fuera cierta. De estas muestras, especificamos un subconjunto de posibles muestras que sean tan inconsistentes con H_0 (o tan extremas), que la probabilidad de que la muestra observada esté realmente entre ellas, cuando H_0 sea cierta, sea muy pequeña. Entonces, si en nuestra investigación realmente observamos una muestra que esté incluida en ese subconjunto, rechazamos H_0.

En otras palabras, nuestro procedimiento es rechazar H_0 en favor de H_1, si una prueba estadística proporciona un valor cuya probabilidad de ocurrencia asociada de acuerdo con H_0 sea igual o menor que alguna probabilidad pequeña, generalmente denotada por α. A esa probabilidad se le conoce como el *nivel de significación*. Los valores comunes de α son 0.05 y 0.01.[1] Reiteramos: si la probabilidad asociada con la ocurrencia de acuerdo con H_0 (esto es, cuando la hipótesis nula es

[1] Con base en la exposición acerca de los niveles de significación presentada en este libro, el lector no podría inferir que creemos en una aproximación rígida e inflexible al colocar los niveles de significación. Más que esto, es por razones heurísticas que se destacan dichos niveles de significación; tal exposición parece ser el mejor método de clarificar el papel que la información contenida en la distribución muestral desempeña en el proceso de toma de decisiones.

cierta) de un valor particular proporcionado por una prueba estadística (y valores más extremos) es igual o menor que α, rechazamos H_0 en favor de H_1, la declaración operacional de la hipótesis de investigación. El propósito de colocar un nivel de significancia es definir un evento raro de acuerdo con H_0 cuando la hipótesis nula sea verdadera. Así, si H_0 fuera cierta y si el resultado de una prueba estadística en un conjunto de datos observados tuviera una probabilidad menor o igual a α, es la ocurrencia de un evento raro lo que nos conduciría, sobre una base probabilística, a rechazar H_0.

Entonces, se puede ver que α proporciona la probabilidad de rechazar equivocada o falsamente a H_0. El error de rechazar H_0 equivocadamente se conoce como *error de tipo I*, el cual se examinará posteriormente en este capítulo.

Ya que la probabilidad de α determina que H_0 sea o no rechazada, el requerimiento de objetividad exige que α sea especificada antes de que se recaben los datos. El nivel en el cual el investigador elige colocar a α puede ser determinado por una estimación de la importancia o de la significación práctica del resultado que será obtenido. En el estudio de un posible efecto terapéutico de cirugía cerebral, por ejemplo, el investigador bien puede elegir un nivel de significación bastante riguroso, debido a que las consecuencias de rechazar de manera inadecuada la hipótesis nula (y, por tanto, abogar o recomendar injustificadamente una técnica clínica drástica) son ciertamente grandes. Al presentar los resultados, el investigador debe indicar el nivel de probabilidad real asociado con los resultados obtenidos, de modo que el lector pueda usar su propio juicio para decidir si la hipótesis nula debe o no ser rechazada. Un investigador puede decidir trabajar en el nivel 0.05, pero un lector tal vez se niegue a aceptar cualquier resultado a menos que sea significativo en el nivel 0.01, 0.005 o 0.001, mientras que otro lector puede estar interesado en cualquier resultado que alcance, por ejemplo, el nivel 0.08 o 0.10. Estas diferencias a menudo reflejan las apreciaciones subjetivas percibidas de la aplicación de los resultados por diferentes individuos. Siempre que sea posible, el investigador debe proporcionar a los lectores la información del nivel de probabilidad realmente asociado con los datos.

Existen dos tipos de errores que se pueden cometer al tomar una decisión acerca de H_0. El primero, el *error de tipo I*, se refiere a rechazar la hipótesis H_0 cuando de hecho es verdadera. El segundo, el *error de tipo II*, se refiere a aceptar la hipótesis nula H_0 cuando de hecho es falsa.

La probabilidad de cometer el error de tipo I se denota por α. Mientras más grande sea la probabilidad α, más probable será que H_0 sea rechazada equivocadamente, esto es, existe mayor probabilidad de que se cometa el error de tipo I. El error de tipo II generalmente se denota por β. α y β se usan para indicar tanto el tipo de error como la probabilidad de cometerlo. Esto es:

$$P[\text{error de tipo I}] = \alpha$$

$$P[\text{error de tipo II}] = \beta$$

Idealmente, los valores particulares de α y β deben ser elegidos por el investigador antes de empezar el estudio. Estos valores deben determinar el tamaño de la muestra N que será necesario utilizar para usar la prueba estadística que se ha elegido.

Sin embargo, en la práctica es más común que α y N se especifiquen con anticipación. Una vez que α y N han sido especificadas, se determina β. Ya que existe una relación inversa entre la probabilidad de cometer los dos tipos de errores, para cualquier N dada, a decrementos en α corresponderán incrementos en β. Por otro lado, si deseamos reducir la posibilidad de ambos tipos de errores, debemos incrementar el tamaño de la muestra N.

Debe quedar claro que en cualquier inferencia estadística existe el peligro de cometer uno de los dos tipos de errores y, por tanto, el investigador debe comprometerse para tratar de mejorar el equilibrio entre las probabilidades de cometer ambos errores. Las diversas pruebas estadísticas ofrecen la posibilidad de diferentes balances entre estos factores. Lograr tal balance es importante para la potencia de una prueba estadística.

La *potencia de una prueba* se define como la probabilidad de rechazar H_0 cuando de hecho es *falsa*. Esto es,

$$\text{Potencia} = 1 - P[\text{error de tipo II}] = 1 - \beta$$

Las curvas en la figura 1.1 muestran que, para una prueba en particular, la probabilidad de cometer un error de tipo II β disminuye al incrementar el tamaño de la muestra N y, por tanto, se incrementa la potencia de la prueba. Se puede considerar que $1 - \beta$ es "la fuerza de la evidencia". Así mismo, la potencia de una prueba paramétrica se incrementa con la diferencia entre el parámetro "real" de la población, por ejemplo, μ, y el valor especificado por H_0, por ejemplo, μ_0. En la figura 1.1 se ilustra el incremento en la potencia de una prueba de la media de dos colas con incrementos en el tamaño de las muestras $N = 4, 10, 20, 50, 100$. Estas muestras se extraen de poblaciones que tienen distribuciones normales con varianza σ^2.[2] Cuando la hipótesis nula es verdadera, la media es μ_0, esto es, $\mu = \mu_0$.

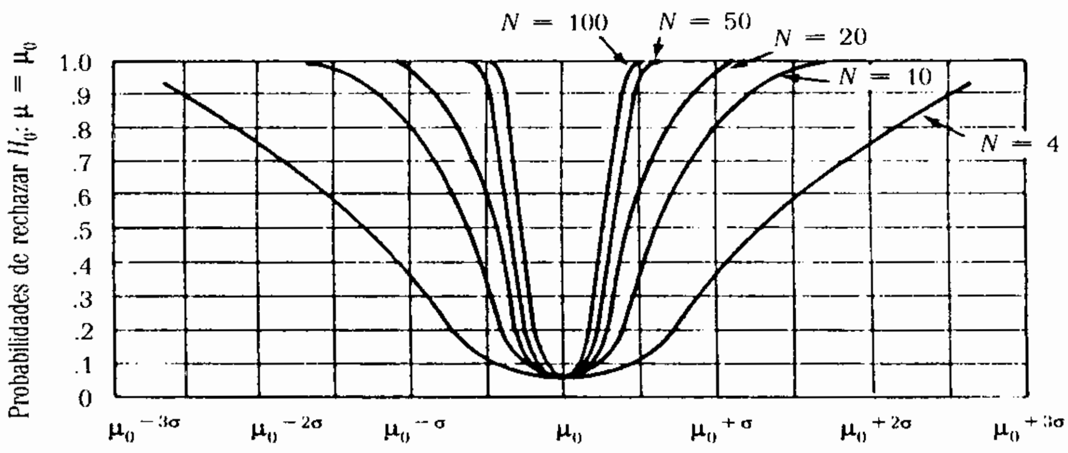


Figura 1.1. Curvas de potencia de la prueba bidireccionales con $\alpha = 0.05$, con diferentes tamaños de muestra.

[2] La distribución normal es la distribución de una variable aleatoria x que tiene la siguiente forma:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-1/2[(x-\mu)/\sigma]^2}$$

donde μ es la media y σ es la desviación estándar de la distribución. Ésta es la distribución familiar de "forma de campana".

En la figura 1.1 se representan las curvas de potencia para pruebas con $\alpha = 0.05$. Esto es, las curvas se trazaron suponiendo que cuando H_0 es verdadera —cuando la media verdadera es μ_0—, la probabilidad de rechazar H_0 es igual a 0.05.

En esta exposición es importante que el lector comprenda cabalmente los siguientes cinco puntos, que resumen lo que hemos dicho acerca de la selección del nivel de significación y del tamaño de la muestra:

1. El nivel de significación α de una prueba es la probabilidad de que, cuando la hipótesis H_0 es verdadera, una prueba estadística proporcionará un valor que conducirá al rechazo de H_0; es decir, el nivel de significación indica la probabilidad de cometer un error de tipo I.
2. β es la probabilidad de que una prueba estadística proporcione un valor según el cual la hipótesis nula pudiera ser aceptada cuando de hecho es falsa: es decir, β es la probabilidad de cometer un error de tipo II.
3. La potencia de una prueba. $1 - \beta$, es la probabilidad de rechazar la hipótesis nula cuando es falsa (y, por tanto, debería ser rechazada).
4. La potencia es una función de la prueba estadística elegida.[3]
5. Generalmente, la potencia de una prueba estadística se incrementa al incrementarse el tamaño de la muestra.

LA DISTRIBUCIÓN MUESTRAL

Después de que un investigador ha elegido usar cierta prueba estadística con un conjunto de datos, se debe determinar la distribución muestral del estadístico de la prueba.

La distribución muestral es una distribución teórica. Es ésta la distribución que podríamos obtener si tomáramos *todas las posibles* muestras del mismo tamaño de la misma población, extraídas cada una de ellas aleatoriamente. En otras palabras: la distribución muestral es la distribución de todos los posibles valores que algún estadístico (por ejemplo, la media de la muestra $\overline{X}$) puede tomar, siendo H_0 verdadera, cuando ese estadístico es computado de muchas muestras de igual tamaño extraídas de la misma población.

La distribución muestral nula de algún estadístico consiste en las probabilidades bajo H_0 asociadas con varios valores numéricos posibles del estadístico. La *probabilidad asociada* con la ocurrencia de un valor particular del estadístico cuando H_0 es verdadera, *no* es la probabilidad exacta de ese valor. En lugar de esto, "la probabilidad asociada con la ocurrencia según H_0" es usada para referirse a la probabilidad de un valor particular *más* las probabilidades de todos los valores posibles que son más extremos o más inconsistentes con H_0. Esto es, la "probabilidad asociada" o "la probabilidad asociada con la ocurrencia bajo H_0", es la probabilidad de ocurrencia según H_0 de un valor "tan extremo o más extremo que" el valor particular del estadístico de la prueba. En este libro tendremos frecuentes ocasiones de

[3] La potencia también está relacionada con la naturaleza de H_1. Si H_1 tiene dirección, se usa una prueba unidireccional, que es más potente que una prueba bidireccional. Las pruebas uni y bidireccional se describen en la sección denominada *La región de rechazo*. La potencia está también relacionada con el tamaño de la muestra N, la varianza σ^2, el nivel de significación α y otras variables, dependiendo de la prueba que se esté usando.

usar las frases anteriores, y en cada caso, cada una de ellas tiene el significado anteriormente dado.

Supongamos que nos interesa en la probabilidad de que cuando sean lanzadas al aire simultáneamente tres monedas "normales", caigan caras. La distribución muestral del número de caras podría trazarse a partir de la lista de todos los posibles resultados de lanzar al aire tres monedas normales, los cuales se presentan en la tabla 1.1. El número total de eventos posibles (combinaciones posibles de caras y cruces) es ocho; sólo uno de ellos es el evento en el que estamos interesados: la ocurrencia simultánea de tres caras. Así, la probabilidad de ocurrencia bajo H_0 de las tres caras en el lanzamiento de tres monedas es 1/8. Aquí H_0 es la aseveración de que las monedas son "normales", lo que significa que para cada moneda la probabilidad de que caiga una cara es igual a la probabilidad de que caiga una cruz.

Tabla 1.1. Resultados posibles en el lanzamiento de tres monedas.

	Monedas		
Resultados	1	2	3
1	C	C	C
2	C	C	X
3	C	X	C
4	C	X	X
5	X	C	C
6	X	C	X
7	X	X	C
8	X	X	X

La distribución muestral de todos los posibles eventos proporciona la probabilidad de ocurrencia del evento en el que estamos interesados, cuando H_0 es verdadera.

Es obvio que para nosotros sería esencialmente imposible usar este método de imaginar todos los posibles resultados con el propósito de enumerar la distribución muestral de muestras, aun cuando las poblaciones no fueran muy grandes. Si éste es el caso, dependemos de la autoridad de los teoremas matemáticos establecidos. Estos teoremas invariablemente incluyen suposiciones, y al aplicar los teoremas debemos tenerlas presentes. Por lo general, tales suposiciones conciernen a la distribución de la población, al tamaño de la muestra o a ambos. Un ejemplo de tal teorema es el *teorema del límite central.*

Cuando una variable está normalmente distribuida, su distribución está por completo caracterizada por su media y su desviación estándar. Si éste es el caso, sabemos, con base en el análisis de la distribución, que la probabilidad de que un valor observado de la variable difiera de la media de la población en más de 1.96

desviaciones estándar, es menor que 0.05. (Las probabilidades asociadas con cualquier diferencia en las desviaciones estándar de la media de una variable normalmente distribuida, se proporcionan en la tabla A del Apéndice I.)

Supóngase que queremos conocer, antes de que la muestra sea extraída, la probabilidad asociada con la ocurrencia de un valor particular de $\overline{X}$ (la media de la muestra), esto es, la probabilidad según H_0 de la ocurrencia de un valor al menos tan grande como un valor particular de $\overline{X}$ cuando la muestra se extrae aleatoriamente de alguna población, la media μ y la desviación estándar σ de lo que conocemos. Una versión del teorema del límite central establece que:

> Si una variable se distribuye con media $= \mu$ y desviación estándar $= \sigma$, y se extraen medias aleatorias de tamaño N, entonces las medias de estas muestras, las $\overline{X}$, estarán en forma aproximada normalmente distribuidas con media μ y desviación estándar $\sigma/\sqrt{N}$ cuando N sea grande.[4]

En otras palabras, sabemos que la distribución muestral de $\overline{X}$ tiene una media igual a la media poblacional μ, una desviación estándar igual a la desviación estándar de la población dividida por la raíz cuadrada del tamaño de la muestra, esto es, $\sigma_{\bar{x}} = \sigma/\sqrt{N}$; y si N es lo suficientemente grande, aquélla es aproximadamente normal.

Por ejemplo, supongamos que sabemos que en una población de estudiantes, algún atributo psicológico, al ser medido por alguna prueba, está distribuido con $\mu = 100$ y $\sigma = 16$. Ahora queremos saber la probabilidad de extraer una muestra aleatoria de $N = 64$ casos de esta población y encontrar que la puntuación media en esa muestra, $\overline{X}$, sea tan grande como 104. La distribución muestral de las $\overline{X}$ de todas las muestras posibles de tamaño 64 tendrán una media igual a 100 ($\mu = 100$) y una desviación estándar igual a $\sigma/\sqrt{N} = 16/\sqrt{64} = 2$, y el teorema del límite central nos dice que la distribución de $\overline{X}$ será aproximadamente normal al incrementarse N. (Si la variable X tiene una distribución normal al empezar, $\overline{X}$ podría tener una distribución normal independientemente del tamaño de la muestra.) Podemos ver que 104 difiere de 100 por dos errores estándar.[5] La tabla A del Apéndice I revela que la probabilidad asociada con la ocurrencia según H_0 de un valor tan grande como el valor observado de $\overline{X}$, esto es, de una $\overline{X}$ que esté al menos dos errores estándar por encima de la media ($z \geq 2.0$), es $p < 0.023$. Esta computación puede representarse en la forma siguiente:

$$
\begin{aligned}
z &= \frac{\overline{X} - \mu}{\sigma/\sqrt{N}} \\
&= \frac{104 - 100}{16/\sqrt{64}} \\
&= 2
\end{aligned}
$$

[4] Aunque decimos que la distribución se vuelve *aproximadamente* normal al incrementarse N, el teorema del límite central establece que si $N \rightarrow \infty$, la distribución se vuelve normal. Sin embargo, ya que todas las muestras son finitas, el término *aproximada* es adecuado.

[5] La desviación estándar de una distribución muestral de la media de la muestra frecuentemente se denota como el *error estándar* de la distribución.

A partir de esta exposición y de este ejemplo debe quedar claro que al conocer la distribución muestral de algún estadístico, podemos hacer declaraciones acerca de la probabilidad de ocurrencia de ciertos valores numéricos de un estadístico. En las siguientes secciones se mostrará cómo usar tales declaraciones de probabilidad al tomar una decisión acerca de H_0.

LA REGIÓN DE RECHAZO

La región de rechazo es una región de la distribución muestral nula. La distribución muestral incluye *todos* los valores posibles que un estadístico de prueba puede adoptar. La región de rechazo consiste en un subconjunto de estos valores posibles, y se elige tal que la probabilidad de ocurrencia de un estadístico de prueba según H_0, tenga un valor que en ese subconjunto sea α. En otras palabras, la región de rechazo consiste en un conjunto de valores posibles que son tan extremos que cuando H_0 es verdadera, la probabilidad es muy pequeña (es decir, igual a α), de manera que la muestra que observamos realmente proporcione un valor que esté entre esos valores. La probabilidad *asociada* con cualquier valor en la región de rechazo es igual o menor que α.

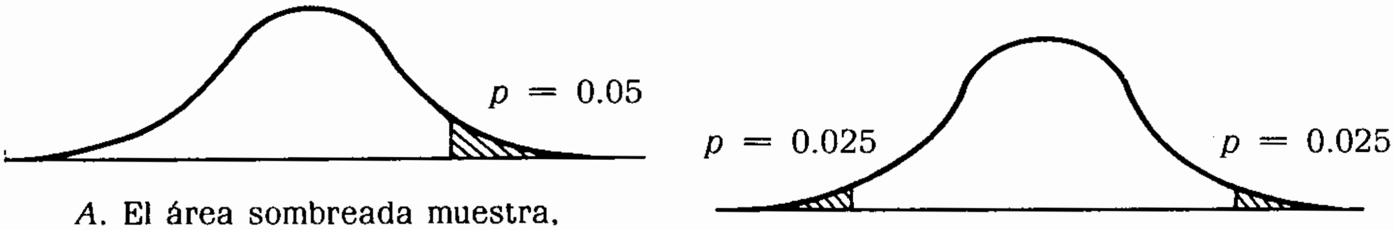
La naturaleza de la región de rechazo es afectada por la forma de la hipótesis alterna H_1. Si H_1 también indica la dirección predicha de la diferencia, entonces se usa una prueba unidireccional. Si H_1 no indica la dirección de la diferencia predicha, se usa una prueba bidireccional. Las pruebas uni y bidireccional difieren en la localización (pero no en el tamaño) de la región de rechazo; es decir, en una prueba unidireccional la región de rechazo está enteramente en un extremo (o cola) de la distribución muestral. En una prueba bidireccional, la región de rechazo se localiza en ambos extremos (o colas) de la distribución muestral.

Como ejemplo, supongamos que un investigador quiere determinar si un régimen particular de entrenamiento tiene algún efecto sobre la habilidad de recordar nombres de lugares geográficos. La hipótesis nula podría ser que la ejecución de un grupo control que no recibió entrenamiento especial, no difiere de la ejecución de un grupo entrenado. Si el investigador únicamente quiere saber si existe una diferencia, entonces los grandes incrementos o decrementos en la ejecución podrían conducir al rechazo de H_0, se debe usar una prueba bidireccional. Sin embargo, si el investigador estuviera interesado en determinar si el régimen de entrenamiento puede conducir a una mejor ejecución, sólo los grandes incrementos en la ejecución podrían conducir al rechazo de H_0 y se debe usar una prueba unidireccional.

El tamaño de la región de rechazo es expresado por α, el nivel de significancia. Si $\alpha = 0.05$, entonces el tamaño de la región de rechazo comprende el 5 % del área total incluida bajo la "curva" de la distribución muestral. En la figura 1.2 se ilustran las regiones de rechazo, unidireccional y bidireccional, para $\alpha = 0.05$. Nótese que estas dos regiones difieren en la localización, pero no en el tamaño total.

LA DECISIÓN

Si la prueba estadística proporciona un valor que cae en la región de rechazo, rechazamos H_0.



A. El área sombreada muestra, para un contraste unidireccional, la región de rechazo cuando $\alpha = 0.05$

B. El área sombreada muestra, para un contraste bidireccional, la región de rechazo cuando $\alpha = 0.05$

Figura 1.2. Regiones de rechazo para pruebas unidireccionales y bidireccionales.

El razonamiento que subyace a este proceso de decisión es muy simple. Si la probabilidad asociada con la ocurrencia de un valor particular en la distribución muestral, según la hipótesis nula, es muy pequeña, podemos explicar la ocurrencia real de ese valor en una de dos formas: 1. diciendo que la hipótesis nula es falsa, o 2. diciendo que un evento raro e improbable ha ocurrido. En el proceso de decisión elegimos la primera de estas explicaciones. Naturalmente, en forma ocasional la segunda explicación puede ser la correcta. De hecho, la probabilidad asociada con la segunda explicación está dada por α, ya que rechazar H_0 cuando de hecho es verdadera, es un error de tipo I.

Cuando la probabilidad asociada con un valor observado de una prueba estadística es igual o menor que el valor de α previamente determinado, concluimos que H_0 es falsa. Tal valor observado es llamado *significativo*. La hipótesis sometida a prueba H_0 es rechazada siempre que un resultado significativo ocurre. Un valor significativo es aquel que se encuentra en la región de rechazo y cuya probabilidad asociada de ocurrencia según H_0 (como es mostrada por la distribución muestral) es igual o menor que α.

EJEMPLO ILUSTRATIVO

En este libro se proporcionarán numerosos ejemplos de toma de decisiones en el examen de las variadas pruebas estadísticas no paramétricas. A continuación presentamos sólo un ejemplo de cómo se toma una decisión estadística, con el propósito de ilustrar los puntos enunciados en este capítulo.

Ejemplo. Supóngase que sospechamos que una moneda particular está sesgada porque al ser lanzada, con frecuencia cae "cara". Para probar esta sospecha (que llamaremos nuestra *hipótesis de investigación*), decidimos lanzar la moneda 12 veces y observar la frecuencia de ocurrencia de "caras".

i. *Hipótesis nula.* H_0: $P[C] = P[X] = 1/2$. Para esta moneda no existe diferencia entre la probabilidad de ocurrencia de una cara, esto es, $P[C]$, y la probabilidad de una cruz, esto es, $P[X]$. Dicho de otro modo, la moneda es "normal". La hipótesis alterna H_1: $P[C] > 1/2$, es una representación de la hipótesis de investigación.

ii. *Prueba estadística.* La prueba estadística adecuada para probar nuestra hipótesis es la prueba binomial, la cual está basada en la distribución binomial. (El número de caras observadas cuando se lanza una moneda al aire tiene una distribución binomial. Sabemos que si el número de caras observadas es muy grande, rechazaremos H_0. Sin embargo, necesitamos conocer las probabilidades de diferentes resultados posibles para el experimento. La distribución binomial nos proporciona estas probabilidades. Esta distribución y la prueba asociada se examinan en el capítulo 3.)

iii. *Nivel de significación.* Anticipadamente decidimos usar $\alpha = 0.01$ como nuestro nivel de significación. $N = 12$ es el número de lanzamientos independientes de la moneda.

iv. *Distribución muestral.* La distribución muestral que proporciona la probabilidad de obtener caras #C y cruces $N - \#C$ según la hipótesis nula (la hipótesis de que la moneda es de hecho normal), es la función de distribución binomial:

$$P(\#C) = \frac{N!}{(\#C)!\ (N - \#C)!}\left(\frac{1}{2}\right)^N \qquad \#C = 0, 1, 2, \ldots, N$$

En la tabla 1.2 se presenta la distribución muestral de #C, el número de caras cuando una moneda normal es lanzada al aire 12 veces.[6] Esta distribución muestra que el resultado más probable cuando se lanza una moneda 12 veces, es seis caras y seis cruces. El obtener siete caras y cinco cruces es poco menos probable, pero completamente factible. Sin embargo, la obtención de 12 caras en 12 lanzamientos es ciertamente muy improbable. La ocurrencia de cero caras (12 cruces) es igualmente improbable.

Tabla 1.2. Distribución muestral de # C (número de caras), para dos muestras de tamaño $N = 12$.

Número de caras	*Distribución muestral**	*Probabilidad*
12	1	0.00024
11	12	0.0029
10	66	0.0161
9	220	0.0537
8	495	0.1208
7	792	0.1936
6	924	0.2256
5	792	0.1936
4	495	0.1208
3	220	0.0537
2	66	0.0161
1	12	0.0029
0	1	0.00024
	4 096	1.000

* Frecuencia de ocurrencia esperada de las 4 096 muestras posibles (12^{12}) al realizar 12 lanzamientos de una moneda.

[6] Los detalles y la racionalización de la distribución binomial se examinan en detalle en la sección dedicada a la prueba binomial (cap. 3). Para el ejemplo aquí presentado es necesario entender que la distribución muestral del # C puede determinarse analíticamente.

v. *Región de rechazo.* Ya que H_1: $p > 1/2$ especifica una dirección de diferencia, se usará una prueba de una cola y la región de rechazo estará enteramente en un extremo de la distribución muestral, es decir, cuando el número de caras es grande. La región de rechazo consiste en todos los valores de #C que son tan grandes que la probabilidad asociada con su ocurrencia según H_0 es igual o menor que $\alpha = 0.01$.
La probabilidad de obtener 12 caras es $1/4096 = 0.00024$. Ya que $p = 0.00024$ es menor que $\alpha = 0.01$, la ocurrencia de 12 caras estará claramente en la región de rechazo.
La probabilidad de obtener ya sea 12 o 11 caras es

$$1/4096 + 12/4096 = 13/4096 = 0.003$$

Ya que $p = 0.003$ es menor que $\alpha = 0.01$, la ocurrencia de 11 caras también estaría en la región de rechazo.
La probabilidad de obtener 10 caras (o un valor más extremo) es

$$1/4096 + 12/4096 + 66/4096 = 79/4096 = 0.019$$

Ya que $p = 0.019$ es mayor que $\alpha = 0.01$, la ocurrencia de 10 caras no estaría en la región de rechazo.[7] Vale decir, si caen 10 o menos caras en nuestra muestra de 12 lanzamientos, no podemos rechazar H_0 en el nivel de significancia $\alpha = 0.01$.

vi. *Decisión.* Supongamos que en la muestra de 12 lanzamientos obtenemos 11 caras. La probabilidad asociada con una ocurrencia tan extrema como ésta es $p = 0.003$. Ya que tal probabilidad es más pequeña que nuestro nivel de significación fijado previamente ($\alpha = 0.01$), nuestra decisión sería rechazar H_0 en favor de H_1. Podríamos concluir que la moneda está sesgada para los resultados de "cara".

En este capítulo hemos examinado el procedimiento de decidir si una hipótesis particular, definida operacionalmente, pudiera ser aceptada o rechazada en términos de la información proporcionada por los datos obtenidos en la investigación. En el capítulo siguiente se complementa la exposición general al profundizar en la cuestión de cómo elegir la prueba estadística más apropiada para usar con nuestros datos de investigación (esta elección es el paso 2 del procedimiento que ya hemos delineado). El análisis del capítulo 2 esclarece las condiciones en las que las pruebas paramétricas son óptimas e indica las condiciones en las cuales las pruebas no paramétricas son más adecuadas.

Referencias bibliográficas

El lector que desea tener una mejor comprensión de los temas resumidos en el sencillo esquema de este capítulo, puede consultar libros de estadística de las ciencias de la conducta y sociales. Especialmente dignos de atención son los libros de Bailey (1971) y Hays (1981).

[7] Debido a que las distribuciones muestrales para muchos estadísticos no paramétricos son discretas, podría no ser posible seleccionar la región de rechazo tal que α sea exactamente igual a un valor predeterminado. Por tanto, el punto de corte que divide la distribución podría ser elegido de tal manera que la probabilidad asociada con la región de rechazo sea tan grande como sea posible, *pero menor* que el nivel de significación elegido α. Estos resultados en una prueba conservadora proporcionan una prueba simple para usar en una prueba de hipótesis.

Elección de la prueba estadística adecuada

Cuando se dispone de pruebas estadísticas alternativas y válidas para una hipótesis de investigación en particular, es necesario emplear algunas racionalizaciones para elegir entre ellas. En el capítulo 1 presentamos un criterio para elegir entre pruebas estadísticas alternativas válidas: el criterio de potencia. En este capítulo presentaremos otros criterios.

El lector recordará que la *potencia* de un análisis estadístico es, en parte, una función de la prueba estadística que se emplee para el análisis. Una prueba estadística es válida si la probabilidad de rechazar H_0 cuando H_0 es verdadera, es igual al valor elegido para α; es una prueba potente si tiene gran probabilidad de rechazar H_0 cuando H_0 es falsa. Supongamos que encontramos dos pruebas estadísticas, A y B, las cuales tienen la misma probabilidad de rechazar H_0 cuando ésta es verdadera. Esto significa que ambas pruebas son igualmente válidas. Podría parecer que nosotros simplemente deberíamos seleccionar aquella que tiene la probabilidad más grande de rechazar H_0 cuando sea falsa.

Sin embargo, existen otras consideraciones además de la potencia, que determinan la elección de la prueba estadística. En esta elección debemos considerar la manera en que se obtuvo la muestra de puntuaciones o datos, la naturaleza de la población de la cual fue extraída la muestra, las hipótesis particulares que deseamos probar y el tipo de medición o escala que se empleó en las definiciones operacionales de las variable implicada; esto es, en las puntuaciones. Todas estas cuestiones determinan qué prueba estadística es óptima o más apropiada para analizar un conjunto particular de datos de investigación.

EL MODELO ESTADÍSTICO

Cuando hemos identificado la naturaleza de la población y la forma del muestreo, hemos establecido un modelo estadístico. Para cada prueba estadística se asocia un modelo y un requisito de medida. La prueba es válida en ciertas condiciones, y el modelo y el requisito de medida especifican esas condiciones. Algunas veces somos capaces de probar si se encuentran las condiciones de un modelo estadístico particular, pero la mayoría de las veces *suponemos* que se encuentran. Debemos examinar la situación y determinar si es razonable o no suponer que el modelo es correcto. Todas las decisiones tomadas por el uso de cualquier prueba estadística deben llevar consigo esta fórmula: "Si el modelo usado fue correcto y si los requisitos de medida fueron satisfechos, entonces. . ."

Es obvio que mientras más pobres o débiles sean las suposiciones que definen un modelo particular, necesitaremos simplificar más la decisión alcanzada por la prueba estadística asociada con ese modelo; es decir, mientras más pobres o débiles sean las suposiciones, más generales serán las conclusiones.

Sin embargo, las pruebas más potentes son aquellas que tienen las suposiciones más fuertes o extensas. Las pruebas paramétricas, por ejemplo la prueba t o la prueba F, tienen una variedad de fuertes suposiciones que subyacen a su uso. Si esas suposiciones son válidas, las pruebas basadas en las mismas son las que tienen mayor probabilidad de rechazar H_0 cuando H_0 es falsa; esto es, cuando los datos de investigación pueden ser analizados de manera adecuada mediante una prueba paramétrica, ésta será más potente que cualquier otra. Sin embargo, nótese que los requerimientos de los datos de investigación deben ser adecuados para la prueba. ¿Qué constituye ser adecuado? ¿Cuáles son las condiciones asociadas con el modelo estadístico y el requisito de medida, por ejemplo, para la prueba t? Las condiciones que se deben satisfacer para hacer la prueba t la más potente y aceptar con bastante confianza las conclusiones de probabilidad obtenidas por el uso de la prueba t, son al menos las siguientes:

1. Las observaciones deben ser independientes, es decir, la selección de un caso de la población para su inclusión en la muestra, no debe sesgar las oportunidades de cualquier otro caso para su inclusión, y la puntuación que se asigna a cualquier caso no debe sesgar la puntuación que es asignada a cualquier otro caso.
2. Las observaciones deben ser derivadas de poblaciones normalmente distribuidas.
3. En el caso de análisis concerniente a dos grupos, las poblaciones deben tener la misma varianza (o, en casos especiales, deben tener una razón conocida de varianzas).
4. Las variables deben haber sido medidas por lo *menos* en una escala de intervalo, de modo que sea posible *interpretar* los resultados.

Todas las condiciones anteriores (incluida la condición 4, que establece el requisito de medición) son elementos del *modelo estadístico paramétrico* asociado con la distribución normal. Con la posible excepción de la suposición de varianzas iguales, estas condiciones de ordinario no son probadas en el curso de la ejecución de un análisis estadístico. En lugar de esto, son presunciones aceptadas y su certeza o falsedad determinan la exactitud y significatividad de la probabilidad establecida mediante la prueba paramétrica. Como puede verse, las pruebas paramétricas prueban hipótesis acerca de *parámetros* específicos, tales como la media. Se supone que las hipótesis acerca de tales parámetros son idénticas a nuestras hipótesis de investigación.

Cuando se tienen razones para creer que estas condiciones se encuentran en los datos que se están analizando, entonces ciertamente es posible elegir una prueba estadística paramétrica, tal como t o F, para analizar esos datos. Tal elección es adecuada debido a que la prueba paramétrica es una prueba válida y más potente.

Pero, ¿qué ocurre si estas condiciones no se encuentran? ¿Qué sucede cuando la población *no* está normalmente distribuida? ¿Qué pasa cuando la medición *no* es tan fuerte como una escala de intervalo? Si existen múltiples medidas o grupos, ¿qué ocurre cuando las poblaciones *no* tienen igual varianza?

Cuando no se encuentran las suposiciones que constituyen el modelo estadístico de una prueba, entonces ésta no puede ser válida; esto es, un estadístico de prueba puede caer en la región de rechazo con una probabilidad más grande que α. Es aún difícil estimar la extensión en la cual un juicio de probabilidad varía debido a la aplicación inadecuada de la prueba. Aunque se ha reunido evidencia empírica para mostrar que encontrar ligeras desviaciones en las suposiciones que subyacen a las pruebas paramétricas puede no tener efectos radicales en los niveles de probabilidad obtenidos, no existe un acuerdo general de lo que constituye una desviación "ligera". Más aún, desviaciones ligeras en más de un factor o suposición pueden tener consecuencias mayores.

EFICACIA

Ya hemos señalado que mientras más pobres o más débiles sean las suposiciones que constituyen un modelo particular, menos potentes serán las pruebas válidas disponibles. Esta aseveración es generalmente cierta para cualquier tamaño de muestra. Pero puede no ser cierta cuando se comparan dos pruebas estadísticas que se aplican a dos muestras de tamaño diferente; esto es, si $N = 30$ en ambos casos, la prueba A puede ser más potente que la prueba B. Pero la misma prueba B puede ser más potente con $N = 30$ que la prueba A con $N = 20$. Recuérdese que la potencia de una prueba se incrementa al incrementarse N. Así, podemos usar una prueba menos potente con un tamaño de muestra más grande. En otras palabras, podemos evitar el dilema de tener que elegir entre potencia y generalidad al seleccionar una prueba estadística que tenga una amplia generalidad, y entonces incrementar su potencia como la prueba disponible más potente, al aumentar el tamaño de la muestra.

El concepto de *potencia-eficacia* se relaciona con el incremento en el tamaño de la muestra que es necesario hacer para lograr que la prueba B sea tan potente como la prueba A cuando el nivel de significación y el tamaño de la muestra de la prueba A se mantienen también constantes. Si la prueba A es la prueba conocida más potente de su tipo (cuando se usa con datos que cumplen sus condiciones), y si la prueba B es otra prueba para el mismo diseño de investigación que es justamente tan poderosa con N_B casos que la prueba A con N_A casos, entonces:

$$\text{Potencia-eficacia de la prueba } B = \frac{100\ N_A}{N_B}\ \%$$

Por ejemplo, si la prueba B requiere una muestra de $N = 25$ casos para tener la misma potencia que la prueba A con $N = 20$ casos cuando el nivel de significación es α, entonces la prueba B tiene una potencia-eficacia de (100) (20/25) = 80 %. Una potencia-eficacia de 80 % significa que con el propósito de igualar la potencia de la prueba A y la prueba B (cuando se encuentran todas las condiciones de ambas pruebas y cuando la prueba A es la más potente), necesitamos tener 10 casos para la prueba B por cada ocho casos para la prueba A.

Los estudiosos de la estadística también comparan modelos al calcular la *eficacia relativa asintótica* de un estadístico. Como la potencia-eficacia, la eficacia relativa asintótica es un modo de determinar el tamaño de muestra necesario para que la prueba B tenga la misma potencia que la prueba A. Sin embargo, a diferencia de la potencia-eficacia, esta razón es expresada independientemente del tamaño de la muestra de la prueba A. La razón es asintótica ya que es la razón de los tamaños de muestra requeridos para una α fija al incrementarse el tamaño de la muestra de la prueba A hasta el límite ($N_A \rightarrow \infty$). Esto puede expresarse como sigue:

$$\text{Eficacia relativa asintótica de la prueba } B = 100 \lim_{N_A \rightarrow \infty} \frac{N_A}{N_B} \%$$

La eficacia relativa asintótica tiene algunas ventajas sobre la potencia-eficacia. Una de ellas es que el límite por lo general se vuelve independiente de α. Por otra parte, una desventaja de la eficacia relativa asintótica es que el límite está basado en grandes muestras, mientras que muchas de las pruebas de interés en este libro son aplicadas a muestras pequeñas. Afortunadamente para algunas pruebas, la eficacia relativa asintótica se alcanza con muestras ligeramente pequeñas. La potencia-eficacia y la eficacia relativa asintótica son características importantes de las pruebas estadísticas. En cierto sentido, son conceptos complementarios debido a que nos dan información acerca de qué tan bien se comporta una prueba válida con respecto a otra.

En suma, podemos evitar la pérdida de potencia simplemente eligiendo una prueba diferente y escogiendo una muestra más grande. En otras palabras, al elegir otra prueba estadística con menores suposiciones en su modelo y así una generalidad más grande que las pruebas t y F, e incrementando además N, podemos evitar tener que satisfacer las condiciones 2 y 3 dadas en la sección dedicada al modelo estadístico, y aún tener la potencia equivalente para rechazar H_0. Esto es especialmente importante cuando creemos que las suposiciones de un modelo estadístico son inadecuadas. El investigador tiene la responsabilidad de estudiar apropiadamente la situación y hacer sólo suposiciones razonables.

Otras dos condiciones, la 1 y la 4 de la sección ya citada, subyacen al uso y a la interpretación de las pruebas estadísticas basadas en la distribución normal. La condición 1, que las puntuaciones sean obtenidas de manera independiente de la población, es una suposición que subyace a todas las pruebas estadísticas. Pero la condición 4, que se refiere a la fuerza de medida requerida para una *interpretación* adecuada de las pruebas paramétricas basadas en la distribución normal, no es compartida por todas las pruebas estadísticas. Diferentes pruebas suponen distintos tipos de medición. Es indispensable entender el requisito de medida para una interpretación significativa de varias pruebas estadísticas; el lector debe familiarizarse con algunas de las nociones básicas de la teoría de la medida. El siguiente análisis de la medición proporciona una noción general de algunos aspectos importantes de la medición.

MEDICIÓN

Cuando un físico habla acerca de medición, generalmente quiere expresar la asignación de números a observaciones de modo tal que los números sean factibles de análisis por la manipulación u operación de acuerdo con ciertas reglas. El propósito de este análisis por manipulación es revelar nueva información acerca de los objetos que están siendo medidos. En otras palabras, la relación entre las cosas que están siendo observadas y los números asignados a las observaciones es tan directa, que al manipular los números el físico obtiene nueva información acerca de los objetos. Por ejemplo, el científico puede determinar cuánto podría pesar una masa homogénea de material al ser cortada por la mitad, simplemente dividiendo su peso por dos.

El científico social o de la conducta, tomando a la física como modelo, generalmente intenta hacer lo mismo al medir variables sociales o conductuales. Pero al escalar tales datos, el científico con frecuencia pasa por alto un hecho fundamental en la teoría de la medida: se soslaya el hecho de que, con el propósito de ejecutar ciertas operaciones con los números que han sido asignados a las observaciones, la estructura del método de mapear números (asignar puntuaciones) a las observaciones, debe ser *isomórfico* a la estructura de la aritmética que incluye estas operaciones. Si dos sistemas son isomórficos, sus estructuras son las mismas en las relaciones y operaciones que permiten.

Por ejemplo, si un investigador recaba datos, les asigna puntuaciones numéricas y después manipula esas puntuaciones; por ejemplo, sumando y obteniendo la raíz cuadrada (que son operaciones necesarias para encontrar medias y desviaciones estándar), está suponiendo que la estructura de la medición es isomórfica a la estructura numérica conocida como *aritmética*; esto es, él supone que se ha logrado un alto nivel de medición.

La teoría de la medición consiste en un conjunto de teorías separadas o distintas, cada una de las cuales concierne a un distinto *nivel* de medición. Las operaciones interpretables en un conjunto dado de puntuaciones dependen del nivel de medición alcanzado.

Aquí examinaremos cuatro tipos o niveles de medición —nominal, ordinal, de intervalo y de razón— y las implicaciones de cada uno de ellos para la interpretación de las pruebas estadísticas.[1]

La escala nominal o categórica

DEFINICIÓN

La medición en su nivel más débil existe cuando los números u otros símbolos se usan simplemente para clasificar un objeto, una persona o una característica. Cuando se emplean números u otros símbolos para identificar los grupos a los cua-

[1] Existen muchos modos de describir y categorizar la medición. Se han propuesto numerosas escalas, subescalas y generalizaciones de escalas. Los niveles de medición descritos aquí son aquellos que tienen las implicaciones más prácticas para la mayoría de los investigadores.

les pertenecen varios objetos, estos números o símbolos constituyen una escala nominal o categórica. Esta escala se conoce como *escala clasificatoria*.

Ejemplos

El sistema psiquiátrico de diagnóstico constituye una escala nominal. Cuando un diagnosticador identifica a una persona como "esquizofrénica", "paranoica", "maniacodepresiva" o "neurótica", usa un símbolo para representar el tipo de gente al cual pertenece la persona, y de esta manera está empleando una escala nominal o categórica.

Los números de las placas de los automóviles constituyen una escala nominal. Si la asignación de los números de las placas es puramente arbitraria, entonces cada placa es un miembro de una subclase única. Pero, si un cierto número o conjunto de letras en la placa indican la ciudad en la cual está registrado el vehículo, entonces cada subclase en la escala nominal consta de varias entidades: autos registrados en una ciudad en particular. Aquí la asignación de números debe ser tal que el mismo código de números (o código de letras) sea dado a todos los automóviles registrados en la misma ciudad, y diferentes números (o letras) sean asignados a automóviles registrados en diferentes ciudades. Vale decir, el número o la letra en la placa debe indicar claramente a qué conjunto de subclases mutuamente excluyentes pertenece el auto.

Los números en las camisetas de los futbolistas y los números de los policías son otros ejemplos del uso de números en la escala nominal o categórica.

Propiedades formales

Todas las escalas tienen ciertas propiedades formales, las cuales proporcionan definiciones casi exactas de las características de la escala; definiciones más exactas que las que pueden darse en términos verbales. Estas propiedades pueden ser formuladas de manera más abstracta de lo que hemos hecho aquí, por un conjunto de axiomas que especifican las operaciones de la escala y las relaciones entre los objetos que han sido escalados.

En una escala nominal, las operaciones de la escala dividen a una clase dada en un conjunto de subclases mutuamente excluyentes. La única relación implicada es la de *equivalencia*; esto es, los miembros de cualquier subclase deben ser equivalentes en la propiedad que está siendo escalada. Esta relación se simboliza por el signo familiar de "igual" ($=$). La relación de equivalencia es reflexiva, simétrica y transitiva.[2]

Considérese un conjunto de objetos $x_1, x_2, \ldots, x_N$. Supóngase que el objeto x_i tiene algún atributo *verdadero*, $A(x_i)$. Entonces, para cualquier par de atributos en el conjunto

[2] Reflexiva: $x = x$ para todos los valores de x. Simétrica: si $x = y$, entonces $y = x$. Transitiva: si $x = y$ y $y = z$, entonces $x = z$.

$$A(x_i) = A(x_j) \qquad \text{si } x_i \text{ y } x_j \text{ están en la misma clase}$$

$$\text{y} \quad A(x_i) \neq A(x_j) \qquad \text{si } x_i \text{ y } x_j \text{ están en diferentes clases}$$

Una *escala nominal* es un sistema de *clasificación* de los objetos $L(x)$ tal que

$$L(x_i) = L(x_j) \qquad \text{si y sólo si } A(x_i) = A(x_j)$$

$$\text{y} \quad L(x_i) \neq L(x_j) \qquad \text{si y sólo si } A(x_i) \neq A(x_j)$$

Operaciones admisibles

Ya que en una escala nominal la clasificación puede estar igualmente bien representada por cualquier conjunto de símbolos, se dice que la escala nominal es "única hasta una transformación de uno a uno". Los símbolos que designan las variadas subclases en la escala pueden ser intercambiados si esto se hace de manera cabal y consistentemente. Por ejemplo, cuando se emiten nuevas placas para automóviles, el código que previamente pertenecía a una ciudad puede ser intercambiado con el de otra ciudad. La escala nominal podría preservarse si este cambio se ejecutara cabal y consistentemente en la emisión de todas las placas.

Ya que los símbolos que designan los variados grupos de una escala nominal pueden ser intercambiados sin alterar la información esencial en la escala, el único tipo de estadísticos descriptivos admisibles son aquellos que pueden ser incambiables por tal transformación: la moda, la cuenta de frecuencias, etc. En ciertas condiciones, podemos probar hipótesis considerando la distribución de casos entre las categorías, usando pruebas no paramétricas tales como la *ji* cuadrada o una prueba basada en la distribución binomial. Estas pruebas son adecuadas para datos escalados nominalmente debido a que se enfocan sobre la frecuencia en las categorías, es decir, sobre datos enumerativos. En suma, cuando los datos en una escala nominal, podemos rotular las categorías "1", "2", "3",. . ., *en cualquier orden que elijamos*. En una muestra podemos contar el número de "1", el número de "2", etc. (Éstas son cuentas de frecuencia.) Podemos calcular el porcentaje de "1" en la muestra, el porcentaje de "2", etc. (Ésta es la distribución de frecuencia relativa.) Y podemos registrar qué categoría tiene la frecuencia más grande. (Ésta es la moda.) Pero en general, no podemos "sumar" las categorías "1" y "2" para formar la categoría "3", ya que podríamos violar las suposiciones de un sistema de clasificación nominal. En capítulos posteriores estudiaremos diferentes técnicas estadísticas adecuadas para datos categóricos o escalados nominalmente.

La escala ordinal o de rangos

Definición

Puede suceder que los objetos en una categoría de una escala no sean tan sólo diferentes de los objetos en otras categorías de esa escala, sino que también exista algún tipo de *relación* entre ellos. Las relaciones típicas entre las clases son: más alto, más preferido, más difícil, más perturbador, más maduro, etc. Tales relaciones se denotan por medio del símbolo $>$, el cual en general significa "mayor que". En referencia a escalas particulares, $>$ puede ser usado para designar que es *preferido a*, *es más alto que*, *es más difícil que*, etc. Su significado específico depende de la naturaleza de la relación que define la escala.

Dado un grupo de clases de equivalencia (esto es, dado una escala nominal), si la relación $>$ se sostiene entre algunos pero no todos los pares de clases, tenemos una *escala parcialmente ordenada*. Si la relación $>$ se sostiene para todos los pares de clases, de manera que es posible un rango completo ordenado de clases, tenemos una *escala ordinal*.

Ejemplos

El estatus socioeconómico, tal como se concibe comúnmente, constituye una escala ordinal. En el prestigio o la aceptación social, todos los miembros de la clase media superior son mayores que ($>$) todos los miembros de la clase media inferior. A su vez, los miembros de la clase media inferior son mayores que los miembros de la clase baja. La relación $=$ se sostiene en todos los miembros de la misma clase, y la relación $>$ se sostiene entre cualquier par de clases.

El sistema de grados en el servicio militar es otro ejemplo de una escala ordinal: sargento $>$ cabo $>$ soldado raso.

Muchos inventarios de personalidad y pruebas de habilidades y aptitudes dan como resultado puntuaciones que tienen la fuerza de los rangos. Aunque las puntuaciones parecen ser más precisas que los rangos, por lo general esas escalas no cumplen los requisitos de cualquier nivel de medición más alto y pueden considerarse de manera adecuada como ordinales.

Un ejemplo final de una escala ordinal serían las calificaciones asignadas a un curso. Las calificaciones asignadas por medio de letras son generalmente A, B, C, D y E. Estas letras constituyen un ordenamiento de ejecución: $A > B > C > D > E$. Por varias razones, se pueden asignar números a estos grados de letras: $A = 4$, $B = 3$, $C = 2$, $D = 1$, $E = 0$. Estas asignaciones numéricas son arbitrarias: se pueden hacer cualesquiera otras asignaciones numéricas que preserven el orden intentado (por ejemplo, $A = 10$, $B = 7$, $C = 5$, $D = 3$, $E = 0$).

Propiedades formales

Axiomáticamente, la diferencia fundamental entre una escala nominal y una ordinal es que esta última incorpora no sólo la relación de equivalencia ($=$), sino también la relación "mayor que" ($>$). Esta última relación es irreflexiva, asimétrica y transitiva.[3]

Considérese un conjunto de objetos $x_1, x_2, \ldots, x_N$. Supóngase que existe alguna relación en el atributo *verdadero* entre los objetos de cada categoría, además de la equivalencia dentro de las categorías. Esto es,

$$A(x_i) = A(x_j) \quad \text{si } x_i \text{ y } x_j \text{ están en la misma clase}$$

$$A(x_i) \neq A(x_j) \quad \text{si } x_i \text{ y } x_j \text{ están en diferentes clases}$$

y $$A(x_i) > A(x_j) \quad \text{si } x_i \text{ escede a } x_j \text{ en la "cantidad" que tiene del atributo}$$

Entonces, una escala *ordinal es un sistema de clasificación* $L(x)$ de los objetos tal que

$$L(x_i) = L(x_j) \quad \text{si y sólo si } A(x_i) = A(x_j)$$

y $$L(x_i) \neq L(x_j) \quad \text{si y sólo si } A(x_i) \neq A(x_j)$$

Además, $$L(x_i) > L(x_j) \quad \text{si y sólo si } A(x_i) > A(x_j)$$

Es decir, la función de clasificación ordena los objetos en el mismo modo en que de hecho están ordenados los atributos.

Operaciones admisibles

Ya que cualquier transformación que preserve el orden no cambia la información contenida en la escala ordinal, se dice que la escala es "única hasta una transformación monotónica". Una transformación monotónica es aquella que preserva el orden de los objetos. Esto es, no importa qué números demos a un par de clases o a los miembros de esas clases, siempre que les sea asignado un número mayor a los miembros de la clase que es "mayor que" o "más preferida". (Naturalmente, se pueden usar números menores para las clases "más preferidas". Así nos referimos generalmente a una ejecución excelente como "primera clase", y a ejecuciones progresivamente inferiores como "segunda clase" y "tercera clase". Siempre que seamos consistentes, no importa si se usan números mayores o menores para denotar "mayor que" o "más preferido".)

Por ejemplo, en el ejército un cabo usa dos bandas en su manga y un sargento usa tres. Estas insignias denotan que el sargento $>$ el cabo, y el símbolo $>$ denota "mayor rango que". Esta relación podría ser igualmente bien expresada si el cabo usara cuatro bandas y el sargento siete. Vale decir, una transformación que no cambia el orden de las clases es completamente admisible ya que *no implica pérdida alguna de información*. Cualesquiera o todos los números que se aplican a las clases en una escala ordinal pueden ser cambiados de cualquier forma que no alte-

[3] Irreflexiva: no es cierto para cualquiera x que $x > x$. Asimétrica: si $x > y$ entonces $y \not> x$. Transitiva: si $x > y$ y $y > z$, entonces $x > z$.

re el orden (rango) de los objetos. Puede aplicarse *cualquier* transformación montónica y aún preservarse las propiedades de la escala, esto es, preservar la relación entre los objetos.

El estadístico más apropiado para describir la tendencia central de las puntuaciones en una escala ordinal es la mediana, ya que en relación con la distribución de puntuaciones, la mediana no es afectada por los cambios en cualesquiera de las puntuaciones que están por arriba o por abajo de ella, siempre que el número de puntuaciones por arriba y por debajo permanezca constante.[4] Con el escalamiento ordinal, las hipótesis pueden ser probadas usando el gran grupo de pruebas estadísticas no paramétricas que en ocasiones se llaman *estadísticos de rango* o *estadísticos de orden*.

Además de la suposición de independencia, la única suposición hecha por algunas pruebas de rango es que las puntuaciones que observamos se obtengan de una distribución lineal continua. Las pruebas paramétricas también hacen esta suposición, pero además hacen suposiciones específicas acerca de la *forma* de la distribución continua por ejemplo, que es normal. Una *variable continua* es aquella que puede tomar cualquier valor dentro de cierto intervalo; por ejemplo, cualquier valor entre 0 y 100. Por otra parte, una *variable discreta* es aquella que sólo puede tomar un número finito (contable) de valores, por ejemplo, 0, 10, 20, . . . , 100. Además, una variable continua es aquella que puede tomar un número infinito (incontable) de diferentes valores, así como valores *entre* cualesquiera dos valores.

Para algunas técnicas, estadísticas que requieren medidas ordinales, es necesario que las puntuaciones observadas estén sobre una *línea* continua, aunque las puntuaciones reales que observamos puedan caer en categorías discretas. Por ejemplo, en una prueba de salón de clases, las puntuaciones reales registradas pueden ser para un reactivo particular "aprobado" o "reprobado". Podemos suponer que existe un continuo de posibles resultados subyacente a tal dicotomía; es decir, que algunos individuos que fueron categorizados como reprobados hayan estado más cerca de aprobar que otros que también reprobaron. De manera similar, algunos aprobaron sólo mínimamente, mientras que otros aprobaron con mucha facilidad. La suposición es que "aprobado" y "reprobado" representan un continuo dicotomizado dentro de dos intervalos. Por ejemplo, las puntuaciones reales pudieron haber sido 0, 1, 2, . . . , 100, y "aprobado" significa cualquier puntuación ≥ 70 y "reprobado" incluye cualquier valor < 70.

Del mismo modo, en materia de opinión, aquellas que son clasificadas como "de acuerdo", "ambivalentes" y "en desacuerdo", puede pensarse que caen en un continuo que refleja la fuerza del acuerdo/desacuerdo. Aquellas opiniones que son clasificadas como "de acuerdo", realmente pudieran no estar muy interesadas con la cuestión, mientras que otras pueden ser fuertemente convincentes de su posición. Aquellas opiniones que están en "desacuerdo" incluyen las que están sólo ligeramente en desacuerdo, así como los oponentes más intransigentes.

A menudo, lo imperfecto de nuestros dispositivos de medición oscurece la continuidad subyacente que puede existir. Si una variable está distribuida de ma-

[4] Es necesario destacar que si se cambian las asignaciones numéricas a las puntuaciones, la mediana cambiará en relación con el cambio en las asignaciones, pero aún permanecerá a la mitad de la distribución. No se puede hacer una afirmación similar acerca de la media.

nera continua verdaderamente, entonces la probabilidad de un empate es cero. Sin embargo, las puntuaciones empatadas ocurren con frecuencia. Tales puntuaciones casi invariablemente son un reflejo de la carencia de sensibilidad de nuestros instrumentos de medición, esto es, de la inhabilidad de los mismos para distinguir las pequeñas diferencias que existen entre las observaciones que se registran consecuentemente como empates. Por tanto, aun cuando se observen empates, es posible que una distribución continua subyazca a nuestras mediciones gruesas.

La escala de intervalo

Definición

Cuando una escala tiene todas las características de una escala ordinal y cuando además tienen sentido las *distancias* o *diferencias* entre cualesquiera dos números de la escala, se ha logrado una medición considerablemente más fuerte que la ordinal. En tal caso, la medición ha sido lograda en el sentido de una escala de intervalo. Esto es, si nuestro mapeo de varias clases de objetos es tan preciso que conocemos cuán grandes son los intervalos (distancias) entre todos los objetos de la escala, y estos intervalos tienen significado sustantivo, entonces hemos logrado una medida de intervalo. Una escala de intervalo está caracterizada por una unidad común y constante de medida que asigna un número a todos los pares de objetos en el orden establecido. En esta clase de medición, la razón de cualesquiera dos intervalos es independiente de la unidad de medida y del punto cero. En la escala de intervalo, el punto cero y la unidad de medida son arbitrarios.

Ejemplos

Medimos la temperatura en una escala de intervalo. De hecho, comúnmente se usan dos diferentes escalas: Celsius y Fahrenheit. Al medir la temperatura, la unidad de medida y el punto cero son arbitrarios; son diferentes en ambas escalas. Sin embargo, las dos escalas contienen la misma cantidad y la misma clase de información. Esto es así debido a que están linealmente relacionadas. Es decir, una lectura en una escala puede ser transformada en la lectura equivalente de la otra por medio de una transformación lineal.[5]

$$°F = 9/5\ °C + 32$$

donde

°F = número de grados en la escala Fahrenheit
°C = número de grados en la escala Celsius

Se puede mostrar que las razones de las diferencias de temperatura (interva-

[5] Matemáticamente, tales transformaciones son referidas como ajustes; sin embargo, en la bibliografía de estadística aplicada, el referente más común es la transformación lineal.

los) son independientes de la unidad de medida y del punto cero. Por ejemplo, el punto de "congelación" ocurre en 0° en la escala Celsius, y el punto de "ebullición" ocurre en los 100°. En la escala Fahrenheit, la "congelación" ocurre en los 32° y la "ebullición" en 212°. Algunas otras lecturas de la misma temperatura en las dos escalas son las siguientes:

Celsius	− 18	0	10	30	100
Fahrenheit	0	32	50	86	212

Nótese que la razón de las *diferencias* entre las lecturas de temperatura en una escala, es igual a la razón entre las diferencias equivalentes en la otra escala. Por ejemplo, en la escala Celsius la razón de las diferencias entre 30 y 10, y 10 y 0 es $(30 - 10) / (10 - 0) = 2$. Para las lecturas comparables en la escala Fahrenheit, la razón es $(86 - 50) / (50 - 32) = 2$. En ambos casos las razones son las mismas; a saber, 2. En otras palabras, en una escala de intervalo, la razón de cualesquiera dos intervalos es independiente de la unidad usada y del punto cero, siendo ambos arbitrarios.

Muchos científicos de la conducta aspiran a crear escalas de intervalo, y en pocas ocasiones tienen éxito. Sin embargo, generalmente lo que es tomado como éxito son suposiciones no probadas que el constructor de la escala voluntariamente cree. Una suposición frecuente es que la variable que está siendo escalada está normalmente distribuida entre los individuos a los que se evalúa con base en esta suposición, el constructor de la escala manipula las unidades de la escala hasta que se encuentre la *supuesta* distribución normal de las puntuaciones de los individuos. Naturalmente, el procedimiento es sólo tan bueno como la intuición del investigador al elegir la distribución que supone.

Otra suposición que se hace a menudo para crear una escala de intervalo aparente es la suposición de que las respuestas "afirmativas" de las personas en cualquier reactivo son exactamente equivalentes a responder de manera afirmativa en cualquier otro reactivo. Esta suposición se hace para satisfacer el requisito de que una escala de intervalo debe tener una unidad de medida común y constante. En escalas de habilidades o de aptitudes, la suposición de equivalencia consiste en que dar la respuesta correcta a cualquier reactivo es exactamente equivalente (en la cantidad de habilidad mostrada) a dar la respuesta correcta a cualquier otro reactivo.

Propiedades formales

Axiomáticamente, se puede mostrar que las operaciones y relaciones que dan origen a la estructura de una escala de intervalo son tales que las *diferencias* en la escala son isomórficas a la estructura de la aritmética. Los números pueden ser asociados con las posiciones de los objetos en una escala de intervalo tal que las operaciones de la aritmética pueden ser significativamente ejecutadas con las *diferencias* entre los números.

Al construir una escala de intervalo no sólo se deben especificar equivalencias, como en la escala nominal, y relaciones "mayor que", como en la escala ordinal, sino también se debe ser capaz de especificar la razón entre dos intervalos cualesquiera.

Considérese un conjunto de objetos $x_1, x_2, \ldots, x_N$. Supóngase que los atributos *verdaderos* de los objetos existen en alguna relación unos con otros, además de sus equivalencias dentro de las categorías. Esto es:

$A(x_i) = A(x_j)$ si x_i y x_j están en la misma clase

$A(x_i) \neq A(x_j)$ si x_i y x_j están en diferentes clases

y $A(x_i) > A(x_j)$ si x_i excede a x_j en la "cantidad" que tiene del atributo

Entonces, una escala de *intervalo* es un sistema *clasificatorio* de los objetos $L(x)$ que tienen las propiedades de una escala ordinal y, además

$$L(x) = cA(x) + b \qquad c > 0$$

Nótese que en este caso, la diferencia entre los atributos de los dos objetos es proporcional a la diferencia entre las asignaciones de clasificación:

$$L(x_i) - L(x_j) = c\,[A(x_i) - A(x_j)]$$

El lector debe ser capaz de verificar que la razón de las diferencias entre los atributos verdaderos será igual a la razón de las diferencias entre las asignaciones de clasificación hechas a los objetos.

Operaciones admisibles

Cualquier cambio en los números asociados con las posiciones de los objetos medidos en una escala de intervalo debe preservar no sólo el orden de los objetos, sino también las diferencias relativas entre los objetos. Esto es, la escala de intervalo es "única hasta una transformación lineal". Así, como hemos señalado, la información proporcionada por la escala no es afectada si cada número se multiplica por una constante positiva y después se le suma a este producto una constante, esto es, $f(x) = cx + b$. (En el ejemplo de la temperatura, $c = 9/5$ y $b = 32$.)

Ya hemos notado que en una escala de intervalo el punto cero es arbitrario. Esto es inherente al hecho de que la escala está sujeta a transformaciones que consisten en agregar una constante a los números que constituyen la escala.

La escala de intervalo es la primera escala verdaderamente "cuantitativa" que hemos encontrado. Todos los estadísticos paramétricos comunes (medias, desviaciones estándar, correlaciones producto-momento, etc.) son aplicables a los datos en una escala de intervalo. Si de hecho se ha logrado una medida en una escala de intervalo y si se han encontrado adecuadamente todas las suposiciones del modelo estadístico paramétrico (dadas en la sección "El modelo estadístico"), entonces el investigador puede utilizar pruebas estadísticas paramétricas tales como la prueba t o la prueba F. En tal caso, los métodos no paramétricos no aprovechan toda la información contenida en los datos de investigación. Puede notarse que una escala de intervalo es una condición necesaria, pero no suficiente, para usar una prueba estadística paramétrica que incluya la distribución normal.

La escala de razón

Definición

Cuando una escala tiene todas las características de una escala de intervalo y, además, tiene un punto cero verdadero en su origen, se llama *escala de razón*. En una escala de razón, la razón de cualesquiera dos puntos es independiente de la unidad de medida.

Ejemplo

Medimos la masa o el peso en una escala de razón. La escala de onzas y libras tiene un punto cero verdadero, al igual que la escala de gramos. La razón entre cualesquiera dos pesos es independiente de la unidad de medida. Por ejemplo, si determinamos los pesos de dos objetos diferentes no sólo en libras sino también en gramos, encontraremos que la razón de los dos pesos en libras es idéntica a la razón de los dos pesos en gramos.

Aunque es difícil identificar ejemplos significativos en las ciencias sociales y de la conducta, los contraejemplos abundan. Consideramos dos. Notamos anteriormente que las calificaciones se miden en una escala ordinal. Considérese a dos estudiantes, uno de los cuales recibe una *A* y el otro una *C*; y supóngase que las asignaciones numéricas fueron 4 y 2, respectivamente. Aunque la razón de las dos calificaciones es dos ($4/2 = 2$), no tiene sentido decir que el estudiante con una *A* posee el doble de "algo" del estudiante que recibe la *C*. (El estudiante puede obtener el doble de ciertos puntos, pero no es claro si esto tiene algún significado sustantivo en conocimiento, habilidad o perseverancia.) Finalmente, en el caso de la temperatura, considérese un cambio en la temperatura de 10° a 30 °C. No podemos decir que el incremento representa que el calor se incrementó al triple. Para ver esto, nótese que el cambio en la temperatura es equivalente a un cambio de 50° a 86 °F. Debido a que las razones de las temperaturas en las dos escalas son claramente diferentes, la razón no tiene sentido interpretable alguno.

Propiedades formales

Las operaciones y relaciones que dan origen a los valores numéricos en una escala de razón son tales que la escala es isomórfica a la estructura de la aritmética. Por tanto, las operaciones de la aritmética son permisibles con los valores numéricos asignados a los objetos, así como a los intervalos entre los números, como en el caso de la escala de intervalo.

Las escalas de razón, que se encuentran más comúnmente en las ciencias físicas, se logran sólo cuando son operacionalmente posibles de alcanzar todas las siguientes cuatro relaciones: 1. equivalencia; 2. mayor que; 3. razón conocida entre cualesquiera dos intervalos, y 4. razón conocida entre cualesquiera dos valores de la escala.

Considérese un conjunto de objetos $x_1, x_2, \ldots, x_N$. Supóngase que el atributo *verdadero* de los objetos existe con alguna relación entre cada uno de ellos, además de la equivalencia dentro de las categorías. Esto es

$$A(x_i) = A(x_j) \qquad \text{si } x_i \text{ y } x_j \text{ están en la misma clase}$$

$$A(x_i) \neq A(x_j) \qquad \text{si } x_i \text{ y } x_j \text{ están en diferentes clases}$$

$$\text{y} \quad A(x_i) > A(x_j) \qquad \text{si } x_i \text{ excede a } x_j \text{ en la “cantidad” que tiene del atributo}$$

Entonces, una escala de *razón* es un sistema *clasificatorio* de los objetos $L(x)$ si

$$L(x_i) = cA(x_i) \qquad c > 0$$

Así,

$$\frac{L(X_i)}{L(x_j)} = \frac{A(X_i)}{A(x_j)}$$

y la razón de las clasificaciones asignadas es igual a la razón de los atributos verdaderos.

Operaciones admisibles

Los números asociados con los valores de la escala de razón son números “verdaderos” con un cero verdadero: sólo la unidad de medida es arbitraria. Así, la escala de razón es única hasta la multiplicación por una constante positiva. Esto es, las razones entre cualesquiera dos números se preservan cuando los valores de la escala son todos multiplicados por una constante positiva y, además, tal transformación no altera la información contenida en la escala.

Cualquier prueba estadística paramétrica puede usarse cuando se han logrado medidas de razón y se encuentran las suposiciones adicionales concernientes a la distribución. Más aún, existen algunos estadísticos que se aplican sólo a datos que descansan en una escala de razón; debido a la fuerza de las suposiciones que subyacen a la escala, la mayoría de estas pruebas son paramétricas.

Resumen

La medición es el proceso de mapear o asignar números a objetos u observaciones. La clase de medición alcanzada es una función de las reglas según las cuales los números se asignan a los objetos. Las operaciones y relaciones empleadas en obtener las puntuaciones definen y limitan las manipulaciones y operaciones que son permisibles al manipular las puntuaciones: las manipulaciones y operaciones deben ser aquellas de la estructura numérica a la cual la medida particular es isomórfica.

Se examinaron cuatro de las escalas más generales: nominal, ordinal, de intervalo y de razón; en la tabla 2.1 se resumen estas escalas de medición. Las medidas nominales y ordinales son los tipos más comunes alcanzados en las ciencias sociales y de la conducta. Los datos medidos en las escalas nominal u ordinal deben ser analizados por métodos no paramétricos, si el modelo estadístico es válido para esos datos. Poder usar pruebas paramétricas depende de las suposiciones que sostiene el modelo estadístico paramétrico particular. Como hemos señalado, estas suposiciones nunca se encuentran a menos que tengamos datos en escalas de intervalo o de razón.

Tabla 2.1. Cuatro niveles de medición.

Escala	*Relaciones que la definen*
Nominal	1. Equivalencia
Ordinal	1. Equivalencia 2. Mayor que
De intervalo	1. Equivalencia 2. Mayor que 3. Razón conocida entre cualesquiera de dos intervalos
De razón	1. Equivalencia 2. Mayor que 3. Razón conocida entre cualesquiera de dos intervalos 4. Razón conocida entre cualesquiera de dos valores de la escala

Aun con el riesgo de ser excesivamente repetitivos, deseamos destacar que algunas pruebas estadísticas paramétricas que suponen que las puntuaciones tienen una distribución normal y que usan medias y desviaciones estándar (esto es, que requieren las operaciones de la aritmética en las puntuaciones originales), no deben ser usadas con datos que no están en una escala de intervalo. Las propiedades de una escala ordinal *no* son isomórficas al sistema numérico conocido como *aritmética*. Cuando sólo se conoce el orden de rango de las puntuaciones, obtener medias y desviaciones estándar con las puntuaciones es un *error* o *equivocación* en la extensión en que los intervalos sucesivos (distancias entre las clases) de la escala no son iguales y *no tienen significado sustancial*. Cuando se usan técnicas paramétricas de la inferencia estadística con tales datos, cualesquiera decisiones acerca de las hipótesis son dudosos. Las conclusiones de probabilidad obtenidas de la aplicación de pruebas estadísticas paramétricas con datos ordinales, puede ser un error cuando las variables no satisfacen las suposiciones paramétricas. Ya que la mayoría de las medidas hechas por los científicos de la conducta culminan en escalas nominales u ordinales, este tema merece un énfasis mayor.

Se debe destacar que estamos hablando acerca de asignaciones numéricas usadas en nuestra investigación. Debe ser obvio que una media y una desviación estándar pueden ser computadas para cualquier conjunto de números. Sin embargo, los estadísticos computados de estos números sólo "tienen sentido" si el procedimiento de asignación original imparte interpretaciones "aritméticas" a las asignaciones. Éste es un punto sutil y crítico al cual retornaremos posteriormente.

Puesto que este libro está dirigido a científicos sociales y de la conducta, y ya que las escalas usadas por ellos son típicamente como máximo no más fuertes que la escala ordinal, la principal porción de este libro está dedicada a los métodos que resultan adecuados para probar hipótesis con datos medidos en una escala ordinal. Estos métodos, que están basados en suposiciones menos circunscritas o restrictivas en sus modelos estadísticos que las pruebas paramétricas, proporcionan el volumen de las pruebas no paramétricas.

Referencias bibliográficas

El lector puede encontrar otros análisis sobre medición en Bailey (1971), Hays (1983), Davidson, Siegel y Suppes (1955), y un informe, que se recomienda especialmente, de Townsend y Ashby (1984).

PRUEBAS ESTADÍSTICAS PARAMÉTRICAS Y NO PARAMÉTRICAS

Una prueba estadística paramétrica especifica ciertas condiciones acerca de la distribución de respuestas en la población de la cual se ha obtenido la muestra investigada. Ya que estas condiciones no son ordinariamente evaluadas, sólo se suponen. La significación de los resultados de la prueba paramétrica depende de la validez de estas suposiciones. Una adecuada interpretación de las pruebas paramétricas basadas en la distribución normal también supone que las puntuaciones que están siendo analizadas resultan de medidas en por lo menos una escala de intervalo.

Una prueba estadística no paramétrica está basada en un modelo que especifica sólo condiciones muy generales y ninguna acerca de la forma específica de la distribución de la cual fue obtenida la muestra. Ciertas suposiciones están asociadas con la mayoría de las pruebas no paramétricas, a saber: que las observaciones son independientes y quizá que la variable en estudio es continua; pero estas suposiciones son menores y más débiles que aquellas asociadas con las pruebas paramétricas. Más aún, como veremos, los procedimientos no paramétricos prueban diferentes hipótesis acerca de la población, que los procedimientos paramétricos no hacen. Por último, a diferencia de las pruebas paramétricas, existen pruebas no paramétricas que pueden aplicarse apropiadamente a datos medidos en una escala ordinal, y otras pruebas para datos en una escala nominal o categórica.

En este capítulo hemos examinado los diversos criterios que deben considerarse en la elección de la prueba estadística que se va a usar para decidir acerca de una hipótesis de investigación. Estos criterios son los siguientes: 1. la aplicabilidad o validez de la prueba (que incluye el nivel de medición y otras suposiciones de la prueba), y 2. la potencia y eficacia de la prueba. Se ha establecido que una prueba estadística paramétrica es más potente cuando se encuentran todas las suposiciones de su modelo estadístico. Sin embargo, aun cuando se satisfagan todas las suposiciones de las pruebas paramétricas acerca de la población y los requisitos mínimos del nivel de medición, sabemos del concepto de eficacia (ya sea potencia-eficacia o eficacia relativa asintótica) que podemos usar una prueba no paramétrica incrementando el tamaño de la muestra en una pequeña cantidad, y aún obtener la misma potencia para rechazar H_0.

Debido a que la potencia de cualquier prueba puede ser incrementada simplemente incrementando N, y los científicos de la conducta raras veces tienen datos que satisfacen las suposiciones de las pruebas paramétricas que incluyen alcanzar la clase de medida que permite la interpretación significativa de las pruebas paramétricas, las pruebas estadísticas no paramétricas desempeñan un papel prominente en la investigación en las ciencias sociales y de la conducta. En este libro se presenta una variedad de pruebas no paramétricas. El uso de las pruebas paramétricas basadas en la distribución normal en la investigación ha sido presentada en una variedad de fuentes[6] y, por tanto, no las examinaremos aquí.

En muchas de las pruebas estadísticas no paramétricas que se estudian en este libro, los datos han sido cambiados de puntuaciones a rangos y aun a signos. Tales métodos pueden despertar la crítica de que "no se está usando toda la información proporcionada por la muestra" o que "se está perdiendo información". La réplica a esta objeción se encuentra en las respuestas a las siguientes preguntas:

1. De los métodos disponibles, paramétricos y no paramétricos, ¿en cuál de ellos se usa adecuadamente la información contenida en la muestra? Esto es, ¿qué prueba es válida?
2. ¿Han sido satisfechas las suposiciones que subyacen a un modelo o una prueba estadística en particular?
3. ¿Las hipótesis probadas por el modelo estadístico son apropiadas para la situación?

La respuesta a la primera pregunta depende del nivel de medición alcanzado en la investigación y del conocimiento de la población por parte del investigador. Si la medición es más débil que aquella de una escala de intervalo, al usar una prueba paramétrica el investigador podría "agregar información" y, por tanto, crear distorsiones que pueden ser tan grandes y dañinas como aquellas introducidas por la "pérdida de información" que ocurre cuando las puntuaciones son convertidas a rangos. Más aún, las suposiciones que deben hacerse para justiciar el uso de las pruebas paramétricas, por lo general descansan la conjetura y la fe, pero el conocimiento acerca de los parámetros de la población casi invariablemente es inexistente. Por último, para algunas distribuciones de población, una prueba estadística no paramétrica es claramente superior en potencia a la prueba paramétrica.

Las respuestas a la segunda y tercera preguntas pueden ser dadas sólo por el investigador al considerar los aspectos sustantivos del problema de investigación y al examinar los datos.

La relevancia de la exposición en este capítulo para elegir entre pruebas estadísticas paramétricas y no paramétricas, puede ser reforzada con la lectura de las dos siguientes secciones, en las que se enumeran las ventajas y desventajas de dichas pruebas.

[6] Entre las diversas fuentes de las pruebas estadísticas paramétricas, las siguientes son especialmente útiles: Hays (1983), Bailey (1971), Edwards (1967).

Ventajas de las pruebas estadísticas no paramétricas

1. Si el tamaño de la muestra es muy pequeño, puede no haber otra opción que usar una prueba estadística no paramétrica, a menos que la naturaleza de la distribución de la población se *conozca con exactitud*.
2. Las pruebas no paramétricas típicamente hacen menos suposiciones acerca de los datos y pueden ser más relevantes a una situación particular. Además, las hipótesis probadas por una prueba no paramétrica pueden ser más adecuadas para la investigación.
3. Las pruebas estadísticas no paramétricas están disponibles para analizar datos que son inherentes a los rangos, así como datos cuyas puntuaciones numéricas tienen aparentemente la fuerza de los rangos. Esto es, el investigador puede sólo ser capaz de decir que algunos sujetos de investigación tienen más o menos de la característica en cuestión que otros, sin ser capaces de determinar *qué tanto* más o menos. Por ejemplo, al estudiar variables tales como la ansiedad, podemos ser capaces de establecer que el sujeto *A* es más ansioso que el sujeto *B*, sin conocer con exactitud cuánto más ansioso es *A*. Si los datos están inherentemente en rangos, o aun si pueden ser categorizados sólo como mayor o menor (más o menos, mejor o peor), pueden ser tratados por métodos no paramétricos, a menos que se hagan suposiciones precarias y quizá irreales acerca de las distribuciones.
4. Los métodos no paramétricos están disponibles para tratar datos que son simplemente clasificatorios o categóricos, es decir, que son medidos en una escala nominal. Ninguna técnica paramétrica se aplica a tales datos.
5. Existen pruebas estadísticas no paramétricas que son adecuadas para tratar muestras obtenidas de observaciones de *diferentes* poblaciones. Las pruebas paramétricas a menudo no pueden manipular tales datos sin exigirnos hacer suposiciones aparentemente irreales o requisitos pesados de computación.
6. Las pruebas estadísticas no paramétricas típicamente son más fáciles de aprender y aplicar que las pruebas paramétricas. Además, su interpretación suele ser más directa que la interpretación de las pruebas paramétricas.

Supuestas desventajas de las pruebas estadísticas no paramétricas

Si se encontraran en los datos todas las suposiciones del modelo estadístico paramétrico, y si las hipótesis de investigación pudieran ser probadas mediante una prueba paramétrica, entonces las pruebas estadísticas no paramétricas serían inútiles. Este grado de falta de utilidad es expresado por la potencia-eficacia de la prueba no paramétrica. (Se recordará que si una prueba estadística no paramétrica tiene una potencia-eficacia, por ejemplo, de 90 %, esto significa que *cuando todas las condiciones de la prueba estadística paramétrica son satisfechas*, la prueba paramétrica adecuada es tan efectiva con una muestra que es 10 % más pequeña, que la usada en el análisis no paramétrico.)

Otra objeción a las pruebas estadísticas no paramétricas es que no son *sistemáticas*, mientras que las pruebas estadísticas paramétricas han sido sistematizadas y diferentes pruebas son simplemente variaciones de un tema central. Aunque esto es parcialmente verdadero, no nos parece que el valor de la aproximación sistemática justifique su costo. Más aún, un examen cuidadoso de las pruebas no paramétricas revela temas comunes: las pruebas para datos categóricos *son* sistemáticas, como lo son muchas de las pruebas aplicadas a datos ordenados. Las diferencias están en la superficie, es decir, las fórmulas computacionales algunas veces oscurecen las relaciones subyacentes entre las pruebas.

Una objeción más a las pruebas estadísticas no paramétricas se relaciona con la conveniencia. Las tablas necesarias para aplicar las pruebas no paramétricas están muy difundidas y aparecen en diferentes formatos. (Lo mismo es cierto para numerosas pruebas paramétricas.) En este libro hemos tratado de proporcionar juntas muchas de las tablas necesarias para probar hipótesis conveniente al usar pruebas estadísticas no paramétricas y presentarlas en un formato sistemático.

En esta obra también hemos procurado presentar la mayoría de las técnicas no paramétricas de inferencia estadística y medidas de asociación que los científicos de la conducta y sociales probablemente necesiten, y proporcionamos las tablas necesarias para aplicar estas técnicas. Aunque este texto no es exhaustivo en el tema de las pruebas no paramétricas —no podría serlo sin ser excesivamente redundante y voluminoso—, se incluyen suficientes pruebas en los siguientes capítulos, que proporcionan a los científicos de la conducta un rango amplio para elegir la técnica no paramétrica útil para probar sus hipótesis de investigación y adecuada a sus diseños de investigación.

3

El caso de una muestra simple

En este capítulo presentamos varias pruebas estadísticas no paramétricas que pueden utilizarse para probar una hipótesis derivada de una muestra únicamente. Las pruebas nos dicen si la muestra particular proviene de alguna población especificada. Estas pruebas son distintas a las pruebas para dos muestras, que comparan dos muestras y prueban si es probable que las dos provengan de la misma población. Las pruebas de dos muestras pueden resultar más familiares a algunos lectores.

Las pruebas de una muestra con frecuencia son pruebas de bondad de ajuste. En el caso típico, extraemos una muestra aleatoria de alguna población y probamos la hipótesis de que la muestra se extrajo de una población con una distribución específica o con características específicas. Las pruebas de una muestra responden a preguntas como las siguientes:

1. ¿Existe una diferencia significativa en la localización (tendencia central) entre la muestra y la población?
2. ¿Existe una diferencia significativa entre las frecuencias observadas y las frecuencias que podríamos esperar en base a los postulados de alguna teoría?
3. ¿Existe una diferencia significativa entre las proporciones observadas y esperadas en una serie de observaciones dicotómicas?
4. ¿Es razonable creer que la muestra fue extraída de una población con una forma específica (por ejemplo, normal o uniforme)?
5. ¿Es razonable creer que la muestra es una muestra aleatoria de alguna población conocida?
6. En una serie de observaciones, ¿existe un cambio en el modelo teórico subyacente que se supone genera los datos?

En el caso de una muestra, una técnica paramétrica común es aplicar una prueba t a la diferencia entre la media observada (de la muestra) y la media esperada (de la población). En términos estrictos, la prueba t supone que las observaciones o puntuaciones en la muestra provienen de una población normalmente distribuida. La interpretación apropiada de la prueba t supone que las variables están medidas como mínimo en una escala de intervalo.

Existen muchas clases de datos para los cuales la prueba t puede ser inadecuada. El investigador puede encontrar que:

1. Las suposiciones y los requisitos para una apropiada interpretación de la prueba t no son realistas para los datos.
2. Es preferible evitar hacer las suposiciones de la prueba t y así ganar una generalidad mayor en las conclusiones.
3. Los datos están inherentemente en rangos (esto es, en una escala ordinal) y, por tanto, las pruebas paramétricas estándar pueden ser inadecuadas.
4. Los datos pueden ser categóricos o clasificatorios.
5. No existe una prueba paramétrica útil para la hipótesis particular que va a ser probada.

En tales circunstancias, el investigador debe elegir una de las pruebas estadísticas no paramétricas para una muestra, descritas en este capítulo.

También se presentarán varias pruebas para el caso de una muestra. En el siguiente capítulo se describen pruebas adicionales de una muestra basadas en observaciones múltiples o repetidas. El capítulo concluye con una comparación y un contraste de las pruebas, que ayudará al investigador a seleccionar la que mejor se ajuste a una hipótesis en particular.

PRUEBA BINOMIAL

Función y racionalización

Existen muchas poblaciones que son concebidas como compuestas de sólo dos clases. Ejemplos de tales clases son: hombre y mujer; alfabeto y analfabeto; miembro y no miembro; soltero y casado; internado y ambulatorio. Para tales casos, todas las posibles observaciones de la población caerán en una de dos categorías discretas. Tal población generalmente se denomina *población binaria* o *población dicotómica.*

Supóngase que una población consta de sólo dos categorías o clases. Entonces, cada observación (X) muestreada de la población puede tomar uno de dos valores, dependiendo de la categoría muestreada. Podemos denotar los posibles valores de la variable aleatoria usando cualquier par de valores, pero es conveniente denotar cada resultado como 1 o 0. Asumiremos posteriormente que la probabilidad de muestrear un objeto de la primera categoría es p y la probabilidad de muestrear un objeto de la otra categoría es $q = 1 - p$. Esto es,

$$P[X = 1] = p \qquad \text{y} \qquad P[X = 0] = 1 - p = q$$

También se supone que cada probabilidad es constante, sin considerar el número de sujetos muestreados u observados.

Aunque el valor de p puede variar de población a población, es un valor fijo para una determinada población. Sin embargo, aun si conocemos (o suponemos) el valor de p para alguna población, no podemos esperar que una muestra aleatoria de observaciones de la población contenga exactamente las proporciones p y $1 - p$ para cada una de las dos categorías. El muestreo aleatorio generalmente impide que la muestra duplique precisamente los valores de la población de p y q. Por ejemplo, de los registros oficiales podemos conocer que los votantes de cierta ciudad están divididos por mitades entre los partidos republicano y demócrata. Pero una muestra aleatoria de los votantes registrados en esa ciudad puede contener 47 % de demócratas y 53 % de republicanos, o quizá 56 % de demócratas y 44 % de republicanos. Tales diferencias entre los valores de la población y los observados se originan debido a las fluctuaciones al azar o aleatorias en las observaciones. No debemos sorprendernos por desviaciones pequeñas de los valores poblacionales; sin embargo, desviaciones grandes —aunque posibles— son poco probables.

La distribución binomial se usa para determinar las probabilidades de los posibles resultados que podemos observar al muestrear una población binomial. Si nuestra hipótesis es $H_0 : p = p_0$, podemos calcular las probabilidades de varios resultados cuando suponemos que H_0 es cierta. La prueba nos dirá si es razonable creer que las proporciones (o frecuencias) de las dos categorías en nuestra muestra han sido extraídas de una población con los valores hipotéticos de p_0 y $1 - p_0$. Por conveniencia, al hablar de la distribución binomial, denotamos el resultado $X = 1$ como "éxito" y el resultado $X = 0$ como "fracaso". Además, en una serie de N observaciones,

$$Y = \sum_{i=1}^{N} X_i$$

es el número de "éxitos" o el número de resultados de tipo $X = 1$.

Método

En una muestra de tamaño N, la probabilidad de obtener k objetos en una categoría y $N - k$ objetos en la otra categoría, está dada por

$$P[Y = k] = \binom{N}{k} p^k q^{N-k} \qquad k = 0, 1, \ldots, N \tag{3.1}$$

donde

p = la proporción de observaciones esperadas cuando $X = 1$
q = la proporción de observaciones esperadas cuando $X = 0$

y
$$\binom{N}{k} = \frac{N!}{k!(N-k)!}$$ [1]

[1] $N!$ es "N factorial", el cual es definido como

$$N! = N(N-1)(N-2)\ldots(2)(1)$$

La tabla E del Apéndice I proporciona los valores de $P[Y = k]$ para diferentes valores de N y p.

Un ejemplo aclarará la ecuación (3.1). Supongamos que un dado es lanzado cinco veces. ¿Cuál es la probabilidad de que exactamente dos de las tiradas muestren un "seis"? En este caso, Y es la variable aleatoria (el resultado de los cinco lanzamientos del dado), N = al número de lanzamientos (5), k = el número observado de seises (2), p = la proporción esperada de seises (1/6) y $q = 5/6$. La probabilidad de que exactamente en dos de los cinco lanzamientos aparezca un seis está dada por la ecuación (3.1):

$$P[Y = k] = \binom{N}{k} p^k (1 - p)^{N-k}$$

$$P[Y = 2] = \frac{5!}{2!3!}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3 = 0.16$$

La aplicación de la fórmula al problema nos muestra que la probabilidad de obtener exactamente dos "seises" cuando se lanza un dado normal cinco veces es $p = 0.16$.

Ahora bien, cuando probamos hipótesis, la cuestión *no* es generalmente "¿cuál es la probabilidad de obtener *exactamente* los valores que fueron observados?", sino que más bien, por lo regular preguntamos "¿cuál es la probabilidad de obtener valores *tan extremos* o *más extremos que* los valores observados, cuando suponemos que los datos son generados por un proceso particular?" Para responder a preguntas de este tipo, la probabilidad deseada es

$$P[Y \geq k] = \sum_{i=k}^{N} \binom{N}{i} p^i q^{N-i} \tag{3.2}$$

En otras palabras, sumamos la probabilidad de los resultados observados con la probabilidad de resultados que son aún más extremos.

Supóngase ahora que queremos conocer la probabilidad de obtener dos *o menos* seises cuando se lanza cinco veces un dado normal. Aquí, de nuevo $N = 5$, $k = 2$, $p = 1/6$ y $q = 5/6$. Ahora la probabilidad de obtener dos o menos seises se denota $p[Y \leq 2]$. De la ecuación (3.1), la probabilidad de obtener 0 seises es $P[Y = 0]$, la probabilidad de obtener un seis es $P[Y = 1]$, etc. Usando la ecuación (3.2), tenemos

$$P[Y \leq 2] = P[Y = 0] + P[Y = 1] + P[Y = 2]$$

Por ejemplo, $4! = (4)(3)(2)(1) = 24$ y $5! = 120$. Por definición, $0! = 1$. En la tabla W del Apéndice I se proporcionan los factoriales para valores de N hasta 20. En la tabla X del Apéndice I se proporcionan coeficientes factoriales

$$\binom{N}{x} = \frac{N!}{x!\,(N - x)!}$$

para valores de N hasta 20.

Esto es, la probabilidad de obtener dos o menos seises es la suma de las tres probabilidades. Si usamos la ecuación (3.1) para determinar estas probabilidades, tenemos

$$P[Y = 0] = \frac{5!}{0!5!}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^5 = 0.40$$

$$P[Y = 1] = \frac{5!}{1!4!}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^4 = 0.40$$

$$P[Y = 2] = \frac{5!}{2!3!}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3 = 0.16$$

y así,

$$\begin{aligned} P[Y \leq 2] &= P[Y = 0] + P[Y = 1] + P[Y = 2] \\ &= 0.40 + 0.40 + 0.16 \\ &= 0.96 \end{aligned}$$

Hemos determinado que la probabilidad según H_0 (la suposición de un dado normal) de obtener dos o menos seises cuando un dado es lanzado cinco veces, es $p = 0.96$.

Muestras pequeñas

En el caso de una muestra, cuando se están usando categorías binarias, una hipótesis común es H_0: $p = 1/2$. En la tabla D del Apéndice I se proporcionan las probabilidades de una cola asociadas con la ocurrencia de diferentes valores tan extremos como k, según la hipótesis nula H_0: $p = 1/2$. Cuando se consulte dicha tabla, tómese a k como la más pequeña de las frecuencias observadas. Esta tabla es útil cuando $N \leq 35$. Aunque se puede usar la ecuación (3.2), la tabla es más conveniente. En la tabla D se proporcionan las probabilidades asociadas con la ocurrencia de diferentes valores tan pequeños como k para diferentes N. Por ejemplo, supongamos que observamos siete éxitos y tres fracasos. Aquí, $N = 10$ y $k = 7$. En la tabla D se muestra que la probabilidad de ocurrencia de una cola según H_0: $p = 1/2$ para $Y \leq 3$ cuando $N = 10$ es 0.172. Debido a la simetría de la distribución binomial, cuando $p = 1/2$, $P[Y \geq k] = P[Y \leq \mathrm{N} + k]$. Así, en este ejemplo, $P[Y \leq 3] = P[Y \geq 7] = 0.172$.

Las probabilidades proporcionadas en la tabla D son unidireccionales. Se usa una prueba unidireccional cuando hemos predicho con anterioridad cuál de las dos categorías contendrá el número más pequeño de casos. Cuando la predicción es simplemente que las dos frecuencias difieren, se usa una prueba bidireccional. Para una prueba bidireccional, los valores de probabilidad en la tabla D deben ser duplicados. Así, para $N = 10$ y $k = 7$, la probabilidad asociada con la ocurrencia bidireccional según H_0 es 0.344.

El siguiente ejemplo ilustra el uso de la prueba binomial en un estudio donde H_0: $p = 1/2$.

Ejemplo. En un estudio de los efectos del estrés,[2] un investigador enseñó a 18 estudiantes dos métodos diferentes para hacer el mismo nudo. La mitad de los sujetos (seleccionados aleatoriamente) aprendieron primero el método *A* y la otra mitad aprendió primero el método *B*. Posteriormente —a medianoche y después de un examen final de cuatro horas de duración—, a cada sujeto se le pidió que hiciera el nudo. La predicción fue que el estrés induciría regresión, esto es, que los sujetos regresarían al primer método aprendido para hacer el nudo. Cada sujeto fue categorizado conforme a si usó el primer o el segundo método aprendido de hacer nudos cuando se le pedía que hiciera el nudo bajo estrés.

i. *Hipótesis nula.* H_0: $p = q = 1/2$, esto es, no existen diferencias entre la probabilidad de usar el primer método aprendido bajo estrés (p) y la probabilidad de emplear el segundo método bajo estrés (q). Cualquier diferencia entre las frecuencias que pueda ser observada es de tal magnitud que pudiera esperarse en una muestra de la población de posibles resultados según H_0. H_1: $p > q$; es decir, cuando se está bajo estrés, la probabilidad de usar el primer método aprendido es mayor que la probabilidad de usar el segundo método aprendido.
ii. *Prueba estadística.* Se elige la prueba binomial debido a que los datos están en dos categorías discretas y el diseño es del tipo de una muestra. Ya que los métodos *A* y *B* se asignaron aleatoriamente para ser enseñados en primer y en segundo lugares, no hay razón para pensar que el primer método aprendido debería ser preferido al segundo método aprendido según H_0, y así $p = q = 1/2$.
iii. *Nivel de significación.* Sea $\alpha = 0.01$ y N es el número de casos $= 18$.
iv. *Distribución muestral.* La distribución muestral está proporcionada por la ecuación (3.2). Sin embargo, cuando $N \leq 35$ y cuando $p = q = 1/2$, la tabla D proporciona las probabilidades asociadas con la ocurrencia según H_0 de valores observados tan pequeños como k, y así en este ejemplo no es necesario calcular directamente la distribución muestral.
v. *Región de rechazo.* La región de rechazo consta de todos los valores de Y (donde Y es el número de sujetos que usó el segundo método aprendido bajo estrés), que son tan pequeños que la probabilidad asociada con su ocurrencia según H_0 es igual o menor que $\alpha = 0.01$. Ya que la dirección de la preferencia se predijo con anterioridad, la región de rechazo unidireccional.
vi. *Decisión.* En el experimento, todos los sujetos menos dos usaron el primer método aprendido cuando se les pidió que hicieran el nudo bajo estrés (tarde en la noche y después de un largo examen final). Estos datos se muestran en la tabla 3.1. En este caso, N es el número de observaciones independientes $= 18$; k es la frecuencia más pequeña $= 2$. En la tabla D del Apéndice I se muestra que para $N = 18$, la probabilidad asociada con $k \leq 2$ es 0.001. Ya que esta probabilidad es más pequeña que $\alpha = 0.01$, la decisión es rechazar H_0 en favor de H_1. Así, concluimos que $p > q$, esto es, la gente bajo estrés, regresa al primero de los dos métodos aprendidos.

Tabla 3.1. Método de atar nudos escogido bajo estrés.

	Método escogido		
	Aprendido antes	*Aprendido después*	*Total*
Frecuencia	16	2	18

[2] Barthol, R. P. y Ku, N. D., "Specific regression under a non—related stress situation", en *American Psychologist*, núm. 10, 1953, pág. 482.

Muestras grandes

La tabla D del Apéndice I no se puede utilizar cuando N es más grande que 35. Sin embargo, puede demostrarse que al incrementar el tamaño de N, la distribución binomial tiende a la distribución normal. Más precisamente, al incrementarse N la distribución de la variable Y se aproxima a la distribución normal. La tendencia es rápida cuando p se aproxima a 1/2, pero es más lenta cuando p está cercana a 0 o a 1. Vale decir, mientras más grande sea la disparidad entre p y q, más grande debe ser N antes de que la aproximación sea cercanamente útil a la distribución normal. Cuando p está cercana a 1/2, la aproximación puede ser usada para una prueba estadística cuando $N > 25$. Cuando p está cercana a 0 o a 1, una regla común es que Npq debe ser más grande que 9 antes de que la prueba estadística basada en la distribución normal sea suficientemente exacta para ser usada. Dentro de estas limitaciones, la distribución muestral de Y es aproximadamente normal, con media Np y varianza Npq y, por tanto, H_0 puede ser probada por

$$z = \frac{x - \mu_x}{\sigma_x} = \frac{Y - Np}{\sqrt{Npq}} \tag{3.3}$$

donde z en forma aproximada está normalmente distribuida con media 0 y desviación estándar 1.

La aproximación a la distribución normal es mejor si se usa una corrección por "continuidad". La corrección es necesaria debido a que la distribución normal es continua, mientras que la distribución binomial implica variables discretas. Para corregir por continuidad, consideramos la frecuencia observada Y de la ecuación (3.3) como ocupando un intervalo, el límite inferior del cual está media unidad abajo de la frecuencia observada, mientras que el límite superior está media unidad arriba de la frecuencia observada. La corrección por continuidad consiste en reducir, por 0.5, la diferencia entre los valores observados de Y y su valor esperado $\mu_Y = Np$. Por tanto, cuando $Y < \mu_Y$ agregamos 0.5 a Y, y cuando $Y > \mu_Y$ sustraemos 0.5 de Y. Esto es, la diferencia observada es reducida por 0.5. Así, z se calcula

$$z = \frac{(Y \pm 0.5) - Np}{\sqrt{Npq}} \tag{3.4}$$

donde $Y + 0.5$ se usa cuando $Y < Np$, y $Y - 0.5$ se usa cuando $Y > Np$. El valor de la z obtenido por la aplicación de la ecuación (3.4) está distribuido en forma normal asintóticamente con media 0 y varianza 1. Por tanto, la significancia de una z obtenida puede ser determinada por referencia a la tabla A del Apéndice I. Dicha tabla proporciona las probabilidades de una cola asociadas con la ocurrencia según H_0 de valores tan extremos como una z observada. (Si se requiere una prueba de dos colas, la probabilidad mostrada por la tabla A debe ser duplicada.)

Para mostrar cuán buena es esta aproximación cuando $p = 1/2$ aun para $N < 25$, podemos aplicarla a los datos de hacer nudos que hemos presentado. En ese caso, $N = 18$, $Y = 2$ y $p = q = 1/2$. Para esos datos, $Y < Np$, esto es, $2 < 9$, y por la ecuación (3.4)

$$z = \frac{(2 + 0.5) - (18)(1/2)}{\sqrt{(18)\,(1/2)\,(1/2)}}$$

$$= -3.06$$

En la tabla A del Apéndice I se muestra que un valor de z tan extremo como -3.06 tiene una probabilidad de una cola asociada con su ocurrencia según H_0 de 0.0011. Ésta es esencialmente la misma probabilidad que encontramos por el otro análisis, que usó una tabla de probabilidades exactas. Sin embargo, recuérdese que en este ejemplo $p = 1/2$, tal que la aproximación lo hizo bien.

Resumen del procedimiento

En síntesis, éstos son los pasos para usar la prueba binomial con H_0: $p = 1/2$:

1. Determinar $N =$ el número total de casos observados.
2. Determinar las frecuencias de las ocurrencias observadas en cada una de las dos categorías.
3. El método para encontrar la probabilidad de ocurrencia de los valores observados según H_0, o valores aún más extremos, depende del tamaño de la muestra:

   a) Si $N \leq 35$, la tabla D del Apéndice I proporciona las probabilidades de una cola según H_0 de diferentes valores tan pequeños como una Y observada. Especificar H_1 y determinar si la prueba debe ser uni o bidireccional.

   b) Si $N > 35$, probar H_0 usando la ecuación (3.4). La tabla A del Apéndice I proporciona las probabilidades asociadas con la ocurrencia según H_0 de valores tan grandes como una z observada. Esta tabla A proporciona probabilidades unidireccionales; para una prueba bidireccional, duplique la probabilidad obtenida.

4. Si la probabilidad asociada con el valor observado de Y o valores aún más extremos es igual o menor que α, rechazar H_0. De otro modo, no rechazar H_0.

Potencia-eficacia

Ya que no existe técnica paramétrica aplicable a datos medidos como una variable dicotómica, no tiene sentido inquirir acerca de la potencia-eficacia de la prueba binomial cuando se usa con tales datos.

Si una variable continua es dicotomizada y se emplea la prueba binomial con los datos resultantes, la prueba puede perder datos. En tales casos, la prueba binomial tiene una potencia-eficacia (en el sentido definido en el capítulo 2) de 95 % para $N = 6$, decrementándose a una eficacia asintótica de $2/\pi = 63$ % al incrementarse N. Sin embargo, si los datos son básicamente dicotómicos, aun sabiendo que la variable tiene una distribución subyacente continua, la prueba binomial puede ser la única opción práctica.

Referencias bibliográficas

Para otros detalles acerca de la distribución binomial y sus aplicaciones, consúltese Hays (1981) o Bailey (1971).

PRUEBA JI CUADRADA DE LA BONDAD DE AJUSTE

Función y racionalización

Frecuentemente, en el estudio que un investigador lleva a cabo es necesario conocer el número de sujetos, objetos o respuestas que caen en varias categorías. Por ejemplo, un grupo de pacientes puede ser clasificado de acuerdo con su tipo preponderante de respuestas en la prueba de Rorschach, y el investigador puede predecir que ciertos tipos serán más frecuentes que otros. O los niños pueden ser categorizados de acuerdo con sus modalidades de juego más frecuentes, siendo la hipótesis que esas modalidades diferirán en frecuencia de una manera prescrita. O las personas pueden ser categorizadas con base en si están "en favor de", "indiferentes a" u "opuestas a" una opinión que facilite al investigador probar la hipótesis de que esas respuestas difieren en frecuencia.

La prueba ji cuadrada es adecuada para analizar datos como éstos. El número de categorías puede ser dos o más. La técnica es del tipo de bondad de ajuste en que puede ser usada para probar si existe una diferencia significativa entre un número *observado* de objetos o respuestas que caen en cada categoría y un número *esperado* basado en la hipótesis nula. Es decir, la prueba ji cuadrada evalúa el grado de correspondencia entre las observaciones observadas y esperadas en cada categoría.

Método

Para comparar un grupo de frecuencias observado con uno esperado, debemos ser capaces de establecer qué frecuencias deben ser esperadas. La hipótesis H_0 establece la proporción de objetos que caen en cada una de las categorías en la población supuesta. Esto es, de la hipótesis nula podemos deducir cuáles son las frecuencias esperadas. La técnica ji cuadrada proporciona la probabilidad de que las frecuencias observadas pudieran haber sido muestreadas de una población con los valores esperados proporcionados.

La hipótesis nula H_0 puede probarse mediante el siguiente estadístico:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{3.5}$$

donde

O_i = el número observado de casos en la categoría iésima
E_i = el número esperado de casos en la categoría iésima cuando H_0 es verdadera
k = el número de categorías

Así, la ecuación (3.5) nos indica sumar sobre k categorías el cuadrado de las diferencias entre cada frecuencia observada y esperada, dividido por la frecuencia esperada correspondiente.

Si el acuerdo entre las frecuencias observadas y esperadas es cercano, la diferencia $(O_i - E_i)$ será pequeño y, consecuentemente, X^2 será pequeña. Sin embargo, si la divergencia es grande, el valor de X^2 computado por la ecuación (3.5) también será grande. En términos generales, mientras mayor sea el valor de X^2, menor será la probabilidad de que las frecuencias observadas provengan de la población en la cual están basadas la hipótesis H_0 y las frecuencias esperadas.

Aunque la ecuación (3.5) es útil para entender el estadístico X^2, con frecuencia es molesto de calcular debido al número de sustracciones implicadas. Después de alguna manipulación, se encuentra una fórmula un poco más conveniente:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{3.5}$$

$$= \sum_{i=1}^{k} \frac{O_i^2}{E_i} - N \tag{3.5a}$$

donde N es el número total de observaciones.

Se puede demostrar que la distribución muestral de X^2 según H_0, al ser calculada por la ecuación (3.5), sigue la distribución ji cuadrada[3] con grados de libertad $gl = k - 1$. La noción de grados de libertad se examinará con más detalle posteriormente. La tabla C del Apéndice I contiene la distribución muestral de ji cuadrada y proporciona la probabilidad asociada con ciertos valores. En la parte superior de cada columna en la tabla C están seleccionadas probabilidades de ocurrencia de valores de ji cuadrada cuando H_0 es verdadera. Los valores en cualquier columna con los valores de ji cuadrada que tienen la probabilidad asociada de ocurrencia según H_0 proporcionados en la parte superior de esa columna. Existe un valor diferente de ji cuadrada para cada gl. Por ejemplo, cuando $gl = 1$ y H_0 es verdadera, la probabilidad de observar un valor de ji cuadrada tan grande como 3.84 (o mayor) es 0.05. Esto es, $P[\chi^2 \geq 3.84] = 0.05$.

Existen un número de diferentes valores muestrales para ji cuadrada, uno para cada valor de gl, los grados de libertad. El tamaño de gl refleja el número de "observaciones" que son libres de variar después de que se han colocado ciertas restricciones en los datos. Por ejemplo, si los datos de 50 casos se clasifican en dos categorías, tan pronto como sabemos que, digamos, 35 casos caen en una categoría, sabemos también que 15 deben caer en la otra. Para este ejemplo, $gl = 1$, porque con dos categorías y cualquier valor fijo de N, tan pronto como se ha averiguado el número de casos en una categoría, entonces se determina el número de casos en la otra categoría.

En general, para una prueba de una muestra de bondad de ajuste, cuando H_0

[3] En algunos textos se usa el símbolo griego χ^2 para designar tanto la distribución ji cuadrada y el estadístico X^2. Sin embargo, hay una diferencia. El estadístico X^2 *asintóticamente* tiene una distribución ji cuadrada o χ^2. Nosotros mantendremos una distinción entre el estadístico y su distribución muestral.

especifica completamente las E_i, $gl = k - 1$, donde k es el número de categorías en la clasificación.

Para usar la ji cuadrada a fin de probar una hipótesis en una situación de una muestra de bondad de ajuste, se debe colocar cada observación dentro de cada una de las k celdas. El número total de tales observaciones debe ser N, el número de casos en la muestra. Es decir, cada observación debe ser independiente de cualquier otra; así, no podemos hacer varias observaciones de la misma persona y contarlas como independientes. Hacer esto produce una N "inflada". Para cada una de las k celdas, la frecuencia esperada también debe ser calculada. Si H_0 es que existe una proporción igual de casos en cada categoría de la población, entonces $E_i = N/k$. Con los diferentes valores conocidos de E_i y O_i se pueden calcular los valores de X^2 aplicando la ecuación (3.5). La significancia de este valor obtenido de X^2 puede ser determinada con referencia a la tabla C del Apéndice I. Si la probabilidad asociada con la ocurrencia según H_0 de la X^2 obtenida para $gl = k - 1$ es igual o menor que el valor previamente determinado de α, entonces H_0 puede ser rechazada. En caso contrario, H_0 no puede rechazarse.

Ejemplo. Los aficionados a las carreras de caballos sostienen que en una carrera alrededor de una pista circular, los caballos tienen ventajas significativas acumuladas al ser colocados en ciertas posiciones. Cualquier posición del caballo se asigna en la línea de salida. La posición 1 es la más cercana al carril del interior de la pista; la posición 8 está en el extremo, más alejada del carril en una carrera de ocho caballos. Podemos probar el efecto de la posición analizando los resultados de la carrera, proporcionados de acuerdo con la posición, durante el primer mes de la temporada, en una pista circular en particular.[4]

i. *Hipótesis nula.* H_0: no existe diferencia en el número esperado de ganadores comenzando en cada una de las posiciones, y cualesquiera diferencias observadas son meramente variables casuales que pueden esperarse en una muestra aleatoria de una distribución uniforme. H_1: las frecuencias teóricas no son iguales.
ii. *Prueba estadística.* Ya que estamos comparando los datos de una muestra con alguna población supuesta, la prueba ji cuadrada de bondad de ajuste es apropiada. Se elige la prueba ji cuadrada debido a que la hipótesis que se va a probar concierne a la comparación de frecuencias observadas y esperadas en categorías discretas. En este ejemplo, las categorías comprenden las ocho posiciones.
iii. *Nivel de significación.* Sea $\alpha = 0.01$ y $N = 144$, el número total de ganadores en 18 días de carreras.
iv. *Distribución muestral.* La distribución muestral del estadístico X^2 calculado de la ecuación (3.5) sigue la distribución ji cuadrada con $gl = k - 1 = 8 - 1 = 7$.
v. *Región de rechazo.* H_0 será rechazada si el valor observado de X^2 es tal que la probabilidad asociada con el valor calculado según H_0 para $gl = 7$ es ≤ 0.01.
vi. *Decisión.* La muestra de 144 ganadores rindieron los datos que se presentan en la tabla 3.2. Las frecuencias observadas de ganadores están ubicadas en el centro de cada celda; las frecuencias esperadas están en cursivas en la esquina de cada celda. Por ejemplo, 29 ganadores resultaron de caballos colocados en la posición 1, mientras que según H_0 deberían haber sido esperados sólo 18 ganadores. Sólo resultaron 11 ganadores de caballos colocados en la posición 8, mientras que según H_0 deberían haber sido 18.

[4] Estos datos fueron publicados en el *New York Post*, el 30 de agosto de 1955, página 42.

Tabla 3.2. Resultados de los caballos ganadores, de acuerdo con ocho posiciones, en una pista circular.

	Posiciones								
	1	2	3	4	5	6	7	8	*Total*
Número de ganadores Esperados	29 *18*	19 *18*	18 *18*	25 *18*	17 *18*	10 *18*	15 *18*	11 *18*	144

El cálculo de X^2 es directo:

$$X^2 = \sum_{i=1}^{8} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(29-18)^2}{18} + \frac{(19-18)^2}{18} + \frac{(18-18)^2}{18}$$

$$+ \frac{(25-18)^2}{18} + \frac{(17-18)^2}{18} + \frac{(10-18)^2}{18}$$

$$+ \frac{(15-18)^2}{18} + \frac{(11-18)^2}{18}$$

$$= \frac{121}{18} + \frac{1}{18} + 0 + \frac{49}{18} + \frac{1}{18} + \frac{64}{18} + \frac{9}{18} + \frac{49}{18}$$

$$= 16.3$$

La tabla C del Apéndice I muestra que $P[X^2 \geq 16.3]$ para $gl = 7$ tiene una probabilidad de ocurrencia entre $p = 0.05$ y $p = 0.02$. Esto es, $0.05 > p > 0.02$. Ya que esta probabilidad es más grande que el nivel de significación establecido previamente, $\alpha = 0.01$, no podemos rechazar H_0 en ese nivel de significación. Notamos que la hipótesis nula podría haber sido rechazada si hubiéramos colocado $\alpha = 0.05$. Podría parecer que son necesarios más datos antes de que se llegue a cualquier conclusión definitiva concerniente a H_1.

Ejemplo. Un investigador aplica una prueba de vocabulario a un grupo de niños de $N = 103$. Con base en una investigación previa y la teoría que subyace a la prueba, la distribución de las puntuaciones debería tener una distribución normal. La media de la muestra fue de 108 y la distribución estándar 12.8. Con el propósito de aplicar la prueba ji cuadrada de la bondad de ajuste para una muestra, se deben definir las categorías y determinar las frecuencias esperadas. Escogemos $k = 10$ intervalos para las frecuencias. Los valores de corte (denotados X_{corte}) corresponderán a los deciles de la distribución normal con media y desviación estándar proporcionadas por los datos.

Los deciles de la distribución normal unitaria (denotados por z_{corte}) puede obtenerse de la tabla A del Apéndice I:

Categoría	Z_{corte}	*p acumulativa*	X_{corte}
1	−1.2816	0.10	91.60
2	−0.8416	0.20	97.23
3	−0.5244	0.30	101.29
4	−0.2534	0.40	104.76
5	0.0000	0.50	108.00
6	0.2534	0.60	111.24
7	0.5244	0.70	114.71
8	0.8418	0.80	118.77
9	1.2816	0.90	124.40
10	∞	1.00	*Ningún límite*

Estos valores deben ser transformados a los puntos de corte en la distribución observada. Esto puede hacerse mediante la siguiente fórmula general:

$$X_{\text{corte}} = \overline{X} + s_x z_{\text{corte}} \qquad \text{en general}$$

y

$$X_{\text{corte}} = 108 + 12.8 z_{\text{corte}} \qquad \text{para este ejemplo}$$

Para el problema dado, estos valores están resumidos en la tabla anterior. Así, si un dato observado es menor que 91.60, puede ser contado en la categoría 1, mientras que si el dato observado fuera 103, podría ser contado en la categoría 4. El investigador clasificó todas las puntuaciones en categorías y obtuvo las siguientes frecuencias: 8, 10, 13, 15, 10, 14, 12, 8, 7, 6. La frecuencia esperada en cada categoría es $N/k = 103/10 = 10.3$. El investigador desea probar la hipótesis usando $\alpha = 0.05$. El valor obtenido de X^2 es

$$\begin{aligned}
X^2 &= \frac{(8-10.3)^2}{10.3} + \frac{(10-10.3)^2}{10.3} + \frac{(13-10.3)^2}{10.3} \\
&\quad + \frac{(15-10.3)^2}{10.3} + \frac{(10-10.3)^2}{10.3} + \frac{(14-10.3)^2}{10.3} \\
&\quad + \frac{(12-10.3)^2}{10.3} + \frac{(8-10.3)^2}{10.3} + \frac{(7-10.3)^2}{10.3} + \frac{(6-10.3)^2}{10.3} \\
&= 8.36
\end{aligned}$$

Al calcular los valores esperados, usamos dos partes de información de la muestra. Esto se debe a que no podemos especificar las probabilidades asociadas con una distribución normal sin estimar la media y la desviación estándar (o varianza) de la población usando los datos de la muestra. Para cada parámetro estimado de los datos otorgamos un grado de libertad. Para este ejemplo, el número de parámetros estimados fue $n_p = 2$. Así, los gl para la distribución ji cuadrada son $gl = k - n_p - 1 = 10 - 2 - 1 = 7$. Ahora, al probar H_0 en el nivel 0.05, el valor crítico de X^2 es 14.07. Ya que el valor obtenido de X^2 fue 8.36, no podemos rechazar la hipótesis H_0 de que los datos fueron muestreados de una población normal.

Frecuencias esperadas pequeñas

Cuando $gl = 1$, esto es, cuando $k = 2$, cada frecuencia esperada debe ser por lo menos de 5. Cuando $gl > 1$, es decir, cuando $k > 2$, la prueba de la bondad de ajuste de una muestra ji cuadrada no debería ser usada si más del 20 % de las frecuencias esperadas son menores de 5, o cuando cualquier frecuencia esperada es menor que 1. Esto se debe a que la distribución muestral de X^2 es sólo asintóticamente ji cuadrada, es decir, la distribución muestral de X^2 es la misma que la distribución ji cuadrada al volverse más grandes las frecuencias esperadas (infinitas). Para propósitos prácticos, la aproximación es buena cuando las frecuencias esperadas son mayores que 5. Cuando las frecuencias esperadas son pequeñas, las probabilidades asociadas con la distribución ji cuadrada pudieran no ser lo suficientemente cercanas a las probabilidades en la distribución muestral de X^2 para poder hacer inferencias apropiadas. Las frecuencias esperadas algunas veces pueden ser incrementadas al combinar categorías adyacentes dentro de una categoría combinada. Esto es deseable sólo si las combinaciones de las categorías pueden hacerse de manera significativa (y, naturalmente, si para empezar existen más de dos categorías).

Por ejemplo, una muestra de gente puede ser categorizada conforme a si su respuesta ante un juicio de opinión es "fuertemente de acuerdo", "de acuerdo", "indiferente", "opuesto" o "fuertemente opuesto". Para incrementar las E_i, las categorías adyacentes pudieran ser combinadas, y la gente categorizada como "apoyo", "indiferente" u "opuesto", o posiblemente como "apoyo", "indiferente" y "fuertemente opuesto". Sin embargo, si las categorías son combinadas, se advierte que los significados de los rótulos asignados a las categorías restantes pueden ser diferentes de los significados originales.

Si empezamos con sólo dos categorías y tenemos una frecuencia esperada menor de 5, o si después de combinar categorías adyacentes finalizamos con sólo dos categorías y aún tenemos una frecuencia esperada menor de 5, entonces se puede usar la prueba binomial (véase la sección correspondiente) en lugar de la prueba ji cuadrada, para determinar la probabilidad asociada con la ocurrencia de las frecuencias observadas según H_0.

Resumen del procedimiento

En esta descripción del método para usar la prueba de la bondad de ajuste ji cuadrada para el caso de una muestra, hemos señalado que el procedimiento para usar la prueba incluye estos pasos:

1. Coloque las frecuencias observadas dentro de k categorías. La suma de las frecuencias debe ser N, el número de observaciones independientes.
2. A partir de H_0, determine las frecuencias esperadas (las E_i) para cada una de las k celdas. Cuando $k > 2$, y más del 20 % de las E_i son menores que 5, combínense categorías adyacentes cuando esto sea razonable, reduciendo por tanto el valor de k e incrementando los valores de algunas de las E_i. Cuando $k = 2$, la prueba de la bondad de ajuste ji cuadrada para una muestra es exacta sólo si cada frecuencia esperada es 5 o más grande.
3. Use la ecuación (3.5) para computar el valor de X^2.

4. Determine los grados de libertad, $gl = k - n_p - 1$, donde n_p es el número de parámetros estimados de los datos y usados al calcular las frecuencias esperadas.
5. Por referencia a la tabla C del Apéndice I, determine la probabilidad asociada con X^2 según H_0 como un valor tan grande como el valor observado para X^2 para los grados de libertad apropiados para los datos. Si la probabilidad es menor que o igual a α, rechace H_0.

Potencia

Ya que esta prueba es la más usada cuando no tenemos claramente una alternativa disponible, por lo general no estamos en una posición para computar la potencia exacta de la prueba. Cuando se usa la medición nominal o categórica o cuando los datos consisten en frecuencias en categorías inherentemente discretas, entonces la noción de potencia-eficacia no es significativa y en tales casos no existe una prueba paramétrica que sea adecuada.

En los casos en los que se ha estudiado la potencia de la prueba de la bondad de ajuste ji cuadrada, existe una interacción entre el número de categorías k y el número de observaciones N. Aunque las recomendaciones específicas dependen de la distribución teórica que va a ser ajustada, las siguientes reglas resultan adecuadas:

1. Elíjanse categorías y límites de intervalo tales que las frecuencias esperadas sean iguales a N/k.
2. Se debe escoger el número de categorías tal que las frecuencias esperadas estén entre 6 y 10, con el valor más bajo apropiado para N grandes (mayor que 200).

También debe considerarse que cuando $gl > 1$, la prueba ji cuadrada es insensible a los efectos del ordenamiento de categorías y, por tanto, cuando una hipótesis tiene el orden en cuenta, la prueba ji cuadrada no puede ser la mejor prueba. Para métodos que fortalecen las pruebas ji cuadrada comunes cuando H_0 es probada contra alternativas específicas, véase Cochran (1954) o Everitt (1977). En el apartado "Potencia" de la siguiente sección se proporciona mayor información concerniente a la prueba de la bondad de ajuste ji cuadrada.

Referencias bibliográficas

En Cochran (1954), Dixon y Massey (1983), McNemar (1969) y Everitt (1977) se encuentran análisis útiles de la prueba de la bondad de ajuste ji cuadrada.

LA PRUEBA DE KOLMOGOROV-SMIRNOV DE UNA MUESTRA

Función y racionalización

La prueba de Kolmogorov-Smirnov de una muestra es otra prueba de la bondad de ajuste. Es decir, está interesada en el grado de acuerdo entre la distribución de un conjunto de valores muestreados (puntuaciones observadas) y alguna distribución teórica específica. Esta prueba determina si las puntuaciones en una muestra pueden razonablemente provenir de una población que tiene una distribución teórica.

Brevemente, la prueba incluye la especificación de la distribución de frecuencias *acumuladas* que pudieran ocurrir dada la distribución teórica y comparándola con la distribución de frecuencias acumuladas observadas. La distribución teórica representa lo que podría ser esperado según H_0. La prueba permite mostrar en estas dos distribuciones, la teórica y la observada, la mayor divergencia. La referencia a la distribución muestral indica si una divergencia tan grande es probable que ocurra sobre la base del azar. Esto es, la distribución muestral indica la probabilidad de que una divergencia de la magnitud observada pudiera ocurrir si las observaciones fueran realmente una muestra aleatoria de una distribución teórica. La prueba de Kolmogorov-Smirnov supone que la distribución de las variables subyacentes que van a ser probadas es continua, como es especificada por la distribución de frecuencias acumuladas. Así, la prueba es adecuada para probar la bondad de ajuste para variables que son medidas en al menos una escala ordinal.

Método

Sea $F_0(X)$ una función de distribución de frecuencias relativas acumuladas completamente especificada por la distribución teórica según H_0. Esto es, para cualquier valor de X, el valor de $F_0(X)$ es la proporción de casos esperados que tienen puntuaciones iguales o menores que X.

Sea $S_N(X)$ la distribución de frecuencias relativas acumuladas observadas de una muestra aleatoria de N observaciones. Si X_i es una puntuación posible, entonces $S_N(X_i) = F_i/N$, donde F_i es el número de observaciones que son iguales o menores que X_i. $F_0(X_i)$ es la proporción esperada de observaciones que son menores o iguales a X_i.

Ahora, según la hipótesis nula de que la muestra ha sido extraída de la distribución teórica especificada, se espera que para cada valor X_i, $S_N(X_i)$ sea ligeramente cercano a $F_0(X_i)$. Esto es, cuando H_0 es verdadera, podemos esperar que las diferencias entre $S_N(X_i)$ y $F_0(X_i)$ sean pequeñas y dentro de los límites del error aleatorio. La prueba de Kolmogorov-Smirnov se enfoca sobre las desviaciones *más grandes*. El valor absoluto más grande de $F_0(X_i) - S_N(X_i)$ se llama *máxima desviación de D*:

$$D = \max | F_0(X_i) - S_N(X_i) | \quad i = 1, 2, \ldots, N \tag{3.6}$$

La distribución muestral de D según H_0 es conocida. La tabla F del Apéndice I proporciona ciertos valores críticos para esa distribución muestral. Nótese que la significancia de un valor dado D depende de N.

Por ejemplo, supóngase que al aplicar la ecuación (3.6) se encuentra, que $D = 0.325$ cuando $N = 15$. La tabla F del Apéndice I muestra que la probabilidad de $D \geq 0.325$ está entre 0.05 y 0.10.

Si N es más grande que 35, los valores críticos de D pueden ser determinados en la última hilera de la tabla F del Apéndice I. Por ejemplo, supóngase que un investigador tiene una muestra de tamaño $N = 43$ y elige $\alpha = 0.05$. La tabla F muestra que cualquier $D \geq 1.36/ N$ será significativo. Esto es, cualquier D, como es definida por la ecuación (3.6), que es igual o más grande que $1.36/ 43 = 0.207$, será significativa en el nivel 0.05 (prueba bidireccional).

Ejemplo. Durante los últimos años los investigadores han estado estudiando la duración de una variedad de eventos tales como trabajos, huelgas y guerras. Como parte de tal investigación, las suposiciones concernientes a acciones individuales y el curso de los acontecimientos, ha conducido a modelos matemáticos de los mismos que hacen predicciones acerca de su distribución.[5] Ya que los detalles de los modelos matemáticos no son de especial interés en esta obra, la evaluación del acuerdo entre los datos y las predicciones del modelo proporciona una buena ilustración de la prueba de la bondad de ajuste Kolmogorov-Smirnov de una muestra. Los datos concernientes a la duración de las huelgas que empezaron en 1965 en el Reino Unido fueron recabados, analizados y se hicieron predicciones con el uso del modelo matemático. La tabla 3.3 contiene la distribución de frecuencias acumuladas para las $N = 840$ duraciones de huelga. También se proporcionan en la tabla ias frecuencias acumuladas predichas por el modelo matemático.

i. *Hipótesis nula,* H_0: la distribución de las duraciones de huelga sigue las predicciones del modelo matemático. Es decir, la diferencia entre las duraciones de huelga observadas y predichas no excede la diferencia que podría ser esperada si ocurrieran al azar. H_1: las duraciones de huelga observadas no coinciden con aquellas predichas por el modelo matemático.

ii. *Prueba estadística.* Se elige la prueba de Kolmogorov-Smirnov de una muestra debido a que el investigador desea comparar una distribución de puntuaciones observadas de una escala ordinal con una distribución teórica de puntuaciones.

iii. *Nivel de significación.* Sea $\alpha = 0.05$ y N es el número de huelgas que empezaron en el Reino Unido en $1965 = 840$.

iv. *Distribución muestral.* Los valores críticos de D, la desviación máxima absoluta entre las distribuciones acumulativas observadas y predichas, están presentados en la tabla F del Apéndice I, junto con sus probabilidades asociadas de ocurrencia cuando H_0 es verdadera.

v. *Región de rechazo.* La región de rechazo consista de todos los valores de D [computados de la ecuación (3.6)], que son tan grandes que la probabilidad asociada con su ocurrencia cuando H_0 es verdadera, es menor o igual a $\alpha = 0.05$.

vi. *Decisión.* En este estudio, la diferencia entre la distribución de frecuencias relativas acumuladas observadas $S_N(X)$ y la distribución de frecuencias relativas acumuladas predichas $F_0(X)$ es calculada. Estas diferencias se resumen en la tabla 3.3. El valor de D, la diferencia máxima entre las frecuencias acumuladas, es $F_0(X) - S_N(X) = 510.45/840 - 523/840 = 0.015$. Ya que $N > 35$, debemos usar la aproximación de muestras grandes. Con $N = 840$ el valor crítico de D es $1.36/ 840 = 0.047$. Puesto que el valor observado de D, 0.015, es menor que el valor crítico, no podemos rechazar H_0, la hipótesis de que los datos observados provienen de una población especificada por el modelo teórico resumido en la tabla 3.3.

[5] Morrison, D. G. y Schmittlein, D. C. "Jobs, strikes, and wars: Probability models for duration", en *Organizational Behavior and Human Performance*, núm. 25, 1980, págs. 224-251.

Tabla 3.3. Datos de huelgas en el Reino Unido ($N = 840$).

Duración máxima (días)	*Frecuencia acumulada*		*Frecuencia acumulada relativa*		
	Observada	*Predicha*	*Observada*	*Predicha*	$\lvert F_0(X) - S_N(X)\rvert$
1 – 2	203	212.81	0.242	0.253	0.011
2 – 3	352	348.26	0.419	0.415	0.004
3 – 4	452	442.06	0.538	0.526	0.012
4 – 5	523	510.45	0.623	0.608	0.015
5 – 6	572	562.15	0.681	0.669	0.012
6 – 7	605	602.34	0.720	0.717	0.003
7 – 8	634	634.27	0.755	0.755	0.000
8 – 9	660	660.10	0.786	0.786	0.000
9 – 10	683	681.32	0.813	0.811	0.002
10 – 11	697	698.97	0.830	0.832	0.002
11 – 12	709	713.82	0.844	0.850	0.006
12 – 13	718	726.44	0.855	0.865	0.010
13 – 14	729	737.26	0.868	0.878	0.010
14 – 15	744	746.61	0.886	0.889	0.003
15 – 16	750	754.74	0.893	0.899	0.006
16 – 17	757	761.86	0.901	0.907	0.006
17 – 18	763	768.13	0.908	0.914	0.006
18 – 19	767	773.68	0.913	0.921	0.008
19 – 20	771	778.62	0.918	0.927	0.009
20 – 25	788	796.68	0.938	0.948	0.010
25 – 30	804	807.86	0.957	0.962	0.005
30 – 35	812	815.25	0.967	0.971	0.004
35 – 40	820	820.39	0.976	0.977	0.001
40 – 50	832	826.86	0.990	0.984	0.006
> 50	840	840.01	1.000	1.000	0.000

Resumen del procedimiento

En la aplicación de la prueba de Kolmogorov-Smirnov, los pasos son los siguientes:

1. Especifique la distribución acumulativa teórica, esto es, la distribución acumulativa esperada según H_0.
2. Arregle las puntuaciones observadas en una distribución acumulativa y convierta las frecuencias acumulativas en frecuencias relativas acumuladas

[$S_N(X_i)$]. Para cada intervalo, encuéntrense las frecuencias relativas acumuladas esperadas $F_0(X_i)$.

3. Con el uso de la ecuación (3.6), encuentre D.
4. Con base en la tabla F del Apéndice I, encuentre la probabilidad asociada (dos colas) con la ocurrencia según H_0 de valores tan grandes como los valores observados de D. Si esta probabilidad es igual o menor que α, se debe rechazar H_0.

Potencia

La prueba de la bondad de ajuste de Kolmogorov-Smirnov de una muestra trata las observaciones individuales por separado y, por tanto, a diferencia de la prueba ji cuadrada ya examinada, no necesariamente pierde información al hacer la combinación de categorías, aunque puede ser conveniente usar agrupaciones de variables. Cuando las muestras son pequeñas y las categorías adyacentes deben combinarse para usar adecuadamente el estadístico X^2, la prueba ji cuadrada es definitivamente menos potente que la prueba de Kolmogorov-Smirnov. Más aún, para muestras muy pequeñas, la prueba ji cuadrada no puede ser usada, pero la prueba de Kolmogorov-Smirnov sí. Estos hechos sugieren que esta última puede ser en todos los casos más potente que su prueba alternativa, la ji cuadrada.

Sin embargo, es posible que las pruebas rindan resultados similares, particularmente cuando el tamaño de la muestra es grande. Si aplicamos la prueba de Kolmogorov-Smirnov a los datos de las carreras de caballos que hemos examinado, encontramos que $D = \text{máx } S_N(X) - F_0(X) = 91/144 - 72/144 = 0.132$. Si probamos en $\alpha = 0.05$, entonces podemos rechazar H_0 si $D \geq 1.36/\ 144 = 0.113$. Como con la prueba ji cuadrada, podemos rechazar H_0.

La prueba ji cuadrada supone que las distribuciones son nominales, mientras que la prueba de Kolmogorov-Smirnov supone una distribución continua. En principio, ambas pruebas pueden aplicarse a datos ordinales; sin embargo, el agrupamiento que es necesario para la aplicación de la prueba ji cuadrada la hace menos precisa que la prueba de Kolmogorov-Smirnov.

La elección entre ellas no es fácil. Es difícil comparar la potencia de las dos pruebas debido a que cada una de ellas depende de diferentes cantidades. Cuando pueda aplicarse cualquier prueba, la elección depende de la facilidad de computación o de otra preferencia. Sin embargo, con muestras pequeñas, la prueba de Kolmogorov-Smirnov es exacta, mientras que la prueba de la bondad de ajuste ji cuadrada es sólo aproximadamente (asintóticamente) exacta. En tales casos, la preferencia debe darse a la prueba de Kolmogorov-Smirnov.

Referencias bibliográficas

En Gibbons (1976) y Hays (1981) se encuentran detalles de la prueba de Kolmogorov-Smirnov y otras pruebas de la bondad de ajuste.

PRUEBA PARA EVALUAR LA SIMETRÍA DE LA DISTRIBUCIÓN

Función y racionalización

Las pruebas que hemos examinado en este capítulo han tratado con dos aspectos de una distribución. La prueba binomial trata con la cuestión de si los datos dicotómicos pueden razonablemente ser generados por una distribución binomial hipotética. Las siguientes dos pruebas consideraron el ajuste de una distribución empírica a una distribución hipotética. Otro tipo de hipótesis acerca de un conjunto de datos puede ser acerca de la *forma* de una distribución. La prueba descrita en esta sección es una prueba para evaluar la simetría de la distribución. Esto es, ¿podemos inferir que un conjunto de datos fue generado por una distribución desconocida, pero *simétrica*? La hipótesis H_0 es que las observaciones fueron extraídas de la misma distribución simétrica con una mediana desconocida. La hipótesis alterna es que la distribución no es simétrica.

La prueba incluye el examen de subconjuntos de tres variables (o triadas) para determinar la probabilidad de que la distribución sea sesgada a la izquierda o a la derecha, así como una ligera cantidad de cómputo, pero es relativamente directa.

Método

Para aplicar la prueba debe ser examinado y codificado cada subconjunto de tamaño tres de la muestra. Cada tríada X_i, X_j, X_k se codifica como un tríada derecha o izquierda (o como ninguna). Aunque es posible clasificar las tríadas por inspección, se dará una especificación más formal. La siguiente tabla proporciona el código para las tríadas:

Tríada derecha	X — X ———— X	$(X_i + X_j + X_k/3 > \text{med}\ (X_i, X_j, X_k)$
Tríada izquierda	X ———— X — X	$(X_i + X_j + X_k/3 < \text{med}\ (X_i, X_j, X_k)$
Ninguna	X ——— X ——— X	$(X_i + X_j + X_k/3 = \text{med}\ (X_i, X_j, X_k)$

Cada una de los $N(N - 1)(N - 2)/6$ posibles tríadas deben ser codificadas como izquierdas, derechas o ninguna. El estadístico de interés es

$$T = \#\ \text{tríadas derechas} - \#\ \text{tríadas izquierdas} \tag{3.7}$$

Ahora, cuando H_0 es verdadera, $\mu_T = 0$, esto es, las X son simétricas alrededor de la mediana. Para completar la prueba, necesitamos definir los siguientes estadísticos:

$$B_i = \#\ \text{de tríadas derechas que incluyen } X_i$$
$$- \#\ \text{de tríadas izquierdas que incluyen } X_i$$

$$B_{jk} = \#\ \text{de tríadas derechas que incluyen tanto } X_j \text{ como } X_k$$
$$- \#\ \text{de tríadas izquierdas que incluyen tanto } X_j \text{ como } X_k$$

Entonces H_0 puede ser probada usando el estadístico $z = T/\sigma_T$, donde

$$\sigma_T^2 = \frac{(N-3)(N-4)}{(N-1)(N-2)} \sum_{i=1}^{N} B_i^2 + \frac{N-3}{N-4} \sum_{1 \leq j < k \leq N} B_{jk}^2 + \frac{N(N-1)(N-2)}{6}$$

$$- \left[1 - \frac{(N-3)(N-4)(N-5)}{N(N-1)(N-2)} \right] T^2 \qquad (3.8)$$

El estadístico z está distribuido normalmente de manera asintótica con media cero y varianza uno. La significancia de z puede determinarse usando la tabla A del Apéndice I, y el valor crítico para una prueba bidireccional usando $\alpha/2$. Comparándola con procedimientos alternativos, esta prueba es satisfactoria para N mayor que 20; es decir, mantiene el nivel de significación elegido al mismo tiempo que conserva una buena potencia para detectar distribuciones asimétricas.

Ejemplo. En un estudio de supresión de la sal,[6] los sujetos probaron una mezcla de sal y sacarosa con el propósito de hacer un escalamiento de juicios de salinidad, como una función de la concentración de sal en la solución. Hubo diferencias individuales sustanciales en los juicios acerca de la salinidad. El interés de la investigación era la distribución de los juicios de salinidad. Se usaron cuatro diferentes concentraciones y se asignaron sujetos separados a cada una de ellas. Los datos están resumidos en la tabla 3.4. Con el propósito de ilustrar la prueba que evalúa la simetría de la distribución, se analizarán los datos para una razón de 0.5 de concentración salina.

Tabla 3.4. Juicio de salinidad para un nivel de concentración de sal.

13.53
28.42
48.11
48.64
51.40
59.91
67.98
79.13
103.05

[6] Kroeze, J. H. A., "The influence of relative frequencies of pure and mixed stimuli on mixture suppression in taste", en *Perception & Psychophysics*, núm. 31, 1982, págs. 276-278.

i. *Hipótesis nula.* H_0: la distribución de los juicios de salinidad es simétrico. La hipótesis alterna es que la distribución de los juicios es asimétrica. Esto es, la hipótesis nula es que las desviaciones de la simetría son tales que podría esperarse que ocurrieran al azar.

ii. *Prueba estadística.* El número de observaciones es $N = 9$. (En términos estrictos, nuestro ejemplo viola la recomendación de que la prueba es apropiada cuando $N > 20$. Se escogió un ejemplo con muestra pequeña para ilustrar el procedimiento.) El primer paso incluye calcular las tríadas y determinar si son tríadas derechas, tríadas izquierdas o ninguna de ellas. El número total de tríadas para $N = 9$ es $N(N - 1)(N - 2)/6 = 84$. Para los primeros tres puntos (13.53, 28.42, 48.11) la mediana es 28.42 y la media es 30.03. Ya que la media es más grande que la mediana, la tríada (X_1, X_2, X_3) se clasifica como una tríada derecha. La tríada (X_1, X_3, X_4) es una tríada izquierda, ya que la mediana es 48.11 y es más grande que la media $(13.53 + 48.11 + 48.64)/3 = 36.76$. El número de tríadas derechas es 44 y el de tríadas izquierdas es 40. Así, el valor de T es $44 - 40 = 4$.
En seguida se debe encontrar la varianza de T. Para esto, deben calcularse las cantidades intermedias B_i y B_{jk}. Después estas cantidades se emplean en la ecuación (3.8) para determinar la varianza. (Las dos sumas de cuadrados de B_i y B_{jk} son 320 y 364, respectivamente.) La varianza es, entonces, 680.04. Finalmente, se calcula el estadístico $z = T/\sigma_T = 4/\ 680.04 = 0.154$.

iii. *Nivel de significación y decisión.* Sea $\alpha = 0.05$. El nivel de significación para z puede determinarse con referencia a la tabla A del Apéndice I, la tabla de la distribución normal unitaria. No podemos rechazar la hipótesis de simetría en un nivel de significancia de 0.05 (e incluso en una mayor).

Debe recordarse que la prueba es razonablemente buena para $N \geq 20$. Al incrementarse el tamaño de la muestra, el cómputo de las tríadas, aunque es directo, consume relativamente un poco más de tiempo. Por tanto, esta técnica quizá se use mejor cuando se dispone de un algoritmo de computación. El programa 1 (véase el Apéndice II) proporciona el código para un programa general para computar T y σ_T en cualquier tamaño de muestra. Para este estadístico, se recomienda el uso de un programa.

Resumen del procedimiento

Los siguientes son los pasos en la aplicación de la prueba de simetría para una secuencia de observaciones:

1. Para cada subconjunto de tamaño 3 en la secuencia de observaciones, determine si es una tríada derecha o izquierda (o ninguna).
2. Calcule las cantidades B_i y B_{jk} para cada variable X_i y par de variables X_j y X_k.
3. Calcule T, el número de tríadas derechas menos el número de tríadas izquierdas, y la varianza de T usando la ecuación (3.8).
4. Pruebe H_0 usando el estadístico $z = T/\sigma_T$, que está distribuido normalmente de manera asintótica con media 0 y desviación estándar 1. La significancia de T puede encontrarse usando la tabla A del Apéndice I. Ya que la hipótesis alternativa es de dos colas, el valor crítico de T se determina usando $\alpha/2$. Debido al número relativamente grande de cálculos implica-

dos, es conveniente emplear un programa de computadora como el programa 1 del Apéndice II.

Potencia

La potencia de la prueba de la simetría se ha estudiado por medio de procedimiento Monte Carlo con el uso de un gran número de muestras simuladas de varias distribuciones. Con base en tales estudios, la prueba tiene una potencia razonable para muestras mayores que 20. Se han propuesto otras pruebas, pero la mayoría de ellas tienen muy poca potencia.

Referencias bibliográficas

Existen varias pruebas para evaluar la simetría de la distribución. La única que se presenta aquí es de Randles, Fligner, Policello y Wolfe (1980).

LA PRUEBA DE UNA MUESTRA DE SERIES ALEATORIAS

Función y racionalización

Si un investigador desea llegar a alguna conclusión acerca de una población usando la información contenida en una muestra extraída de esa población, entonces la muestra debe ser aleatoria; es decir, las observaciones sucesivas deben ser independientes. Se han desarrollado varias técnicas para facilitarnos probar la hipótesis de que una muestra es aleatoria. Estas técnicas están basadas en el *orden* o la *secuencia* en que se obtuvieron originalmente las puntuaciones u observaciones individuales.

Las técnicas que presentaremos están basadas en el número de series que exhibe una muestra. Una *serie* se define como una sucesión de símbolos idénticos que son seguidos y precedidos por diferentes símbolos o por ningún símbolo.

Por ejemplo, supóngase una serie de eventos binarios (indicados por signos más y menos), ocurrieron en este orden:

$$+\ +\ -\ -\ -\ +\ -\ -\ -\ -\ +\ +\ -\ +$$

Esta muestra de puntuaciones empieza con una serie de dos más. Sigue una serie de tres menos. Después viene otra serie que consiste en un más. Es seguida por una serie de cuatro menos, después viene una serie de dos más, etc. Podemos agrupar estas puntuaciones en series subrayando y numerando cada sucesión de símbolos idénticos:

$$\underset{1}{\underline{+\ +}}\quad \underset{2}{\underline{-\ -\ -}}\quad \underset{3}{\underline{+}}\quad \underset{4}{\underline{-\ -\ -\ -}}\quad \underset{5}{\underline{+\ +}}\quad \underset{6}{\underline{-}}\quad \underset{7}{\underline{+}}$$

Observamos varias series en el proceso: r es el número de series $= 7$.

El número total de series en una muestra de cualquier tamaño dado, proporciona una indicación de si esa muestra es o no aleatoria. Si ocurren muy pocas series, se sugiere una tendencia en el tiempo o alguna agrupación que carece de independencia. Si ocurren muchas series, es posible suponer que ciertas fluctuaciones sistemáticas cíclicas de corta duración parecen estar influyendo en las puntuaciones.

Por ejemplo, supóngase que una moneda se lanzó al aire 20 veces y se observó la siguiente secuencia de caras (C) y cruces (X):

C C C C C C C C C C X X X X X X X X X X

Sólo dos series ocurrieron en 20 lanzamientos. Esto podría ser demasiado poco para una moneda "normal" (o para un lanzador normal). Esto sugiere alguna carencia de independencia en los eventos. Por otra parte, supongamos que ha ocurrido la siguiente secuencia:

C X C X C X C X C X C X C X C X C X C X

Aquí se observan demasiadas series. En este caso, con $r = 20$, cuando $N = 20$, también podría parecer razonable rechazar la hipótesis de que la moneda es "normal". Ninguna de las secuencias anteriores parece ser una serie aleatoria de C y X. Esto es, las observaciones sucesivas no parecen ser independientes.

Nótese que nuestro análisis, que está basado en el *orden* de los eventos, nos proporciona información que no está indicada por la *frecuencia* de los eventos. En los anteriores casos ejemplificados, ocurrieron 10 cruces y 10 caras. Si las puntuaciones se analizaran de acuerdo con sus frecuencias, por ejemplo, usando la prueba ji cuadrada o la prueba binomial, no tendríamos razón para sospechar de la "normalidad" de la moneda. Es sólo una prueba de series, que se centra en el orden de los eventos, que revela fuerte carencia de aleatorización de las puntuaciones y, por tanto, la posible carencia de "normalidad" en la moneda.

La distribución muestral de los valores de r que podríamos esperar de muestras aleatorias repetidas, es conocida. Usando esta distribución muestral, podemos decidir si una muestra observada tiene más o menos series que las que podrían esperarse que ocurrieran al azar en una muestra aleatoria.

Método

Sea m el número de elementos de una clase, y n el número de elementos de la otra clase en una secuencia de $N = m + n$ eventos binarios. Esto es, m puede ser el número de caras y n el número de cruces en una serie de lanzamientos de moneda; o m puede ser el número de signos "más" y n el número de signos "menos" en una serie de respuestas a un cuestionario.

Para usar la prueba de una muestra de series, primero observe los eventos m y n en la secuencia en la cual ocurrieron y determine el valor de r, el número de series.

Muestras pequeñas

Si tanto m como n son menores o iguales que 20, entonces la tabla G del Apéndice I proporciona los valores críticos de r según H_0 para $\alpha = 0.05$. Éstos son valores críticos para la distribución muestral de r según H_0 cuando se supone que la secuencia es aleatoria. Si el valor observado de r cae entre los valores críticos, no podemos rechazar H_0. Si el valor observado de r es igual o más extremo que uno de los valores críticos, rechazamos H_0.

Hay dos entradas para cada valor de m y n en la tabla G del Apéndice I. La primera entrada proporciona el máximo de aquellos valores de r que son tan *pequeños* que la probabilidad asociada con su ocurrencia según H_0 es $p = 0.025$ o menos. La segunda entrada proporciona el mínimo de los valores de r que son tan *grandes* que la probabilidad asociada con su ocurrencia según H_0 es $p = 0.025$ o menos.

Cualquier valor observado de r que es *igual* o *menor* que el valor superior mostrado en la tabla G, o es *igual* o *más grande* que el valor inferior mostrado en la misma tabla, está en la región de rechazo para $\alpha = 0.05$.

Por ejemplo, en el primer lanzamiento de moneda que hemos mencionado, observamos dos series: la serie de 10 caras seguida por la serie de 10 cruces. Aquí $m = 10$, $n = 10$ y $r = 2$. En la tabla G del Apéndice I se muestra que para estos valores de m y n una muestra aleatoria podría contener entre siete y 15 series el 95 % de las veces. Cualquier r observada de seis o menos o de 16 o más se encuentra en la región de rechazo para $\alpha = 0.05$. La $r = 2$ observada es menor que 6, tal que en el nivel de significación de 0.05 podemos rechazar la hipótesis nula de que la moneda está produciendo una serie aleatoria de caras y cruces.

Si se desea una prueba unidireccional, esto es, si se predice con anterioridad la dirección de la desviación, entonces sólo se necesita usar una de las dos entradas. Si la predicción es que se observarán muy pocas series, en la tabla G del Apéndice I se proporcionan los valores críticos de r. Si la r observada de acuerdo con tal prueba unidireccional es igual o menor que el valor superior mostrado en la tabla G, H_0 puede ser rechazada en $\alpha = 0.025$. Si la predicción es que se observarán demasiadas series, los valores menores en la tabla G son los valores críticos de r que resultan significativos en el nivel 0.025.

Por ejemplo, tómese el caso de la segunda secuencia de lanzamientos de moneda ya mencionado. Supóngase que habíamos predicho, por alguna razón, que la moneda podría producir demasiadas series. Observamos que $r = 20$ para $m = 10$ y $n = 10$. Ya que nuestro valor observado de r es igual o mayor que el valor inferior mostrado en la tabla G del Apéndice I, rechazamos H_0 en $\alpha = 0.025$, y concluimos que la moneda "está cargada" en la dirección predicha.

Al desarrollar la hipótesis alterna para la prueba de series, un investigador podría concluir que los datos están agrupados o reunidos. En ese caso, la hipótesis alterna sería que podrían haber menos series que las esperadas si los datos fueran aleatorios. Por otra parte, el investigador podría conjeturar que los datos son más variables que lo que se espera sobre la base de asignación aleatoria. En este caso, la hipótesis alterna sería que podría haber más series que las esperadas si los datos fueran aleatorios. En cada uno de estos casos, la prueba de H_0 podría ser unidireccional.

Ejemplo para muestras pequeñas. En un estudio de la dinámica de la agresión en niños pequeños, un investigador observó pares de niños en una situación de juego controlada.[7] La mayoría de los 24 niños que sirvieron como sujetos en el estudio provenían de la

[7] Siegel, Alberta E., "The effect of film-mediated fantasy aggression on strength of aggresive drive in young children", tesis doctoral inédita, Stanford University, 1955.

misma guardería y, por tanto, jugaban juntos diariamente. Ya que el observador fue capaz de ingeniarse para observar sólo dos niños en cualquier día, estaba interesado en que podrían haberse introducido sesgos en el estudio por discusiones entre aquellos niños que ya habían servido como sujetos y aquellos que sirvieron posteriormente. Si tales discusiones tenían algún efecto en el nivel de agresión en las sesiones de juego, este efecto podría mostrarse como una carencia de aleatoridad en las puntuaciones de agresión en el orden en que fueron colectadas. Después de concluido el estudio, la aleatoriedad de la secuencia de puntuaciones fue probada al convertir la puntuación de agresión de cada niño a un signo más o un signo menos, dependiendo de si se encontraba por arriba o por abajo de la mediana del grupo, y aplicando entonces la prueba de una muestra de series para la secuencia observada de signos "más" y signos "menos".

i. *Hipótesis nula.* H_0: los signos "más" y los signos "menos" ocurren en un orden aleatorio. Esto es, la hipótesis nula consiste en que las puntuaciones de agresión ocurren aleatoriamente a través del experimento por arriba y por debajo de la mediana. H_1: El orden de los signos "más" y "menos" depende de la aleatoriedad.
ii. *Prueba estadística.* Ya que la hipótesis concierne a la aleatoriedad de una secuencia simple de observaciones, se elige la prueba de una muestra de series.
iii. *Nivel de significación.* Sea $\alpha = 0.05$ y N el número de sujetos $= 24$. Ya que las puntuaciones estarán caracterizadas como un signo "más" o un signo "menos", dependiendo de si se encuentran por arriba o por abajo de la puntuación mediana en el grupo, $m = n = 12$.
iv. *Distribución muestral.* En la tabla G del Apéndice I se proporcionan los valores críticos de r de la distribución muestral.
v. *Región de rechazo.* Puesto que H_1 no predice la dirección de la desviación de la aleatoriedad, se usó una prueba bidireccional. Ya que $m = n = 12$, la referencia de la tabla G muestra que H_0 podría ser rechazada en el nivel 0.05 de significación, si la r observada es igual o menor que 7 o igual o mayor que 19.
vi. *Decisión.* En la tabla 3.5 se muestran las puntuaciones de agresión para cada niño en el orden en que fueron obtenidas. La mediana del conjunto de puntuaciones es 25.5. En la tabla 3.5, todas las puntuaciones que se encuentran por debajo de la mediana están designadas por un signo "menos"; todas las puntuaciones que se encuentran por arriba de la mediana están denotadas por un signo "más". En la columna que muestra los signos $+$ y $-$, se puede ver rápidamente que ocurrieron 10 series en el conjunto de observaciones, esto es, $r = 10$.
La referencia de la tabla G del Apéndice I revela que $r = 10$ para $m = n = 12$, no se encuentra en la región de rechazo. Así, no podemos rechazar la hipótesis de que la serie de observaciones ocurrió en un orden aleatorio.

Muestras grandes

Si m o n son mayores que 20, no se puede usar la tabla G del apéndice I. Para tales muestras grandes, una buena aproximación a la distribución muestral de r es la distribución normal con

$$\text{media} = \mu_r = \frac{2mn}{N} + 1$$

$$\text{y desviación estándar} = \sigma_r = \sqrt{\frac{2mn\ (2mn - N)}{N^2(N - 1)}}$$

Tabla 3.5. Puntuaciones de agresión de acuerdo al orden de ocurrencia.

Niño	*Puntuación*	*Posición de la puntuación con respecto a la mediana*
1	31	+
2	23	−
3	36	+
4	43	+
5	51	+
6	44	+
7	12	−
8	26	+
9	43	+
10	75	+
11	2	−
12	3	−
13	15	−
14	18	−
15	78	+
16	24	−
17	13	−
18	27	+
19	86	+
20	61	+
21	13	−
22	7	−
23	6	−
24	8	−

Por tanto, cuando m o n son mayores que 20, H_0 puede ser probada por

$$z = \frac{r - \mu_r}{\sigma_r} = \frac{\dfrac{r + h - 2mn}{N - 1}}{\sqrt{\dfrac{[2mn(2mn - N)]}{[N^2 (N - 1)]}}} \tag{3.9}$$

donde $h = +0.5$ si $r < 2mn/N + 1$, y $h = -0.5$ si $r > 2mn/N + 1$. Ya que los valores de z que se obtienen usando la ecuación (3.9) están de manera aproximada normalmente distribuidos con media 0 y desviación estándar 1 cuando H_0 es verdadera, la significancia de cualquier valor observado de z usando la ecuación, puede determinarse a partir de una tabla de distribución normal como la tabla A del Apéndice I, la cual proporciona las probabilidades de una cola asociadas con la ocurrencia, según H_0, de valores tan extremos como la z observada.

En el siguiente ejemplo de muestras grandes se usa esta aproximación de distribución normal para la distribución muestral de r.

Ejemplo para muestras grandes. Un investigador estaba interesado en averiguar si la disposición de hombres y mujeres en una fila enfrente de la taquilla de un teatro era un arreglo aleatorio. Los datos se obtuvieron simplemente anotando el sexo de cada una de 50 personas al aproximarse a la taquilla.

i. *Hipótesis nula.* H_0: el orden de hombres y mujeres en la fila es aleatorio. H_1: el orden de hombres y mujeres en la fila no es aleatorio.
ii. *Prueba estadística.* Se elige la prueba de una muestra de series ya que la hipótesis concierne a la aleatoriedad en una secuencia de observaciones. Puesto que el tamaño de la muestra es grande, se usa una prueba para muestras grandes.
iii. *Nivel de significación.* Sea $\alpha = 0.05$ y N el número de personas observadas $= 50$. Los valores de m y n pueden determinarse sólo después de que se recaben los datos.
iv. *Distribución muestral.* Para muestras grandes, los valores de z calculados de la ecuación (3.9) cuando H_0 es verdadera, están distribuidos de manera aproximada normalmente con media 0 y desviación estándar 1. En la tabla A del Apéndice I se proporcionan las probabilidades de una cola asociadas con la ocurrencia, cuando H_0 es verdadera, de valores tan extremos como una z observada.
v. *Región de rechazo.* Ya que H_1 no predice la dirección de la desviación de la aleatoriedad, se usa una región de rechazo bidireccional. Ésta consta de todos los valores de z, calculados mediante la ecuación (3.9), que son tan extremos que la probabilidad asociada con su ocurrencia cuando H_0 es verdadera, es menor o igual a $\alpha = 0.05$. Así, la región de rechazo incluye todos los valores de z más extremos que 1.96.
vi. *Decisión.* Los hombres (M) y las mujeres (F) se formaron enfrente de la taquilla en el orden que se observa en la tabla 4.6. El lector puede verificar que hubo $m = 30$ hombres y $n = 20$ mujeres en la muestra. La cuenta del número de series es $r = 35$.

Tabla 3.6. Orden en la fila de la oficina de un teatro, de 30 hombres (M) y 20 mujeres (F).*

M	F	M	F	MMM	FF	M	F	M	F
M	F	MMMM	F	M	F	M	F	MM	
FFF	M	F	M	F	M	F	MM	F	
MM	F	MMMM	F	M	F	MM			

* Las series se indican por el subrayado.

Para determinar si $r \geq 35$ pudiera realmente haber ocurrido según H_0, calculamos el valor de z usando la ecuación (3.9):

$$z = \frac{r - \mu_r}{\sigma_r} = \frac{r + h - 2mn/N - 1}{\sqrt{[2mn(2mn - N)]/[N^2(N - 1)]}}$$

$$z = \frac{r - \mu_r}{\sigma_r} = \frac{35 - 0.5 - 2(30)(20)/50 - 1}{\sqrt{\{2(30)(20)[2(30)(20) - 50]\}/[50^2(50 - 1)]}}$$

$$= 2.83$$

Ya que 2.83 es mayor que el valor crítico de z (1.96), podemos rechazar la hipótesis de aleatoriedad. Ciertamente, la probabilidad de obtener un valor de $z \geq 2.83$ cuando H_0 es verdadera, es $p = 2(.0023) = 0.0046$. (La probabilidad obtenida de la tabla A se multiplica por dos debido a que estamos usando una prueba bidireccional.) Como resultado de la prueba, podemos concluir que el orden de hombres y mujeres en la fila de la taquilla no es aleatorio.

Resumen del procedimiento

Los siguientes son los pasos para el uso de la prueba de una muestra de series:

1. Disponga las observaciones m y n en su orden de ocurrencia.
2. Cuente el número de series r.
3. Determine la probabilidad de p según H_0 asociada con un valor tan extremo como el valor observado de r. Si esa probabilidad es igual o menor que α, rechace H_0.

La técnica para determinar el valor de p depende del número de observaciones, m y n, en los dos grupos:

a) Si m y n son ambos 20 o menos, consulte la tabla G del Apéndice I. Para una prueba de dos colas con $\alpha = 0.05$, si el número observado de series es menor o igual a la entrada superior o igual o mayor que la entrada inferior, rechace H_0. Para una prueba unidireccional con $\alpha = 0.025$, rechace H_0 si el número de series es menor o igual a (o más grande o igual a) la entrada de la tabla.

b) Si m o n es mayor que 20, determine el valor de z usando la ecuación (3.9). En la tabla A del Apéndice I se proporcionan las probabilidades unidireccionales asociadas con la ocurrencia según H_0 de valores tan extremos como una z observada. Para una prueba bidireccional, multiplique por dos la probabilidad obtenida de la tabla.

Si la probabilidad asociada con el valor observado de r es igual o menor que α, rechace H_0.

Potencia-eficacia

Debido a que no existen pruebas paramétricas para la aleatoriedad de una secuencia de eventos en una muestra, el concepto de potencia-eficacia no es significativo en el caso de la prueba de una muestra de series. La prueba de series se usa para probar la hipótesis nula de que la secuencia de observaciones es aleatoria. A diferencia de las técnicas que se examinarán en los siguientes dos capítulos, esta forma de la prueba de series no es útil para estimar diferencias entre grupos. Sin embargo, para la hipótesis particular de interés la prueba es útil y directa.

PRUEBA DEL MOMENTO DEL CAMBIO

Función y racionalización

Existen muchas situaciones experimentales en las cuales un investigador observa una secuencia de eventos y, como una de las hipótesis de investigación, quiere determinar si ha habido un cambio en el proceso subordinado que genera la secuencia de eventos. Sin embargo, por cualquiera de un número de posibles razones, el investigador no conoce el momento en el cual el cambio realmente ocurre. Aunque él podría haber inducido un cambio en la situación experimental en un tiempo particular, quizá no haya veracidad acerca de cuándo un cambio correspondiente realmente ocurre en la conducta observada. Otro ejemplo sería una tarea de aprendizaje conceptual en la que un sujeto tiene una ejecución en un determinado nivel hasta que se produce un tipo de consolidación cognoscitiva, después de lo cual se presenta un cambio en el nivel de ejecución. En tales casos, la variación muestral normal en la tarea puede oscurecer el momento en que el cambio real ocurre.

Las pruebas que describiremos en esta sección suponen que las observaciones forman una secuencia ordenada, que inicialmente la distribución de respuestas tiene una mediana y en algún punto existe un cambio en la mediana de la distribución. La hipótesis alterna podría ser unidireccional, por ejemplo, que existe un cambio ascendente en la distribución; o bidireccional, por ejemplo, que hubo un cambio en la distribución, pero no se hace predicción alguna acerca de la dirección del cambio. En otras palabras, H_0 es la hipótesis de que no existe un cambio en la dirección del parámetro, es decir, la mediana, de la secuencia de observaciones; y H_1 es la hipótesis de que hay un cambio en la localización del parámetro de la secuencia.

Se presentarán dos pruebas. Una es adecuada cuando los datos son binarios y constituyen observaciones de algún proceso binomial. La segunda prueba supone que los datos son continuos. La lógica de las pruebas es similar, aunque las fórmulas computacionales son diferentes.

Método para variables binomiales

En una serie de N observaciones binarias, $X_1, X_2, \ldots, X_N$, el dato para cada observación X_i se codifica como $X_i = 1$ para un valor de la variable (un éxito) y

$X_i = 0$ para el otro valor (un fracaso). De las N observaciones, sea m el número de éxitos (o eventos de un tipo) y sea n el número de fracasos (o eventos del otro tipo). Entonces

$$m = \sum_{i=1}^{N} X_i \qquad \text{y} \qquad n = N - m$$

Entonces, el número acumulado de éxito ($X = 1$) se determina en cada punto en la secuencia. Esta frecuencia se designa como

$$S_j = \sum_{i=1}^{j} X_i \qquad j = 1, 2, \ldots, N$$

El estadístico para probar la hipótesis de cambio es

$$D_{m,n} = \text{máx} \left| \frac{N}{mn} \left(S_j - \frac{jm}{N} \right) \right| \tag{3.10}$$

La expresión se evalúa para todos los valores de j desde 1 hasta $N - 1$. $D_{m,n}$ es la diferencia absoluta más grande observada en la frecuencia. La distribución muestral de $D_{m,n}$ ha sido tabulada y algunos valores se proporcionan en la tabla L_{II} del Apéndice I y es una forma de la prueba de Kolmogorov-Smirnov. Si $D_{m,n}$ es igual o excede al valor de la tabla, podemos rechazar H_0 en el nivel especificado de significación y concluir que ha habido un cambio en la distribución.

Si el tamaño de la muestra es grande, los valores críticos pueden determinarse de la tabla L_{III} del Apéndice I. Por ejemplo, si $N = 60$ y $m = 45$, $n = 15$, podemos rechazar H_0 el nivel 0.05 si $D_{m,n} \geq 1.36\sqrt{N/mn} = 1.36\,(0.298) = 0.41$.

Ejemplo. En un estudio del efecto del cambio en el pago en una tarea de aprendizaje con probabilidad de dos elecciones,[8] el pago o la recompensa proporcionado a un sujeto se cambiaba (o no se cambiaba) después de que la ejecución individual se había estabilizado en una asíntota (o un nivel de ejecución constante). La hipótesis fue que un cambio en el pago por respuestas correctas, podía afectar el nivel de respuestas dadas por el sujeto. El experimento constó de 300 ensayos en cada uno de los cuales el sujeto daba una respuesta binaria. Ya que no se puede pensar que un patrón de respuestas del sujeto se ha estabilizado hasta que se verifica algún aprendizaje, sólo se analizan aquí los últimos 240 ensayos. En el ensayo 120 (ensayo 180 en la secuencia original), la mitad de los sujetos experimentaba un cambio en el pago. El investigador deseaba determinar si hubo un cambio en el parámetro de la secuencia binaria de respuestas sobre los últimos 240 ensayos. Si hubo un cambio para aquellos sujetos que experimentaban un cambio en el pago, entonces se podría concluir que el cambio en el pago indujo un cambio en el nivel de respuestas.

Para ilustrar la prueba, se analizarán las secuencias de respuesta para dos sujetos. El sujeto A recibió 10 centavos por cada respuesta correcta durante todo el experimento. El sujeto B recibió 10 centavos hasta el ensayo 120, después de lo cual el pago se redujo a 1 centavo por cada respuesta correcta. Los datos se resumen en la tabla 3.7.

[8] Castellan, N. J., Jr. "Effect of change of payoff in probability learning", en *Journal of Experimental Psychology*, núm. 79, 1969, págs. 178-182.

Tabla 3.7. Datos de dos sujetos en un experimento de probabilidad de aprendizaje.

Secuencia de respuestas para el sujeto A —ningún cambio en el pago
111100111100111111111110110011100111110111100111111101110011
01101111001011110111001111111000011110111111011100001111011
011011110011111111111011011111111111110011100111001111101101
001111010111111111001111111000111111111110111100111111110011
Secuencia de respuestas para el sujeto B —cambio en el pago
001101111111111111111111111111111111111111111111111110111110
110110011110000111110111001011111001111011011100111100000101
110110111100000011111011111101111111110111111001100111100111
100001111011011000011100011111100001111011010010000011100 11

i. *Hipótesis nula.* H_0: no existe un cambio en p, la probabilidad de que $X_i = 1$, a través de la secuencia de ensayos. H_1: existe un cambio en P a través de la secuencia de ensayos.
ii. *Prueba estadística.* Se usará la prueba del momento del cambio para variables binomiales, debido a que el investigador desea determinar si ocurrió un cambio en la distribución observada de respuestas binarias durante los últimos 240 ensayos.
iii. *Nivel de significación.* Sea $\alpha = 0.05$ y N es el número de observaciones $= 240$.
iv. *Distribución muestral.* Los valores críticos de $D_{m,n}$ de la distribución muestral se presentan en las tablas L_{II} y L_{III} del Apéndice I, junto con sus probabilidades de ocurrencia asociadas cuando H_0 es verdadera.
v. *Región de rechazo.* La región de rechazo consista de todos los valores de $D_{m,n}$ calculados con la ecuación (3.10), que son tan grandes que la probabilidad asociada con su ocurrencia cuando H_0 es verdadera, es menor o igual a $\alpha = 0.05$.
vi. *Decisión.* Ya que la hipótesis en este ejemplo concierne a sujetos individuales, cada uno de ellos se analizará por separado. Para el sujeto A las diferencias.

$$\left| \frac{N}{mn} \left(S_j - \frac{jm}{N} \right) \right|$$

se calcularon para cada ensayo, j. S_j es el número de $X_i = 1$ respuestas e incluyen el ensayo j, m es el número de $X_i = 1$ respuestas a través del total de ensayos N, y $n = N - m$ es el número de $Xi = 0$ respuestas. Para este sujeto, $N = 240$, $m = 178$ y $n = 62$. La diferencia máxima fue $D_{178,62} = 0.096$.
Ya que m y n son grandes, debemos usar los valores de grandes muestras de la tabla L_{III} del Apéndice I. Los valores críticos de $D_{m,n}$ para $\alpha = 0.05$, $m = 178$, $n = 62$ es $1.36\sqrt{N/mn} = 1.36\sqrt{240/(178)(62)} = 0.201$. Puesto que el valor observado de $D(0.096)$ es menor que el valor crítico (0.201), no rechazamos H_0 y, así, concluimos que no hubo un cambio de punto en la secuencia de respuestas a través de los últimos 240 ensayos para el sujeto A.
Para el sujeto B las diferencias

$$\left| \frac{N}{mn} \left(S_j - \frac{jm}{N} \right) \right|$$

se calcularon para cada ensayo j. Para este sujeto, $N = 240$, $m = 167$ y $n = 73$. La diferencia máxima fue $D_{167,73} = 0.275$.

Ya que m y n son grandes, debemos usar los valores de muestras grandes de la tabla L_{III} del Apéndice I. El valor crítico de $D_{m,n}$ para $\alpha = 0.05$, $m = 167$, $n = 73$ es $1.36\sqrt{N/mn} = 1/.36\sqrt{240/(167)(173)} = 0.191$. Puesto que el valor observado de $D(0.191)$ es más grande que el valor crítico (0.191), podemos rechazar H_0 y concluir que hubo un cambio de punto en la secuencia de respuestas a través de los últimos 240 ensayos para el sujeto B.

Así, para el sujeto que no tuvo cambio en el nivel de pago durante el experimento, podemos concluir que no hubo cambio en el nivel de ejecución; mientras que para el sujeto que sufrió un decremento en el pago, podemos concluir que hubo un cambio en el nivel de ejecución.

Resumen del procedimiento

Éstos son los pasos en la aplicación de la prueba del momento del cambio a una secuencia de variables binomiales:

1. Codifique cada una de las N observaciones como 1 o 0 para "éxito" o "fracaso", respectivamente.
2. Calcule el número total de éxitos, m, en las N observaciones. Sea $n = N - m$.
3. Calcule el estadístico $D_{m,n}$ usando la ecuación (3.10), que es la diferencia máxima entre los éxitos acumulados observados y "predichos" en cada punto en la secuencia.
4. Consulte la tabla L_{II} del Apéndice I (para muestras pequeñas) o la tabla L_{III} del Apéndice I (para muestras grandes), a fin de determinar si H_0 (no existe un cambio en la secuencia) puede ser rechazada en favor de H_1 (existe un cambio en la secuencia).

Método para variables continuas

Primero, cada una de las observaciones $X_1, X_2, \ldots, X_N$ deben ser puestas en orden de rango desde 1 hasta N. Sea r_i el rango asociado con el dato X_i. Entonces, en cada lugar j en las series, calculamos

$$W_j = \sum_{i=1}^{j} r_i \qquad j = 1, 2, \ldots, N - 1$$

que es la suma de los rangos de las variables en o antes del punto j.

En seguida, para cada punto en la secuencia, calculamos $2W_j - j(N + 1)$. Entonces

$$K_{m,n} = \text{máx}\,|2W_j - j(N + 1)| \qquad j = 1, 2, \ldots, N - 1 \tag{3.11}$$

El valor de j donde el máximo en la ecuación (3.11) ocurre, es el cambio de punto estimado en la secuencia y se denota por m. $N - m = n$ es el número de observaciones después del cambio de punto. Así, $K_{m,n}$ es el estadístico que divide la secuencia en m y n observaciones que ocurren antes y después del cambio, respectivamente.

Si este valor de $K_{m,n}$ es más grande de lo que podríamos esperar por azar cuando no hay cambio en la secuencia, puede ser probado consultando una tabla de la distribución muestral de W_j, la suma de los rangos. La distribución muestral de W se resume en la tabla J del Apéndice I para varios valores de m y n. S W excede el valor tabulado de W en el nivel de significación apropiado, podemos rechazar H_0 de que no hay cambio en la distribución.

Empates

La prueba supone que las puntuaciones provienen de una población con una distribución continua. Si las medidas son precisas, la probabilidad de un empate es cero. Sin embargo, con las medidas que por lo general se usan en las ciencias conductuales, pueden ocurrir puntuaciones empatadas. Cuando ocurren rangos empatados, dé a cada una de las observaciones empatadas el promedio de los rangos que deberían tener si no hubieran ocurrido empates. Así, si dos observaciones son iguales y están empatadas para los rangos 3 y 4, a cada una se le debe asignar el rango promedio $(3 + 4)/2 = 3.5$.

Muestras grandes

De acuerdo con la suposición de no cambio en la distribución, la media de W es $m(N + 1)/2$ y su varianza es

$$\text{Varianza de } W = \sigma_w^2 = \frac{mn\ (N + 1)}{12}$$

y, como N se vuelve grande, W está de manera aproximada normalmente distribuida con media y varianza mencionadas anteriormente. Así, cuando la serie es larga, se puede hacer la prueba para el cambio y probarla usando la tabla A del Apéndice I, transformando W a z:

$$z = \frac{W + h - m(N + 1)/2}{\sqrt{mn\ (N + 1)/12}} \tag{3.12}$$

donde $h = \frac{1}{2}$ si $W > m(N + 1)/2$ y $h = +\frac{1}{2}$ es $W < m(N + 1)/2$. Si hay empates, la varianza debe ajustarse usando la ecuación (5.12) del capítulo 5.

Ejemplo. En un estudio de los efectos de las anfetaminas sobre la actividad neuronal,[9] dos investigadores midieron la tasa de descarga de neuronas en el núcleo caudado, como una

[9] Rebec, G. V. y Groves, P. M. "Differential effects for the optical isomers of amphetamine on neuronal activity in the reticular formation and caudate nucleus of the rat" en *Brain Research*, núm. 83, 1975, págs. 301-318.

Tabla 3.8. Tasa de descarga neuronal como un porcentaje de la línea base para los 25 periodos siguientes a la inyección de anfetamina.

Periodo de tiempo	*Tasa de descarga*	*Rango*	W_j	$\|2W_j - j(N+1)\|$
1	112	23.5	23.5	21
2	102	14.5	38.0	24
3	112	23.5	61.5	45
4	120	25	86.5	69
5	105	19	105.5	81
6	105	19	124.5	93
7	100	11	135.5	89
8	105	19	154.5	101
9	97	6	160.5	87
10	102	14.5	175.0	90
11	91	4	179.0	72
12	97	6	185.0	58
13	89	3	188.0	38
14	85	1	189.0	14
15	101	12	201.0	12
16	98	8.5	209.5	3
17	102	14.5	224.0	6
18	99	10	234.0	0
19	102	14.5	248.5	3
20	110	22	270.5	21
21	97	6	276.5	7
22	88	2	278.5	15
23	107	21	299.5	1
24	98	8.5	308.0	8
25	104	17	325.0	0

función del tiempo, después de la inyección de varios isómeros de anfetaminas. Los datos en la tabla 3.8 resumen la tasa de descarga neuronal como un porcentaje de una tasa base como una función del tiempo, ya que la inyección es una condición. Los investigadores querían saber si hubo un cambio en la tasa de descarga durante el tiempo que se estuvieron tomando las medidas. Si ocurría un cambio, sería la evidencia de la acción de la droga en el lugar donde se tomaron dichas medidas.

i. Hipótesis nula. H_0: no existe cambio en la tasa de descarga neuronal como una función del tiempo. H_1: existe un cambio en la tasa de descarga.

ii. Prueba estadística. Se usará la prueba del momento del cambio para variables continuas, debido a que los investigadores desean detectar un cambio en la distribución observada de las tasas de descarga neuronal durante los 25 periodos de tiempo.

iii. Nivel de significación. Sea $\alpha = 0.01$ y N es el número de observaciones o periodos de tiempo $= 25$.

iv. *Distribución muestral.* Los valores críticos de la distribución muestral de W están presentados en la tabla J del Apéndice I para niveles de significación seleccionados y valores de m y n seleccionados. Sin embargo, ya que para este experimento $m > 10$, no se puede usar la tabla J y debe utilizarse la aproximación de grandes muestras (y, por tanto, la tabla A de dicho Apéndice).

v. *Región de rechazo.* La región de rechazo consta de todos los valores de W calculados con la ecuación (3.11), que son tan grandes que la probabilidad asociada con su ocurrencia cuando H_0 es verdadera, es menor o igual a 0.01.

vi. *Decisión.* Primero fueron puestas en orden de rango del 1 al 25 las tasas de descarga. Estos rangos están resumidos en la tabla 3.8, junto con W_j, la suma acumulativa de rangos hasta el periodo de tiempo j. En seguida se calcularon los valores $|2W_j - j(N + 1)|$ para cada periodo de tiempo. El examen de estos valores (también enumerados en la tabla 3.8), muestra que el máximo es $K_{8,17} = 101$. Esto es, el máximo ocurrió en el tiempo 8. La prueba estadística es W, la suma de los rangos donde la función K es maximizada, $W = 154.5$. Ya que la distribución de W para $m = 8$, $n = 17$ no está proporcionada en la tabla J del Apéndice I, la aproximación normal debe encontrarse usando la ecuación (3.12):

$$z = \frac{W + h - m(N + 1)/2}{\sqrt{mn\ (N + 1)/12}} \tag{3.12}$$

$$= \frac{154.5 - 0.5 - 8(25 + 1)/2}{\sqrt{8(17)(25 + 1)/12}}$$

$$= 50/17.166$$

$$= 2.91$$

Usando la tabla A del Apéndice I y $\alpha = 0.01$, encontramos que el valor crítico de z es 2.58. Puesto que el valor observado es más grande que el valor crítico, podemos rechazar H_0 y concluir que hubo un cambio en la tasa de descarga neuronal durante el periodo de medida.

Resumen del procedimiento

En la aplicación de la prueba del momento del cambio para variables continuas, se siguen los pasos que se detallan a continuación:

1. Disponga en orden de rango las observaciones en la secuencia de N observaciones.
2. Calcule la suma de rangos W para cada punto j en la secuencia de observaciones.
3. Para cada punto en la secuencia use la ecuación (3.11), a fin de calcular la diferencia entre la suma de rangos observada y "predicha". $K_{m,n}$ es el máximo y divide la secuencia dentro de las m observaciones antes del cambio y dentro de las n observaciones después del cambio.
4. Dependiendo de los valores de m y n, el método para probar varía.

   a) *Muestras pequeñas.* En el punto m en el que ocurre el máximo, use los valores W_j, m y n para entrar en la tabla J del Apéndice I, para determinar si se rechaza la hipótesis nula H_0 de que no existe cambio en la secuencia, en favor de H_1, de que sí existe un cambio en la secuencia de observaciones.

b) *Muestras grandes* ($m > 10$ o $n > 10$). Use el valor observado de W_j, m y n para calcular el valor de z usando la ecuación (3.12). Si el valor observado de z excede al valor crítico de z encontrado en la tabla A del Apéndice I, rechace la hipótesis nula H_0 de que no existe cambio en la secuencia.

Potencia-eficacia

Para la prueba del momento del cambio binomial, el concepto de eficacia no es significativo cuando la variable es binomial. Sin embargo, los comentarios concernientes a la prueba de la bondad de ajuste de Kolmogorov-Smirnov (véase el capítulo 5) son relevantes a esta prueba cuando una variable continua ha sido dicotomizada para formar una variable binaria con el propósito de aplicar la prueba.

Para la prueba del momento del cambio para variables continuas, los procedimientos de Monte Carlo sugieren que la prueba es poderosa respecto a los cambios en la forma de la distribución. La eficacia del procedimiento no se ha analizado explícitamente. Sin embargo, la relación entre esta prueba y la prueba Wilcoxon-Mann-Whitney (capítulo 5) sugiere que la prueba puede ser altamente eficaz.

Referencias bibliográficas

Las pruebas descritas han sido presentadas por Pettitt (1979). Una prueba anterior para secuencias binomiales perteneciente a Page (1955) se ha utilizado ampliamente, pero se hacen suposiciones adicionales acerca de los parámetros iniciales de la distribución binomial.

ANÁLISIS

En este capítulo hemos presentado seis pruebas estadísticas no paramétricas para usar en diseños de una muestra. Tres de estas pruebas son del tipo de la bondad de ajuste, una es la prueba para la simetría de la distribución contra la no simetría; otra es una prueba de aleatoriedad de la secuencia de eventos en una muestra; la tercera es una prueba para el cambio en una distribución. Este análisis, que brevemente compara y contrasta dichas pruebas, ayudará al lector a seleccionar la que mejor se ajuste a los datos de un estudio determinado.

Al probar hipótesis acerca de si una muestra fue extraída de una población con una distribución específica, el investigador puede usar una de las tres pruebas de la bondad de ajuste: la prueba binomial, la prueba ji cuadrada de una muestra o la prueba de Kolmogorov-Smirnov de una muestra. La elección entre estas tres pruebas debe estar determinada por: 1. el número de categorías en las medidas; 2. el nivel de medición usado; 3. el tamaño de la muestra, y 4. la potencia de la prueba estadística.

La prueba binomial es adecuada cuando hay justo dos categorías en la clasificación de los datos. Es útil únicamente cuando el tamaño de la muestra es tan pequeño que la prueba ji cuadrada resulta inapropiada.

La prueba ji cuadrada debe utilizarse cuando los datos están en categorías discretas y cuando las frecuencias esperadas son suficientemente grandes. Cuando $k = 2$, cada E_i debe ser 5 o más. Cuando $k > 2$, no más que cerca del 20 % de los E_j deben ser menores que 5 y ninguno debe ser menor que 1.

Tanto la prueba binomial como la prueba ji cuadrada deben usarse con datos medidos en una escala nominal u ordinal.

La prueba ji cuadrada examinada en este capítulo es insensible a los efectos de orden cuando $gl > 1$ y, por tanto, pudiera no ser la mejor prueba cuando una hipótesis supone que las variables están ordenadas.

La prueba de Kolmogorov-Smirnov debe emplearse cuando se puede suponer que la variable en consideración tiene una distribución continua. Sin embargo, si esta prueba se usa cuando la distribución de la población $F_0(X)$ no es continua, el error que ocurre en el juicio de probabilidad resultante está en la dirección "segura" (Goodman, 1954). Esto es, si las tablas que suponen que $F_0(X)$ es continua, se usan para probar una hipótesis acerca de una variable no continua, la prueba es una prueba cautelosa; si H_0 es rechazada por esta prueba, podemos tener confianza real en esa decisión.

Ya hemos mencionado que la prueba de Kolmogorov-Smirnov trata observaciones individuales por separado y, por tanto, no pierde información debido al agrupamiento, como algunas veces ocurre con la prueba ji cuadrada. Con una variable continua, si la muestra es pequeña y, por consiguiente, las categorías adyacentes deben ser combinadas para la prueba ji cuadrada, ésta es definitivamente menos potente que la prueba de Kolmogorov-Smirnov. Puede parecer que en todos los casos en los que es aplicable la prueba de Kolmogorov-Smirnov, ésta es la prueba más potente de todas las de bondad de ajuste presentadas.

En los casos en que los parámetros deben ser estimados de la muestra, la prueba de la bondad de ajuste ji cuadrada se modifica fácilmente para su uso al reducir los grados de libertad. Sin embargo, para la prueba de Kolmogorov-Smirnov, la distribución de D no es conocida para el caso en que ciertos parámetros de la población hayan sido estimados de la muestra. Existe alguna evidencia que sugiere que si la prueba de Kolmogorov-Smirnov se aplica en tales casos (por ejemplo, para probar la bondad de ajuste para una distribución normal con media y desviación estándar estimadas de la muestra), el uso de la tabla F del Apéndice I conducirá a una prueba cautelosa. Esto es, si el valor crítico de D (como se muestra en la tabla F) es excedido por el valor observado en esas circunstancias, podemos rechazar H_0 con confianza considerable.

La prueba para evaluar la simetría de una distribución es útil para determinar la forma de una distribución. La forma (o el sesgo) de una distribución es de especial interés cuando sospechamos que, debido a que algunas observaciones son "extremas", la distribución no es simétrica alrededor de su mediana.

La prueba de una muestra de series se interesa en la aleatoriedad de la ocurrencia o secuencia temporal de las puntuaciones en una muestra. Así, también podría emplearse para probar hipótesis concernientes al agrupamiento o la dispersión de observaciones dicotómicas. Ningún juicio general acerca de la eficacia de las pruebas de aleatoriedad basadas en las series puede ser significativo; en este caso, la cuestión de eficacia tiene significado sólo en el contexto de un problema específico.

La prueba del momento del cambio es útil cuando se desea probar la hipótesis de que ha habido un cambio en la distribución de una secuencia de eventos. Para usar la prueba adecuadamente, no es necesario conocer *a priori* cuándo ocurrió el cambio. La prueba evalúa la probabilidad de que ocurra realmente un cambio en la secuencia de observaciones y si el cambio observado excede la fluctuación esperada debida al azar. Se describieron dos pruebas del momento del cambio: una para observaciones basadas en un proceso binomial o binario y la otra para muestras de una distribución continua.

4

El caso de una muestra medida dos veces y obtenida por medio de pares replicados

Las pruebas estadísticas de una sola muestra que implican dos medidas o pares replicados, se utilizan cuando el investigador desea establecer si dos tratamientos son diferentes o si un tratamiento es mejor que otro. El tratamiento puede ser cualquiera de una amplia variedad de condiciones: aplicación de una droga, cierto entrenamiento, "aculturación", propaganda, separación familiar, trastornos quirúrgicos, introducción de un nuevo elemento en la economía, etc. En cada caso, el grupo al cual se le aplica el tratamiento es comparado con uno al cual no se le aplicó, o bien, se le aplicó un tratamiento diferente.

En las comparaciones entre dos grupos, en ocasiones las diferencias significativas que se observan no son el resultado del tratamiento. Por ejemplo, un investigador puede intentar comparar dos métodos de enseñanza utilizando dos grupos de estudiantes, un grupo que está siendo enseñado con un método y un grupo diferente al cual se le enseña con un procedimiento distinto. Ahora bien, si uno de los grupos incluye estudiantes más capaces o más motivados, la ejecución de los dos grupos después de las experiencias de aprendizaje puede no reflejar con precisión la efectividad relativa de los diferentes métodos utilizados, porque otras variables son las que produjeron las diferencias observadas en la ejecución.

Una manera de resolver la dificultad impuesta por las diferencias extrañas entre los grupos es utilizar dos muestras relacionadas en la investigación. Esto es, se pueden "igualar" o relacionar las dos muestras estudiadas. Esta igualación se obtiene utilizando a cada sujeto como su propio control o pareando a los sujetos, y entonces se asigna a los miembros del par a una de las dos condiciones. El sujeto que sirve como su propio control es expuesto a ambas condiciones en diferentes ocasiones (tiempos). Cuando se utiliza el método de apareamientos, el objetivo es seleccionar pares de sujetos los cuales sean lo más semejantes posible en lo que respecta a cualquier variable extraña que pueda influir en el resultado de la investigación. En el ejemplo mencionado, el método de apareamiento requeriría que fuera seleccionado un cierto número de pares de estudiantes, cada par compuesto por dos estudiantes lo más similares en cuanto a capacidad o motivación. Un miembro de cada par, escogido por algún procedimiento al azar, sería asignado a uno de los métodos de enseñanza y su pareja, asignada al método restante.

Como se advierte, el método de utilizar a cada sujeto como su propio control (utilizando un diseño contrabalanceado en el cual presentar secuencialmente los tratamientos) es preferible al de los sujetos apareados, debido a que nuestra capacidad de igualar sujetos (personas) está limitada por nuestra ignorancia (o poco conocimiento) acerca de las variables relevantes que subyacen a la conducta que está siendo estudiada. Más aún, aunque conocemos las variables que son importantes y que, por tanto, pueden ser controladas por nosotros, nuestros instrumentos de medición de tales variables son más bien gruesos e inexactos y así, nuestro apareamiento basado en tales mediciones será defectuoso. Un diseño de igualación (o de sujetos apareados) es sólo tan bueno como la capacidad del investigador para determinar cuán "igualados" están los pares, y esta habilidad con frecuencia está severamente limitada. Este problema se resuelve cuando cada sujeto se utiliza como su propio control; no existe igualación más precisa que la proporcionada por la propia identidad.

La técnica estadística paramétrica usual para analizar los datos de dos muestras relacionadas es la aplicación de una prueba *t* a las diferencias en las puntuaciones obtenidas. Las diferencias se obtienen entre las puntuaciones obtenidas por los dos miembros de cada par o bien, de las dos puntuaciones obtenidas por el mismo sujeto en cada condición. La prueba *t* supone que las diferencias en las puntuaciones obtenidas pertenecen a (fueron extraídas de) una distribución normal, lo cual implica que las variables pueden medirse al menos en una escala de intervalo.

En ocasiones la prueba *t* no es adecuada. El investigador puede encontrar que:

1. Las suposiciones y los requisitos de la prueba *t* no son aplicables a los datos.
2. Es conveniente evitar hacer las suposiciones o probar los requisitos de la prueba *t* y así dar una mayor generalidad a sus conclusiones.
3. Las diferencias entre los pares igualados no se presentan como puntuaciones, sino más bien como signos (por ejemplo, podemos decir que cualquier miembro del par es "más grande" que el otro, pero no decir cuán *grande* es).
4. Las puntuaciones son simplemente clasificatorias: los dos miembros del par pueden responder de la misma manera o de maneras diferentes, lo cual no afirma o propone alguna relación cuantitativa a cada uno.

Para tales circunstancias el investigador debe seleccionar alguna de las pruebas estadísticas no paramétricas, para las dos mediciones de una sola muestra o para los pares replicados, de las que se presentan en este capítulo. Adicionalmente, a fin de ser aplicables a los ejemplos mencionados, estas pruebas tienen la ventaja de que no requieren que todos los pares sean escogidos de la misma población. Se presentan cuatro pruebas; el análisis al final del capítulo indica los rasgos y usos característicos de cada una de ellas. Esta exposición ayudará al lector en la selección de la técnica más adecuada para su situación en particular.

LA PRUEBA DEL CAMBIO DE McNEMAR

Función

La prueba de McNemar para la significación de los cambios es particularmente aplicable a los diseños "antes-después", en los cuales cada sujeto se utiliza como su propio control y en los que las mediciones se realizan ya sea en escala nominal u ordinal. En estas condiciones puede emplearse para probar la efectividad de un tratamiento particular (reuniones, editoriales en los diarios, discursos en campaña, visitas personales, etc.) sobre las preferencias de los votantes acerca de los candidatos a puestos públicos, o para probar el efecto de la migración del campo a la ciudad sobre la filiación política de las personas. Nótese que en estos estudios las personas pueden servir como su propio control y que la escala nominal (o de categorización) se utiliza de manera adecuada para evaluar el cambio "antes-después".

Racionalización y método

Con este método para probar la significación de cualquier cambio observado, se utiliza una tabla de 2×2 para representar el primero y el segundo conjuntos de respuestas de los mismos individuos. Los rasgos generales de dicha tabla se muestran en la tabla 4.1, en donde $+$ y $-$ se usan para denotar diferentes respuestas. Nótese que todos aquellos casos que muestran cambios entre la primera y segunda respuestas aparecen en las celdillas superior izquierda (de $+$ a $-$) e inferior derecha (de $-$ a $+$) de la tabla. Las entradas en la tabla corresponden a las frecuencias (ocurrencias) de los resultados asociados. Así, A denota el número de individuos cuyas respuestas fueron $+$ en la primera medición y $-$ en la segunda medición. De manera similar, D es el número de individuos quienes cambiaron de $-$ a $+$. B es la frecuencia de individuos que respondieron $+$ en ambas ocasiones, y C es el número de personas que respondieron $-$ en la primera y la segunda evaluaciones.

Así, $A + D$ es el total de personas cuyas respuestas cambiaron. La hipótesis nula es que el número de cambios en cada dirección es el mismo. Así es que de $A + D$ individuos que cambiaron, nosotros esperaríamos que $(A + D)/2$ individuos cambiaran de $+$ a $-$ y $(A + D)/2$ personas cambiaran de $-$ a $+$. En otras palabras, cuando H_0 es verdadera, la frecuencia esperada en cada una de las dos celdillas es $(A + D)/2$.

Tabla 4.1. Tabla de 2×2 utilizada en la prueba de significación de los cambios.

		Después	
		$-$	$+$
Antes	$+$	A	B
	$-$	C	D

Como se recordará del capítulo 3

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{3.5}$$

donde

O_i = número de casos observados en la iésima categoría
E_i = número de casos esperados en la iésima categoría cuando H_0 es verdadera
k = número de categorías

En la prueba de McNemar para la significación de los cambios, estamos interesados sólo en las celdillas en las cuales pueden ocurrir cambios. Así, si A es el número de casos observados cuyas respuestas cambiaron de $+$ a $-$, D es el número observado de casos que cambiaron de $-$ a $+$, y $(A + D)/2$ es el número de casos esperado en las celdillas A y D. Entonces

$$X^2 = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{[A - (A + D)/2]^2}{(A + D)/2} + \frac{[D - (A + D)/2]^2}{(A + D)/2}$$

Desarrollando y reduciendo términos, tenemos que

$$X^2 = \frac{(A - D)^2}{A + D} \qquad \text{con } gl = 1 \tag{4.1}$$

La distribución muestral de X^2 calculada por medio de la ecuación (4.1) cuando H_0 es verdadera, se distribuye asintóticamente como ji cuadrada con grados de libertad igual a uno.

Corrección por continuidad

La aproximación por medio de la distribución ji cuadrada de la distribución muestral de X^2 llega a ser más precisa si se hace una correción por continuidad. La corrección es necesaria porque una distribución continua (ji cuadrada) se utiliza para aproximarse a una distribución discreta X^2. Cuando todas las frecuencias esperadas son pequeñas, la aproximación puede ser muy pobre. El propósito de la corrección por continuidad (Yates, 1934) es eliminar esta fuente de imprecisión.

Con la corrección por continuidad incluida,

$$X^2 = \frac{(|A - D| - 1)^2}{A + D} \qquad \text{con } gl = 1 \tag{4.2}$$

La evaluación del numerador en la ecuación (4.2) indica que se debe restar una unidad del valor absoluto de la diferencia entre *A* y *D* (es decir, independientemente del signo de la diferencia entre *A* y *D*), antes de elevar el cuadrado. La significación de cualquier valor observado de X^2 computado de la ecuación (4.2), se determina haciendo referencia a la tabla C del Apéndice I, en la cual se proporcionan algunos valores críticos de la distribución ji cuadrada con grados de libertad de 1 a 30. Así, si se observa un valor de X^2 mayor o igual al valor crítico proporcionado por la tabla para un cierto valor de significación y $gl = 1$, podemos rechazar la hipótesis acerca de que los cambios en cada dirección son los mismos.

Ejemplo. Durante las campañas presidenciales (y algunas otras campañas para puestos públicos) de 1980 en Estados Unidos se realizaron debates televisivos entre dos o más candidatos. Un investigador en técnicas de comunicación estaba interesado —tanto como los candidatos— en determinar si los debates entre los candidatos presidenciales en las elecciones de 1980 eran efectivos o no en cuanto a cambiar las preferencias de los televidentes hacia los distintos candidatos. Se predijo que si los candidatos (Jimmy Carter y Ronald Reagan) eran igualmente efectivos, habría cambios comparables en las preferencias a cada candidato por parte de los televidentes. Por otro lado, si un candidato era más efectivo o persuasivo durante el debate, entonces habría un cambio diferencial de un candidato a otro. Para evaluar la efectividad del debate, el investigador seleccionó 70 adultos al azar antes del debate y les pidió que indicaran sus preferencias hacia ambos candidatos. Después de la conclusión del debate, les volvió a preguntar acerca de su predilección. Así, en cada caso el conocía las preferencias de las personas antes del debate y después del mismo. Los resultados obtenidos pueden presentarse de acuerdo con la tabla 4.2.

i. *Hipótesis nula.* H_0: entre los televidentes *que cambiaron* sus preferencias, la probabilidad de que hayan cambiado de Reagan a Carter será la misma de los que cambiaron de Carter a Reagan.[1] La hipótesis alterna es H_1: Hay un cambio diferencial en la preferencia. Las hipótesis pueden resumirse como sigue:

$$H_0\text{: } P[\text{Reagan} \rightarrow \text{Carter}] = P[\text{Carter} \rightarrow \text{Reagan}]$$

$$H_1\text{: } P\,[\text{Reagan} \rightarrow \text{Carter}] \neq P[\text{Carter} \rightarrow \text{Reagan}]$$

Tabla 4.2. Tabla de 2 × 2 utilizada para mostrar los cambios en las preferencias acerca de los candidatos presidenciales.

Preferencia antes del debate televisivo	*Preferencia después del debate televisivo*	
	Reagan	*Carter*
Carter	*A*	*B*
Reagan	*C*	*D*

[1] La proposición de esta H_0 sugiere una aplicación íntegra de la prueba binomial (véase la sección correspondiente en el capítulo 3). La relación entre la prueba de McNemar y la prueba binomial está delineada en la exposición anterior acerca de las frecuencias esperadas pequeñas.

Tabla 4.3. Preferencias de los sujetos acerca de los candidatos presidenciales antes y después del debate televisivo.

Preferencia antes del debate televisivo	*Preferencia después del debate televisivo*	
	Reagan	*Carter*
Carter	13	28
Reagan	27	7

ii. *Prueba estadística.* Se selecciona la prueba de McNemar para la significación de los cambios, ya que el estudio utiliza dos muestras relacionadas (los mismos sujetos medidos en dos ocasiones); esta prueba es del tipo "antes-después" y utiliza medidas nominales (categorías).

iii. *Nivel de significación.* $\alpha = 0.05$ y $N = 70$ (número de personas a las cuales se les pidió su opinión antes del debate y después de éste).

iv. *Distribución muestral.* La tabla C del Apéndice I nos proporciona los valores críticos de la distribución ji cuadrada para varios niveles de significancia. La distribución muestral de X^2 calculada por medio de la ecuación (4.2) se distribuye asintóticamente como ji cuadrada con $gl = 1$.

v. *Región de rechazo.* Puesto que H_1 no especifica la dirección de la diferencia en cuanto a la preferencia, la región de rechazo es bidireccional. La región de rechazo consiste en todos los valores de X^2 que sean mayores que aquellos que tienen una probabilidad no direccional asociada con su ocurrencia cuando H_0 es verdadera para $\alpha = 0.05$ o menor.

vi. *Decisión.* Los datos de este estudio se presentan en la tabla 4.3. Ésta nos muestra que $A = 13$ (los televidentes que cambiaron de Carter a Reagan) y $D = 7$ (los televidentes que cambiaron de Reagan a Carter). $B = 28$ y $C = 27$ son los televidentes que no cambiaron su preferencia a pesar del debate. Nosotros estamos interesados en aquellos que cambiaron su preferencia, es decir, los representados por A y D.

Con los datos anteriores tenemos:

$$X^2 = \frac{(|A - D| - 1)^2}{A + D} \quad \text{con } gl = 1 \tag{4.2}$$

$$= \frac{(|13 - 7| - 1)^2}{13 + 7}$$

$$= 5^2/20$$

$$= 1.25$$

Recurriendo a la tabla C del Apéndice I, tenemos que cuando H_0 es verdadera y $gl = 1$, la probabilidad de $X^2 \geq 3.84$ es 0.05.

Como el valor observado de X^2 (1.25) es menor que el valor crítico de ji cuadrada (3.84), no podemos rechazar la hipótesis de que los candidatos fueron igualmente efectivos para cambiar las preferencias de los televidentes.

Nótese que en este ejemplo el investigador estaba interesado en si había o no cambios en la preferencia de los televidentes. Los candidatos podían estar interesados en lo mismo; sin embargo, la hipótesis alterna adecuada para ellos hubiera sido que el debate fuera efectivo en una dirección particular. Es decir, la H_1 hubiera sido de una cola; en ese caso, la tabla C del Apéndice I se hubiera utilizado con los valores de probabilidad compartida, de aquí que las entradas de la tabla están basadas en una prueba bidireccional.

FRECUENCIAS PEQUEÑAS ESPERADAS

Como se pudo notar, la distribución muestral de X^2 en la prueba ji cuadrada (y, por tanto, en la prueba de McNemar) se aproxima a la distribución ji cuadrada sólo cuando el tamaño de la muestra es grande. Para muestras pequeñas, la aproximación es pobre. Sin embargo, existe un procedimiento alternativo cuando N es pequeña. Si la frecuencia esperada para la prueba de McNemar $(A + D)/2$ es muy pequeña — menor a 5 — , se debe utilizar la prueba binomial (véase la sección correspondiente en el cap. 3), en lugar de la prueba de McNemar. Para emplear la prueba binomial, N deberá ser la suma de las celdillas A y D $(N = A + D)$, x será más pequeña que ambas frecuencias observadas (A y D) y se utilizará la tabla D del Apéndice I para evaluar la significación de x.

Como se advierte, pudimos haber analizado los datos de la tabla 4.3 utilizando la prueba binomial. En este caso, la hipótesis hubiera sido que los casos de la muestra de $N = A + D$ pertenecerían a una población binomial donde $p = q = 1/2$. Para los datos mencionados, $N = 20$ y $x = 7$ (x es la más pequeña de las frecuencias observadas). La tabla D del Apéndice I nos proporciona la probabilidad según H_0 de observar siete o menos cambios en una dirección. La probabilidad es 0.132, la cual, cuando se duplica, produce la probabilidad asociada con la prueba del cambio bidireccional, que para este ejemplo es 0.264. Así, el resultado es esencialmente el mismo que el obtenido utilizando la prueba de McNemar. La diferencia entre las dos se debe principalmente al hecho de que la distribución ji cuadrada no incluye los valores de la probabilidad entre 0.20 y 0.30. Aun teniendo la tabla de la distribución ji cuadrada (tabla C del Apéndice I) más completa, la probabilidad sería la misma que la obtenida mediante la prueba binomial, ya que la distribución muestral de X^2 es sólo la asíntota de la distribución ji cuadrada. Por supuesto, con muestras pequeñas no debemos esperar una correspondencia estrecha en las probabilidades cuando utilizamos ambas pruebas.

Resumen del procedimiento

Los pasos para el cálculo de la prueba del cambio de McNemar son los siguientes:

1. Presente las frecuencias observadas en una tabla de 2×2, como se ilustra en la tabla 4.1.
2. Determine el número total de cambios, $A + D$. Si el total es menor a 10, utilice la prueba binomial (véase el capítulo 3) en lugar de la prueba de McNemar.
3. Si el número total de cambios es mayor de 10, proceda a calcular el valor de X^2 utilizando la ecuación (4.2).

4. Determine la probabilidad asociada con el valor tan grande como el valor de X^2 recurriendo a la tabla C del Apéndice I. Si utiliza la prueba de una cola, divida el valor de la probabilidad que proporciona la tabla. Si el valor de la probabilidad de la tabla para el valor observado de X^2 con $gl = 1$ es menor o igual a X^2, rechace H_0 y acepte H_1.

Potencia-eficacia

Cuando la prueba de McNemar se utiliza con medidas nominales, el concepto potencia-eficacia no tiene mucho sentido porque no existen alternativas con las cuales comparar la prueba. Sin embargo, cuando las medidas y otros aspectos de los datos son tales que es posible aplicar la prueba paramétrica t, tanto la prueba de McNemar como la prueba binomial tienen una potencia-eficacia de alrededor del 95 % para $A + D = 6$; la potencia-eficacia va decrementando conforme $A + D$ es más pequeña y se vuelve asintótico en el nivel del 63 %.

Referencias bibliográficas

Los análisis relativos a esta prueba se presentan en McNemar (1969) y Everitt (1977).

PRUEBA DE LOS SIGNOS

Función

La prueba de los signos adquiere su nombre del hecho que está basada en la dirección de las diferencias entre dos mediciones, más que en medidas cuantitativas (los datos de donde proceden las diferencias). Es particularmente aplicable a investigaciones en las cuales las mediciones cuantitativas son imposibles o no son viables, pero en las que sí se puede determinar, para cada par de observaciones, cuál es la "más grande" (en algún sentido).

La prueba de los signos es aplicable al caso de dos muestras relacionadas cuando el investigador desea establecer que dos condiciones son diferentes. La única suposición que subyace a esta prueba es que la variable estudiada tiene una distribución continua. La prueba no hace suposiciones acerca de la forma de la distribución y tampoco supone que los sujetos pertenecen a la misma población. Los diferentes pares pueden pertenecer a diferentes poblaciones en cuanto a edad, sexo, inteligencia, etc.; el único requisito es que dentro de cada par, el investigador haya igualado respecto a las variables extrañas relevantes. Como se mencionó al principio de este capítulo, o bien utilizar a cada sujeto como su propio control.

Método

La hipótesis nula evaluada por la prueba de los signos es si

$$P[X_i > Y_i] = P[X_i < Y_i] = 1/2$$

donde X_i es el juicio o puntuación de acuerdo con una condición (o antes del tratamiento) y Y_i es el juicio o puntuación de acuerdo con la otra condición (o después del tratamiento). Esto es, X_i y Y_i son las dos puntuaciones obtenidas por cada miembro de la pareja. Otra manera de plantear la H_0 es la siguiente: la mediana de las diferencias entre X y Y es cero.

Durante la aplicación de la prueba de los signos debemos prestar especial atención a la dirección de la diferencia de cada X_i y Y_i, notando dónde el signo de la diferencia es positivo o negativo ($+$ o $-$). Cuando H_0 es verdadera, debemos esperar que el número de pares donde $X_i > Y_i$ sea igual al número de pares donde $X_i < Y_i$. Vale decir, si la hipótesis nula fuera verdadera, esperaríamos que alrededor de la mitad de diferencias fuera positiva y la otra mitad fuera negativa. Se rechaza H_0 si ocurren pocas diferencias con el mismo signo.

Muestras pequeñas

La probabilidad asociada a la ocurrencia de un número particular de positivos ($+$) y negativos ($-$) puede determinarse recurriendo a la distribución binomial con $p = q = 1/2$, donde N es el número de pares. Si algún(os) par(es) no muestran diferencia por tanto, no existe signo, dichos datos son excluidos del análisis y N se reduce, respectivamente. La tabla D del Apéndice I nos proporciona las probabilidades asociadas a la ocurrencia de acuerdo con valores de H_0 tan pequeños como x para $N \leq 35$. Para utilizar esta tabla, x será el número de signos menor.

Por ejemplo, supongamos que observamos 20 pares, de los cuales 16 muestran diferencias en una dirección ($+$) y los otros 4 muestran diferencias en la otra dirección ($-$). En este caso, $N = 20$ y $x = 4$. Al remitirnos a la tabla D del Apéndice I, ésta revela que la probabilidad de estos pocos signos negativos cuando H_0 es verdadera (esto es, que $p = 1/2$) es 0.006 (unidireccional).

La prueba de los signos puede ser tanto unidireccional como bidireccional. En la prueba unidireccional, la hipótesis alterna (H_1) plantea que un signo ($+$ o $-$) ocurrirá más frecuentemente. En la prueba bidireccional, la predicción es simplemente que las frecuencias de los signos diferirán significativamente. Para la prueba bidireccional, los valores de probabilidad de la tabla D del Apéndice I deberán duplicarse.

Ejemplo para muestras pequeñas. Un investigador estaba estudiando el proceso de toma de decisión esposo-esposa.[2] Se estudió exhaustivamente una muestra de parejas esposo-esposa para determinar el papel percibido de cada uno de ellos respecto de mejorar las adquisiciones domésticas. En cada ocasión, una pareja (cada uno por separado) contestaba un cuestionario concerniente a la influencia que creía ejercer cuando el matrimonio enfrentaba una situación en la que tenía que decidirse la adquisición de enseres para el hogar. Las respuestas a las preguntas se evaluaban mediante una escala que iba de esposo dominante a esposa dominante. Para cada pareja, la diferencia entre sus "percepciones" era determinada y codificada como $+$ si a juicio del esposo, la esposa no debería tener una mayor influencia que él y esto no coincidía con lo informado por la esposa (esposo: "mi opinión debería tener mayor peso que la de ella", y esposa: "ambos deberíamos ponernos de acuerdo para

[2] Este ejemplo es propuesto por Qualls, W. J. (1982), y consiste en un estudio acerca de las decisiones de pareja esposo-esposa en cuanto a las adquisiciones domésticas. Tesis doctoral inédita, Universidad de Indiana.

decidir"). La diferencia se codificaba como (−) cuando ocurría el caso contrario. La diferencia se codificaba como 0 (cero) si la pareja estaba en completo acuerdo acerca del grado de influencia ejercida en la decisión.

i. *Hipótesis nula*. H_0: los esposos y esposas están de acuerdo en el grado de influencia que ambos deben tener cuando deciden sobre las adquisiciones domésticas. H_1: los esposos juzgan que ellos deben tener mayor influencia que sus esposas acerca de las decisiones de adquirir enseres para el hogar.
ii. *Prueba estadística*. La escala utilizada en este estudio es una escala parcialmente ordenada. La información contenida en los juicios se mantiene si las diferencias entre las parejas se puede expresar por medio de un signo (+ o −). Cada pareja en este estudio constituye un par igualado; están igualados en el sentido de que cada uno de ellos respondió a la misma pregunta. La prueba de los signos es apropiada para la clase o el tipo de medidas descritas y, por supuesto, para el caso de muestras relacionadas o igualadas.
iii. *Nivel de significación*. $\alpha = 0.05$ y N es el número de parejas en una de las condiciones $= 17$ (N puede reducirse si ocurren empates).
iv. *Distribución muestral*. La probabilidad asociada a la ocurrencia de los valores tan grandes como x, es proporcionada por la distribución binomial para $p = q = 1/2$. La distribución binomial para los valores seleccionados de N se presenta en la tabla D del Apéndice I.
v. *Región de rechazo*. Puesto que H_1 predice la dirección de las diferencias, la región de rechazo es unidireccional. Ésta consiste en todos los valores de x (donde x es

Tabla 4.4. Juicios acerca de la influencia en la toma de decisiones.

	Tasa de influencia			
Pareja	*Esposo*	*Esposa*	*Dirección en la diferencia*	*Signo*
A	5	3	$X_H > X_W$	+
B	4	3	$X_H > X_W$	+
C	6	4	$X_H > X_W$	+
D	6	5	$X_H > X_W$	+
E	3	3	$X_H = X_W$	0
F	2	3	$X_H < X_W$	−
G	5	2	$X_H > X_W$	+
H	3	3	$X_H = X_W$	0
I	1	2	$X_H < X_W$	−
J	4	3	$X_H > X_W$	+
K	5	2	$X_H > X_W$	+
L	4	2	$X_H > X_W$	+
M	4	5	$X_H < X_W$	−
N	7	2	$X_H > X_W$	+
O	5	5	$X_H = X_W$	0
P	5	3	$X_H > X_W$	+
Q	5	1	$X_H > X_W$	+

el número de signos positivos, dada la predicción para H_1 de que predominarán los signos positivos) para los valores de la probabilidad de ocurrencia (unidireccional); H_0 es verdadera cuando éstos son iguales o menores que $\alpha = 0.05$.

vi. *Decisión.* Los juicios acerca de la influencia de los esposos varían en una escala de 1 a 7. En esta escala, el 1 representa al juicio en que la esposa tiene la autoridad completa sobre la decisión; una puntuación de 7 representa al juicio donde es el esposo quien tiene la autoridad completa; los valores intermedios representan el juicio de diferentes niveles de acuerdo o influencia. En la tabla 4.4 se muestran las puntuaciones asignadas para cada esposo (*H*) y esposa (*W*) de las 17 parejas. Los signos de las diferencias entre las puntuaciones de los pares se presentan en la última columna de la tabla. Nótese que tres parejas mostraron diferencias opuestas a las predichas; éstas se codificaron como ($-$). Otras tres parejas estuvieron completamente de acuerdo en el nivel de influencia de los miembros de la pareja; por tanto, se declaró empate y se redujo la N a 14 ($N = 17 - 3$). Las parejas restantes mostraron las diferencias en la dirección predicha.

Para los datos de la tabla 4.4, x es el número de signos positivos $= 11$ y N el número de pares iguales $= 14$. En la tabla D del Apéndice I se muestra que para $N = 14$ la probabilidad de observar $x \geq 11$ (de una cola) es de 0.029. Puesto que este valor está en la región de rechazo para $\alpha = 0.05$, nuestra decisión es rechazar H_0 en favor de H_1. Así, podemos concluir que los esposos creen que son ellos los que deben tener una mayor influencia al momento de tomar decisiones acerca de la adquisición de enseres domésticos, en comparación con la que deben tener las esposas.

Empates

Para la prueba de los signos ocurre un empate cuando no es posible discriminar entre los valores de un par igualado o ambos valores son iguales. En el ejemplo anterior de las parejas, ocurrieron tres empates: el investigador consideró que esas parejas coincidieron totalmente en sus juicios.

Todos los casos que representan empates son excluidos del análisis en la prueba de los signos y entonces la N se decrementa el mismo número de empates que existan. N es el número de pares igualados para quienes la diferencia tiene un signo ($+$ o $-$). En el ejemplo anterior, 14 de las 17 parejas tenían diferencias en las puntuaciones, de tal forma que para este estudio N fue igual a 14 ($N = 14$).

Relación con la expansión binomial

En el estudio presentado anteriormente, deberíamos esperar que cuando la H_0 es verdadera, la frecuencia en los signos positivos y negativos fuera la misma que las caras y cruces de 14 lanzamientos de moneda. (De manera más exacta, que de los 17 lanzamientos, tres monedas cayeran de canto, las cuales no contarían para el análisis posterior.) La probabilidad de obtener 11 caras y tres cruces en 14 lanzamientos nos la proporciona la distribución binomial como

$$\sum_{i=x}^{N} \binom{N}{i} p^i q^{N-i}$$

donde

$$N = \text{número de monedas lanzadas} = 14$$
$$x = \text{número de caras obtenidas} = 11$$

y
$$\binom{N}{i} = \frac{N!}{i!(N-i)!}$$

En el caso de 11 o más caras cuando se lanzan 14 monedas

$$P[x \geq 11] = \frac{\binom{14}{11} + \binom{14}{12} + \binom{14}{13} + \binom{14}{14}}{2^{14}}$$

$$= \frac{364 + 91 + 14 + 1}{16\ 284}$$

$$= 0.029$$

La probabilidad encontrada por el método anterior es, por supuesto, idéntica al valor encontrado en el ejemplo de las parejas.

Muestras grandes

Si N es mayor que 35 ($N \geq 35$), puede utilizarse la aproximación normal a la distribución binomial. Esta distribución tiene una

$$\text{Media} = \mu_x = Np = \frac{N}{2}$$

y una
$$\text{Varianza} = \sigma_x^2 = Npq = \frac{N}{4}$$

Esto es, el valor de z está dado por:

$$z = \frac{x - \mu_x}{\sigma_x} = \frac{x - N/2}{0.5\sqrt{N}} \tag{4.3}$$

$$= \frac{2x - N}{\sqrt{N}} \tag{4.3a}$$

Esta expresión esta distribuida normalmente (de una manera aproximada) con una media igual a cero y la varianza igual a uno. La ecuación (4.3*a*) es más conveniente para realizar los cálculos; sin embargo, complica un poco la forma de la prueba.

La aproximación llega a ser mejor cuando se emplea la corrección por continuidad. Esta corrección se efectúa reduciendo la diferencia entre el número observado de signos positivos (o negativos) y el número esperado (la media) cuando H_0 es verdadera al 0.5 (para un análisis más completo de este tema, véase la pág. 65). Así, con la corrección de la continuidad queda:

$$z = \frac{(x \pm 0.5) - N/2}{0.5\sqrt{N}} \tag{4.4}$$

donde $X + 0.5$ se utiliza cuando $X < N/2$, y $x - 0.5$ se usa cuando $x > N/2$. La siguiente ecuación en una forma simplificada de la anterior que facilita los cálculos:

$$z = \frac{2x \pm 1 - N}{\sqrt{N}} \tag{4.4a}$$

Aquí usamos $+ 1$ cuando $x < N/2$, y $- 1$ cuando $x > N/2$. El valor obtenido de z mediante la aplicación de la ecuación (4.4) puede considerarse normalmente distribuida, con media igual a cero y varianza igual a uno. Por tanto, la significación de la z obtenida se determina haciendo referencia a la tabla A del Apéndice I, la cual nos proporciona los valores de probabilidad (unidireccional) asociada con la ocurrencia cuando H_0 es verdadera, con valores extremos observados de x. Si se requiere la prueba bidireccional, el valor de probabilidad obtenido en la tabla A debe duplicarse.

Ejemplo para muestras grandes. Supongamos que un investigador estuviera interesado en determinar si una cierta película acerca de delincuencia juvenil puede cambiar las opiniones de ciertos miembros de alguna comunidad particular, en relación con la severidad de las medidas punitivas aplicadas a menores infractores. Él obtiene una muestra al azar de 100 adultos de dicha comunidad y lleva a cabo un diseño "antes-después", teniendo a cada sujeto como su propio control. Pide a cada sujeto que dé su opinión acerca de la cantidad o el grado de las medidas punitivas que deberían aplicarse a menores infractores. Después les muestra la película a los 100 adultos y posteriormente les repite la pregunta.

i. *Hipótesis nula.* H_0: la película no tiene efectos sistemáticos sobre las opiniones; es decir, las posibles diferencias observadas se deberán más bien a lo esperado de una muestra tomada al azar de una población en la cual la película no tiene efectos sistemáticos. H_1: la película tiene efectos sistemáticos en las opiniones.
ii. *Prueba estadística.* Para este estudio se escogió la prueba de los signos por tratarse de una muestra relacionada y porque se utilizan medidas ordinales y, por tanto, las diferencias pueden ser representadas, adecuadamente, por signos positivos y negativos.
iii. *Nivel de significación.* $\alpha = 0.01$ y N es el número de adultos (probablemente menor que 100) que muestran cambio en su opinión.
iv. *Distribución muestral.* Cuando H_0 es verdadera, z calculada mediante las ecuaciones (4.4a) o (4.4) se encuentra aproximadamente distribuida de manera normal para $N > 35$. La tabla A del Apéndice I nos proporciona la probabilidad asociada de ocurrencia de valores tan extremos como la z obtenida.
v. *Región de rechazo.* Ya que H_1 no plantea la dirección de las diferencias predichas, la región de rechazo es bidireccional. Ésta consiste en todos los valores de z cuya

probabilidad de ocurrencia asociada sea extrema; cuando H_0 es verdadera, es menor o igual que $\alpha = 0.01$.

vi. *Decisión*. Los resultados del estudio anterior acerca del efecto de la película sobre la opinión se presentan en la tabla 4.5. ¿Tuvo algún efecto la película? Los resultados nos muestran que sólo 15 adultos no presentan cambio en su opinión y 85 que sí cambiaron. El análisis se fundamenta sólo en aquellos sujetos que cambiaron. Si la película no hubiera tenido un efecto sistemático, habríamos esperado que alrededor de la mitad de las personas que mostraron cambios en su opinión se repartiera equitativamente entre "incremento en la severidad" y "decremento en la severidad". Esto es, que de las 85 personas, 42.5 estuvieran en una categoría y 42.5 estuvieran en la categoría contraria. Podemos observar en la tabla 4.5 que 59 de ellas están en una categoría (decrementar la severidad) y 26 en la categoría contraria. Podemos determinar la probabilidad de que H_0 sea verdadera utilizando la educación (4.4), notando que $X > N/2$ $(59 > 42.5)$. Entonces tenemos

$$z = \frac{2x \pm 1 - N}{\sqrt{N}} \tag{4.4a}$$

$$= \frac{118 - 1 - 85}{\sqrt{85}}$$

$$= 3.47$$

La tabla A del Apéndice I nos revela que la probabilidad de $|z| \geq 3.47$ cuando H_0 es verdadera es $2x(0.0003) = 0.0006$. (El valor de la probabilidad se duplica porque la tabla de valores es unidireccional.) Puesto que 0.0006 es más pequeño que $\alpha = 0.01$, la decisión es rechazar la hipótesis nula en favor de la hipótesis alterna. Podemos concluir, a partir del análisis de los resultados, que la película tuvo efectos sistemáticos sobre la opinión de 100 adultos en relación con la severidad deseable de las medidas punitivas aplicables a menores infractores.

Este ejemplo se incluyó no sólo porque demuestra una aplicación exitosa de la prueba de los signos, sino porque a menudo los datos de este tipo se analizan incorrectamente. Los datos en la tabla 4.5 se nos presentan en forma de variables de interés. Se puede construir una tabla de 2×2 que contenga la misma información, pero requiere que conozcamos las frecuencias de las celdillas B y C.[3] Es demasiado común que los investigadores analicen tales datos como si representaran

Tabla 4.5. Opiniones de los adultos respecto al grado de severidad del castigo aplicado a menores infractores.

Opinión	*Número*
Incremento en la severidad	26
Decremento en la severidad	59
No hubo cambio	15

[3] Se exhorta al lector a que construya la tabla de 2×2 usando los valores $B = 7$ y $C = 8$.

muestras independientes. Éste no es el caso; los totales por renglón y columna están separados, pero no son representaciones independientes de los mismos datos.

Este ejemplo también pudo ser analizado con la prueba de McNemar para la significación de los cambios (véase la sección correspondiente). Usando los datos de la tabla 4.5 tenemos,

$$X^2 = \frac{(|A - D| - 1)^2}{A + D} \quad \text{con } gl = 1 \tag{4.2}$$

$$= \frac{(|59 - 26| - 1)^2}{59 + 26}$$

$$= 12.05$$

La tabla C del Apéndice I muestra que $X^2 \leq 12.05$ con gl $= 1$ tiene una probabilidad de ocurrencia cuando H_0 es verdadera menor que 0.001. Este resultado no se contrapone con el de la prueba de los signos. La pequeña diferencia entre ambos resultados se debe a las limitaciones de la tabla utilizada en la distribución ji cuadrada. Debe notarse que si z se calcula utilizando la ecuación (4.3) y si X^2 se calcula con la ecuación (4.1) (es decir, no se hace la corrección por continuidad en ambos casos), entonces z^2 será idéntica a X^2 para cualquier conjunto de datos. Lo mismo se cumple si los cálculos se realizan utilizando la corrección por continuidad [ecuaciones (4.2) y (4.4)].

Resumen del procedimiento

Los siguientes son los pasos que hay que dar para utilizar la prueba de los signos:

1. Determine el signo de las diferencias entre los dos miembros de cada par.
2. Determine el valor de N, que debe ser igual al número de pares cuya diferencia muestra un signo (los empates se excluyen del análisis).
3. El método para determinar la probabilidad de ocurrencia de los datos cuando H_0 es verdadera depende del tamaño de N:

   a) Si $N \leq 35$, la tabla D del Apéndice I muestra la probabilidad asociada (una cola) con valores tan pequeños observados de $x =$ el número de signos menor. Para una región de rechazo de dos colas, duplíquese la probabilidad proporcionada por dicha tabla.

   b) Si $N > 35$, calcule el valor de z utilizando la ecuación (4.2*a*). La tabla A del Apéndice I nos proporciona las probabilidades asociadas (unidireccional) a los valores de z. Para una región de rechazo bidireccional, duplíquese la probabilidad proporcionada por dicha tabla.

4. Si la probabilidad mostrada por la prueba es menor o igual a α, rechácese H_0.

Potencia-eficacia

La potencia-eficacia de la prueba de los signos es de alrededor del 95 % para $N = 6$, pero decrementa respecto al incremento del tamaño de la muestra hasta una eventual eficacia del 63 % (asíntota) (mientras más grande sea la muestra, más pequeña es la eficacia). El lector encontrará en Lehmann (1975) una exposición respecto a la potencia-eficacia de la prueba de los signos para muestras grandes.

Referencias bibliográficas

Para otros análisis respecto a la prueba de los signos, el lector debe consultar Dixon y Massey (1983), Lehmann (1975), Moses (1952) y Randles y Wolfe (1979).

PRUEBA DE RANGOS ASIGNADOS DE WILCOXON

La prueba de los signos examinada en la sección anterior utiliza información sólo en términos de la dirección de las diferencias en cada uno de los pares analizados. Si se consideran tanto la magnitud relativa como la dirección de las diferencias, se puede utilizar una prueba más poderosa. La prueba de rangos asignados de Wilcoxon adjudica mayor peso a los pares que muestran mayores diferencias entre las dos condiciones, más que a los pares cuya diferencia es pequeña.

La prueba de Wilcoxon se aplica con bastante éxito en las ciencias de la conducta. Con datos conductuales, es común que el investigador pueda: 1. determinar cuál miembro del par es "más grande que" (determinar el signo de la diferencia entre cualquier par) y 2. establecer rangos en las diferencias en orden de tamaño absoluto. Esto es, el investigador puede hacer juicios de "mayor que" entre los valores de cualquier par, tanto como acerca de las diferencias entre dos pares cualesquiera. Con esta información, el investigador puede utilizar la prueba de Wilcoxon.

Racionalización y método

La diferencia de las puntuaciones entre los miembros del par igualado (d_i) representa la diferencia entre las puntuaciones del par en los dos tratamientos (X y Y). Esto es, $d_1 = X_i - Y_i$. Para utilizar la prueba de Wilcoxon, se deben poner en columna todas las diferencias sin tener en cuenta el signo: adjudique el rango 1 a la d_i más pequeña, el rango 2 a la siguiente menos pequeña, etc. Cuando se tiene que decidir el rango entre un -1 y un $+2$ o -2, el más pequeño será -1.

Entonces, a cada *rango* se debe añadir el signo de la diferencia. Así podemos indicar e identificar los rangos de las diferencias positivas, de los rangos de las diferencias negativas.

La hipótesis nula es que los tratamientos X y Y son equivalentes, esto es, son muestras de la misma población, con la misma mediana y la misma distribución continua. Si H_0 es verdadera, deberíamos encontrar algunas diferencias en favor del tratamiento X y otras diferencias en favor del tratamiento Y. Es decir, si sumamos los rangos que tienen signo positivo y aquellos con signo negativo, esperaríamos que ambas sumas fueran iguales (siempre que H_0 sea verdadera). Pero si la suma de los rangos positivos es muy diferente de la suma de los rangos negativos, inferiríamos que el tratamiento X difiere del tratamiento Y y, por tanto, rechazaríamos la H_0. Es decir, rechazamos H_0 siempre que cualquiera de las sumas de las diferencias (positivas o negativas) sea demasiado pequeña.

Para desarrollar esta prueba definiremos dos estadísticos:

$$T^+ = \text{suma de los rangos de las diferencias positivas}$$

$$T^- = \text{suma de los rangos de las diferencias negativas}$$

De lo anterior, la suma de todos los rangos es

$$N(N+1)/2, T^- = N(N+1)/2 - T^+$$

Empates

Ocasionalmente las dos puntuaciones de cualquier par son iguales. Es decir, no existe diferencia entre los miembros de ese par, así que $X_i - Y_i = d_i = 0$. Tales pares son excluidos del análisis y el tamaño de N se reduce respectivamente. Es lo mismo que se hizo con la prueba de los signos. Así, N es el número de pares igualados menos el número de pares donde $X = Y$.

Puede ocurrir otro tipo de empate cuando dos o más diferencias son de la misma magnitud. A estos casos se les asigna el mismo rango, el cual se calcula de la siguiente manera: imaginemos que tres pares presentan diferencias de la misma magnitud, -1, -1 y $+1$, a cada par se le asigna el rango 2, ¿por qué? Porque los rangos que les corresponderían se promediaron $(1 + 2 + 3)/3 = 2$; el rango que correspondería al par siguiente sería 4, porque los rangos 1, 2 y 3 ya fueron asignados. Si el ejemplo se realizara con dos pares, el rango sería 1.5, ya que $(1 + 2)/2 = 1.5$ y la diferencia siguiente recibiría el rango 3. Lo anterior tiene como objetivo que la prueba de Wilcoxon sea más adecuada.

Para la aplicación de los principios de manejo de empates, el lector debe consultar el ejemplo para muestras grandes en la sección correspondiente.

Muestras pequeñas

Sea T^+ la suma de los rangos para los cuales las diferencias (d_i) fueron positivas. La tabla H del Apéndice I nos proporciona varios valores de T^+ y sus probabilidades de ocurrencia asociadas, en la suposición de que no existen diferencias en los grupos X y Y. Esto es, si una T^+ observada es igual al valor presentado en la tabla H para un tamaño de muestra (N) particular, la probabilidad de un valor de T^+ tan grande es tabulada. Si la probabilidad es menor o igual al nivel de significación obtenido, la hipótesis nula puede rechazarse en ese nivel de significación.

La tabla H del Apéndice I se utiliza para pruebas tanto unidireccionales como bidireccionales. Una prueba unidireccional es adecuada si el investigador ha predicho alguna dirección particular de las diferencias. Para pruebas bidireccionales, se tiene que duplicar el valor proporcionado por la tabla. Por ejemplo, si $T^+ = 42$ es la suma de los rangos positivos cuando $N = 9$, podemos rechazar H_0 en el nivel de $\alpha = 0.02$ si la prueba ha sido (bidireccional) y se puede rechazar H_0 en el nivel de 0.01 si H_1 plantea que la mediana de X es mayor a la mediana de Y (unidireccional).

Ejemplo para muestras pequeñas. Existe considerable evidencia acerca de que los adultos son capaces de utilizar señales visuales en el procesamiento de información auditiva. En una conversación normal, las personas pueden utilizar los movimientos de los labios en el procesamiento de la charla. La congruencia entre los movimientos de los labios y los sonidos del habla son particularmente benéficos en ambientes ruidosos. La investigación ha demostrado que el procesamiento del habla se deteriora cuando las señales auditivas y visuales no son congruentes. En los niños, la habilidad para discriminar y localizar la fuente de estímulos auditivos y visuales complejos se establece alrededor de los seis meses de edad.

Se diseñó un experimento para determinar si los niños de 10 a 16 semanas de edad se dan cuenta de la sincronía entre los movimientos de los labios y los sonidos del habla en una conversación normal.[4] Los niños se colocaron en una habitación a prueba de ruido, que tenía una ventana a través de la cual podían ver a una persona hablando. La persona hablaba en un micrófono y el sonido era dirigido directamente al cuarto (en sincronía) o con una demora de 400 milisegundos (fuera de sincronía). En cada condición se midió el tiempo que el niño miraba la cara de la persona que hablaba. Se argumentó que si el pequeño es capaz de discriminar ambas condiciones, la cantidad de tiempo de ver a la persona sería diferente, aunque *a priori* no se planteó en cuál de las dos condiciones el tiempo sería mayor (en términos de la poca experiencia en sincronía y lo novedoso que era fuera de sincronía).

Existen considerables diferencias individuales entre los infantes respecto al tiempo que pasaron atendiendo al estímulo. Sin embargo, la diferencia en el tiempo que pasaron viendo en la condición en sincronía y el tiempo que pasaron viendo en la condición fuera de sincronía podría ser un indicador confiable de la capacidad de discrminar. Si el niño pasa más tiempo atendiendo al estímulo en la presentación sincrónica, la diferencia sería positiva; y si el niño pasa más tiempo atendiendo al estímulo en la presentación asincrónica, la diferencia sería negativa. Si el pequeño es capaz de discriminar, las diferencias deberían tender hacia una dirección; más aún, cualquier diferencia en dirección contraria debería ser relativamente pequeña.

Aunque el investigador confía en que las diferencias en el tiempo promedio que se pasaron mirando indican las diferencias en la atención, no está seguro de que las puntuaciones sean suficientemente precisas para que sean representadas en una escala que no sea ordinal. Esto es, sólo puede afirmar que las grandes diferencias reflejan incrementos en la atención; por ejemplo, una diferencia de 30 indica una mayor diferencia en atención que una diferencia de 20. Así, aunque la interpretación de las direrencias en las magnitudes numéricas en el tiempo de mirar no reflejan directamente las diferencias en la atención, el establecer los *rangos* de las diferencias en el mirar reflejará el orden de las diferencias en el atender al estímulo.

i. *Hipótesis nula.* H_0: la cantidad de tiempo que pasan los niños viendo a través de la ventana no depende del tipo de presentación. En términos de la prueba de Wilcoxon, la suma de los rangos positivos no difiere de la suma de los rangos negativos. La hipótesis alterna es H_1: la cantidad de tiempo que los niños pasan viendo depende del tipo de presentación; la suma de los rangos positivos difiere de la suma de los rangos negativos.

[4] Dodd, B., "Lip reading in infants: Attention to speech presented in- and out-of-synchrony" en *Cognitive Psychology*, núm. 11, 1979, págs. 478-484.

ii. *Prueba estadística.* Se selecciona la prueba de Wilcoxon porque en el estudio se emplean dos muestras relacionadas y las diferencias en las puntuaciones pueden ser ordenadas por medio de rangos.

iii. *Nivel de significación.* Plantearemos que $\alpha = 0.01$ y N es el número de pares utilizados (12) menos el número de pares cuyas diferencias sean $d_i = 0$.

iv. *Distribución muestral.* La tabla H del Apéndice I nos proporciona los valores de probabilidad de la distribución muestral de T^+ para $N \leq 15$.

v. *Región de rechazo.* Como no se predice la dirección de las diferencias, una región de rechazo bidireccional es la apropiada. La región de rechazo consiste en todos los valores de T^+ (suma de rangos positivos) cuya probabilidad asociada cuando H_0 es verdadera, es menor o igual a $\alpha = 0.01$ para una prueba bidireccional.

vi. *Decisión.* En este estudio se utilizaron 12 niños como sujetos. El porcentaje del tiempo que pasaron viendo a través de la ventana se muestra en la tabla 4.6, donde se advierte que sólo en dos niños (*RH* y *CW*) se observan diferencias en la dirección de las presentaciones en sincronía. Las diferencias para las puntuaciones de esos dos niños son las más pequeñas; sus rangos son 1 y 4.
La suma de los rangos positivos es $T^+ = 10 + 12 + 6 + 3 + 8 + 5 + 11 + 9 + 2 + 7 = 73$.
La tabla H del Apéndice I nos muestra que con $N = 12$ y $T^+ = 73$, debemos rechazar la hipótesis nula en $\alpha = 0.01$ para una prueba bidireccional, puesto que el valor de tablas es (0.0024) y para una prueba bidireccional se duplica (0.0048). Se rechaza H_0 en favor de H_1 y concluimos que los niños son capaces de discriminar entre la sincronía o asincronía en los movimientos de los labios y los sonidos del habla.

Es importante notar que los datos de la tabla 4.6 pudieron ser analizados con la prueba de los signos (véase la sección correspondiente), que es una prueba menos poderosa. Para esta prueba, $x = 2$ y $N = 12$. La tabla D del Apéndice I nos proporciona la probabilidad asociada con tal ocurrencia cuando H_0 es verdadera

Tabla 4.6. Porcentaje de falta de atención en presencia de sincronía y sin ella.

Sujeto	*En sincronía*	*Fuera de sincronía*	*d*	*Rango de d*
DC	20.3	50.4	30.1	10
MK	17.0	87.0	70.0	12
VH	6.5	25.1	18.6	6
JM	25.0	28.5	3.5	3
SB	5.4	26.9	21.5	8
MM	29.2	36.6	7.4	5
RH	2.9	1.0	− 1.9	−1
DJ	6.6	43.8	37.2	11
JD	15.8	44.2	28.4	9
ZC	8.3	10.4	2.1	2
CW	34.0	29.9	− 4.1	−4
AF	8.0	27.7	19.7	7

$N = 12$, $T^+ = 73$, $T^- = 5$.

de 2(0.019) = 0.038 para una prueba bidireccional. Así, utilizando la prueba de los signos la decisión sería no rechazar H_0 con $\alpha = 0.01$, aunque la prueba de Wilcoxon nos permite rechazarla. Esta diferencia no es sorprendente, ya que la prueba de Wilcoxon considera que las dos diferencias negativas observadas son las más pequeñas, mientras que la prueba de los signos no es afectada por la magnitud relativa de las diferencias (d_i).

Muestras grandes

Cuando N es mayor que 15, la tabla H del Apéndice I no puede utilizarse; sin embargo, se puede demostrar que en tales casos la suma de los rangos T^+ se distribuye aproximadamente de manera normal con

$$\text{Media} = \mu_{T^+} = \frac{N(N+1)}{4}$$

y

$$\text{Varianza} = \sigma^2_{T^+} = \frac{N(N+1)(2N+1)}{24}$$

Por tanto

$$z = \frac{T^+ - \mu_{T^+}}{\sigma_{T^+}} = \frac{T^+ - N(N+1)/4}{\sqrt{N(N+1)(2N+1)/24}} \tag{4.5}$$

también se distribuye aproximadamente de manera normal con una media igual a cero y una varianza igual a uno. Así, la tabla A del Apéndice I puede utilizarse para encontrar la probabilidad asociada a los valores tan extremos de z, calculada mediante la ecuación (4.5); cuando H_0 es verdadera.

Aunque la prueba para muestras grandes parece ser una mejor aproximación aun para muestras relativamente pequeñas, la correspondencia entre la probabilidad exacta y aproximada para una muestra de tamaño determinado depende del valor de T^+. En tanto el tamaño de la muestra sea mayor la probabilidad aproximada será mejor.

Ejemplo para muestras grandes. Los internos de una prisión federal sirvieron como sujetos en un estudio de toma de decisiones.[5] Primero se midió de manera individual la utilidad de los cigarrillos para los reclusos (valor subjetivo), ya que los cigarrillos son el objeto más negociado en las prisiones. Haciendo uso de la función de utilidad de cada sujeto, el investigador intentó predecir las decisiones que cada hombre haría en un juego donde repetidamente tendría que escoger entre dos opciones y en las cuales pudiera perder o ganar cigarrillos.

La hipótesis evaluada fue que los investigadores podrían predecir las decisiones de los sujetos en términos del valor de utilidad, en lugar de suponer que la utilidad de los cigarrillos es equivalente al valor objetivo de los mismos y, por tanto, predecir la elección "racional" en términos del valor objetivo. Esta hipótesis fue confirmada.

[5] Hurst, P.M. y Siegel, S., "Prediction of decisions from a higher ordered metric scale", en *Journal of Experimental Psychology*, núm. 52, 1956, págs. 138-144.

Sin embargo, como era de esperarse, algunas respuestas no se predijeron con éxito, por la hipótesis de maximización de la utilidad esperada. Los investigadores habían conjeturado que tales errores en la predicción se deberían a la probable indiferencia de los sujetos hacia las dos opciones disponibles. Esto es, un recluso podía encontrar igualmente atractivas ambas opciones o no parecerle atractiva ninguna y, por tanto, ser indiferente a la elección entre ambas opciones. Tales elecciones (de indiferencia) son difíciles de predecir. Pero en tales casos, se razonó que el sujeto vacilaría en plantear una apuesta y tardaría más en decidir. Es decir, la latencia entre el ofrecimiento de las opciones y el aceptar alguna de ellas sería mayor. La segunda hipótesis fue que las latencias o tiempos de respuesta que no se predijeran satisfactoriamente por la maximización de la utilidad esperada, serían mayores que las elecciones predichas con éxito.

i. *Hipótesis nula.* H_0: no hay diferencia entre las latencias de las decisiones predichas correctamente y las predichas incorrectamente. H_1: las latencias de las elecciones predichas incorrectamente serán mayores que las predichas correctamente.
ii. *Prueba estadística.* Se seleccionó la prueba de Wilcoxon puesto que los datos representan diferencias entre las puntuaciones de dos muestras relacionadas (decisiones predichas correctamente y decisiones predichas incorrectamente en el mismo recluso), donde cada sujeto sirve como su propio control.
iii. *Nivel de significación.* $\alpha = 0.01$ y N es el número de reclusos que sirvieron como sujetos $= 30$ (el tamaño de N se puede reducir si se encuentran empates).
iv. *Distribución muestral.* Cuando H_0 es verdadera, los valores calculados de z mediante la ecuación (4.5) se distribuyen asintóticamente de manera normal, con una media igual a cero y una varianza igual a uno. Así, la tabla A del Apéndice I nos proporciona la probabilidad asociada a la ocurrencia según H_0 de valores extremos como la z obtenida.
v. *Región de rechazo.* Puesto que se predice la dirección, la prueba es unidireccional. T^+, la suma de los rangos positivos, será la suma de los rangos de los reclusos cuyas diferencias se encuentran en la dirección predicha. La región de rechazo consiste en todos los valores de z (obtenidos de T^+), tan extremos como la probabilidad asociada cuando H_0 es verdadera cuyo valor es igual o menor que $\alpha = 0.01$.
vi. *Decisión.* Las diferencias para cada sujeto se calcularon restando su tiempo promedio en tomar la opción correctamente predicha (Y_i) del tiempo promedio en tomar la opción incorrectamente predicha (X_i) ($d_i = X_i - Y_i$). En la tabla 4.7 se muestran los valores de las diferencias de los 30 sujetos, así como los valores necesarios para completar la prueba de Wilcoxon. Una diferencia negativa indica que el tiempo promedio de las decisiones correctamente predichas fue mayor al tiempo promedio de las decisiones incorrectamente predichas.

Para los datos de la tabla 4.7, $T^+ = 298$, aplicando la ecuación (4.5) tenemos:

$$z = \frac{T^+ - \mu_{T^+}}{\sigma_{T^+}} = \frac{T^+ - N(N+1)/4}{\sqrt{N(N+1)(2N+1)/24}} \tag{4.5}$$

$$= \frac{298 - (26)(27)/4}{\sqrt{(26)(27)(53)/24}}$$

$$= 3.11$$

Nótese que tenemos una $N = 26$, ya que las mediciones de cuatro reclusos muestran empates. Nótese, además, que el valor de la suma T^+ de los rangos de los reclusos cuyas diferencias se encuentran en la dirección predicha, justifica utili-

Tabla 4.7. Diferencia en el tiempo promedio entre las decisiones correcta e incorrectamente predichas de los reclusos.

Reclusos	*d*	*Rango de d*
1	−2	−11.5
2	0	—
3	0	—
4	1	4.5
5	0	—
6	0	—
7	4	20.
8	4	20.
9	1	4.5
10	1	4.5
11	5	23.
12	3	16.5
13	5	23.
14	3	16.5
15	−1	−4.5
16	1	4.5
17	−1	−4.5
18	5	23.
19	8	25.5
20	2	11.5
21	2	11.5
22	2	11.5
23	−3	−16.5
24	−2	−11.5
25	1	4.5
26	4	20.
27	8	25.5
28	2	11.5
29	3	16.5
30	−1	−4.5

$N = 26$, $T^+ = 298$, $T^- = 53$.

zar la prueba unidireccional. La tabla A del Apéndice 1 muestra que el valor de z + 3.11 tiene una probabilidad asociada, cuando H_0 es verdadera, de 0.0009. En vista de que su probabilidad es menor que $\alpha = 0.01$ y el valor de z está en la zona de rechazo, nuestra decisión es rechazar H_0 en favor de H_1. Podemos concluir que las latencias para las decisiones incorrectamente predichas fueron significativamente mayores que las latencias de las decisiones correctamente predichas.

Esta conclusión apoya a la idea de que las decisiones correctamente predichas en relación con las apuestas son equivalentes, o aproximadamente equivalentes, a la utilidad esperada por los sujetos.

Rangos empatados y muestras grandes

Si existieran rangos con empates, es necesario ajustar la prueba estadística para considerar el decremento en la variabilidad de T. La corrección requiere contar los empates y reducir la varianza, respectivamente. En caso de existir rangos empatados, entonces

$$\sigma^2_{T^+} = \frac{N(N+1)(2N+1)}{24} - \frac{1}{2}\sum_{j=1}^{g} t_j(t_j - 1)(t_j + 1) \qquad (4.6)$$

donde

g = número de agrupamientos de diferentes rangos empatados
t_j = número de rangos empatados agrupados en j

En los datos del ejemplo anterior existe un gran número de empates. Hay $g = 6$ grupos de empates; 8 empates en el rango 4.5, 6 empates en el rango 11.5, etc. El factor de corrección de la varianza es 414. Este factor se calculó mediante la ecuación (4.6), de la siguiente manera:

Agrupamiento	*Rango*	t_j
1	4.5	8
2	11.5	6
3	16.5	4
4	20	3
5	23	3
6	25.5	2

La varianza sin corrección es 1 550.25, el valor de la varianza corregida es 1 550.25 − 414 = 1 136.25 [se calculó mediante la ecuación (4.6)]. El valor corregido de z es $z = 3.63$, mientras que el valor no corregido es $z = 3.11$. La corrección de la prueba de Wilcoxon siempre incrementa el valor de z cuando hay empates; por tanto, si una H_0 es rechazada sin corrección, será rechazada con corrección. Debe notarse, además, que el uso de la corrección cuando no hay empates no produce cambios en la varianza (todos los agrupamientos de empates serán de tamaño igual a 1).

Resumen del procedimiento

Para la aplicación de la prueba de Wilcoxon se deben observar los pasos siguientes:

1. Para cada par igualado de observaciones, X_i y Y_i, determine la diferencia con signo entre las dos variables ($d_i = X_i - Y_i$).
2. Ordene los rangos sin tener en cuenta su signo. A las diferencias que tengan el mismo valor, asígneles el rango promedio.
3. A cada rango asigne el signo ($+$ o $-$) de la diferencia correspondiente.
4. Determine N, que es el número de las diferencias que no son iguales a cero.
5. Determine T^+, que es la suma de los rangos que tienen signo positivo.
6. El procedimiento para determinar la significancia del valor observado de T^+ depende del tamaño de N:

   a) Si $N \leq 15$, la tabla H del Apéndice I nos proporciona la probabilidad asociada a los valores de T^+. Si la probabilidad es menor o igual que en nivel de significación (α) seleccionado, rechace H_0.
   b) Si $N > 15$, entonces calcule el valor de z utilizando la ecuación (4.5) y en caso de existir rangos con empates, corrija la varianza por medio de la ecuación (4.6). Determine la probabilidad asociada cuando H_0 es verdadera recurriendo a la tabla A del Apéndice I.

Para una prueba bidireccional, multiplique por dos el valor de tabla. Si la probabilidad obtenida de esta manera es menor o igual que α, rechace H_0.

Potencia-eficacia

La eficacia asintótica conforme a H_0 de la prueba de Wilcoxon comparada con la prueba t es $3\pi = 95.5$ % (Mood, 1954). Esto quiere decir que $3/\pi$ es la razón (proporción) límite del tamaño necesario de la muestra para que la prueba de Wilcoxon y la prueba t tengan el mismo poder. Para muestras pequeñas, la eficacia es cercana al 95 %.

Referencias bibliográficas

El lector puede encontrar comentarios adicionales acerca de la prueba de Wilcoxon en Wilcoxon (1945; 1947; 1949), Lehmann (1975) y Randles y Wolfe (1979).

PRUEBA DE LAS PERMUTACIONES PARA PARES REPLICADOS

Función

Las pruebas de las permutaciones son pruebas no paramétricas que no sólo tienen valor práctico en el análisis de los resultados, sino que además poseen valor heurístico ya que ayudan a exponer la naturaleza subyacente a las pruebas no paramétricas en general. Mediante una prueba de permutación podemos obtener la probabilidad exacta, cuando H_0 es verdadera, de la ocurrencia del dato observado, sin hacer ninguna suposición acerca de la normalidad, de la homogeneidad de la varianza o la forma precisa que subyace a la distribución. Las pruebas de permutación, en ciertas condiciones, son las técnicas no paramétricas más poderosas y resultan adecuadas en el momento en que las medidas sean tan precisas que los valores de las puntuaciones tengan significado numérico.

Racionalización y método

La prueba de las permutaciones supone que cuando realizamos observaciones pareadas para cada sujeto o bien en pares replicados, las puntuaciones observadas están asignadas al azar a las dos condiciones; es decir, suponemos que el sujeto (o el par) nos dio esas dos puntuaciones *sin considerar la condición*. Esto es lo que esperaríamos si la hipótesis nula (H_0) fuera verdadera. Así, si medimos a los sujetos en las dos ocasiones, se supone que las puntuaciones (X y Y) pudieron ser observadas en el orden $X \longrightarrow Y$ o en el orden $Y \longrightarrow X$. Si calculáramos las diferencias de las puntuaciones entre las condiciones, de acuerdo con la suposición de la asignación al azar, serían las mismas ya sea positivas o negativas. Sea $d_i = X_i - Y_i$ la diferencia para el sujeto iésimo; ésta es una medida de la diferencia entre las condiciones. Así, si H_0 fuera verdadera, suponemos que el signo de la diferencia d_i es positivo ($+$) en lugar de negativo ($-$), simplemente porque observamos las puntuaciones en un orden particular. Es como si supiéramos que el sujeto nos va a dar ciertas puntuaciones (X, Y) y lanzando al aire una moneda determináramos qué puntuación ocurrió primero. Si aplicáramos este razonamiento a todos los sujetos y si H_0 fuera verdadera, entonces cada diferencia que observáramos sería exactamente la misma, pero con el signo opuesto.

Supongamos que nuestra muestra consiste de $N = 8$ pares y que las diferencias en las puntuaciones ocurrieron de esta manera:

$$+19, \ +27, \ -1, \ +6, \ +7, \ +13, \ -4, \ +3$$

Cuando H_0 es verdadera, si los lanzamientos fueron diferentes, los valores pudieron ocurrir de la siguiente manera:

$$-19, \ -27, \ +1, \ -6, \ -7, \ -13, \ +4, \ -3$$

o si las monedas cayeron de otra manera, las observaciones pudieron ocurrir así:

$$+19, \ -27, \ +1, \ -6, \ -7, \ -13, \ -4, \ +3$$

De hecho, si la hipótesis nula fuera verdadera habría $2^N = 2^8 = 256$ resultados igualmente posibles y todas las asignaciones dependerían de la manera en que cayera la moneda. Esto significa que asociadas a la muestra de las puntuaciones observadas, habría otros resultados igualmente posibles en términos de $2^8 = 256$. Cuando H_0 es verdadera, cualquiera de los 256 resultados son posibles.

Para cada resultado posible hay una suma de las diferencias (Σd_i). Muchas de las posibles Σd_i son cercanas a cero; esto debería esperarse si la hipótesis nula fuera verdadera. Pocas Σd_i están lejos del cero. Éstas son aquellas combinaciones que esperaríamos si la H_0 fuera *falsa*; es decir, si la media de la población en cierta condición excediera a la otra.

Si deseamos evaluar H_0 en contra de H_1, debemos plantear una región de rechazo que consista de las combinaciones donde Σd_i sea mayor. Supongamos que $\alpha = 0.05$; entonces, la región de rechazo consistiría en el 5 % de las combinaciones posibles, las cuales contendrían los valores más extremos de Σd_i.

En el ejemplo que estamos manejando, 256 resultados son igualmente posibles si H_0 fuera verdadera. La región de rechazo consistiría en 12 resultados posibles $(0.05)(256) = 12.81$. Cuando la hipótesis nula es verdadera, la probabilidad de que podamos observar cualquiera de estos 12 resultados es $12/256 = 0.047$. Si llegamos a observar alguno de estos resultados extremos incluidos en la región de rechazo, debemos rechazar la H_0 en favor de H_1. Básicamente, si llegara a ocurrir alguno de estos resultados, rechazaríamos H_0 argumentando que la probabilidad del resultado observado es demasiada pequeña, de tal forma que la hipótesis debe ser incorrecta.

Cuando es adecuada una prueba bidireccional, como en el ejemplo que sigue, la región de rechazo consiste de los valores extremos posibles, ya sea en el extremo *positivo* o *negativo* de la distribución de las diferencias. Esto es, en el ejemplo, los 12 resultados de la región de rechazo incluirían 6 en el extremo positivo Σd_i y 6 en el extremo negativo Σd_i (o las sumas menores).

Ejemplo. Supongamos que un psicólogo infantil desea evaluar si la atención en la escuela de enfermería tiene algún efecto en la percepción social del niño. Las puntuaciones de percepción social se obtendrán presentando a los niños una serie de fotografías en las que se muestran una variedad de situaciones sociales y se les pedirá que contesten preguntas respecto a cada foto. Así, se asignarán puntuaciones que variarán en una escala de 0 a 100 para cada niño.

Por medio de un cuidadoso procedimiento de estandarización, el investigador está razonablemente seguro de que el índice de percepción social se encuentra en una escala de intervalo. Así, el investigador es capaz de interpretar las magnitudes numéricas de las diferencias observadas.

Para evaluar el efecto de la atención en la escuela de enfermería sobre las puntuaciones de la percepción social, el psicólogo obtuvo ocho pares de gemelos idénticos que servirán como sujetos. Al azar, uno de cada par de gemelos es asignado para que sean atendidos en la escuela de enfermería durante un cierto periodo. Al final de ese periodo se realizará la evaluación de los 16 niños respecto a la percepción social.

i. *Hipótesis nula.* H_0: los dos tratamientos son equivalentes. Esto es, no hay diferencia en la percepción social en las dos condiciones (en la escuela de enfermería y en el hogar). En la percepción social las 16 observaciones (ocho pares) pertenecen a una misma población. H_1: los dos tratamientos no son equivalentes.
ii. *Prueba estadística.* Se seleccionó la prueba de las permutaciones para pares replicados porque es apropiada para este diseño (dos muestras igualadas o pares replicados) y porque consideramos los datos pueden ser representados por una escala de intervalo.
iii. *Nivel de significación.* $\alpha = 0.05$ y N es el número de pares (8).
iv. *Distribución muestral.* La distribución muestral consiste en todas las permutaciones de los signos de las diferencias que incluyan todas las ocurrencias posibles (2^N) de Σd_i. En este caso, $2^N = 2^8 = 256$.

v. *Región de rechazo*. Puesto que H_1 no predice la dirección de las diferencias, la prueba es bidireccional. La región de rechazo consiste en los 12 resultados que contengan la Σd_i más extremas, las seis mayores y las seis más pequeñas.

vi. *Decisión*. Los datos de este estudio se presentan en la tabla 4.8. Las diferencias observadas en orden de magnitud absoluta fueron:

$$+27, \ +19, \ +13, \ +7, \ +6, \ -4, \ +3, \ -1$$

Para estas diferencias la suma es $+70$. A fin de facilitar el cálculo de la distribución de las permutaciones, las diferencias se enumeran en orden decreciente en la tabla 4.9. El primer renglón de esta tabla muestra cada diferencia con valor positivo, resultando la Σd_i mayor. Empezando con el lado derecho de la lista (con el valor más pequeño), comenzamos a alternar los signos. Así, los signos en la última columna para los sucesivos renglones serían $+ \ - \ + \ - \ + \ - \ldots$ Para la siguiente columna el modelo sería $+ \ + \ - \ - \ + \ + \ldots$ Para la siguiente columna el modelo sería $+ \ + \ + \ + \ - \ - \ - \ - \ + \ldots$ El modelo continuaría tantas veces como fuera necesario. Si sumáramos las diferencias de cada modelo, encontraríamos que se presentan en orden de magnitud de Σd_i decreciente. En este ejemplo, las primeras seis diferencias están en la región de rechazo en el 0.05 de nivel de significación bidireccional. Puesto que las Σd_i observadas están en la región de rechazo, debemos rechazar la H_0 de que no hay diferencia entre los grupos (nótese que el resultado 6 es, de hecho, el resultado observado). La probabilidad de su ocurrencia o de la ocurrencia de una Σd_i es tan extrema como 0.047 cuando H_0 es verdadera. Ya que la probabilidad es menor que 0.05, debemos rechazar la hipótesis nula.

En la aplicación de la prueba de las permutaciones, presentar los datos en orden como en la tabla 4.9 facilita los cálculos, ya que es fácil obtener la suma crítica sin enumerar todas las sumas. Conocer el número de permutaciones (2^N) y el nivel de significancia escogido, capacita al investigador para determinar cuál suma

Tabla 4.8. Puntuaciones de percepción social de los niños en la escuela de enfermería y en el hogar.

	Percepción social de los gemelos en		
Par	*Escuela de enfermería*	*Hogar*	*d*
a	82	63	19
b	69	42	27
c	73	74	−1
d	43	37	6
e	58	51	7
f	56	43	13
g	76	80	−4
h	85	82	3

Tabla 4.9. Los seis resultados positivos más extremos posibles para las diferencias mostradas en la tabla 4.8.

	Resultados								Σd_i
(1)	+27	+19	+13	+7	+6	+4	+3	+1	80
(2)	+27	+19	+13	+7	+6	+4	+3	−1	78
(3)	+27	+19	+13	+7	+6	+4	−3	+1	74
(4)	+27	+19	+13	+7	+6	+4	−3	−1	72
(5)	+27	+19	+13	+7	+6	−4	+3	+1	72
(6)*	+27	+19	+13	+7	+6	−4	+3	−1	70

* Resultados observados.

(mas no su valor) está en el nivel crítico. Una vez que este dato se especifica, la suma asociada puede calcularse como el valor crítico.[6]

MUESTRAS GRANDES

Si el número de pares es mayor que 12, el cálculo a mano de la prueba de las permutaciones se vuelve muy tedioso. Por ejemplo, si $n = 13$, el número de posibles resultados es $2^{13} = 8\,192$. La región de rechazo para $\alpha = 0.05$ consistiría en (0.05) (8 192) = 410 posibles resultados extremos. Aunque sólo las sumas extremas necesitan ser calculadas, el procedimiento puede volverse tedioso. El programa para computadora presentado en el Apéndice II facilita la utilización de la prueba de permutaciones.

Dado lo engorroso de los cálculos en la prueba de permutaciones cuando N es demasiado grande, se sugiere utilizar la prueba de Wilcoxon en tales casos. En dicha prueba, los rangos se sustituyen por números. La prueba de Wilcoxon se constituye en una alternativa muy eficaz a la prueba de las permutaciones, ya que esta última prueba está fundamentada en rangos.[7]

Resumen del procedimiento

Cuando N es pequeña y las medidas están al menos en una escala de intervalo, puede emplearse la prueba de las permutaciones para pares replicados o pares igualados. Éstos son los pasos que hay que seguir:

[6] Debido a que pueda haber valores duplicados de Σd_i para diferentes resultados cercanos a la región de rechazo, el valor de Σd_i, para subsecuentes resultados fuera de la región crítica, deberá calcularse para asegurar que los resultados duplicados no cruzan el límite crítico. En caso de darse la situación anterior, debe ajustarse la región de rechazo.

[7] En una prueba de permutación de rangos se consideran todas las 2^N permutaciones de los signos de los rangos y los más extremos posibles constituyen la región de rechazo. Para los datos en la tabla 4.6 existen $2^{12} = 4\,096$ combinaciones igualmente posibles de rangos cuando H_0) es verdadera. El lector curioso podría determinar que la muestra de los rangos dentro de los (0.05)(4 096) = 204 resulta-

1. Observe los valores de las diferencias y sus signos.
2. Haga una lista con las diferencias observadas en orden de magnitud decreciente.
3. Determine el número de posibles resultados cuando H_0 es verdadera.
4. Determine el número posible de resultados en la región de rechazo [(α) (2^N)].
5. Identifique los posibles resultados que se encuentran en la región de rechazo escogiendo aquellos con las mayores Σd_i; para ello utilice el método descrito en el ejemplo o un programa para computadora. Para la prueba unidireccional, los resultados en la región de rechazo están al final de la distribución. Para la prueba bidireccional, la mitad de los resultados se encuentra en el extremo positivo (las Σd_i mayores) y la otra mitad se encuentra en el extremo negativo (las Σd_i menores).
6. Determine si el resultado observado es uno de los que se encuentran en la región de rechazo. Si así ocurriera, rechace H_0 en favor de H_1.

Cuando N es grande ($N > 12$), se sugiere aplicar la prueba de Wilcoxon en lugar de la prueba de permutación.

Potencia–eficacia

Ya que utiliza toda la información de la muestra, la prueba de las permutaciones para pares replicados o pares igualados tiene una potencia–eficacia del 100 %. Es una de las pruebas estadisticas más poderosas.

Referencias bibliográficas

Los lectores encontrarán análisis de la prueba de las permutaciones en Fisher (1973), Moses (1952), Pitman (1937*a*, 1937*b*, 1937*c*) y Schefeé (1943). Moses presenta un método alternativo para determinar la significación de Σd_i cuando N es grande.

ANÁLISIS

En este capítulo hemos presentado cuatro pruebas estadísticas no paramétricas para el caso de una muestra con dos mediciones, ya sea pares replicados o pares igualados. La comparación y contrastación de dichas pruebas auxiliarán al lector en la selección de alguna de ellas, que puede ser la más adecuada a las características de los datos de algún experimento en particular.

Todas estas pruebas, excepto la prueba de McNemar para la significación de los cambios, suponen que la variable en consideración muestra una distribución

---

dos extremos posibles que nos permitieran rechazar H_0 al $\alpha = 0.05$, la cual fue nuestra decisión que se basó en la tabla H del Apéndice I. Efectivamente, por medio del método de la permutación, se puede construir la tabla H (que es la tabla de la distribución muestral de T^+).

continua subyacente a las observaciones. Nótese que no existe el requisito de que la medida misma sea continua; el requisito se debe cumplir en la variable aunque se dé en forma gruesa o aproximada.

La prueba de McNemar para la significación de los cambios se utiliza cuando una o ambas condiciones en estudio han sido medidas en el sentido de una escala nominal. Para el caso de pares igualados, la prueba de McNemar es única en su aplicabilidad a este tipo de datos. Es decir, esta prueba debería emplearse cuando los datos son frecuencias que pueden ser clasificadas por categorías separadas, las cuales no tienen relación una con otra del tipo "mayor que". No necesita hacerse ninguna suposición acerca de la continuidad de la variable, porque esta prueba es equivalente a la prueba que utiliza la distribución binomial con $p = q = 1/2$ y N es el número de cambios.

La prueba de los signos es aplicable cuando utilizando pares es posible obtener mediciones en escala ordinal (si la puntuación de un miembro de un par puede ser ordenado como "mayor que" en comparación con la puntuación del otro miembro del mismo par). Esto es, la prueba de los signos es aplicable a los datos de una variable que es continua, pero que puede ser medida sólo de una manera gruesa. Cuando la prueba de los signos se aplica a datos, los cuales cumplen las condiciones de alternativas paramétricas (la prueba t), su potencia–eficacia llega a ser alrededor del 95 % para $N = 6$, pero se decrementa conforme se incrementa N, hasta un 63 %.

Cuando las mediciones se encuentran en una escala ordinal tanto *intra* como *entre* observaciones, se puede usar la prueba de Wilcoxon; es decir, es aplicable cuando el investigador puede ordenar por rangos las diferencias observadas en varios pares igualados. Es común que los científicos conductuales sean capaces de ordenar por rangos las diferencias en las puntuaciones en el orden de tamaño absoluto sin ser capaces de dar, verdaderamente, puntuaciones numéricas a las observaciones dentro de cada par. Cuando se aplica la prueba de Wilcoxon a datos que encuentran, de hecho, las condiciones para la prueba t, su potencia–eficacia es de alrededor del 95 % en muestras grandes y no mucho menor para muestras pequeñas.

La prueba de las permutaciones debe utilizarse cuando N es suficientemente pequeña para permitir cálculos no tediosos y cuando las mediciones de la variable se encuentran, al menos, en una escala de intervalo. Esta prueba emplea toda la información de la muestra y así su potencia–eficacia es de 100 % en datos que, por sus características, pueden ser analizados adecuadamente por la prueba t. Un programa de computadora facilita los cálculos en muestras de tamaño moderado.

En suma, podemos concluir que la prueba de McNemar para la significación de los cambios debe ser aplicada en muestras tanto pequeñas como grandes, si la medición de al menos una variable es nominal. A los datos crudos del tipo ordinal se les debe aplicar la prueba de los signos; para mediciones más refinadas, la prueba de Wilcoxon puede emplearse en todos los casos. Si se obtuvieron medidas en escala de intervalo, se puede aplicar la prueba de las permutaciones para muestras de tamaño moderado.

5

Dos muestras independientes

Al estudiar las diferencias entre dos grupos, primero debemos determinar si ambos grupos están relacionados o si son independientes. El capítulo 4 contiene pruebas estadísticas para ser utilizadas en diseños que contienen dos grupos relacionados o pares replicados. En este capítulo se presentan pruebas estadísticas para utilizar en diseños que consisten en dos grupos independientes. Como las pruebas que se presentaron en el capítulo anterior, las que aparecen aquí determinan si las diferencias en las muestras constituyen una evidencia convincente producto de una diferencia en los procesos aplicados a ellos.

Aunque los méritos para utilizar dos muestras relacionadas o pares replicados en un diseño de investigación son grandes, hacerlo no siempre es factible. A menudo la naturaleza de la variable dependiente impide la utilización de sujetos como su propio control, como es el caso cuando la variable dependiente es la cantidad de tiempo que pasa un sujeto resolviendo un problema desconocido. Un problema puede ser desconocido sólo una vez. Además, puede resultar imposible diseñar un estudio que utilice pares igualados, tal vez por la incapacidad del investigador de descubrir variables para igualar, o por su incapacidad para obtener medidas adecuadas (para usar en la selección de los pares igualados) de algunas variables que se sabe que son relevantes o, por último, no siempre es posible realizar buenas "igualaciones".

Cuando el uso de dos muestras relacionadas no es factible o apropiado, se puede utilizar dos muestras independientes. En este diseño, las dos muestras deben ser obtenidas por uno de los siguientes dos métodos: 1. pueden obtenerse al azar de dos poblaciones, o 2. pueden originarse asignando al azar un sujeto a uno de dos tratamientos de los miembros de una misma muestra cuyo origen sea arbitrario. En cualquiera de los dos casos anteriores no es necesario que las muestras sean del mismo tamaño.

Un ejemplo del muestreo al azar de dos poblaciones sería obtener cada décimo republicano y cada décimo demócrata de las listas de orden alfabético de los votantes registrados.[1] Esto resultaría en una muestra al azar de los demócratas y republicanos registrados en sus áreas de votación cubiertos en las listas, y el número de demócratas sería igual al de republicanos, sólo si el registro de ambos partidos fuera sustancialmente igual en cada área. Otro ejemplo sería la obtención de cada octavo estudiante de nuevo ingreso por cada duodécimo estudiante en su último año, del mismo colegio.

Un ejemplo del método de asignación al azar podría ocurrir en un estudio de la efectividad de dos instructores en la enseñanza del mismo curso. Se obtendría una tarjeta de registro de los estudiantes inscritos en el curso y la mitad de esas tarjetas serían asignadas al azar, a un instructor y la otra mitad se asignaría al otro instructor.

La técnica paramétrica usual para analizar los datos de dos muestras independientes consiste en aplicar una prueba t a las medias de los dos grupos. La prueba t supone que las puntuaciones en las muestras son observaciones independientes de poblaciones "normalmente distribuidas" con (generalmente) las mismas varianzas. La prueba t supone, además, que las observaciones corresponden, al menos, a una escala de intervalo.

Para un tipo de investigación dado, la prueba t puede no ser aplicable por una gran variedad de razones. El investigador debe encontrar: 1. que los supuestos de la prueba t no son aplicables a sus datos; 2. que prefiere evitar hacer suposiciones y así dar a sus conclusiones mayor generalidad, o 3. que las puntuaciones no sean numéricas y realmente, por tanto, no cubran los requisitos de las mediciones para la prueba t. En casos como éstos, el investigador debe analizar los datos de las pruebas estadísticas no paramétricas para dos muestras independientes, las cuales se presentan en este capítulo. La comparación y el contraste de estas pruebas en el análisis de la conclusión de este capítulo ayudará al investigador en la selección de las pruebas que se presentarán, a una de las cuales puede recurrir para el manejo de sus datos.

PRUEBA EXACTA DE FISHER PARA TABLAS DE 2 $\times$ 2

Función

La prueba de la probabilidad exacta de Fisher para tablas de 2 $\times$ 2 es una técnica extremadamente satisfactoria para analizar datos discretos (tanto nominales como ordinales) cuando dos muestras independientes son pequeñas. Se usa cuando dos puntuaciones, de dos muestras independientes al azar caen dentro de una de dos clases mutuamente excluyentes. En otras palabras, cada sujeto en cada grupo obtiene una de dos puntuaciones posibles, las cuales son representadas por frecuencias en una tabla de contingencia de 2 $\times$ 2, como en la tabla 5.1. Los grupos I y II pueden ser una de dos variables independientes, tales como experimen-

[1] Técnicamente, para que la muestra se considere "una muestra realmente al azar", deberíamos tomar muestras de 10 votantes sucesivos de demócratas (o republicanos) y seleccionar al azar a uno de los 10 elegidos.

tal y control, hombres y mujeres, empleados y desempleados, demócratas y republicanos, padres y madres, etc. Los encabezados de los renglones, aquí indicados arbitrariamente como más (+) y menos (−), pueden tener cualquiera de dos clasificaciones: por arriba y por abajo de la media, acertaron y erraron, ciencias mayores y artes mayores, acuerdos y desacuerdos, etc. La prueba determina si los dos grupos difieren en las proporciones en donde caen dentro de cualquiera de las clasificaciones. Para los datos de la tabla 5.1 (donde A, B, C y D denotan frecuencias), determinaría si los grupos I y II difieren significativamente en la proporción de signos más (+) y signos menos (−) pertenecientes a cada grupo.

Tabla 5.1. Tabla de contingencias de 2 × 2.

	Grupo		
Variable	*I*	*II*	*Combinación*
+	A	B	$A + B$
−	C	D	$C + D$
Total	$A + C$	$B + D$	N

Método

La probabilidad exacta de observar un conjunto particular de frecuencias en una tabla de 2 × 2, cuando los totales marginales se consideran fijos, está dada por la distribución hipergeométrica:

$$p = \frac{\binom{A+C}{A}\binom{B+D}{B}}{\binom{N}{A+B}}$$

$$= \frac{[(A+C)!/A!C!]\,[(B+D)!/B!D!]}{N!/[(A+B)!(C+D)!]}$$

y así

$$p = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{N!A!B!C!D!} \tag{5.1}$$

La tabla W del Apéndice I puede ser útil para calcular los factoriales.

Para ilustrar el uso de la ecuación (5.1), supóngase que observamos los datos que se muestran en la tabla 5.2. En dicha tabla, $A = 5$, $B = 4$, $C = 0$ y $D = 10$. Los totales marginales son $A + B = 9$, $C + D = 10$, $A + C = 5$, y $B + D$

Tabla 5.2.

Variable	Grupo I	Grupo II	Combinación
+	5	4	9
−	0	10	10
Total	5	14	19

= 14. N, el número total de observaciones independientes, es 19. La probabilidad exacta de que esos 19 casos cayeran en las cuatro celdillas tal como si hubieran sido asignados al azar, puede determinarse mediante la ecuación (5.1), donde sustituyendo los valores de la tabla 5.1, tenemos:

$$p = \frac{9!\ 10!\ 5!\ 14!}{19!\ 5!\ 4!\ 0!\ 10!}$$

$$= 0.0108$$

Nosotros determinamos que la probabilidad de obtener un resultado como ése cuando H_0 es verdadera (que viene a ser la asignación al azar), es $p = 0.0108$.

El ejemplo anterior fue comparativamente simple de calcular debido a que una de las celdillas (la inferior izquierda) tuvo una frecuencia de cero. Pero si ninguna de las frecuencias de las celdillas es cero, debemos identificar las desviaciones más extremas de la distribución supuesta por H_0 que podrían ocurrir con los mismos totales marginales, y debemos tener en cuenta esas posibles desviaciones extremas, para el establecimiento de la hipótesis nula: "¿Cuál es la probabilidad, cuando H_0 es verdadera, de la ocurrencia del resultado obtenido o *uno más extremo*?"

Por ejemplo, supongamos que los datos de un estudio particular fueron los que se proporcionan en la tabla 5.3*a*. Con los totales marginales sin modificar, una ocurrencia más extrema sería la que corresponde a la tabla 5.3*b*. Así, si deseamos aplicar una prueba estadística a la hipótesis nula para los datos correspondientes a la tabla 5.3*a*, debemos sumar la probabilidad de su ocurrencia con la probabilidad del resultado más extremo mostrado en la tabla 5.3*b*. Calculamos cada p utilizando la ecuación (5.1). Así, tenemos:

$$P = \frac{5!\ 7!\ 5!\ 7!}{12!\ 4!\ 1!\ 1!\ 6!}$$

$$= 0.04419$$

y

$$P = \frac{5!\ 7!\ 5!\ 7!}{12!\ 5!\ 0!\ 0!\ 7!}$$

$$= 0.00126$$

Tabla 5.3.

	Grupo				*Grupo*		
	I	*II*			*I*	*II*	
	4	1	5		5	0	5
	1	6	7		0	7	7
	5	7	12		5	7	12
	(*a*)				(*b*)		

para las tablas 5.3*a* y 5.3*b*, respectivamente. Entonces, la probabilidad de ocurrencia de la tabla 5.3*a* o de una que sea más extrema (tabla 5.3*b*), es

$$P = 0.04419 + 0.00126$$
$$P = 0.04545$$

Esto es, $p = 0.04545$ es la probabilidad que utilizaríamos para decidir si los datos de la tabla 5.3*a* nos permiten rechazar la H_0.

El lector puede ver rápidamente que si el valor de celdilla más pequeño en la tabla de contingencia es aún moderadamente grande, la prueba de Fisher se vuelve, en términos de cálculos, muy tediosa. Por ejemplo, cuando la hipótesis alterna H_1 es unidireccional, y si el valor de celdilla más pequeño es dos, entonces deben determinarse tres probabilidades exactas utilizando la ecuación (5.1) y sumarlas; si el valor de celdilla más pequeño es tres, entonces deben calcularse cuatro probabilidades exactas y sumarse, etcétera.

Para facilitar el cálculo de la probabilidad asociada con tablas de contingencias de 2×2, puede utilizarse la tabla I del Apéndice, la cual es aplicable a tablas de contingencia de 2×2 cuando $N \leq 15$. Dado su tamaño y arreglo, veremos en detalle la utilización de la tabla I.

Éstos son los pasos para utilizar la tabla I:

1. Determine los totales por renglón y columna. Denote el valor total de columna o renglón como S_1. Denote el total que le siga (en forma ascendente) como S_2. La tabla 5.4 puede ayudarnos en la visualización del procedimiento. El lector notará que si S_1 es un total de renglón, S_2 será un total de columna.
2. X es la frecuencia observada en la celdilla donde se cruzan los valores más pequeño y el segundo más pequeño de renglón y columna.
3. Localice el renglón (N, S_1, S_2, X) en la tabla I. Hay tres entradas. La primera de ellas, "Obs.", es la probabilidad unidireccional de observar una diferencia igual o más extrema que la que se observó. La segunda entrada es la probabilidad de observar una diferencia grande o mayor en la dirección *opuesta*. Finalmente, la tercera entrada, "Total", es la probabilidad bidirec-

Tabla 5.4.

X	—	$S_1 \leftarrow$ Frecuencia más pequeña
—	—	
S_2		N

$\uparrow$ Segunda frecuencia más pequeña

cional de observar una diferencia grande o mayor que la observada en cualquier dirección.

4. Oriente y rotule la tabla para asegurar que las entradas son consistentes con la hipótesis.

Aunque los cálculos de probabilidades asociadas unidireccionales y bidireccionales con la prueba exacta de Fisher se facilitan considerablemente con la tabla I, es importante que el usuario entienda la base o el fundamento de la prueba, a fin de que utilice la tabla efectivamente. Usaremos la tabla 5.5 para ilustrar su aplicación.

Supongamos que un investigador ha formado dos grupos a partir de muestras y la hipótesis nula plantea que no existen diferencias entre ambos grupos en la variable dicotómica medida, la cual es codificada, por conveniencia, como $+$ y $-$. La hipótesis alterna plantea que el grupo 1 excede al grupo 2 en la proporción de respuestas $+$. Si planteamos que p_1 sea la probabilidad de que un sujeto seleccionado al azar del grupo 1 responda $+$ y que p_2 sea la probabilidad de que un sujeto elegido al azar del grupo 2 responda $+$, entonces las hipótesis nula y la alterna serían las siguientes:

$$H_0\colon p_1 = p_2$$
$$H_1\colon p_1 > p_2$$

Supongamos que $N = 15$ sujetos muestreados, siete pertenecientes al grupo 1 y ocho al grupo 2, y 5 sujetos del grupo 1 respondieron $+$ mientras un sujeto del grupo 2 respondió $+$. Los datos pueden ser representados como en el arreglo II de la tabla 5.5. Así, en la muestra $p_1 = 5/7 = 0.714$ y $p_2 = 1/8 = 0.125$. Para evaluar la hipótesis H_0, se debe determinar la probabilidad de observar una tabla de contingencia de 2×2 tan extrema o más. En la tabla 5.5 se muestran todos los posibles resultados que tienen los mismos totales marginales. Para cada uno de estos siete posibles resultados se proporcionan p_1 y p_2 junto con la probabilidad de muestrear esos arreglos resultantes cuando H_0 es verdadera [utilizando la ecuación (5.1)]. Nótese que la probabilidad de muestrear el resultado observado es $P[\mathrm{II}] = 0.0336$. La revisión de la tabla 5.5 muestra sólo un arreglo con un resultado más extremo (por ejemplo, $P_1 - P_2 > 0.714 - 0.125 = 0.589$), esto es, el resultado I que tiene probabilidad 0.0014. Así, la probabilidad de observar un resultado o uno más extremo es la siguiente:

$$\begin{aligned} p &= P[\mathrm{II}] + P[\mathrm{I}] \\ &= 0.0336 + 0.0014 \\ &= 0.035 \end{aligned}$$

Tabla 5.5. Ejemplo del cálculo de las probabilidades uni y bidireccional para la prueba de Fisher.

Tabla	P_1	P_2	$P_1 - P_2$	*P(tabla)*	*Obs.*	*Otras*	*Total*
I: $\begin{array}{cccc} & 1 & 2 & \\ + & 6 & 0 & 6 \\ - & 1 & 8 & 9 \\ & 7 & 8 & 15 \end{array}$	0.857	0	0.857	0.0014	0.001	0.000	0.001
II: $\begin{array}{cccc} & 1 & 2 & \\ + & 5 & 1 & 6 \\ - & 2 & 7 & 9 \\ & 7 & 8 & 15 \end{array}$	0.714	0.125	0.589	0.0336	0.035	0.006	0.014
III: $\begin{array}{cccc} & 1 & 2 & \\ + & 4 & 2 & 6 \\ - & 3 & 6 & 9 \\ & 7 & 8 & 15 \end{array}$	0.571	0.250	0.321	0.1958	0.231	0.084	0.315
IV: $\begin{array}{cccc} & 1 & 2 & \\ + & 3 & 3 & 6 \\ - & 4 & 5 & 9 \\ & 7 & 8 & 15 \end{array}$	0.429	0.375	0.054	0.3916	0.622	0.378	1.000
V: $\begin{array}{cccc} & 1 & 2 & \\ + & 2 & 4 & 6 \\ - & 5 & 4 & 9 \\ & 7 & 8 & 15 \end{array}$	0.286	0.500	−0.214	0.2937	0.378	0.231	0.608
VI: $\begin{array}{cccc} & 1 & 2 & \\ + & 1 & 5 & 6 \\ - & 6 & 3 & 9 \\ & 7 & 8 & 15 \end{array}$	0.143	0.625	−0.482	0.0783	0.084	0.035	0.119
VII: $\begin{array}{cccc} & 1 & 2 & \\ + & 0 & 6 & 6 \\ - & 7 & 2 & 9 \\ & 7 & 8 & 15 \end{array}$	0	.750	−0.750	0.0056	0.006	0.001	0.007

Nótese que ésta es la entrada Obs. en la tabla 5.5 y la tabla I del Apéndice para el resultado II.

Supongamos que la hipótesis alterna fue enunciada bidireccionalmente, esto es,

$$H_1\colon p_1 \neq p_2$$

entonces, los arreglos resultantes que muestran diferencias en posibles p mayores que el resultado observado II, son los resultados I y VII. El resultado VII es un valor más extremo que el resultado observado, pero en "otra" dirección. La probabilidad del resultado es $P[\mathrm{VII}] = 0.0056$. Éste es el valor (redondeado) que aparece en la tabla 5.5 en la entrada Otros y en la tabla I del Apéndice asociado con el resultado II. Así, la probabilidad de observar un resultado tan extremo como el resultado II en cualquier dirección es

$$\begin{aligned} P[\mathrm{II}] + P[\mathrm{I}] + P[\mathrm{VII}] &= 0.0336 + 0.0014 + 0.0056 \\ &= 0.041 \end{aligned}$$

Esto es la entrada Total en la tabla 5.5 y en la tabla I del Apéndice. Si aplicamos una prueba de dos colas en los datos observados al nivel de significación $\alpha = 0.05$, rechazaríamos H_0 ya que la probabilidad observada es 0.041.

Supongamos que se ha observado el resultado III. Entonces, las proporciones observadas serían $P_1 = 4/7 = 0.571$ y $P_2 = 2/8 = 0.250$. La diferencia es $P_1 - P_2 = 0.321$. Los resultados más extremos (en la misma dirección) son el I y el II. Por tanto, la probabilidad asociada con la prueba unidireccional es

$$\begin{aligned} P[\mathrm{III}] + P[\mathrm{I}] + P[\mathrm{II}] &= 0.1958 + 0.0014 + 0.0336 \\ &= 0.231 \end{aligned}$$

Para la prueba bidireccional, los resultados VI y VII son los más extremos, pero en la dirección contraria. En este caso, la probabilidad de un resultado tan o más extremo en cualquier dirección es

$$\begin{aligned} &P[\mathrm{III}] + P[\mathrm{I}] + P[\mathrm{II}] + P[\mathrm{VI}] + P[\mathrm{VII}] \\ &= 0.1958 + 0.0014 + 0.0336 + 0.0783 + 0.0056 \\ &= 0.315 \end{aligned}$$

El lector debería verificar su comprensión cabal para calcular las entradas en las últimas tres columnas de la tabla 5.5 (la cual corresponde a la tabla I del Apéndice).

Ejemplo. En un estudio acerca de las situaciones en las cuales las personas amenazan con suicidarse saltando desde un edificio, un puente o una torre, se advirtió que el abucheo o el hostigamiento por parte de una multitud como espectadora ocurría sólo en algunos casos. Varias teorías proponen que un estado psicológico de disminución de la identidad y la autoconciencia, conocido como *deindividuación*, puede contribuir al fenómeno de hostigamiento. Se conocen algunos factores que inducen reacciones en las multitudes, incluidos la temperatura, el ruido y la fatiga. En un esfuerzo por evaluar varias hipótesis concernientes al hostigamiento por parte de las multitudes, Mann[2] revisó 21 artículos publicados acerca de suicidio y examinó la relación entre el hostigamiento por parte de la multitud y el mes del año; esto último se refería más bien al índice de temperatura. La hipótesis es que habría un incremento en el hostigamiento por parte de los espectadores cuando hiciera calor.

i. *Hipótesis nula.* H_0: el hostigamiento por parte de las multitudes no varía como una función de la temperatura. H_1: existe un incremento en el hostigamiento por parte de las multitudes durante los meses calurosos.
ii. *Prueba estadística.* Este estudio requiere una prueba para determinar la significación de las diferencias entre dos muestras independientes: multitudes que hostigaron y multitudes que no hostigaron. La variable dependiente, tiempo (estación) del año, es dicotómica. Puesto que N es pequeña, la prueba exacta de Fisher resulta apropiada.
iii. *Nivel de significación.* $\alpha = 0.10$ y $N = 21$.
iv. *Distribución muestral.* La probabilidad de ocurrencia cuando H_0 es verdadera, de un conjunto de valores observados en una tabla de 2×2, puede encontrarse utilizando la ecuación (5.1). Puesto que $N > 15$, no puede utilizarse la tabla I del Apéndice.
v. *Región de rechazo.* Puesto que H_1 predice la dirección de la diferencia entre los grupos, la región de rechazo es unidireccional. H_0 será rechazada si los valores de celdilla observados difieren en la dirección predicha y si son de tal magnitud que la probabilidad asociada con su ocurrencia (o la ocurrencia de tablas más extremas) cuando H_0 es verdadera, es igual o menor que $\alpha = 0.10$.
vi. *Decisión.* La información de los artículos periodísticos está resumida en la tabla 5.6. En este estudio hubó 10 multitudes que hostigaron a los suicidas y 11 multitudes que no lo hicieron. La revisión de la tabla muestra que existen dos tablas adicionales que producirían un resultado más extremo (unidireccional). Así, la probabilidad de observar un conjunto de frecuencias de celdilla tan extremas o más extremas que la actualmente observada, se determina al utilizar la ecuación (5.1):

$$p = \frac{(A+B)!\,(C+D)!\,(A+C)!\,(B+D)!}{N!\,A!\,B!\,C!\,D!}$$

para cada tabla posible. Así

$$p = \frac{12!\,9!\,10!\,11!}{21!\,8!\,4!\,2!\,7!} + \frac{12!\,9!\,10!\,11!}{21!\,9!\,3!\,1!\,8!} + \frac{12!\,9!\,10!\,11!}{21!\,10!\,2!\,0!\,9!}$$

$$= 0.0505 + 0.0056 + 0.0002$$
$$= 0.0563$$

Puesto que la probabilidad obtenida 0.0563 es menor que el nivel de significación escogido $\alpha = 0.10$, debemos rechazar H_0 en favor de H_1. Concluimos que el hostigamiento por parte de las multitudes que atestiguan una amenaza de suicidio es afectado por la temperatura (medida según el mes del año).

[2] Mann, L., "The baiting crowd in episode of threatened suicide", en *Journal of Personality and Social Psychology*, núm. 41, 1981, págs. 703-709.

Tabla 5.6. Incidencia de hostigamiento en episodios de intento de suicidio.

	Multitud		
Mes	*Hostigamiento*	*No hostigamiento*	*Combinación*
Junio-septiembre	8	4	12
Octubre-mayo	2	7	9
Total	10	11	21

Resumen del procedimiento

Éstos son los pasos que hay que seguir para hacer uso de la prueba exacta de Fisher:

1. Presente las frecuencias observadas en una tabla de 2 × 2.
2. Determine los totales marginales. N será el número total de observaciones, S_1 será el valor total de renglón o la columna más pequeña, S_2 será el valor total de renglón o la columna que siga a S_1, y X será la frecuencia de celdilla donde se cruzan S_1 y S_2.
3. Utilizando los valores N, S_1, S_2 y X, determine la probabilidad unidireccional en la tabla del Apéndice I, de observar datos tan extremos o más que los observados (en la entrada Obs.); o bidireccional, determine la probabilidad utilizando la entrada Total.
4. Si $N > 15$, utilice la ecuación (5.1) para determinar la probabilidad o emplee la prueba ji cuadrada aproximada (véase la siguiente sección).

Potencia

La prueba exacta de Fisher es una de las más poderosas pruebas unidireccionales para datos, los cuales son apropiados a las características de la prueba: variables dicotómicas y en escala nominal.

Referencias bibliográficas

En Cochran (1952) y McNemar (1969) se encuentran otras referencias a la prueba exacta de Fisher.

PRUEBA JI CUADRADA PARA DOS MUESTRAS INDEPENDIENTES

Función

Cuando los datos corresponden a frecuencias de categorías discretas, puede utilizarse la prueba *ji cuadrada* para determinar lo significativo de las diferencias entre dos grupos independientes. La medida implicada puede ser tan "débil" como en escala nominal o categorial.

La hipótesis que está siendo probada generalmente es aquella que plantea que los grupos difieren respecto a algunas características y, por tanto, respecto a la frecuencia relativa con que los miembros de los grupos caen dentro de algunas categorías; por ejemplo, existe un grupo producto de interacción de variables. Para probar esta hipótesis, contamos el número de casos de cada grupo que caen en las distintas categorías y comparamos la proporción de casos de un grupo en las distintas variables, con la proporción de casos del otro grupo en las mismas variables. Si las proporciones son las mismas, entonces no hay interacción; en caso contrario, hay una interacción. El centro de la prueba se ubica en si las diferencias en las proporciones excede a aquellas esperadas por oportunidad o por desviaciones al azar de la proporcionalidad. Por ejemplo, debemos probar si dos grupos políticos difieren en su acuerdo o desacuerdo con algunas opiniones, o debemos probar si las personas de distinto sexo difieren en cuanto a la elección de actividades para aprovechar su tiempo libre, etcétera.

Método

Primero, los datos se presentan en una tabla de frecuencia (o *de contingencia*) en la cual las columnas representan grupos y cada renglón representa una categoría de la variable medida. En la tabla 5.7 se muestra un ejemplo de la tabla anterior. En ésta hay una columna para cada grupo y la variable medida puede adoptar uno de tres valores. La frecuencia observada de la ocurrencia de cada i-ésimo valor o categoría para cada j-ésimo grupo se denota n_{ij}.

Tabla 5.7. Tabla de contingencia de 3 × 2.

	Grupo		
Variable	*1*	*2*	*Combinación*
1	n_{11}	n_{12}	R_1
2	n_{21}	n_{22}	R_2
3	n_{31}	n_{32}	R_3
Total	C_1	C_2	N

La hipótesis nula de que los grupos se muestrean de la misma población puede ser probada por

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{5.2}$$

o

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}^2}{E_{ij}} - N \tag{5.2a}$$

donde

n_{ij} = el número observado de casos categorizados en el iésimo renglón de la j-ésima columna, y

E_{ij} = el número de casos esperado en el iésimo renglón de la j-ésima columna cuando H_0 es verdadera

y la doble suma es sobre todos los renglones y columnas de la tabla (por ejemplo, suma de todas las celdillas). Los valores de X^2 producidos por la ecuación (5.2) se distribuyen asintóticamente (conforme N va siendo mayor) como ji cuadrada, con grados de libertad $gl = (r - 1)(c - 1)$, donde r es el número de renglones y c es el número de columnas de la tabla de contingencia. Aunque el estadístico X^2 es más fácil de calcular utilizando la ecuación (5.2*a*) en este libro recomendaremos el uso de la ecuación (5.2) porque refleja de manera más natural los aspectos intuitivos del estadístico.

Bajo el supuesto de independencia, la frecuencia esperada de observaciones en cada celdilla debería ser proporcional a la distribución de totales de renglón y columna. Esta frecuencia esperada se estima como el producto correspondiente al número total de observaciones, dividido entre los totales de renglón y columna. Comenzamos encontrando los totales de renglón y columna. La frecuencia total en el i-ésimo renglón es

$$R_i = \sum_{j=1}^{c} n_{ij}$$

De manera similar, la frecuencia total en la j-ésima columna es

$$C_j = \sum_{i=1}^{r} n_{ij}$$

Así, tenemos que en la tabla 5.7, $R_1 = n_{11} + n_{12}$, y $C_1 = n_{11} + n_{21} + n_{31}$. Para encontrar la frecuencia esperada en cada celdilla (E_{ij}), se multiplican los dos totales marginales comunes a una celdilla en particular y el producto se divide por el número total de casos (N). Así, tenemos

$$E_{ij} = \frac{R_i C_j}{N}$$

Podemos ilustrar este método de encontrar los valores esperados mediante un ejemplo simple, utilizando para ello datos ficticios. Supongamos que se desea probar si las personas de alta y baja estatura difieren respecto a sus cualidades de liderazgo. En la tabla 5.8 se muestran las frecuencias con las que 43 personas de baja estatura y 52 personas de alta estatura son categorizadas como “líderes”, “seguidores” y “sin clasificación”. La hipótesis nula es que la altura es independiente de la posición líder-seguidor, es decir, que la proporción de personas altas que son líderes es la misma que la proporción de personas de baja estatura que son líderes; que la proporción de personas altas que son seguidores es la misma que la proporción de personas de baja estatura que son seguidores, etc. Con tal hipótesis, podemos determinar la frecuencia esperada para cada celdilla mediante el método sugerido anteriormente. En cada caso multiplicaremos los dos totales marginales comunes a una celdilla partícular y dividiremos este resultado entre N para obtener la frecuencia esperada. Así, por ejemplo, la frecuencia esperada para la celdilla inferior derecha de la tabla 5.8 es $E_{32} = (44)\ (52)\ /\ 95 = 24.1$. En la tabla 5.9 se muestran las frecuencias esperadas para cada una de las seis celdillas correspondientes a la tabla 5.8. En cada celdilla se marcó con números en cursivas cada una de las frecuencias esperadas y con números redondos, las frecuencias observadas.

Tabla 5.8. Altura y liderazgo.

	Bajos	*Altos*	*Combinación*
Seguidores	22	14	36
Sin clasificación	9	6	15
Líderes	12	32	44
Total	43	52	95

Tabla 5.9. Altura y liderazgo: frecuencias observada y esperada.

	Bajos	*Altos*	*Combinación*
Seguidores	22 *16.3*	14 *19.7*	36
Sin clasificación	9 *6.8*	6 *8.2*	15
Líderes	12 *19.9*	32 *24.1*	44
Total	43	52	95

Ahora, si las frecuencias observadas son muy similares a las frecuencias esperadas, las diferencias $(n_{ij} - E_{ij})$, por supuesto, serán pequeñas y, consecuentemente, el valor de X^2 será pequeño. Con un valor de X^2 pequeño no podemos rechazar la hipótesis nula de que las dos variables son independientes una de otra. Sin embargo, si algunas o muchas diferencias son grandes, entonces el valor de X^2 será mayor. Un valor de X^2 mayor será más probable si ambos grupos diferen respecto a las clasificaciones que se utilicen.

La distribución muestral de X^2 definida por medio de la ecuación (5.2) se distribuye aproximadamente como ji cuadrada[3] con grados de libertad

$$gl = (r - 1)(c - 1)$$

[3] En este libro distinguimos entre una variable que asintóticamente tenga una distribución como la ji cuadrada y la distribución misma. Así, el estadístico X^2 tiene una distribución muestral que es asintóticamente como la distribución ji cuadrada.

Las probabilidades asociadas con varios valores de ji cuadrada se proporcionan en la tabla C del Apéndice I. Si un valor observado de X^2 es igual o menor que el valor de dicha tabla para un nivel de significancia particular, con grados de libertad también particulares, la hipótesis nula puede ser rechazada en ese nivel de significancia.

Nótese que existe una diferente distribución muestral de X^2 para cada valor de grados de libertad. Esto es, la significación de cualquier valor particular de X^2 depende del número de grados de libertad en los datos en que fue calculado. El tamaño de los grados de libertad refleja el número de observaciones que tienen la posibilidad de variar después de que se han impuesto ciertas restricciones a los datos. (El tema de los grados de libertad se trata más ampliamente en el capítulo 3.)

Los grados de libertad para una tabla de contingencia $r \times c$ pueden ser encontrados por

$$gl = (r - 1)(c - 1)$$

donde

r = el número de clasificaciones (renglones), y
c = el número de grupos (columnas)

Para los datos de la tabla 5.9, $r = 3$ y $c = 2$, porque tenemos tres clasificaciones (seguidores, líderes y sin clasificación) y dos grupos (personas de baja estatura y personas de alta estatura). Así, los grados de libertad son $gl = (3 - 1)(2 - 1) = (2)(1) = 2$.

El cálculo de X^2 para los datos de la tabla 5.9 es sencillo:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{5.2}$$

$$= \frac{(22 - 16.3)^2}{16.3} + \frac{(14 - 19.7)^2}{19.7} + \frac{(9 - 6.8)^2}{6.8} + \frac{(6 - 8.2)^2}{8.2}$$

$$+ \frac{(12 - 19.9)^2}{19.9} + \frac{(32 - 24.1)^2}{24.1}$$

$$= 1.99 + 1.65 + 0.71 + 0.59 + 3.14 + 2.59$$

$$= 10.67$$

Utilizando la ecuación (5.2*a*), el estadístico anterior se calcularía como sigue:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}^2}{E_{ij}} - N \tag{5.2a}$$

$$= \frac{22^2}{16.3} + \frac{14^2}{19.7} + \frac{9^2}{6.8} + \frac{6^2}{8.2} + \frac{12^2}{19.9} + \frac{32^2}{24.1} - 95$$

$$= 10.67$$

Para determinar la significación de $X^2 = 10.67$ con $gl = 2$, regresemos a la tabla C del Apéndice I, en la cual se muestra que este valor de X^2 es significativo más allá del nivel de 0.01. Por tanto, podemos rechazar la hipótesis nula de "no diferencia entre grupos" con un nivel de significación de $\alpha = 0.01$.

Ejemplo. En un estudio realizado con ex fumadores como sujetos, Shiffman recolectó los datos durante las crisis producto de las recaídas.[4] Las crisis de recaída incluyen el periodo actual de abstinencia y las situaciones en las cuales era inminente un periodo de abstinencia, pero que fue evitado con éxito. Estos episodios críticos fueron recolectados a través de llamadas que los ex fumadores realizaban en el momento de la crisis. Varios datos se recolectaron, incluida la estrategia utilizada en un intento de evitar la recaída. Las estrategias de afrontamiento se categorizaron como conductuales (es decir, abandonar la situación) o cognoscitivas (revisar mentalmente las razones por las cuales la persona había decidido dejar de fumar). Algunos sujetos dijeron haber utilizado una clase de estrategia de afrontamiento, otros informaron que habían empleado ambas clases de estrategias y otros más refirieron no haber usado ninguna de las dos. La hipótesis fue que la utilización de estrategias de afrontamiento diferiría entra aquellas que fueron exitosas y aquellas que no lo fueron en cuanto a evitar la recaída.

i. *Hipótesis nula.* H_0: no existe diferencia en las estrategias de afrontamiento empleadas por los ex fumadores que evitaron exitosamente una recaída en un periodo de abstinencia y aquellos que no tuvieron éxito. H_1: los dos grupos difieren en las estrategias de afrontamiento empleadas durante las crisis.
ii. *Prueba estadística.* Puesto que las conductas de que se informó (afrontamiento conductual, cognoscitivo o ambos, y no afrontamiento) son variables categoriales (nominales), ya que existen dos grupos (aquellos que sufrieron recaída y aquellos que no), y puesto que las categorías son mutuamente excluyentes y exhaustivas, la prueba ji cuadrada para grupos independientes resulta apropiada para evaluar la H_0.
iii. *Nivel de significación.* $\alpha = 0.01$ y N es el número de sujetos quienes presentaron datos $= 159$.
iv. *Distribución muestral.* La distribución muestral de X^2 es aproximadamente la de la ji cuadrada con tres grados de libertad. Los grados de libertad (gl) se determi-

Tabla 5.10. Efecto de las estrategias de afrontamiento en crisis de recaída relacionadas con dejar de fumar.

	Grupo de resultados		
Afrontamiento	*Fumaron*	*No fumaron*	*Combinación*
Conductual	15	24	39
Cognoscitivo	15	21	36
Cognoscitivo y conductual	13	43	56
Ninguno	22	6	28
Total	65	94	159

[4] Shiffman, S., "Relapse following smoking cessation: A situational analysis", en *Journal of Counseling and Clinical Psychology*, núm. 50, 1982, págs. 71-86.

nan mediante $gl = (r - 1)(c - 1)$, donde r es el número de categorías (4) y c es el número de grupos (2). En virtud de esto, tenemos $(4 - 1)(2 - 1) = 3$.

v. *Región de rechazo.* H_1 simplemente predice una diferencia entre los dos grupos; la región de rechazo consiste en aquellos valores de X^2 que excedan el valor crítico de la distribución ji cuadrada con grados de libertad $gl = 3$. En la tabla C del Apéndice I se indica que el valor crítico de X^2 es 11.34 cuando $\alpha = 0.01$.

vi. *Decisión.* En la tabla 5.10 se resumen los datos obtenidos de las llamadas realizadas durante las crisis. Ésta muestra que 65 personas sufrieron una recaída y 94 personas se abstuvieron de fumar durante las crisis. Los valores esperados para cada celdilla se obtuvieron mediante la fórmula $E_{ij} = R_i C_j/N$. Así, tenemos que $E_{11} = (39)(65)/159 = 15.94$, $E_{21} = 14.72$, etc. Utilizando la ecuación (5.2), el resultado obtenido de X^2 fue 23.78.

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{5.2}$$

$$= \frac{(15 - 15.94)^2}{15.94} + \frac{(24 - 23.06)^2}{23.06} + \cdots + \frac{(6 - 16.55)^2}{16.55}$$

$$= 23.78$$

Puesto que el valor de X^2 excede el valor crítico, rechazamos la hipótesis de que la estrategia de afrontamiento es independiente de si la persona sufrió o no una recaída durante la crisis.

Tablas de contingencia de 2 × 2

Tal vez el más común de los usos de la prueba ji cuadrada sea el de evaluar si un conjunto de frecuencias observadas en una tabla de contingencia de 2×2 pudieran ocurrir siendo H_0 verdadera. Para nosotros es familiar la forma de una tabla como ésta; un ejemplo de lo anterior lo constituye la tabla 5.1. Cuando aplicamos la prueba X^2 a datos donde tanto la r como la c son iguales a 2, debe utilizarse la siguiente ecuación:

$$X^2 = \frac{N(|AD - BC| - N/2)^2}{(A + B)(C + D)(A + C)(B + D)} \qquad gl = 1 \tag{5.3}$$

Esta ecuación es un poco más facil de aplicar que la ecuación (5.2), en vista de que sólo se requiere una división en todo el cálculo. Tiene la ventaja adicional de incorporar una correción para la continuidad, la cual mejora notablemente la aproximación de la distribución muestral de la X^2 calculada a la distribución de la ji cuadrada.

Ejemplo. Otra variable registrada en el estudio examinado en la sección anterior sobre la abstinencia en el fumar, fue si el consumo o no de alcohol era un factor durante las crisis de recaída. A los sujetos se les preguntó si consumieron alcohol antes de la crisis o durante su transcurso. La hipótesis fue que el consumo de alcohol estaba relacionado con el hecho de que el sujeto recayera o se abstuviera durante la crisis.

i. *Hipótesis nula.* H_0: el consumo de alcohol no está relacionado con el resultado de la crisis. H_1: el consumo de alcohol está relacionado con el éxito o el fracaso en abstenerse durante la crisis.

ii. *Prueba estadística.* Puesto que ambas variables (grupo y consumo de alcohol) son nominales (categoriales) y ya que las medidas son mutuamente excluyentes y exhaustivas, es apropiada la prueba ji cuadrada. Aún más, puesto que ambas variables son de naturaleza dicotómica, puede emplearse la prueba ji cuadrada para tablas de 2×2.

iii. *Nivel de significación.* $\alpha = 0.01$ y N es el número de sujetos que respondieron $= 177$.

iv. *Distribución muestral.* La distribución muestral de X^2 determinada por la ecuación (5.3) está asintóticamente distribuida como ji cuadrada con $gl = 1$.

v. *Región de rechazo.* La región de rechazo para este ejemplo consta de todos los valores de X^2 para los cuales la probabilidad de observar un valor tan grande o mayor cuando H_0 es verdadera, sea menor que $\alpha = 0.01$.

vi. *Decisión.* En la tabla 5.11 se resumen los datos observados. 20 de las 68 personas que sufrieron una recaída (29 %) consumieron alcohol durante la crisis y 13 de las 109 (12 %) que no sufrieron de recaída, no consumieron alcohol durante la crisis. Se calculó el valor de X^2 utilizando la ecuación (5.3):

$$X^2 = \frac{N(|AD - BC| - N/2)^2}{(A + B)(C + D)(A + C)(B + D)} \qquad gl = 1 \tag{5.3}$$

$$= \frac{177(|(20)(96) - (13)(48)| - 177/2)^2}{(33)144)(68)(109)}$$

$$= 7.33$$

Al hacer referencia a la tabla C del Apéndice I, ésta muestra que $X^2 \geq 7.33$ con $gl = 1$ tiene una probabilidad de ocurrencia, cuando H_0 es verdadera, menor que 0.01. Puesto que el valor observado de X^2 excede el valor crítico de 6.64, rechazamos la hipótesis de que el consumo de alcohol no tiene efecto en la recaída o abstinencia durante una crisis relacionada con dejar de fumar.

Tabla 5.11. Efecto del consumo de alcohol en crisis relacionadas con el dejar de fumar.

	Grupo de resultados		
Consumo de alcohol	*Fumaron*	*No fumaron*	*Combinación*
Sí	20	13	33
No	48	96	144
Total	68	109	177

Partición de los grados de libertad en tablas de $r \times 2$

Una vez que el investigador determina que el valor de X^2 para una tabla particular de contingencia es significativo, debe saber si existe una diferencia entre los dos grupos donde fue medida la variable. Sin embargo, puede no saber *dónde* se dieron esas diferencias. Puesto que las variables medidas pueden tomar distintos valores, es posible que la diferencia se refleje por algunos valores y no por otros. La pregunta de dónde se encuentran las diferencias en la tabla de contingencia puede contestarse al *dividir* la tabla de contingencia en subtablas y analizar cada una de ellas. Se podría considerar construir un cierto número de subtablas de 2×2 que se analizarían mediante la prueba exacta de Fisher; sin embargo, tales tablas no son independientes e interpretarlas resulta difícil. Afortunadamente, es posible construir subtablas de 2×2 que sean independientes, las cuales son interpretables construyéndolas según los métodos que se sugieren más adelante. Cualquier tabla de contingencia puede ser dividida en tantas subtablas de 2×2 como lo permita el número de grados de libertad de la tabla original. El método de construir tablas es relativamente sencillo y se comprende mejor mediante los ejemplos que siguen a continuación. Para la tabla 5.7 de 3×2, existen dos posibles divisiones; éstas se muestran en la tabla 5.12.

Cada una de estas tablas tiene un grado de libertad. Para probar la independencia entre los dos grupos, la X^2 debe ser modificada para reflejar que son *subtablas* obtenidas de una tabla mayor y reflejar también las características de la muestra total. Las fórmulas para las particiones de la tabla 5.12 son las siguientes:

$$X_1^2 = \frac{N^2\,(n_{22}\,n_{11} - n_{21}\,n_{12})^2}{C_1\,C_2 R_2\,R_1(R_1 + R_2)} \tag{5.4a}$$

$$X_2^2 = \frac{N\,[n_{32}\,(n_{11} + n_{21}) - n_{31}\,(n_{12} + n_{22})]^2}{C_1\,C_2 R_3(R_1 + R_2)} \tag{5.4b}$$

Cada uno de estos estadísticos X^2 se distribuye asintóticamente como ji cuadrada con un grado de libertad. El lector notará que estas fórmulas son similares a aquellas para tablas de contingencia de 2×2. Una diferencia importante es que las distribuciones marginales reflejan las distribuciones marginales para la muestra completa, no sólo para la subtabla de 2×2 particular. Además, la primera subtabla parece "colapsarse" dentro de la segunda subtabla.

Tabla 5.12. Particiones aditivas de una tabla de contingencia de 3×2.

n_{11}	n_{12}	R_1
n_{21}	n_{22}	R_2
C_1	C_2	N

(1)

$n_{11} + n_{21}$	$n_{12} + n_{22}$	$R_1 + R_2$
n_{31}	n_{32}	R_3
C_1	C_2	N

(2)

Para las tablas generales $r \times 2$ examinadas en esta sección, pueden formarse particiones del tipo $r - 1$. La ecuación general para la t-ésima partición de una tabla $r \times 2$ es la siguiente:

$$X_t^2 = \frac{N^2\left(n_{t+1,2}\sum_{i=1}^{t} n_{i1} - n_{t+1,1}\sum_{i=1}^{t} n_{i2}\right)^2}{C_1 C_2 R_{t+1}\left(\sum_{i=1}^{t} R_i\right)\left(\sum_{i=1}^{t+1} R_i\right)} \qquad t = 1, 2, \ldots, r - 1 \qquad (5.5)$$

La fórmula para dividir cada tabla ha sido "colapsada" para formar la siguiente tabla. El arreglo procede de la parte superior a la parte inferior. Esto es meramente una conveniencia para escribir la ecuación. El investigador debe arreglar la tabla de manera tal que el "colapso" y la combinación tengan sentido. Por ejemplo, en la tabla 5.8 la atención se centra en los líderes contra los no líderes. Por tanto, debemos comparar primero a las personas de corta y alta estatura quienes son seguidores o no tienen clasificación; entonces, estas dos variables deberían "colapsarse" para formar la segunda división con la cual se compararía a los líderes con los no líderes. Estas particiones se resumen en la tabla 5.13.

Tabla 5.13. Ejemplo de una tabla de contingencia de 3×2 de particiones aditivas (los datos corresponden a la tabla 5.8).

	Estatura		
	Bajos	*Altos*	*Combinados*
Seguidores	22	14	36
Sin clasificación	9	6	15
Total	43	52	95

(1)

	Estatura		
	Bajos	*Altos*	*Combinados*
Seguidores o sin clasificación	31	20	51
Líderes	12	32	44
Total	43	52	95

(2)

Para la primera partición de la tabla 5.13 debemos calcular X^2 utilizando la ecuación (5.4*a*):

$$X_1^2 = \frac{N^2\,(n_{22}\,n_{11} - n_{21}\,n_{12})^2}{C_1\,C_2\,R_2\,R_1(R_1 + R_2)} \tag{5.4a}$$

$$= \frac{95^2\,[(6)(22) - (9)(14)]^2}{(43)(52)(15)(36)(51)}$$

$$= 0.005$$

Ésta se distribuye como ji cuadrada con $gl = 1$ y claramente no es significativa. El investigador puede concluir, sin riesgo, que no existe relación entre la estatura y que las personas sean seguidoras y no tengan clasificación en términos de liderazgo. Así, tenemos que es razonable combinar estas dos categorías para formar el primer renglón de la segunda tabla. Las dos categorías se "colapsan" para formar la segunda partición de la tabla 5.13. El valor de la X^2 dividida se obtiene utilizando la ecuación (5.4*b*):

$$X_2^2 = \frac{N[n_{32}\,(n_{11} + n_{21}) - n_{31}\,(n_{12} + n_{22})]^2}{C_1\,C_2 R_3\,(R_1 + R_2)} \tag{5.4b}$$

$$= \frac{95\,[32(22 + 9) - 12(14 + 6)]^2}{(43)(52)(44)(51)}$$

$$= 10.707$$

Puesto que este valor excede el valor crítico de la distribución ji cuadrada para $\alpha = 0.05$, el investigador puede concluir que la distribución de los líderes y no líderes difiere como una función de la estatura. El lector notará que este resultado es similar al encontrado cuando se analizó la tabla de 3×2 en su conjunto (sin particiones). Sin embargo, lo importante es que hemos sido capaces de concluir que los seguidores y los que no tienen clasificación, esencialmente, son similares. Debería notarse que la sumatoria de los dos valores de X^2 dividida es aproximadamente el mismo que se obtuvo para la ji cuadrada total: $10.707 + 0.005 = 10.71$ para las particiones contra 10.67 para la tabla total (sin particiones). Así, tenemos que en una muestra la suma de los valores de partición de X^2 será aproximadamente el mismo que el valor total, y esto puede servir como una burda comprobación de nuestros cálculos.

Ejemplo. En el estudio de ex fumadores descrito y resumido en la tabla 5.10, se detectó que existían diferencias significativas en las conductas de afrontamiento entre aquellos que fumaron y los que no fumaron en sus crisis de recaída. En esa sección se encontró que $X^2 = 23.78$ con $gl = 3$. Sería deseable determinar cuáles de las conductas resultaron efectivas durante las crisis. Para determinar esto, dividiremos la X^2 obtenida. Es necesario determinar *a priori* en qué orden debemos dividir la tabla. Puesto que $gl = 3$, existe la posibilidad de realizar tres particiones. El examen de los niveles de las variables sugiere las particiones más provechosas. La primera partición contrasta los dos tipos de afrontamiento cuando se emplearon individualmente, es decir, afrontamiento conductual contra afrontamiento cognoscitivo. La segunda partición compara la utilización de una sola conducta de afrontamiento contra la utilización de dos conductas de afrontamiento. La tercera partición compara el uso de cualquier conducta de afrontamiento con el fracaso en el uso de cualquier conducta de afrontamiento. Las tablas resultantes de la partición de la tabla 5.10 se muestran en la tabla 5.14.

Para cada una de estas particiones, el valor de X^2 asociado puede determinarse mediante la ecuación (5.5). Para la primera partición tenemos:

$$X_t^2 = \frac{N^2\left(n_{t+1,2}\sum_{i=1}^{t} n_{i1} - n_{t+1,1}\sum_{i=1}^{t} n_{i2}\right)^2}{C_1C_2R_{t+1}\left(\sum_{i=1}^{t} R_i\right)\left(\sum_{i=1}^{t+1} R_i\right)} \tag{5.5}$$

$$X_1^2 = \frac{159^2\,[(21)(15) - (15)(24)]^2}{(65)(94)(36)(39)(75)} = 0.08$$

Tabla 5.14. Particiones aditivas de una tabla de contingencia para fumadores, para el ejemplo de dejar de fumar (los datos corresponden a la tabla 5.10).

	Grupo de resultados		
Afrontamiento	*Fumaron*	*No fumaron*	*Combinados*
Conductual	15	24	39
Cognoscitivo	15	21	36
Total	65	94	159

(1)

	Grupo de resultados		
Afrontamiento	*Fumaron*	*No fumaron*	*Combinados*
Conductual o cognoscitivo	30	45	75
Conductual y cognoscitivo	13	43	56
Total	65	94	159

(2)

	Grupo de resultados		
Afrontamiento	*Fumaron*	*No fumaron*	*Combinados*
Conductual, cognoscitivo o ambos	43	88	131
Ninguno	22	6	28
Total	65	95	159

(3)

El valor de X^2 para la segunda partición es el siguiente:

$$X_2^2 = \frac{159^2\,[(43)(30) - (13)(45)]^2}{(65)(94)(56)(75)(131)}$$

$$= 3.74$$

Finalmente, para la tercera partición tenemos:

$$X_3^2 = \frac{159^2\,[(6)(43) - (22)(88)]^2}{(65)(94)(28)(131)(159)}$$

$$= 19.98$$

Cada uno de estos valores de X^2 se distribuye asintóticamente como ji cuadrada con $gl = 1$. En la prueba sin particiones se seleccionó un nivel de significación de $\alpha = 0.01$. Utilizando el mismo nivel, el valor crítico de X^2 es 6.64. Así, tenemos que sólo la tercera partición es significativa. El investigador puede concluir que no existen diferencias en la efectividad de las conductas de afrontamiento, y que la diferencia entre los dos grupos depende de si utilizaron o no *cualquier* conducta de afrontamiento. Esto es, las conductas de afrontamiento son igualmente efectivas y resultan más efectivas que no utilizar ninguna durante las crisis.

En la partición de cualquier tabla, el investigador debe examinar *a priori* la variable medida para determinar cuáles variables pueden ser combinadas apropiadamente como parte de un esquema de partición. Una vez que se determinan esas combinaciones, la tabla puede ser arreglada de tal manera que pueda aplicarse la ecuación (5.5) a cada partición. Si la variable original se encuentra en escala nominal, los renglones pueden ser arreglados de una manera adecuada para la partición. Si la variable representa categorías ordenadas, tal rearreglo puede no tener sentido para la variable que se va a estudiar; sin embargo, se puede, aún, arreglar la tabla a fin de comenzar a dividir en cualquier "extremo" de la misma. No obstante, para el investigador es importante utilizar particiones que resultan en tablas de 2×2, que sean interpretables en el contexto de su investigación en particular.

Resumen del procedimiento

Éstos son los pasos que hay que seguir en el uso de la prueba ji cuadrada para dos muestras independientes:

1. Arregle las frecuencias observadas en una tabla de contingencia $r \times c$, utilizando las c columnas para los grupos y los r renglones para las condiciones. Así, para esta prueba $c = 2$.
2. Calcule los totales de renglón R_1 y los totales de columna C_1.
3. Determine la frecuencia esperada para cada celdilla encontrando el producto de los totales marginales en común y dividiendo este resultado entre N (N representa el número total de observaciones *independientes*); así, $E_{ij} = R_iC_j/N$. Nótese que los valores de N "inflados" invalidan la prueba. Los pasos 2 y 3 son innecesarios si los datos se encuentran en una tabla

de 2 × 2, para los cuales puede utilizarse tanto la tabla del Apéndice I si $N \leq 15$ como la ecuación (5.3) si $N \geq 15$. Si r o c es mayor que 2, debe emplearse la ecuación (5.2).

4. Determine la significación de la X^2 observada haciendo referencia de la tabla C del Apéndice I. Si la probabilidad proporcionada por la tabla C es igual o menor ($\leq$) que α, rechace H_0 en favor de H_1.
5. Si la tabla es mayor que 2 × 2 y si H_0 es rechazada, la tabla de contingencia puede ser dividida en subtablas independientes para determinar sólo dónde se encuentra la diferencia en la tabla original. Utilice la ecuación (5.5) (o las ecuaciones (5.4*a*) y (5.4*b*) si la tabla es de 3 × 2) para calcular el valor de X^2 para cada partición. Pruebe la significancia de cada X^2 haciendo referencia a la distribución de la ji cuadrada con $gl = 1$ en la tabla C del Apéndice I. El programa para computadora del Apéndice II le ayudará a agilizar los cálculos.

Cuándo utilizar la prueba ji cuadrada

Como ya se habrá notado, la ji cuadrada requiere que las frecuencias esperadas E_{ij} en cada celdilla no sean demasiado pequeñas. Cuando son demasiado pequeñas, la prueba puede no ser la más apropiada. Cochran (1954) y otros autores hacen algunas recomendaciones al respecto, las cuales se incluyen en los siguientes apartados.

El caso de 2 × 2

Si las frecuencias se encuentran en una tabla de contingencia de 2 × 2, la decisión concerniente al uso de la ji cuadrada debe basarse en las siguientes consideraciones:

1. Cuando $N \leq 20$, siempre se utiliza la prueba exacta de Fisher.
2. Cuando $20 < N < 40$, la prueba X^2 (ecuación 5.3) puede utilizarse si las frecuencias esperadas son cinco o más. Si la frecuencia esperada más pequeña es menor que cinco, use la prueba exacta de Fisher (véase la sección correspondiente).
3. Cuando $N > 40$, utilice la prueba X^2 corregida para la continuidad, es decir, utilice la ecuación (5.3).

Tablas de contingencia con gl mayor que 1

Cuando r es mayor que 2 (y, por tanto, $gl > 1$), puede utilizarse la prueba X^2 si menos del 20 % de celdillas tienen una frecuencia esperada menor que cinco y si no hay celdillas con frecuencia esperada menor que uno. Si estos requisitos no son cubiertos por los datos en la forma en que originalmente fueron recolectados, el investigador deberá combinar categorías adyacentes para incrementar las frecuencias esperadas en las celdillas de que se trate. Sólo después de combinar categorías para cubrir los requisitos anteriores, los valores de tabla para la distribución de la ji cuadrada pueden ser lo suficientemente cercanos a la distribución muestral de la X^2.

Cuando $gl > 1$, las pruebas ji cuadrada son insensibles a los efectos del orden, y así, cuando una hipótesis tiene en cuenta el orden, la ji cuadrada puede no ser la mejor prueba. Las pruebas que se ajustan a este tipo de datos se examinan más adelante en este capítulo y en el 8.

Valores esperados pequeños

La prueba ji cuadrada es aplicable a datos de una tabla de contingencia sólo si las frecuencias esperadas son lo suficientemente grandes. Los requisitos de tamaño para las frecuencias esperadas se estudiaron anteriormente en esta sección. Cuando las frecuencias esperadas no cubren dichos requisitos. Podemos incrementar sus valores combinando clasificaciones adjuntas, es decir, combinando clasificaciones adyacentes y, por tanto, reducir el número de celdillas. Esto puede hacerse con toda propiedad si tales combinaciones no restan significado a los datos. En nuestro ejemplo de estatura y liderazgo, por supuesto, cualquier combinación de categorías hubiera restado "fuerza" a nuestra evaluación de la hipótesis. El investigador, generalmente, debe evitar este problema y reunir el número suficiente de casos relacionados con el número de clasificaciones que se utilizarán en el análisis.

Potencia

Cuando se utiliza la ji cuadrada, generalmente no es clara una prueba alternativa y, por tanto, la potencia exacta de la prueba es difícil de calcular. Sin embargo, Cochran (1952) ha mostrado que la potencia limitante de la distribución de X^2 tiende a ser 1 conforme N va siendo mayor.

Referencias bibliográficas

Para análisis adicionales de la prueba ji cuadrada, el lector puede consultar a Cochran (1952, 1954), Everitt (1977), McNemar (1969), un artículo clásico acerca del uso y abuso de la ji cuadrada escrito por Lewis y Burke (1949), y un artículo escrito por Delucchi (1983). Otros estudios sobre los procedimientos de partición pueden encontrarse en Castellan (1966).

PRUEBA DE LA MEDIANA

Función

La *prueba de la mediana* es un procedimiento para evaluar si dos grupos independientes difieren en sus tendencias culturales. Más precisamente, esta prueba nos proporciona información acerca de qué tan probable es que dos grupos independientes (no necesariamente del mismo tamaño) hayan sido extraídos de la misma población con la misma mediana. La hipótesis nula plantea que los dos grupos son de la misma población y tienen la misma mediana; la hipótesis alterna puede plantear que la mediana de una población es *diferente* de la otra población (si es prueba bidireccional) o que la mediana de una población es superior que la de la otra población (si es prueba unidireccional). La prueba puede utilizarse cuando las puntuaciones de los dos grupos se miden en, al menos, una escala ordinal. Se notará que puede no existir una prueba alterna a la prueba de la mediana, aun para datos en escala de intervalo. Esto podría ocurrir cuando una o más de las observaciones están "fuera de la escala" y truncadas hacia el máximo (o mínimo) de las observaciones previamente asignadas.

Racionalización y método

Para aplicar la prueba de la mediana debemos determinar primero la puntuación de la mediana para el grupo combinado (es decir, la mediana para todas las puntuaciones en ambas muestras). Después, debemos "dicotomizar" (dividir) ambos conjuntos de puntuaciones a partir de la mediana combinada y presentar estos datos en una tabla de 2×2 como en la tabla 5.15.

Tabla 5.15. Prueba de la mediana: formato para los datos.

	Grupo		
	I	*II*	*Combinados*
Número de puntuaciones combinadas por debajo de la mediana	A	B	$A + B$
Número de puntuaciones combinadas por arriba de la mediana	C	D	$C + D$
Total	m	n	$N = m + n$

Ahora, si los grupos I y II son muestras de poblaciones cuyas medianas son las mismas, esperaríamos que alrededor de la mitad de las puntuaciones de cada grupo estuviera por arriba de la mediana combinada y que la otra mitad estuviera por debajo; es decir, esperaríamos que tanto las frecuencias A y C como B y D fueran iguales.

Puede demostrarse (Mood, 1950) que si A es el número de m casos en el grupo I que cae sobre la mediana combinada, y si B es el número de n casos en el grupo II que cae sobre la mediana combinada, entonces la distribución muestral de A y B según la hipótesis nula (H_0 es que las medianas son las mismas) es la distribución hipergeométrica

$$P\,[A,\,B] = \frac{\binom{m}{A}\binom{n}{B}}{\binom{m\,+\,n}{A\,+\,B}} \tag{5.6}$$

Por tanto, si el número total de casos en ambos grupos ($m = n$) es pequeño, se puede utilizar la prueba exacta de Fisher para probar H_0. Si el número total de casos es suficientemente grande, puede utilizarse la prueba ji cuadrada con $gl = 1$ para evaluar la H_0.

Para analizar datos divididos por la mediana, el investigador debe guiarse por las siguientes consideraciones, a fin de realizar una buena elección entre la prueba exacta de Fisher y la ji cuadrada para tablas de 2×2:

1. Cuando $N = m + n$ es mayor que 20, utilice la X^2 corregida para la continuidad [ecuación (5.3)].
2. Cuando $N = m + n = 20$ o menor, utilice la prueba exacta de Fisher.

Puede surgir una dificultad en el cálculo de la prueba de la mediana; varias puntuaciones pueden caer exactamente en la mediana combinada. Si esto sucede, el investigador tiene dos opciones:

1. Los grupos pueden ser "dicotomizados" en aquellas puntuaciones que *exceden* a la mediana y aquellas puntuaciones que no.
2. Si ($m + n$) es grande y sólo pocos casos caen exactamente en la mediana combinada, esos pocos casos pueden ser eliminados del análisis.

La primera opción es la que se prefiere.

Ejemplo. En una prueba intercultural de algunas hipótesis de teoría de la conducta adaptadas de la teoría psicoanalítica,[5] Whiting y Child estudiaron la relación entre la práctica de la crianza de los niños y las costumbres vinculadas con las enfermedades en varias culturas analfabetas. Una hipótesis de su estudio, derivada de la noción de la fijación negativa, era sus explicaciones orales de la enfermedad: la enfermedad resulta de ingerir veneno, de beber ciertos líquidos y de ciertas fórmulas verbales y encantamientos ejecutados por otros. Los juicios de la típica socialización oral de la ansiedad en cualquier sociedad estuvieron fundamentados en la rapidez y severidad de la socialización oral, la frecuencia del castigo típico en la socialización oral y la severidad del conflicto emocional evidenciado por los niños durante el periodo de socialización oral.

Se utilizaron extractos de los informes etnológicos de las culturas analfabetas para recabar los datos necesarios. Utilizando sólo los resúmenes concernientes a las costumbres relacionadas a la enfermedad, los jueces calificaron a las sociedades en dos grupos: aquellas con explicaciones orales de la enfermedad y aquellas con ausencia de explicaciones orales de la enfermedad. Otros jueces, empleando los extractos concernientes a la práctica de crianza de los niños, adjudicaron puntuaciones a cada sociedad en cuanto al grado de la socialización oral típica de sus niños. Las puntuaciones de las 39 sociedades en las cuales fue posible emitir un juicio de ausencia o presencia de explicaciones orales, varían de 6 a 17.

[5] Whiting, J. W. M. y Child, I.L., *Child training and personality*, Yale University Press, Nueva Haven, 1953.

i. Hipótesis nula. H_0: no existe diferencia entre la mediana de socialización oral de la ansiedad en sociedades en que existen explicaciones orales de las enfermedades y aquellas sociedades que no tienen explicaciones de ese tipo. H_1: la mediana de la socialización oral de la ansiedad en sociedades con explicaciones orales es mayor que en aquellas sociedades que no tienen explicaciones orales.

ii. Prueba estadística. Las puntuaciones constituyen, en el mejor de los casos, mediciones en escala ordinal; en virtud de esto, lo más apropiado es utilizar una prueba no paramétrica. Para evaluar la H_0 con los datos pertenecientes a dos grupos de sociedades independientes podemos utilizar la prueba de la mediana.

Tabla 5.16. Socialización oral de la ansiedad y explicaciones orales de las enfermedades.[1,2]

	Sociedades con ausencia de explicaciones orales	*Sociedades con presencia de explicaciones orales*
		17 Marquesans
		16 Dobuans
		15 Baiga
		15 Kwoma
		15 Thonga
		14 Alorese
		14 Chagga
		14 Navaho
Sociedades que presentan puntuaciones de socialización oral de la ansiedad por encima de la mediana	13 Lapp	13 Dahomeans
		13 Lesu
		13 Masai
	12 Chamorro	12 Lepcha
	12 Samoans	12 Maori
		12 Pukapukans
		12 Trobrianders
		11 Kwakiutl
		11 Manus
	10 Arapesh	10 Chiricahua
	10 Balinese	10 Comanche
	10 Hopi	10 Siriono
	10 Tanala	
	9 Paiute	
	8 Chenchu	8 Bena
Sociedades que presentan puntuaciones de socialización oral de la ansiedad por abajo de la mediana	8 Teton	8 Slave
	7 Flathead	
	7 Papago	
	7 Venda	
	7 Warrau	
	7 Wogeo	
	6 Ontong-Javanese	6 Kurtatchi

[1] Reproducción de la tabla 4 de Whiting, J. W. y Child, I. L., *Child Training and personality*, Yale University Press, Nueva Haven, 1953, pág. 156; con autorizacion por cortesía de los autores y el editor.
[2] El nombre de cada sociedad es precedida por su puntuación de socialización oral de la ansiedad.

iii. *Nivel de significación.* $\alpha = 0.01$ y N es el número de sociedades en las cuales fue posible obtener información etnológica en las dos variables medidas = 39; m es el número de sociedades sin explicación oral de la enfermedad = 16; n es el número de sociedades con explicación oral de la enfermedad = 23.

iv. *Distribución muestral.* Puesto que el tamaño de la muestra es grande, se utilizará la X^2 con aproximación a la prueba exacta de Fisher [ecuación (5.3)]. La distribución muestral de X^2 es una ji cuadrada con $gl = 1$ asintóticamente distribuida.

v. *Región de rechazo.* En tanto que H_1 predice la dirección de la diferencia, la región de rechazo es unidireccional. Ésta consiste en todos los valores de tabla de mediana dividida que se encuentran en la dirección predicha y son tan extremos que la probabilidad asociada con su ocurrencia cuando H_0 es verdadera (como lo determina la prueba apropiada), es igual o menor que $\alpha = 0.01$

vi. *Decisión.* En la tabla 5.16 se muestran las puntuaciones asignadas a cada una de las 39 sociedades. Las $m + n$ puntuaciones se encuentran divididas por la mediana combinada. (Utilizamos el valor de mediana de Whiting y Child de las 39 sociedades, que es igual a 10.5.) En la tabla 5.17 se muestran los datos presentados en la forma que se requiere para la aplicación de la prueba de la mediana. Puesto que ninguna de las frecuencias esperadas es menor que cinco y ya que $m + n > 20$, podemos utilizar la prueba de la X^2 para evaluar la H_0.

$$X^2 = \frac{N\,(|AD - BC| - N/2)^2}{(A + B)(C + D)(A + C)(B + D)} \tag{5.3}$$

$$= \frac{N\,(|AD - BC| - N/2)^2}{(A + B)(C + D)(m)(n)}$$

$$= \frac{39\,(|(3)(6) - (17)(13)| - 39/2)^2}{(20)(19)(16)(23)}$$

$$= 9.39$$

Tabla 5.17. Socialización oral de la ansiedad y explicaciones orales de las enfermedades.

	Sociedades con ausencia de explicaciones orales	*Sociedades con presencia de explicaciones orales*	*Combinados*
Sociedades que presentan puntuaciones de socialización oral de la ansiedad por encima de la mediana	3	17	20
Sociedades que presentan puntuaciones de socialización oral de la ansiedad por abajo de la mediana	13	6	19
Total	16	23	39

La tabla C del Apéndice I, nos muestra que $X^2 > 9.39$ con $gl = 1$ tiene una probabilidad de ocurrencia, cuando H_0 es verdadera, de $p < 0.5(0.01) = 0.005$, para pruebas unidireccionales. Así, nuestra decisión es rechazar H_0 con $\alpha = 0.01$. Concluimos que la mediana de la socialización oral de la ansiedad es mayor en sociedades donde existen explicaciones orales de las enfermedades, que en aquellas sociedades en las que no las hay.

Resumen del procedimiento

Éstos son los pasos que hay que seguir para la aplicación de la prueba de la mediana:

1. Determine la mediana para las $m + n$ puntuaciones.
2. Divida las puntuaciones de los grupos utilizando la mediana combinada. Arregle los datos resultantes en una tabla de 2×2 (como la tabla 5.15). Si muchas puntuaciones caen exactamente sobre la mediana combinada, divida las puntuaciones en estas categorías: aquellos datos que exceden a la mediana y aquellos datos que no la exceden.
3. Encuentre la probabilidad de los datos observados ya sea mediante la prueba exacta de Fisher si $(m + n) \leq 20$, o por la aproximación de la ji cuadrada [ecuación (5.3)] si $(m + n) > 20$.
4. Si la probabilidad resultante de los cálculos anteriores es igual o menor que α, rechace H_0.

Potencia–eficacia

Mood (1954) ha demostrado que cuando la prueba de la mediana se aplica a datos medidos, al menos, en escala de intervalo de distribuciones normales con varianza común (es decir, datos que podrían propiamente ser analizados mediante la prueba t), tiene la misma potencia–eficacia que la prueba de los signos. Esto es, su potencia–eficacia es de alrededor del 95 % para $(m + n)$ tan pequeño como 6. Esta potencia–eficacia disminuye conforme se incrementa el tamaño de la muestra, alcanzando una eficacia asintótica de $2/\pi = 63$ %.

Referencias bibliográficas

Varios análisis de la prueba de la mediana se encuentran en Mood (1950), así como en las fuentes mencionadas en la sección prueba exacta de Fisher, de este capítulo.

LA PRUEBA DE WILCOXON–MANN–WHITNEY

Función

La prueba de Wilcoxon–Mann–Whitney[6] puede utilizarse para evaluar si dos grupos independientes fueron extraídos de la misma población, si de las variables en estudio se han obtenido datos en, al menos, escala ordinal. Ésta es una de las pruebas no paramétricas más poderosas y constituye una opción bastante buena a la prueba paramétrica t cuando el investigador desea evitar los supuestos de la prueba t o cuando las mediciones de la investigación se encuentran en una escala inferior a la de intervalo.

Supóngase que tenemos muestras de dos poblaciones, X y Y. La hipótesis nula es que X y Y tienen la misma distribución. La hipótesis alterna es que X es estocásticamente mayor que Y, una hipótesis unidireccional. Podemos aceptar H_1 si la probabilidad de que una puntuación de X sea mayor que una puntuación de Y, es mayor que un medio. Esto es, si X es una observación de la población X y Y es una observación de la población Y, entonces H_1 es que $P[X > Y] > 1/2$. Si las evidencias apoyan a H_1, esto implica que la mayor parte de los elementos de la población X es mayor que la mayor parte de los elementos de la población Y. En este orden de ideas, la hipótesis nula es H_0: $P[X > Y] = 1/2$.

Por supuesto, nuestra hipótesis puede plantear que Y es estocásticamente mayor que X. En tal caso, la hipótesis alterna sería que $P[X > Y] < 1/2$. La confirmación de este planteamiento implicaría que la mayor parte de Y es mayor que el grueso de X.

Para una prueba bidireccional, es decir, para una predicción en donde no se plantea una dirección de las diferencias, H_1 sería que $P[X > Y] \neq 1/2$.

Otra manera de plantear la hipótesis alterna es que la mediana de X es mayor que la mediana de Y, esto es, H_1: $\theta_x > \theta_y$. De manera similar, las otras hipótesis también pueden ser planteadas en términos de medianas.

Método

En una muestra, m es el número de casos del grupo X y n es el número de casos de la muestra del grupo Y. Suponemos que las dos muestras son independientes. Para aplicar la prueba de Wilcoxon, primero debemos combinar las observaciones o puntuaciones de ambos grupos y ordenarlos por rangos de manera ascendente. En este ordenamiento se considera el tamaño algebraico, es decir, los rangos inferiores serán asignados a los valores negativos mayores, en caso de existir.

Enfoquemos nuestra atención a uno de los grupos, digamos, el grupo X con m casos. El valor de W_x (es el estadístico utilizado por esta prueba) es la sumatoria de los rangos del primer grupo.

[6] Mann, Whitney y Wilcoxon (entre otros muchos) independientemente propusieron pruebas no paramétricas, las cuales son en esencia las mismas que las presentadas en esta sección. En la primera edición se presentó la prueba en la forma propuesta por Mann y Whitney. La forma que se emplea en esta edición sigue más bien a la de Wilcoxon. Por conveniencia, a menudo nos referiremos a ésta como la prueba de Wilcoxon.

Por ejemplo, supongamos que teníamos un grupo experimental de tres casos y un grupo control de cuatro casos. Tendríamos que $m = 3$ y $n = 4$. Supongamos que éstas fueran las puntuaciones:

Puntuaciones del grupo experimental X: 9 11 15
Puntuaciones del grupo control Y: 6 8 10 13

Para encontrar el valor de W_x, primero debemos ordenar por rangos de manera ascendente estas puntuaciones, teniendo el cuidado de identificar cada puntuación como correspondiente al grupo X o Y:

Puntuación:	6	8	9	10	11	13	15
Grupo:	Y	Y	X	Y	X	Y	X
Rango:	1	2	3	4	5	6	7

Consideremos ahora al grupo experimental y calculemos la sumatoria de los rangos de ese grupo. Así, tenemos que

$$W_x = 3 + 5 + 7 = 15$$

De la misma manera,

$$W_y = 1 + 2 + 4 + 6 = 13$$

El lector debe recordar que la sumatoria de los primeros N números enteros es la siguiente:

$$1 + 2 + 3 + \ldots + N = \frac{N\,(N + 1)}{2} \tag{5.7}$$

Por tanto, para nuestra muestra de tamaño $N = m + n = 7$, la sumatoria de rangos es $7(7 + 1)/2 = 28$. Además, la sumatoria de rangos para ambos grupos debería ser igual a la sumatoria de los rangos del grupo combinado. Esto es,

$$W_x + W_y = \frac{N\,(N + 1)}{2} \tag{5.8}$$

Si H_0 fuera verdadera, esperaríamos que el promedio de rangos en cada uno de ambos grupos fueran aproximadamente iguales. Si la sumatoria de rangos de un grupo es muy grande (o muy pequeña), entonces tenemos razones para sospechar que las muestras no fueron extraídas de la misma población. La distribución muestral de W_x cuando H_0 es verdadera, es conocida, y con este dato podemos determinar la probabilidad asociada con la ocurrencia según H_0 de cualquier W_x tan extremo como el valor observado.

Muestras pequeñas

Cuando m y n son menores o iguales que 10, puede utilizarse la tabla J del Apéndice I. Para determinar la probabilidad exacta asociada con la ocurrencia cuando H_0 es verdadera de cualquier W_x tan extremo como un valor observado de W_x. El lector observará que la tabla J del Apéndice I está compuesta por subtablas separadas, una para cada valor de m del 1 al 10, y cada una de las subtablas tiene entradas de $n = m$ a 10. (Realmente $n = m$ a 12 para $m = 3$ o 4.) Para determinar la probabilidad según H_0 asociada con el dato, el investigador necesita conocer m (el tamaño del grupo más pequeño), n (el tamaño del grupo más grande) y W_x. Con esta información, la probabilidad asociada con W_x puede ser leída de la subtabla apropiada a la hipótesis H_1.

En nuestro ejemplo, $m = 3$, $n = 4$ y $W_x = 15$. La subtabla para $m = 3$ en la tabla J del Apéndice I muestra que para $n = 4$ la probabilidad de observar un valor de $W_x \geq 15$ cuando H_0 es verdadera es 0.200. Este valor se encuentra al seleccionar el valor crítico superior (c_U) que es 15 y localizando la entrada en la columna para $n = 4$. El valor a la izquierda de (c_U) $= 15$ es la probabilidad requerida. Si desea la probabilidad requerida de que $W_x \leq c_L$ (c_L es el valor crítico inferior), la búsqueda inicia en la entrada correspondiente a la primera columna.

Por conveniencia y economía, la tabla J del Apéndice I está arreglada para $m \leq n$; es decir, el grupo asociado con las puntuaciones X es el más pequeño. Esta restricción no genera problemas en el uso de la prueba de Wilcoxon, ya que las etiquetas de identificación del grupo pueden intercambiarse y la tabla puede utilizarse para los grupos transformados. Sin embargo, el investigador debe tener presente que la hipótesis alterna está planteada correctamente si las etiquetas de las variables fueron intercambiadas.

Ejemplo. Para muestras pequeñas. Solomon y Coles[7] estudiaron si las ratas generalizaban aprendizaje de imitación cuando eran colocadas bajo un nuevo impulso (una nueva condición motivante) y en una nueva situación. Cinco ratas fueron entrenadas en imitar a ratas líderes en un laberinto en T, cuando tenían hambre. Los sujetos se encontraban en un régimen de privación de alimento y debían seguir al líder a través del laberinto para conseguir comida. Después que se logró que las ratas experimentales imitaran a las ratas líderes, fueron cambiadas a una situación de evitación de choques eléctricos, donde la imitación que hicieran del líder les permitiría evitar dicha estimulación aversiva. La conducta en la situación de evitación se comparó con la de cuatro sujetos control que no habían tenido entrenamiento previo en seguir a líderes. La hipótesis era que las primeras cinco ratas que habían sido entrenadas para imitar transferirían ese entrenamiento a la nueva situación y así, alcanzarían el criterio de aprendizaje en la situación de evitación más rápidamente que las cuatro ratas control. Las comparaciones se realizaron en términos de cuántos ensayos le tomó a cada rata alcanzar el criterio de 10 respuestas correctas en 10 ensayos.

i. Hipótesis nula. H_0: el número de ensayos necesarios para alcanzar el criterio en la situación de evitación es el mismo para las ratas que fueron entrenadas para seguir a un líder con el fin de conseguir comida, que para aquellas que no fueron previamente entrenadas. H_1: las ratas que fueron previamente entrenadas para seguir a

[7] Solomon, R. L. y Coles, M. R., "A case of failure of generalization of imitation across drives and across situations", en *Journal of Abnormal and Social Psychology*, núm. 49, 1953, págs. 7-13. En este ejemplo se incluyen sólo dos de los grupos estudiados.

un líder con el fin de conseguir comida alcanzarán el criterio de ejecución en la situación de evitación de choques eléctricos en menos ensayos que las que no tuvieron el entrenamiento previo.

ii. *Prueba estadística.* Se seleccionó la prueba de Wilcoxon ya que en este estudio se emplearon dos muestras independientes, pequeñas, y se utilizaron mediciones (número de ensayos como criterio de índice de velocidad de aprendizaje) que, probablemente, corresponden a una escala ordinal.

iii. *Nivel de significación.* $\alpha = 0.05$, $m = 4$ (ratas control) y $n = 5$ (ratas experimentales).

iv. *Distribución muestral.* Las probabilidades asociadas con la ocurrencia según H_0 de valores tan grandes como una W_x observada para m y n pequeñas, se proporcionan en la tabla J del Apéndice I.

v. *Región de rechazo.* Puesto que H_1 plantea la dirección de la diferencia predicha, la región de rechazo es unidireccional. Ésta consiste en todos los valores de W_x los cuales son tan grandes que la probabilidad asociada con su ocurrencia cuando H_0 es verdadera, sea menor o igual que $\alpha = 0.05$. (Ya que el grupo control es denominado X, la hipótesis alterna es H_1: $\theta_x > \theta_y$, esto es, la mediana del grupo control es mayor que la mediana del grupo experimental.)

vi. *Decisión.* Los números de ensayos como criterio requerido para las ratas experimentales y control fueron los siguientes:

Ratas controles:	110	70	53	51	
Ratas experimentales:	78	64	75	45	82

Al ordenar estas puntuaciones en orden de magnitud, identificando a cada una tenemos:

Puntuación:	45	51	53	64	70	75	78	82	110
Grupo:	*Y*	*X*	*X*	*Y*	*X*	*Y*	*Y*	*Y*	*X*
Rango:	1	2	3	4	5	6	7	8	9

Con estos datos encontramos que la sumatoria de rangos para el grupo control es $W_x = 2 + 3 + 5 + 9 = 19$. En la tabla J del Apéndice I localizamos la subtabla para $m = 4$. Puesto que la hipótesis alterna es que W_x sería la mayor, utilizamos el lado derecho (superior) de la distribución. Cuando H_0 es verdadera, vemos que $P[W_x \geq 19] = 0.6349$. Nuestra decisión es que los datos no proporcionan evidencia que justifique el rechazar H_0 en el nivel de significación seleccionado. La conclusión es que estos datos no apoyan la hipótesis de que el entrenamiento previo en imitación se generalice a otras situaciones y en otras circunstancias motivacionales.[8]

Muestras grandes

La tabla J del Apéndice I, no se puede utilizar si $m > 10$ o $n > 10$ ($n > 12$ si $m = 3$ o 4). Sin embargo, se ha demostrado que conforme se incrementa el tamaño de m y n, la distribución muestral de W_x se aproxima rápidamente a la distribución normal con

[8] Solomon y Coles presentaron las mismas conclusiones. La prueba utilizada en el estudio descrito en dicho artículo no fue señalada por los autores.

$$\text{Media} = \mu_{w_x} = \frac{m\,(N+1)}{2} \tag{5.9}$$

y

$$\text{Varianza} = \sigma^2_{w_x} = \frac{mn\,(N+1)}{12} \tag{5.10}$$

Esto es, cuando $m > 10$ o $n > 10$, podemos determinar la significación de un valor observado de W_x por medio de

$$z = \frac{W_x \pm 0.5 - \mu_{w_x}}{\sigma_{w_x}} = \frac{W_x \pm 0.5 - m\,(N+1)/2}{\sqrt{mn\,(N+1)/12}} \tag{5.11}$$

la cual asintóticamente se distribuye de manera normal con media igual a cero y varianza igual a uno. Esto es, la probabilidad asociada con la ocurrencia cuando H_0 es verdadera de un valor tan extremo como una z observada, puede determinarse mediante la tabla A del Apéndice I. El valor $(+0.5)$ es agregado si deseamos encontrar las probabilidades en el lado *izquierdo* de la distribución y (-0.5) es agregado si deseamos encontrar las probabilidades en el lado *derecho* de la distribución.

Ejemplo. Para muestras grandes. Para nuestro ejemplo reexaminaremos los datos de Whiting y Child que utilizamos para demostrar el uso de la prueba de la mediana.

i. *Hipótesis nula.* H_0: la socialización oral es igualmente severa en ambos tipos de sociedades (con presencia o con ausencia de explicaciones orales de las enfermedades). H_1: las sociedades con explicaciones orales de las enfermedades presentes son (estocásticamente) superiores en la socialización oral de la ansiedad, que aquellas que no tienen explicaciones orales de las enfermedades.
ii. *Prueba estadística.* Los dos grupos de sociedades constituyen dos grupos independientes, y la medida de la socialización oral de la ansiedad (escala de puntuaciones) constituye una medición de escala ordinal (en el mejor de los casos). Por estas razones, la prueba de Wilcoxon es la apropiada para analizar esos datos.
iii. *Nivel de significación.* $\alpha = 0.01$, m es el número de sociedades con explicaciones orales ausentes $= 16$, y n es el número de sociedades con explicaciones orales presentes $= 23$.
iv. *Distribución muestral.* Para $n > 10$ se mantienen los valores de z (ecuación 5.11). La probabilidad asociada con la ocurrencia según H_0 de valores tan extremos como una z observada, puede determinarse utilizando la tabla A del Apéndice I.
v. *Región de rechazo.* Puesto que H_1 predice la dirección de la diferencia, la región de rechazo es unidireccional. Ésta consiste en todos los valores de z que sean tan extremos (en la dirección predicha) que la probabilidad asociada según H_0 es igual o menor que $\alpha = 0.01$.
vi. *Decisión.* Las puntuaciones asignadas a cada una de las 39 sociedades se muestran en la tabla 5.18, junto con el rango de cada una en el grupo combinado. Nótese que a los rangos empatados se les asigna el rango promedio. Para estos datos, $W_x = 200.0$ y $W_y = 580.0$. Podemos encontrar el valor de z al sustituir los valores en la ecuación (5.11):

$$z = \frac{W_x \pm 0.5 - m\,(N + 1)/2}{\sqrt{mn\,(N + 1)/12}} \tag{5.11}$$

$$= \frac{200 + 0.5 - 16\,(39 + 1)/2}{\sqrt{(16)(23)(39 + 1)/12}}$$

$$= -\,3.41$$

Tabla 5.18. Socialización oral de la ansiedad y explicación oral de las enfermedades.

Sociedades con presencia de explicaciones orales	*Puntuación en socialización oral de la ansiedad*	*Rango*	*Sociedades con presencia de explicaciones orales*	*Puntuación en socialización oral de la ansiedad*	*Rango*
Lapp	13	29.5	Marquesans	17	39
Chamorro	12	24.5	Dobuans	16	38
Samoans	12	24.5	Baiga	15	36
Arapesh	10	16	Kwoma	15	36
Balinese	10	16	Thonga	15	36
Hopi	10	16	Alorese	14	33
Tanala	10	16	Chagga	14	33
Paiute	9	12	Navaho	14	33
Chenchu	8	9.5	Dahomeans	13	29.5
Teton	8	9.5	Lesu	13	29.5
Flathead	7	5	Masai	13	29.5
Papago	7	5	Lepcha	12	24.5
Venda	7	5	Maori	12	24.5
Warrau	7	5	Pukapukans	12	24.5
Wogeo	7	5	Trobrianders	12	24.5
Ontong-Javanese	6	1.5	Kwakiutl	11	20.5
		$W_x = 200.00$	Manus	11	20.5
			Chiricahua	10	16
			Comanche	10	16
			Siriono	10	16
			Bena	8	9.5
			Slave	8	9.5
			Kurtatchi	6	1.5
					$W_y = 580.00$

Al recurrir a la tabla A del Apéndice I, ésta revela que $z \leq -3.41$ tiene una probabilidad unidireccional cuando H_0 es verdadera de $p < 0.0003$. Puesto que p es menor que $\alpha = 0.01$, nuestra decisión es rechazar H_0 en favor de H_1. Concluimos que las sociedades con presencia de explicaciones orales de las enfermedades son (estocásticamente) superiores en la socialización oral de la ansiedad, que las sociedades con explicaciones orales ausentes.

Es importante destacar que para estos datos la prueba de Wilcoxon muestra mayor poder para rechazar H_0, que la prueba de la mediana. Al evaluar una hipótesis similar acerca de estos datos, la prueba de la mediana proporcionó un valor que permitió rechazar la H_0 en el nivel de $p < 0.005$ (unidireccional), mientras que la prueba de Wilcoxon proporcionó un valor que permitió en rechazo de H_0 en el nivel de $p < 0.0003$ (unidireccional). El hecho de que la prueba de Wilcoxon sea más poderosa que la prueba de la mediana no es sorprendente, en vista de que considera el valor del rango de cada observación, más que simplemente su localización respecto a la mediana combinada; así, utiliza más la información contenida en los datos. El uso de una prueba más poderosa está justificado si se cubren sus requisitos.

Empates

La prueba de Wilcoxon supone que las puntuaciones se han muestreado de una distribución que es continua. Con una medición precisa de una variable continua, la probabilidad de un empate es cero. Sin embargo, con las mediciones relativamente crudas, las cuales típicamente empleamos en investigación en las ciencias de la conducta, pueden ocurrir empates. Suponemos que las dos observaciones (o más) que resulten empatadas son realmente diferentes, sólo que esa diferencia es muy refinada o diminuta para ser detectada por nuestras mediciones.

Cuando ocurren empates, damos a cada observación empatada el promedio de los rangos que tendrían si no hubieran ocurrido los empates.[9]

Si los empates ocurren entre dos o más observaciones del mismo grupo, el valor de W_x no se ve afectado. Pero si los empates ocurren entre dos o más observaciones en las cuales se impliquen puntuaciones de los dos grupos, se afecta el valor tanto de W_x como de W_y. Aunque el efecto generalmente es despreciable, tenemos disponible una corrección y puede utilizarse cada vez que empleemos la aproximación para muestra grande a la distribución muestral de W_x.

El efecto de los rangos empatados es cambiar la variabilidad del conjunto de rangos. Así, la corrección para los empates debe aplicarse a la varianza de la distribución muestral de W_x. Corregida para los empates, la varianza queda como sigue:

$$\sigma^2_{W_x} = \frac{mn}{N\,(N-1)}\left(\frac{N^3-N}{12} - \sum_{j=1}^{g}\frac{t_j^3-t_j}{12}\right) \tag{5.12}$$

donde $N = m + n$, g es el número de los distintos grupos de rangos empatados, y t es el número de rangos empatados en el j-ésimo grupo. Utilizando esta corrección para los empates, el valor de z se convierte en

$$z = \frac{W_x \pm 0.5 - m\,(N+1)/2}{\sqrt{[mn/N(N-1)]\left[(N^3-N)/12 - \sum_{j=1}^{q}(t_j^3-t_j)/12\right]}} \tag{5.13}$$

[9] Si dos o más puntuaciones se empataron en el mismo rango, el rango que se asigna es el *promedio de los rangos empatados*, los cuales se habrían asignado si las puntaciones hubieran sido claramente diferentes. Así, si tres puntuaciones se empataron en la primera (inferior) posición, a cada puntuación se le asignaría el rango de 2 para $(1 + 2 + 3)/3 = 2$. El siguiente rango que se asignaría sería el 4,

Se puede observar que si no hay empates, la expresión anterior se reduce a la proporcionada originalmente en la ecuación (5.11).

El uso de la corrección para los empates en la prueba de Wilcoxon se ilustra aplicándola a los datos de la tabla 5.18. Para estos datos,

$$m + n = 16 + 23 = 39 = N$$

Nosotros observamos los siguientes grupos de empates:

Agrupación	*Valor*	*Rango*	t_j
1	6	1.5	2
2	7	5	5
3	8	9.5	4
4	10	16	7
5	11	20.5	2
6	12	24.5	6
7	13	29.5	4
8	14	33	3
9	15	36	3

A fin de encontrar la varianza, necesitamos calcular el factor de corrección para los $g = 9$ grupos de empates:

$$\sum_{j=1}^{9} \frac{t_j^3 - t_j}{12} = \frac{2^3 - 2}{12} + \frac{5^3 - 5}{12} + \frac{4^3 - 4}{12} + \ldots + \frac{3^3 - 3}{12}$$

$$= 0.5 + 10.0 + 5.0 + 28.0 + \ldots + 2.0 = 70.5$$

Utilizando este factor de corrección y $m = 16$, $n = 23$, $N = 39$, tenemos

$$z = \frac{W_x \pm 0.5 - m(N + 1)/2}{\sqrt{[mn/N(N - 1)]\left((N^3 - N)/12 - \sum_{j=1}^{g} (t_j^3 - t_j)/12\right)}} \tag{5.13}$$

$$= \frac{200 + 0.5 - 16(39 + 1)/2}{\sqrt{[(16)(23)]/[39(39 - 1)][(39^3 - 39)/12 - 70.5]}} = -3.44$$

porque los rangos 1, 2 y 3 ya fueron asignados. Si dos puntuaciones se empataron en la primera (inferior) posición, a ambas puntuaciones se les asignaría el rango de 1.5 puesto que $(1 + 2)/2 = 1.5$ y la siguiente puntuación recibiría el rango 3.

El valor de z cuando se realizó la corrección para los empates es un tanto mayor que el que se encontró anteriormente cuando no se utilizó la corrección. La diferencia entre $z \leq -3.41$ y $z \leq -3.44$ es despreciable, tanto es así que la probabilidad proporcionada por la tabla A del Apéndice I es la misma. En ambos casos, $p < 0.0003$ (en una prueba unidireccional).

Como se demuestra en este ejemplo, los empates sólo tienen un pequeñísimo efecto. Aun cuando muchas puntuaciones presenten empates (este ejemplo tiene sobre el 90 % de observaciones con empates), el efecto es muy pequeño. Obsérvese, sin embargo, que la magnitud del efecto de corrección depende considerablemente del *número de empates* en cada grupo de ellos. Así, un empate de "tamaño" 4 contribuye en 5.0 puntos al factor de corrección; cualquiera de dos empates de "tamaño" 2 juntos contribuyen sólo en 1.0 puntos (es decir, 0.5 + 0.5), y un empate de "tamaño" 6 contribuye con 17.5 puntos, mientras que dos empates de "tamaño" 3, juntos contribuyen sólo con 2.0 + 2.0 = 4.0.

Cuando se emplea la corrección, siempre *incrementa* levemente la magnitud de z, haciéndola un poco más significativa. Por tanto, cuando no aplicamos la corrección para los empates, nuestra prueba es "conservadora" ya que la probabilidad asociada estará un tanto más "inflada" si se la compara con la que corresponde a la z corregida. Esto es, el valor de la probabilidad asociada con los datos observados cuando H_0 es verdadera, será un poquito mayor que la que encontraríamos si se empleara la corrección. Nuestra recomendación es que se debe aplicar la corrección para los empates, sólo si la proporción de empates es muy grande, si algunas t son grandes, o si la probabilidad obtenida sin corrección se encuentra muy cercana al valor de α previamente seleccionado.

Resumen del procedimiento

Pasos a seguir en la aplicación de la prueba de Wilcoxon:

1. Dar el valor de m y n. El número de casos en el grupo más pequeño (denominado X) es m; y los casos del grupo mayor (denominado Y) es n.
2. Ordene por rangos las puntuaciones de ambos grupos, asignando el rango 1 a la puntuación algebraicamente menor. Los rangos variarán de 1 a $m + n = N$. Asigne a las observaciones empatadas el promedio de los rangos empatados.
3. Determine el valor de W_x sumando los rangos del grupo X.
4. El método para determinar la significación de W_x depende del tamaño de m y n:

   a) Si $m \leq 10$ y $n \leq 10$ (o $n \leq 12$ para $m = 3$ o 4), la probabilidad asociada exacta con un valor tan grande (o tan pequeño) como una W_x, se proporciona en la tabla J del Apéndice I. Los valores de tablas son probabilidades unidireccionales. Para pruebas de dos colas, duplique los valores de tablas.[10]

[10] Puede no ser posible alcanzar un nivel de probabilidad exacta con una prueba bidireccional, debidos a la naturaleza discreta de la distribución muestral. Para alcanzar una mayor precisión, la re-

b) Si $m > 10$ o $n > 10$, la probabilidad asociada con un valor tan extremo como un valor de W_x puede determinarse calculando la aproximación normal mediante la ecuación (5.11) y evaluando la significación de z, con base en la tabla A del Apéndice I. Para una prueba bidireccional, la probabilidad que aparece en tablas debe duplicarse. Si el número de empates es grande o si la probabilidad obtenida está muy cercana al nivel de significación escogido (α), aplique la corrección para los empates, es decir, use la ecuación (5.13).

5. Si el valor observado de W_x tiene una probabilidad asociada igual o menor que α, rechace H_0 en favor de H_1.

Potencia–eficacia

Si la prueba de Wilcoxon se aplica a datos que pueden ser propiamente analizados por la prueba paramétrica más poderosa, la prueba *t*, su potencia–eficacia se aproxima a $3/\pi = 95.5$ %, conforme se incremente el tamaño de *N* y es cercano al 95 % aun con muestras de tamaño moderado. Es, además, una excelente opción a la prueba *t* y, por supuesto, no tiene todas las restricciones en los supuestos y requisitos asociados con dicha prueba.

En algunos casos, se ha demostrado que la prueba de Wilcoxon tiene un poder mayor a 1, vale decir, que es más poderosa que la prueba *t*.

Referencias bibliográficas

Para mayores detalles acerca de la prueba Wilcoxon–Mann–Whitney, el lector puede revisar Mann y Whitney (1947), Whitney (1948), Wilcoxon (1945) y Lehmann (1975).

PRUEBA PODEROSA DE RANGOS ORDENADOS

Función

La prueba de Wilcoxon–Mann–Whitney descrita en la sección anterior se utilizó para evaluar la hipótesis de que dos grupos independientes fueron extraídos de la misma población. Esta prueba supone que las variables *X* y *Y* se obtuvieron de una misma distribución continua. Las variables fueron medidas en (al menos) una escala ordinal. Una manera de plantear la hipótesis nula es H_0: $\theta_x = \theta_y$. Esto es, la mediana de la distribución *X* es igual a la mediana de la distribución *Y*. Cuando suponemos que las distribuciones son las mismas, estamos suponiendo que la variabilidad o las *varianzas* de las distribuciones son iguales. La hipótesis alterna

---

gión de rechazo puede seleccionarse con dos diferentes valores críticos, uno para cada lado, de tal suerte que $\alpha_1 + \alpha_2 = \alpha$.

sólo especifica que existe diferencia entre las medianas y, la hipótesis nula, supone que las variabilidades de las distribuciones son *las mismas*.

En ocasiones deseamos probar la hipótesis H_0: $\theta_x = \theta_y$, sin suponer que las distribuciones subyacentes son las mismas. Tal vez porque son conocidas las diferencias entre los grupos, o por una restricción en el rango, o por algún otro factor, el investigador tiene una razón para creer que las distribuciones subyacentes a X y Y no son iguales, pero aún desea evaluar la H_0. Esta clase de problema de evaluación es conocido entre los estadígrafos como el problema *Behrens-Fisher*. En tales casos, la prueba de Wilcoxon no es apropiada. La prueba poderosa de rangos ordenados, que se examina en esta sección, es una alternativa más adecuada que la prueba de Wilcoxon.

Método

El número de casos de la muestra del grupo X es m, y n es el número de casos de la muestra del grupo Y. Nosotros suponemos que las dos muestras son independientes. Para aplicar la prueba poderosa de rangos ordenados primero debemos combinar las observaciones o puntuaciones de ambos grupos y ordenar los rangos por tamaños, de manera ascendente. En este ordenamiento, el tamaño algebraico es el considerado, es decir, los rangos inferiores son asignados a los valores negativos mayores, si los hubiera.

Enfoquemos nuestra atención a uno de los dos grupos, digamos el grupo X, con m casos. Para ejemplificar el procedimiento, calcularemos un estadístico que es diferente del ordenamiento por rangos, llamado $\grave{U}$. A fin de compararlo con la prueba de Wilcoxon, utilizaremos los mismos datos para ilustrar el cálculo del estadístico. En ese ejemplo, había un grupo experimental de $m = 3$ casos y un grupo control de $n = 4$ casos. Las puntuaciones eran las siguientes:

Puntuaciones del grupo experimental X: 9 11 15
Puntuaciones del grupo control Y: 6 8 10 13

Aunque no emplearemos el estadístico de rangos ordenados, es necesario ordenar por rangos los datos, identificando a cada puntuación como perteneciente a los grupos X o Y:

Puntuación:	6	8	9	10	11	13	15
Grupo:	Y	Y	X	Y	X	Y	X

Para cada X_i contamos el número de observaciones de Y con un rango inferior. Este número representa la *ubicación* de las puntuaciones X y se denominará $U(YX_i)$. Para este ejemplo tenemos

X_i	$U(YX_i)$
9	2
11	3
15	4

Entonces, encontramos la media de la $U(YX_i)$:

$$U(YX) = \sum_{i=1}^{m} \frac{U(YX_i)}{m} \tag{5.14a}$$

$$= \frac{2 + 3 + 4}{3}$$

$$= 3$$

De manera similar, encontramos las *ubicaciones* de cada Y. Esto es, encontramos $U(XY_j)$, el número de observaciones de X que precede a cada Y_j.

Y_j	$U(XY_j)$
6	0
8	0
10	1
13	2

Entonces, encontramos la media:

$$U(XY) = \sum_{j=1}^{n} \frac{U(XY_j)}{n} \tag{5.14b}$$

$$= \frac{0 + 0 + 1 + 2}{4}$$

$$= 0.75$$

El siguiente paso es encontrar un índice de variabilidad de $U(YX_i)$ y $U(XY_j)$. Estos índices se calculan mediante

$$V_x = \sum_{i=1}^{m} [U(YX_i) - U(YX)]^2 \tag{5.15a}$$

y

$$V_y = \sum_{j=1}^{n} [U(XY_j) - U(XY)]^2 \tag{5.15b}$$

Para los datos de este ejemplo, tenemos:

$$V_x = (2 - 3)^2 + (3 - 3)^2 + (4 - 3)^2$$
$$= 1 + 0 + 1$$
$$= 2$$

y

$$V_y = (0 - 0.75)^2 + (0 - 0.75)^2 + (1 - 0.75)^2 + (2 - 0.75)^2$$
$$= 2.75$$

Finalmente, calculamos la prueba estadística $\grave{U}$:

$$\grave{U} = \frac{mU(YX) - nU(XY)}{2\sqrt{V_x + V_y + U(XY)U(YX)}} \tag{5.16}$$

$$= \frac{3(3) - 4(0.75)}{2\sqrt{2 + 2.75 + (0.75)(3)}}$$

$$= 1.13$$

La distribución muestral de $\grave{U}$ se ha tabulado y se encuentra en la tabla K del Apéndice I, para muestras pequeñas ($m \leq n \leq 12$). Conforme se incrementa el tamaño de las muestras, la distribución de $\grave{U}$ se aproxima a la de la distribución normal unitaria, así que puede utilizarse la tabla A del Apéndice I para determinar la significación de los valores del estadístico $\grave{U}$, calculado mediante la ecuación (5.16).

Para los datos del ejemplo anterior, $m = 3$, $n = 4$ y $\grave{U} = 1.13$. En la tabla k del Apéndice I se muestra que la probabilidad de obtener un valor de $\grave{U}$ de la muestra tan grande como 1.13 cuando H_0 es verdadera, es realmente mayor que 0.10. Puesto que los tamaños de la muestra son pequeños, la distribución de $\grave{U}$ es tal que no es posible alcanzar los niveles típicos de significación de 0.05 y 0.01. Por tanto, la tabla consiste en aquellos valores de $\grave{U}$ que son los más cercanos al nivel de significación deseado. Si la hipótesis alterna es bidireccional, las probabilidades de la tabla K deben duplicarse.

Ejemplo. Muchas hipótesis contemporáneas concernientes a la etiología de la esquizofrenia han sugerido que la dopamina desempeña un papel. Se ha comprobado que existe un incremento en la actividad dopaminérgica en algunos centros del sistema nervioso central en pacientes esquizofrénicos, comparados con pacientes no esquizofrénicos. Algunas drogas antipsicóticas parecen bloquear los receptores de la dopamina, y algunas drogas que incrementan la función central de la dopamina, agravan los síntomas esquizofrénicos. Una hipótesis es que los medicamentos neurolépticos actúan decrementando la trasmisión central de dopamina, resultando en una disminución de la actividad esquizofrénica.

En un estudio en el que se utilizaron 25 esquizofrénicos hospitalizados,[11] cada uno de

[11] Sternberg, D. E., Van Kammen, D. P. y Bunney, W. E., "Schizophrenia: Dopamine *b*-hydroxylase activity and treatment response", en *Science*, núm. 216, 1982, págs. 1423-1425.

ellos fue tratado con medicamentos antipsicóticos (neurolépticos), se observaron a lo largo de cierto periodo y fueron clasificados como psicóticos o no psicóticos por enfermeras profesionales del hospital. Quince se juzgaron como no psicóticos y 10 como psicóticos. De cada paciente se extrajeron muestras de líquido cefalorraquídeo y se evaluó la actividad de la dopamina *b*-hidroxilasa (DBH). Los resultados se muestran en la tabla 5.19. Los investigadores deseaban determinar si la diferencia en la actividad DBH entre los dos grupos era significativa.

Tabla 5.19. Actividad de la dopamina *b*-hidroxilasa en el líquido cefalorraquídeo de pacientes esquizofrénicos, después del tratamiento con medicamentos antipsicóticos.

Juzgados como no psicóticos	*Juzgados como psicóticos*
$m = 15$	$n = 10$
0.0252	0.0320
0.0230	0.0306
0.0210	0.0275
0.0200	0.0270
0.0200	0.0245
0.0180	0.0226
0.0170	0.0222
0.0156	0.0208
0.0154	0.0204
0.0145	0.0150
0.0130	
0.0116	
0.0112	
0.0105	
0.0104	

Nota: Las mediciones se encuentran en nmol/(ml)(hr)/(mg) de proteína.

i. *Hipótesis nula.* H_0: la actividad DBH en el líquido cefalorraquídeo de pacientes diagnosticados como psicóticos o no psicóticos durante el tratamiento, es la misma. H_1: la actividad DBH en los dos grupos es diferente.
ii. *Prueba estadística.* Se escogió la prueba poderosa de rangos ordenados porque este estudio abarca muestras independientes y utiliza mediciones (actividad DBH en el líquido cefalorraquídeo como índice de la actividad en centros del sistema nervioso central) que probablemente se encuentran en una escala ordinal. Adicionalmente, ya que estos grupos pueden diferir en términos de la variabilidad, la prueba poderosa de rangos ordenados es la apropiada.
iii. *Nivel de significación.* $\alpha = 0.05$, m son los pacientes juzgados como no psicóticos, y n son los pacientes que permanecieron psicóticos $= 10$.
iv. *Distribución muestral.* Las probabilidades asociadas con la ocurrencia, cuando H_0

es verdadera, de valores tan grandes como $\grave{U}$, pueden determinarse por medio de la distribución normal (tabla A del Apéndice I).

v. *Región de rechazo*. Puesto que H_1 no plantea una dirección para la diferencia, es apropiada una prueba bidireccional. Por tanto, ya que $\alpha = 0.05$, la región de rechazo consiste en todos los valores de $\grave{U}$ que sean mayores que 1.96 o menores que -1.96, utilizando la aproximación a la distribución normal para la distribución muestral de $\grave{U}$.

vi. *Decisión*. Para obtener el estadístico $\grave{U}$, necesitamos calcular las ubicaciones de las puntuaciones X y Y. En la tabla 5.20 se resumen los cálculos de $U(YX_i)$ y $U(XY_j)$. Con los valores encontrados en la tabla 5.20, utilizando las ecuaciones (5.14*a*) y (5.14*b*) se advierte que

$$U(YX) = \sum_{i=1}^{m} \frac{U(YX_i)}{m} \tag{5.14a}$$

$$= \frac{20}{15} = 1.33$$

y

$$U(XY) = \sum_{j=1}^{n} \frac{U(XY_j)}{n} \tag{5.14b}$$

$$= \frac{130}{10} = 13$$

El siguiente paso consiste en utilizar las ecuaciones (5.15*a*) y (5.15*b*) a fin de encontrar el índice de variabilidad para $U(YX_i)$ y $U(XY_j)$

$$V_x = \sum_{i=1}^{m} [U(YX_i) - U(YX)]^2 \tag{5.15a}$$

$$= 49.33$$

y

$$V_y = \sum_{j=1}^{n} [U(XY_j) - U(XY)]^2 = 68 \tag{5.15b}$$

Finalmente, calculamos la prueba estadística $\grave{U}$:

$$\grave{U} = \frac{mU(YX) - nU(XY)}{2\sqrt{V_x + V_y + U(XY)U(YX)}} \tag{5.16}$$

$$= \frac{15(1.33) - 10(13)}{2\sqrt{49.33 + 68 + (1.33)(13)}} = -4.74$$

Puesto que el valor observado de $\grave{U}$ es mayor que el valor crítico (-1.96), podemos rechazar la hipótesis H_0, que plantea que no existe diferencia en la actividad DBH de los dos grupos.

Tabla 5.20. Ubicaciones para las observaciones de la tabla 5.19.

X_i	$U(YX_i)$	$U(YX_i) - U(YX)$	Y_j	$U(XY_j)$	$U(XY_j) - U(XY)$
0.0104	0	$\frac{4}{3}$			
0.0105	0	$\frac{4}{3}$			
0.0112	0	$\frac{4}{3}$			
0.0116	0	$\frac{4}{3}$			
0.0130	0	$\frac{4}{3}$			
0.0145	0	$\frac{4}{3}$			
			0.0150	6	7
0.0154	1	$\frac{1}{3}$			
0.0156	1	$\frac{1}{3}$			
0.0170	1	$\frac{1}{3}$			
0.0180	1	$\frac{1}{3}$			
0.0200	1	$\frac{1}{3}$			
0.0200	1	$\frac{1}{3}$			
			0.0204	12	1
			0.0208	12	1
0.0210	3	$\frac{5}{3}$			
			0.0222	13	0
			0.0226	13	0
0.0230	5	$\frac{11}{3}$			
			0.0245	14	1
0.0252	6	$\frac{14}{3}$			
			0.0270	15	2
			0.0275	15	2
			0.0306	15	2
			0.0320	15	2
	20			130	
$U(YX) = \frac{20}{15} = \frac{4}{3} = 1.33$			$U(XY) = \frac{130}{10} = 13$		

EMPATES

En ocasiones ocurre que los datos observados presentan empates. En el cálculo de la prueba poderosa de rangos ordenados el ajuste para los empates es sencillo: se realiza en el cálculo de las ubicaciones:

$U(YX_i)$ = el número de Y observaciones en la muestra que son menores que X_i + 1/2 del número de Y observaciones en la muestra que son iguales que X_i.
$U(XY_j)$ = el número de X observaciones en la muestra que son menores que Y_j + 1/2 del número de X observaciones en la muestra que son iguales que Y_j.

Los cálculos de $X(YX)$, $U(XY)$, V_x, V_y y $\grave{U}$ se complementan utilizando los ajustes a las ubicaciones.

Resumen del procedimiento

Éstos son los pasos que hay que seguir para la aplicación de la prueba poderosa de rangos ordenados:

1. Ordene las puntuaciones combinando los grupos X y Y. Para cada puntuación en cada grupo, calcule las ubicaciones $U(YX_i)$ y $U(XY_j)$. Si es necesario, realice el ajuste correspondiente a las observaciones empatadas.
2. Calcule la media para las ubicaciones $U(YX)$ y $U(XY)$, los índices de variabilidad V_x y V_y y el estadístico $\grave{U}$ mediante la ecuación (5.16).
3. El método para determinar la significación de $\grave{U}$ depende de el tamaño de m y n:

   a) Si m y n son menores que 12, la significación de las probabilidades asociadas con valores grandes de $\grave{U}$ son proporcionadas en la tabla K del Apéndice I. Los valores de tablas son probabilidades unidireccionales. Para pruebas bidireccionales, duplique estas probabilidades.
   b) Si m y n son mayores que 12, la probabilidad asociada con un valor tan extremo como el valor observado de $\grave{U}$, puede determinarse haciendo uso de la tabla A del Apéndice I que, para m y n grandes, la distribución muestral de $\grave{U}$ es aproximadamente normal.

4. Si el valor observado de $\grave{U}$ tiene una probabilidad asociada igual o menor que α, rechace H_0 en favor de H_1.

Potencia–eficacia

Los procedimientos de la prueba poseen niveles de significación *exactos* iguales a α para evaluar la hipótesis de que X y Y tienen distribuciones idénticas. Los niveles de significación son aproximados cuando probamos la hipótesis H_0: $\theta_x = \theta_y$, sin que se requieran varianzas iguales para ambas poblaciones. En general, la prueba tiene esencialmente la misma potencia que la de Wilcoxon (cuando se cubren los supuestos de la prueba); sin embargo, la prueba parece aproximarse más rápidamente a la distribución normal conforme m y n incrementan su tamaño.

Referencias bibliográficas

El problema de Behrens-Fisher (comparar dos grupos con varianzas distintas) tiene una larga historia en la estadística. Las pruebas no paramétricas son relativamente nuevas y algunas que han sido propuestas resultan difíciles de calcular. Para comentarios adicionales, consúltese Lehmann (1975) y Randles y Wolfe (1979). La prueba descrita en este capítulo se debe a Fligner y Policello (1981).

PRUEBA DE KOLMOGOROV-SMIRNOV PARA DOS MUESTRAS

Función y racionalización

La *prueba de Kolmogorov-Smirnov para dos muestras* es una prueba de si dos muestras independientes se han extraído de la misma población (o de dos poblaciones con la misma distribución). La prueba bidireccional es insensible a cualquier clase de diferencia en las distribuciones de las cuales fueron extraídas las muestras: diferencias en la tendencia central, en la dispersión, en el sesgo, etc. La prueba unidireccional se utiliza para decidir si los datos en la población de donde fue extraída una de las muestras, son estocásticamente mayores que los valores de la población de donde se extrajo la otra muestra, es decir, para probar la predicción de que las puntuaciones de un grupo experimental serán mayores que los correspondientes a un grupo control.

Al igual que la prueba de Kolmogorov-Smirnov para una muestra (véase el capítulo 3), la prueba para dos muestras centra su interés en el acuerdo entre dos distribuciones acumulativas. La prueba de una muestra se interesa en el acuerdo entre la distribución de un conjunto de valores muestra y alguna distribución teórica específica. La prueba para dos muestras centra su interés en el acuerdo entre dos conjuntos de valores muestrales.

Si las dos muestras han sido extraídas de la misma distribución poblacional, entonces las distribuciones acumulativas de ambas tendrían que ser sumamente cercanas, tanto así como si las diferencias sólo mostraran desviaciones al azar de la distribución poblacional. Si las distribuciones acumulativas de las dos muestras están demasiado "lejanas" en cualquier punto, esto sugiere que las muestras provienen de poblaciones distintas. Así, una desviación suficientemente grande entre las distribuciones acumulativas de las dos muestras es evidencia para rechazar H_0.

Método

Para aplicar la prueba de Kolmogorov-Smirnov para dos muestras, debemos determinar la distribución de la frecuencia acumulada[12] para cada muestra de observaciones, utilizando los mismos intervalos para ambas distribuciones. Entonces, en cada intervalo restamos al valor anterior el valor siguiente. La prueba se centra en las mayores de las desviaciones observadas.

Definamos $S_m(X)$ como la distribución acumulativa observada para una muestra (de tamaño m), esto es, $S_m(X) = K/m$, donde K es número de datos iguales o menores que X. Definamos también $S_n(X)$ como la distribución acumulativa observada de la otra muestra, esto es, $S_n(X) = K/n$. Ahora, la prueba estadística de Kolmogorov-Smirnov es

$$D_{m,n} = \max\ [S_m(X) - S_n(X)] \tag{5.17}$$

para pruebas unidireccionales, y

$$D_{m,n} = \max |S_m(X) - S_n(X)| \tag{5.18}$$

para pruebas bidireccionales.

En cada caso, la distribución muestral de $D_{m,n}$ es conocida. Las probabilidades asociadas con la ocurrencia de valores tan grandes como una $D_{m,n}$ observada según la hipótesis nula (de que las dos muestras provienen de la misma población) han sido tabuladas. En realidad, existen *dos* distribuciones muestrales, dependiendo de si la prueba es uni o bidireccional. Nótese que para pruebas unidireccionales debemos encontrar $D_{m,n}$ *en la dirección predicha* [utilizando la ecuación (5.17)]; para pruebas bidireccionales debemos encontrar la diferencia máxima *absoluta* $D_{m,n}$ [utilizando la ecuación (5.18)] indistintamente de su dirección. Esto es porque en las pruebas undireccionales, H_1 significa que los valores poblacionales de donde fue extraída una de las muestras son estocásticamente mayores que los valores poblacionales de donde se extrajo la otra muestra, mientras que en pruebas bidireccionales H_1 simplemente significa que las dos muestras son de poblaciones diferentes.

Al utilizar la prueba de Kolmogorov-Smirnov en datos cuyo tamaño y número de intervalos son arbitrarios, es mejor emplear tantos intervalos como sea posible. Cuando se usan pocos intervalos, la información se puede perder. Esto es, la desviación vertical máxima $D_{m,n}$ de las dos distribuciones acumuladas puede oscurecerse al presentar los datos con tan pocos intervalos.

Por ejemplo, en el ejercicio que se presenta más adelante para el caso de muestras pequeñas, sólo se utilizaron ocho intervalos para simplificar su exposición. Como sucede en el ejemplo, ocho intervalos fueron suficientes para aportar una $D_{m,n}$ que nos posibilitó rechazar la H_0 en el nivel de significación predeterminado. Si hubiera sucedido que con estos ocho intervalos $D_{m,n}$ no hubiera sido lo suficiente-

[12] En esta sección utilizaremos el término *distribución de la frecuencia acumulada* para referirnos a la función de la distribución empírica, que es la *proporción* de observaciones que son menores o iguales a un valor particular. En algunos textos esta función es denominada *distribución de la frecuencia relativa acumulada.*

mente grande para permitirnos rechazar H_0, antes de que pudiéramos aceptar H_0 habría sido necesario aumentar en número de intervalos, a fin de comprobar si la desviación máxima $D_{m,n}$ había sido oscurecida por haber usado tan pocos intervalos. Es conveniente, entonces, utilizar tantos intervalos como sea posible desde el inicio, para no perder información inherente en los datos.

Muestras pequeñas

Cuando m y n son iguales o menores que 25, puede utilizarse la tabla L_I del Apéndice I para evaluar la hipótesis nula en contra de una hipótesis alterna unidireccional, y la tabla L_{II} puede emplearse para evaluar la hipótesis nula en contra de una hipótesis alterna bidireccional. El cuerpo de dichas tablas nos proporciona los valores de $mnD_{m,n}$, los cuales son significativos en varios niveles de significancia. conociendo los valores de m, n, $mnD_{m,n}$ y si la prueba es uni o bidireccional, podemos encontrar los valores críticos del estadístico. Por ejemplo, en una prueba unidireccional donde $m = 6$ y $n = 8$, rechazaríamos H_0 en el nivel de $\alpha = 0.01$ donde $mnD_{m,n} \geq 38$.

Ejemplo. Lepley comparó el aprendizaje serial de 10 estudiantes de undécimo grado con el aprendizaje serial de nueve estudiantes del séptimo grado.[13] Su hipótesis era que el efecto primario debería ser menos prominente en el aprendizaje de los estudiantes más jóvenes. El efecto primario es la tendencia a recordar más eficientemente el material de las primeras series, que el material aprendido en las últimas series. El autor probó esta hipótesis comparando el porcentaje de errores cometidos en el material aprendido por los dos grupos en la primera mitad de las series, con la predicción de que el grupo de estudiantes mayores (los del undécimo grado) cometería relativamente menos errores al repetir la primera mitad de series, que los del grupo de estudiantes de menor edad (los del séptimo grado).

i. *Hipótesis nula*. H_0: no existe diferencia en la proporción de errores cometidos al repetir la primer mitad de series aprendidas entre los estudiantes del undécimo grado y los del séptimo grado. H_1: los alumnos del undécimo grado cometen proporcionalmente menos errores al repetir la primera mitad de las series aprendidas, que los del séptimo grado.
ii. *Prueba estadística*. Puesto que se están comparando dos muestras independientes pequeñas y la hipótesis alterna es unidireccional, la prueba de Kolmogorov-Smirnov para dos muestras, unidireccional, será aplicada a los datos.
iii. *Nivel de significación*. $\alpha = 0.01$, $m = 9$ y $n = 10$.
iv. *Distribución muestral*. La tabla L_I del Apéndice I nos proporciona los valores críticos para la distribución muestral de $mnD_{m,n}$ para m y n menores que 25.
v. *Región de rechazo*. Puesto que H_1 predice la dirección de la diferencia, la región de rechazo es unidireccional H_0 será rechazada si el valor de $D_{m,n}$ (la mayor desviación en la dirección predicha) es tan grande que la probabilidad asociada con su ocurrencia cuando H_0 es verdadera, es menor o igual que $\alpha = 0.01$.
vi. *Decisión*. En la tabla 5.21 se muestra los porcentajes de error de cada estudiante cometidos durante la repetición del material aprendido secuencialmente en la primera mitad de las series. Para analizar los datos mediante la prueba de Kolmogorov-Smirnov, éstos fueron presentados en dos distribuciones de frecuencia acumu-

[13] Lepley, W. M., "Serial reactions considered as conditioned reactions", en *Psychological Monographs*, núm. 46, 1934.

lada que se muestran en la tabla 5.22. Aquí $m = 9$ estudiantes del undécimo grado y $n = 10$ estudiantes del séptimo grado. Obsérvese que la mayor discrepancia entre las dos distribuciones acumuladas es $D_{m,n} = 0.70$. Así, $mnD_{m,n} = (9)(10)(0.70) = 63$. La tabla L_I del Apéndice I revela que el valor crítico para $\alpha = 0.01$ es 61; por tanto, ya que el valor observado es mayor que el valor crítico, rechazamos la H_0 en favor de la H_1. Concluimos que los alumnos del undécimo grado cometen proporcionalmente menos errores que los alumnos del séptimo grado, al repetir la primera mitad de las series aprendidas.

Tabla 5.21. Porcentaje de errores totales en la primera mitad de la serie.

Sujetos del undécimo grado	*Sujetos del séptimo grado*
35.2	39.1
39.2	41.2
40.9	45.2
38.1	46.2
34.4	48.4
29.1	48.7
41.8	55.0
24.3	40.6
32.4	52.1
–	47.2

Tabla 5.22. Datos de la tabla 5.21 arreglados para la prueba de Kolmogorov-Smirnov.

	Porcentaje de errores totales en la primera mitad de la serie							
	24-27	28-31	32-35	36-39	40-43	44-47	48-51	52-55
$S_m(X)$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{5}{9}$	$\frac{7}{9}$	$\frac{9}{9}$	$\frac{9}{9}$	$\frac{9}{9}$	$\frac{9}{9}$
$S_n(X)$	$\frac{0}{10}$	$\frac{0}{10}$	$\frac{0}{10}$	$\frac{1}{10}$	$\frac{3}{10}$	$\frac{6}{10}$	$\frac{8}{10}$	$\frac{10}{10}$
$S_m(X) - S_n(X)$	0.111	0.222	0.556	0.678	0.700	0.400	0.200	0

Muestras grandes: prueba de dos colas

Cuando m y n son mayores que 25, la tabla L_{III} del Apéndice I puede utilizarse para la prueba de Kolmogorov-Smirnov para dos muestras. Para usar esta tabla, determine el valor de $D_{m,n}$ para los datos observados utilizando la ecuación (5.18). Entonces, compare el valor observado con el valor crítico que se obtiene de la tabla empleando como entradas m y n en la expresión proporcionada por la tabla L_{III}. Si la $D_{m,n}$ observada es igual o mayor que el calculado mediante la expresión de la tabla L_{III}, H_0 puede ser rechazada en el nivel de significación (bidireccional) asociado con dicha expresión.

Por ejemplo, supóngase $m = 55$ y $n = 50$, y que un investigador desea aplicar la prueba bidireccional con $\alpha = 0.05$. En el renglón de la tabla L_{III} para $\alpha = 0.05$, el investigador encuentra el valor de $D_{m,n}$ que debe ser igual o mayor a fin de rechazar H_0. Por medio de algunos cálculos, el investigador encuentra que $D_{m,n}$ debe tener, al menos, el valor de 0.254 a fin de rechazar H_0 para

$$1.36\sqrt{m + n/mn} = 1.36\sqrt{55 + 60/(55)(60)} = 0.254$$

Muestras grandes: pruebas de una cola

Cuando m y n son grandes, podemos realizar una prueba unidireccional utilizando

$$D_{m,n} = \max[S_m(X) - S_n(X)] \tag{5.17}$$

Nosotros probamos la hipótesis nula de que dos muestras se han extraído de la misma población, en contra de la hipótesis alterna de que los valores poblacionales de donde fue extraída una de las muestras son estocásticamente mayores que los valores poblacionales de donde fue extraída la otra muestra. Por ejemplo, podemos desear probar no simplemente si un grupo experimental es diferente de un grupo control, sino si el grupo experimental es "superior" al grupo control.

Goodman (1954) ha demostrado que

$$X^2 = 4D_{m,n}^2 \frac{mn}{m + n} \tag{5.19}$$

se aproxima a la distribución ji cuadrada con $gl = 2$, conforme el tamaño de la muestra (m y n) se incrementa. Esto es, podemos determinar la significación de una $D_{m,n}$ observada calculada con la ecuación (5.17), utilizando la ecuación (5.19) y haciendo referencia a la distribución ji cuadrada con $gl = 2$ (tabla C del Apéndice I).

Ejemplo para muestras grandes. En un estudio de correlación de la estructura de la personalidad autoritaria,[14] una hipótesis era que las personas con un nivel alto de autoritarismo mostrarían mayor tendencia a poseer estereotipos acerca de los miembros de varios grupos nacionales y étnicos, que aquellos con bajo nivel de autoritarismo. Esta hipótesis fue probada con un grupo de 98 estudiantes mujeres de un colegio seleccionadas al azar. A cada sujeto se le entregaron 20 fotografías y se les pidió "identificar" (por igualación) tantas fotografías (pocas o muchas) como desearan. Puesto que desconocían a los sujetos, todas las fotografías correspondían a mexicanos radicados en Estados Unidos (tanto candidatos a la legislatura mexicana como ganadoras de concursos de belleza) y ya que la lista de igualación

[14] Siegel, S., "Certain determinants and correlates of authoritarianism", en *Genetic and Psychological Monographs*, núm. 49, 1954, págs. 187-229.

de diferentes grupos nacionales y étnicos no incluía "mexicanos", el número de fotografías que eran "identificadas" por cualquier sujeto constituía un índice de la tendencia de los sujetos a la estereotipia.

El autoritarismo se midió por medio de la escala *F* de autoritarismo y los sujetos fueron agrupados en sujetos de "alta" y "baja" puntuación. Las puntuaciones altas fueron aquellas que se ubicaron en la mediana de la escala *F* o por encima de ella; las puntuaciones bajas fueron las que se ubicaron por abajo de la mediana. La predicción era que estos dos grupos diferirían en el número de fotografías identificadas.

i. *Hipótesis nula.* H_0: las mujeres en esta universidad, independientemente de la puntuación obtenida, tienen el mismo estereotipo de autoritarismo (en términos del número de fotografías identificadas). H_1: las mujeres cuyas puntuaciones sean altos identificarán un mayor número de fotografías que las mujeres de puntuaciones bajas.

ii. *Prueba estadística.* Puesto que las bajas y las altas puntuaciones constituyen dos grupos independientes, se seleccionó una prueba para muestras independientes. En razón de que el número de fotografías identificadas por cada sujeto no puede ser considerado más que un dato en escala ordinal de la tendencia del sujeto a la estereotipia, es deseable una prueba no paramétrica. La prueba de Kolmogorov-Smirnov para dos muestras compara dos distribuciones de frecuencia acumulada y determina si la $D_{m,n}$ observada indica si éstas fueron extraídas de las mismas poblaciones, una de las cuales es estocásticamente superior a la otra.

iii. *Nivel de significación.* $\alpha = 0.01$. Los tamaños de la muestra *m* y *n* pueden determinarse sólo después que se hayan recolectado los datos, porque los sujetos serán agrupados en virtud de si sus puntuaciones están por encima o por debajo de la mediana de la escala *F*.

iv. *Distribución muestral.* La distribución muestral de

$$X^2 = 4D^2_{m,n} \frac{mn}{m + n}$$

donde $D_{m,n}$ se calcula mediante la ecuación (5.17) es aproximada por la distribución ji cuadrada con $gl = 2$. La probabilidad asociada con un valor observado de $D_{m,n}$ cuando H_0 es verdadera, puede ser determinada por medio de la tabla C del Apéndice I.

v. *Región de rechazo.* Puesto que H_1 predice la dirección de la diferencia entre las puntuaciones *F* altas y bajas, se utilizó una prueba unidireccional. La región de rechazo consiste en todos los valores de X^2, calculados con la ecuación (5.19), los cuales son tan grandes que la probabilidad asociada con su ocurrencia cuando H_0, es igual o menor que $\alpha = 0.01$.

vi. *Decisión.* De las 98 estudiantes, 44 obtuvieron puntuaciones por debajo de la mediana de la escala *F*; así, $m = 44$. Las 54 mujeres restantes obtuvieron puntuaciones en la mediana de la escala *F* o por arriba de ella; así, $n = 54$. El número de fotografías identificadas por cada sujeto de los dos grupos se muestra en la tabla 5.23. Para aplicar la prueba de Kolmogorov-Smirnov, tenemos que presentar los datos en dos distribuciones de frecuencias acumuladas, como se observa en la tabla 5.24. Por medio de sustracciones, encontramos las diferencias entre las dos distribuciones de las muestras en varios intervalos. La mayor de estas diferencias en la dirección predicha es 0.406. Esto es,

$$D_{m,n} = \max[S_m(X) - S_n(X)]$$

$$D_{44,54} = \max[S_{44}(X) - S_{54}(X)] = 0.406$$

Con $D_{44,54} = 0.406$, calculamos el valor de X^2 mediante la ecuación (5.19):

$$X^2 = 4D^2_{m,n}\frac{mn}{m+n} \tag{5.19}$$

$$= \frac{4(.406)^2\,(44)(54)}{44+54} = 15.99$$

Tabla 5.23. Número de personas "identificadas" con puntuaciones altas y bajas de autoritarismo mediante el reconocimiento de fotografías.

Número de fotografías identificadas	*Puntuaciones bajas*	*Puntuaciones altas*
0–2	11	1
3–5	7	3
6–8	8	6
9–11	3	12
12–14	5	12
15–17	5	14
18–20	5	6

Tabla 5.24. Datos pertenecientes a la tabla 5.23 arreglados para la prueba de Kolmogorov-Smirnov.

	Número de fotografías "identificadas"						
	0–2	3–5	6–8	9–11	12–14	15–17	18–20
$S_{44}(X)$	$\frac{11}{44}$	$\frac{18}{44}$	$\frac{26}{44}$	$\frac{29}{44}$	$\frac{34}{44}$	$\frac{39}{44}$	$\frac{44}{44}$
$S_{54}(X)$	$\frac{1}{54}$	$\frac{4}{54}$	$\frac{10}{54}$	$\frac{22}{54}$	$\frac{34}{54}$	$\frac{38}{54}$	$\frac{54}{54}$
$S_{44}(X) - S_{54}(X)$	0.232	0.355	0.406	0.252	0.143	0.182	0.0

La tabla C del Apéndice I nos muestra que la probabilidad asociada con $X^2 = 15.99$ con $gl = 2$ es $p = 0.001$ (de una cola). Puesto que este valor es más pequeño que $\alpha = 0.01$, podemos rechazar H_0 en favor de H_1.[15] Concluimos que las mujeres con puntuación alta en la escala de autoritarismo tienen mayor estereotipia (identifican más fotografías) que las mujeres con puntuaciones bajas.

Es interesante notar que la aproximación a la ji cuadrada puede utilizarse además con muestras pequeñas, pero en este caso nos lleva a una prueba más conser-

[15] Con la utilización de una prueba paramétrica, Siegel tomó la misma decisión. Él encontró que $t = 3.55$, $p < 0.001$ (prueba unidireccional).

vadora. Esto es, el error en el uso de la aproximación a la ji cuadrada con muestras pequeñas está siempre en la dirección "a salvo" (Goodman, 1954, pág. 168). En otras palabras, si se rechaza la H_0 mediante el uso de la aproximación a la ji cuadrada con muestras pequeñas, podemos con toda seguridad confiar en la decisión. Así, la aproximación a la ji cuadrada puede utilizarse en muestras pequeñas, pero la prueba es conservadora y se prefiere el uso de la tabla L_I.

Resumen del procedimiento

Éstos son los pasos que hay que seguir para la aplicación de la prueba de Kolmogorov-Smirnov para dos muestras:

1. Arregle las puntuaciones de cada uno de los grupos en una distribución de frecuencias acumuladas, utilizando los mismos intervalos (o clasificaciones) para ambas distribuciones. Emplee tantos intervalos como sea posible.
2. Restando, determine la diferencia entre las dos distribuciones de frecuencias acumuladas en cada punto de la lista.
3. Determine cuál es la mayor de las diferencias, $D_{m,n}$. Para pruebas unidireccionales, $D_{m,n}$ es la diferencia mayor en la dirección predicha. Para pruebas bidireccionales, $D_{m,n}$ es la diferencia mayor en cualquier dirección.
4. El método para determinar la significación de la $D_{m,n}$ observada, depende del tamaño de la muestra y la naturaleza de la H_1:

   a) Cuando m y n son iguales o menores que 25, se utiliza la tabla L_I del Apéndice I si la prueba es unidireccional y la tabla L_{II} si la prueba es bidireccional. En cualquiera de las dos tablas se utiliza la entrada $D_{m,n}$.
   b) Para una prueba bidireccional cuando m o n es mayor que 25, se utiliza la tabla L_{III} del Apéndice I. Los valores críticos de $D_{m,n}$ para cualquier valor mayor de m o n pueden calcularse mediante las fórmulas de esa tabla.
   c) Para pruebas unidireccionales cuando m o n es mayor que 25, el valor de X^2 calculado mediante la ecuación (5.19) se distribuye como ji cuadrada con $gl = 2$. Su significación puede determinarse mediante la tabla C de Apéndice I. (La aproximación a la ji cuadrada también puede emplearse para m y n pequeñas, aunque hacer esto es muy conservador y es preferible utilizar la tabla L_I del Apéndice I.)

5. Si el valor observado es igual o mayor que el proporcionado por la tabla apropiada para un nivel particular de significación, se puede rechazar H_0 en favor de H_1.

Potencia–eficacia

Cuando se comparó con la prueba t, la prueba de Kolmogorov-Smirnov tuvo una potencia–eficacia alta (alrededor del 95 %) para muestras pequeñas. Conforme se incrementa el tamaño de la muestra, la potencia–eficacia disminuye levemente.

La prueba de Kolmogorov-Smirnov es más poderosa en todos los casos que la ji cuadrada y que la prueba de la mediana.

La evidencia parece indicar que mientras para muestras pequeñas la prueba de Kolmogorov-Smirnov es un tanto más eficaz que la prueba de Wilcoxon-Mann-Whitney, para muestras grandes ocurre lo contrario.

Referencias bibliográficas

Para detalles adicionales de la prueba de Kolmogorov-Smirnov para dos muestras, el lector puede consultar Goodman (1954), Kolmogorov (1941) y Smirnov (1948).

PRUEBA DE LAS PERMUTACIONES PARA DOS MUESTRAS INDEPENDIENTES

Función

La *prueba de las permutaciones* para dos muestras independientes es una técnica satisfactoria y poderosa para evaluar la significación de la diferencia entre medias de dos muestras independientes, cuando los tamaños de muestra m y n son pequeños. La prueba emplea los valores numéricos de las puntuaciones y, por tanto, requiere que las mediciones de la variable estudiada, al menos, se encuentren en la escala de intervalo. Con la prueba de las permutaciones podemos determinar la probabilidad exacta asociada con nuestras observaciones, en el supuesto de que H_0 es verdadera, y puede hacerse sin realizar ninguna suposición especial acerca de las distribuciones subyacentes en la población estudiada.

Racionalización y método

Considérese el caso de dos pequeñas muestras independientes, ya sea extraídas al azar de dos poblaciones u originadas por asignaciones aleatorias de dos tratamientos a un grupo de origen arbitrarios. El grupo X incluye cinco sujetos; $m = 5$. El grupo Y está constituido por cuatro sujetos; $n = 4$. Observamos las siguientes puntuaciones:

Puntuaciones del grupo X:	16	19	22	24	29
Puntuaciones del grupo Y:	0	11	12	20	

Con estas puntuaciones[16] deseamos probar la hipótesis nula de no diferencia entre las medias, en contra de la hipótesis alterna de que la media de la población de donde fue extraído el grupo X, es mayor que la media de la población de donde se extrajo el grupo Y. Esto es, H_0: $\mu_x = \mu_y$ y H_1: $\mu_x > \mu_y$.

[16] Este ejemplo se tomó de Pitman E. J. G., "Significance tests wich may be applied to samples from any population", en *Journal of the Royal Statistical Society*, núm. 4, 1937*a*, pág. 122.

Ahora, según la hipótesis nula, todas y cada uno de las $m + n$ observaciones pueden considerarse provenientes de la misma población. Vale decir, que es una cuestión de oportunidad que ciertas observaciones hayan sido catalogadas como X o Y. La asignación de las etiquetas X y Y a las puntuaciones en una manera particular puede considerarse como uno de muchos resultados igualmente posibles si H_0 es verdadera. Cuando H_0 es verdadera, las etiquetas hubieran podido ser asignadas en cualquiera de 126 maneras igualmente posibles:

$$\binom{m+n}{n} = \binom{5+4}{4} = 126$$

Cuando H_0 es verdadera, sólo una vez en 126 "experimentos" sucedería que las cinco puntuaciones mayores de $N = m + n = 9$ tuvieran la etiqueta X, y las cuatro puntuaciones menores tuvieran la etiqueta Y.

Si tal resultado se obtuviera en un experimento, realmente, de un ensayo, podríamos rechazar H_0 en el nivel de significancia $= 1/126 = 0.008$, aplicando el razonamiento de que si los dos grupos fueran realmente de la misma población, es decir, si H_0 realmente fuera verdadera, no hay una buena razón para pensar que los resultados más extremos posibles pudieran ocurrir en el único ensayo que constituye nuestro experimento. Esto es, decidiríamos que la probabilidad de que el evento observado ocurra cuando H_0 es verdadera es muy pequeña y, por tanto, rechazaríamos H_0. Ésta es parte de la lógica familiar de la estadística inferencial.

La prueba de las permutaciones especifica el número de los resultados más extremos posibles que pudieran ocurrir con $N = m + n$ puntuaciones y designa éstos como la región de rechazo. Cuando tenemos $\binom{m+n}{n}$ ocurrencias igualmente posibles según H_0, algunas de estas diferencias entre ΣX (sumatoria de las puntuaciones del grupo X) y ΣY (sumatoria de las puntuaciones del grupo Y) serán extremas. La región de rechazo está constituida por las mayores de estas diferencias.

Si α es el nivel de significación entonces la región de rechazo consiste en $\alpha\binom{m+n}{n}$ ocurrencias de los resultados más extremos posibles. Esto es, el número de resultados posibles que constituyen la región de rechazo es $\alpha\binom{m+n}{n}$ Los resultados escogidos particularmente para constituir ese número, son aquellos para los que la *diferencia* entre la media de X y la media de Y es la mayor. Éstas son las ocurrencias en donde la diferencia entre ΣX y ΣY es mayor. Ahora, si la muestra que obtuvimos se encuentra entre esos casos incluidos en la región de rechazo, rechazamos H_0 en el nivel de significación α.

En el ejemplo anterior de $N = 9$ puntuaciones, existen $\binom{5+4}{4} = 126$ posibles diferencias entre ΣX y ΣY. Si $\alpha = 0.05$, la región de rechazo consiste de $\alpha\binom{m+n}{n} = 0.05(126) = 6.3$ resultados extremos. Puesto que la hipótesis alterna es direccional, la región de rechazo consiste en los seis resultados más extremos posibles en la dirección especificada.

En virtud de que la hipótesis alterna es H_1: $\mu_x > \mu_y$, los seis resultados más extremos posibles que constituyen la región de rechazo de $\alpha = 0.05$ (prueba unidireccional) se presentan en la tabla 5.25. El tercero de los posibles resultados extremos, el que se encuentra marcado con una cruz, es la muestra que obtuvimos. Puesto que nuestro conjunto de puntuaciones observados se encuentra en la región de rechazo, podemos rechazar H_0 en el nivel de significación $\alpha = 0.05$. La probabilidad exacta (unidireccional) de la ocurrencia de las puntuaciones observadas de un conjunto de valores más extremos cuando H_0 es verdadera, es $p = 3/126 = 0.024$.

Tabla 5.25. Los seis resultados más extremos posibles en la dirección predicha.*

Puntuaciones posibles para cinco casos X					*Puntuaciones posibles para cuatro casos Y*				$\Sigma X - \Sigma Y$
29	24	22	20	19	16	12	11	0	114 − 39 = 75
29	24	22	20	16	19	12	11	0	111 − 42 = 69
29	24	22	19	16	20	12	11	0	110 − 43 = 67†
29	24	20	19	16	22	12	11	0	108 − 45 = 63
29	24	22	20	12	19	16	11	0	107 − 46 = 61
29	22	20	19	16	24	12	11	0	106 − 47 = 59

* Esto constituye la región de rechazo para la prueba de permutaciones cuando $\alpha = 0.05$.
† La muestra observada.

Ahora bien, si la hipótesis alterna no hubiera predicho la dirección de la diferencia, por supuesto que una prueba bidireccional habría sido más apropiada. En este caso, los seis conjuntos de resultados posibles en la región de rechazo consistirían en los tres resultados más extremos posibles en una dirección y en otros tres resultados más extremos posibles en la dirección contraria. Así, incluiríamos los seis resultados posibles donde la diferencia entre ΣX y ΣY fueran mayores en cuanto a su valor absoluto. Con propósitos ilustrativos, los seis resultados más extremos posibles para una prueba bidireccional con $\alpha = 0.05$ para las nueve puntuaciones presentadas anteriormente, se muestran en la tabla 5.26. Con nuestras puntuaciones observadas pudimos haber rechazado H_0 en favor de la hipótesis alterna H_1: $\mu_x \neq \mu_y$, porque la muestra obtenida (marcada con una cruz en la tabla 5.26) es uno de los seis resultados más extremos posibles en cualquier dirección. La probabilidad exacta asociada con la ocurrencia cuando H_0 es verdadera de un conjunto tan extremo (o más extremo) como el observado, es $p = 6/126 = 0.048$.

Muestras grandes

Cuando m y n son grandes, el cálculo necesario para la prueba de las permutaciones puede llegar a ser extremadamente tedioso. Se puede desarrollar un programa sencillo para computadora, que calcule los resultados posibles. Sin embargo, como $N = m + n$ llegar a ser grande, el cálculo también consume un tiempo de máquina considerable. Los cálculos se evitan porque se puede demostrar que, si m y n son grandes y la curtosis de las muestras combinadas es pequeña, entonces

Tabla 5.26. Los seis resultados más extremos posibles en cualquier dirección.*

Puntuaciones posibles para cinco casos X					Puntuaciones posibles para cuatro casos Y				$\lvert \Sigma X - \Sigma Y \rvert$
29	24	22	20	19	16	12	11	0	$\lvert 114 - 39 \rvert = 75$
29	24	22	20	16	19	12	11	0	$\lvert 111 - 42 \rvert = 69$
29	24	22	19	16	20	12	11	0	$\lvert 110 - 43 \rvert = 67$†
22	16	12	11	0	29	24	20	19	$\lvert 61 - 92 \rvert = 31$
20	16	12	11	0	29	24	22	19	$\lvert 59 - 94 \rvert = 35$
19	16	12	11	0	29	24	22	20	$\lvert 58 - 95 \rvert = 37$

* Esto constituye la región de rechazo bidireccional para la prueba de permutaciones cuando $\alpha = 0.05$.
† La muestra observada.

la distribución de los $\binom{m+n}{n}$ posibles resultados se aproxima cercanamente a la distribución t. Esto es, si se satisfacen las condiciones anteriores, la prueba t para diferencias entre dos medias puede utilizarse para probar la hipótesis

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\Sigma(X_i - \overline{X})^2/(m - 1) + \Sigma(Y_i - \overline{Y})^2/(n - 1)}} \qquad (5.20)$$

que tiene aproximadamente una distribución t de Student. La expresión para los grados de libertad es complicada,[17] pero una prueba conservadora tendría $gl = m + n - 2$. Por tanto, la probabilidad asociada con la ocurrencia, cuando H_0 es verdadera, de cualquier valor extremo como un valor de t observado, puede determinarse mediante la tabla B del Apéndice I.

El lector notará que aunque la ecuación (5.20) es una forma de la prueba t, la prueba no se utiliza en este caso como una prueba paramétrica, en virtud de que la prueba se fundamenta en el teorema del límite central para la distribución muestral de las medias que tienen una distribución asintóticamente normal cuando las observaciones individuales no la tienen. Sin embargo, su uso supone no sólo que las condiciones mencionadas anteriormente se cubrieron, sino que además las puntuaciones representan mediciones que se encuentran, el menos, en escala de intervalo.

Cuando m y n son grandes, una prueba alternativa a la prueba de las permutaciones es la de Wilcoxon, que puede considerarse una prueba de permutaciones aplicada a rangos de las observaciones y así, constituye una buena aproximación a la prueba de las permutaciones. Se puede demostrar que existen situaciones para las cuales la prueba de Wilcoxon es más poderosa que la prueba t y, por tanto, es una mejor opción.

[17] En el caso de poblaciones con varianzas desiguales, los grados de libertad para la prueba t está en función tanto del tamaño como de las varianzas de las muestras. El valor correcto para los grados de libertad se encuentra entre el tamaño de la muestra menor *menos* 1 y $m + n - 2$. Así, el utilizar $gl = m + n - 2$ da como resultado una prueba conservadora ya que si H_0 es rechazada con el máximo posible de grados de libertad, también será rechazada con menos grados de libertad.

Resumen del procedimiento

Éstos son los pasos que hay que seguir en la aplicación de la prueba de las permutaciones para dos muestras independientes:

1. Determine el número de posibles resultados que se encuentran en la región de rechazo:

$$\alpha \binom{m+n}{n}$$

2. Especifique el número de los resultados más extremos posibles que pertenecen a la región de rechazo. Estos extremos son aquellos que alcanzan la diferencia mayor entre ΣX y ΣY. Para pruebas unidireccionales, todos éstos se encuentran en la dirección predicha. Para pruebas bidireccionales, la mitad del número de resultados más extremos posibles se encuentran en una dirección y la otra mitad se encuentran en la dirección contraria.
3. Si las puntuaciones observadas se encuentran en uno de los resultados comprendidos por la región de rechazo, se rechaza H_0 en ese nivel de significación.

Para muestras que son tan grandes que la enumeración de los posibles resultados en la región de rechazo es demasiado tediosa, puede utilizarse un programa para computadora (véase el Apéndice II) o la ecuación (5.20) como una aproximación si los datos satisfacen las condiciones para su uso. Existen alternativas que no necesitan cubrir tales condiciones, y así, puede resultar más útil utilizar la prueba de Wilcoxon o la prueba poderosa por orden de rangos.

Potencia–eficacia

Debido a que utiliza toda la información de las muestras, la prueba de las permutaciones para dos muestras independientes tiene una potencia–eficacia (en el sentido en que se definió en el capítulo 2) del 100 %.

Referencias bibliográficas

El lector encontrará referencias adicionales acerca de la prueba de las permutaciones para dos muestras independientes, en Moses (1952*a*), Pitman (1937*a*, 1937*b*, 1937*c*) y Lehmann (1975).

PRUEBA DE SIEGEL-TUKEY PARA DIFERENCIAS EN LA ESCALA

Función y racionalización

En las ciencias de la conducta, a veces esperamos que una condición experimental cause que algunos sujetos muestren conductas extremas en una dirección, mientras que esa misma condición ocasiona que otros sujetos muestren conductas extremas en la dirección opuesta. Podemos esperar que la depresión económica y la inestabilidad política ocasionen que algunas personas se tornen más reaccionarias, mientras que otras se vuelvan a la "ala izquierda" en cuanto opiniones políticas. O podemos esperar que la agitación ambiental genere excitación extrema en personas mentalmente enfermas, mientras en otras genere el desinterés total. Dentro de la investigación en psicología utilizando una aproximación perceptual a la personalidad, existen razones teóricas para predecir que la "defensa perceptual" puede manifestarse ya sea en una respuesta perceptual "vigilante" extremadamente rápida o en una respuesta perceptual "represiva" extremadamente lenta.

La prueba de Siegel-Tukey está diseñada específicamente para emplearse con datos que se encuentran, al menos, en escala ordinal. Puede utilizarse cuando se espera que uno de los grupos tenga una variabilidad mayor que otro. Pero las medianas (o medias) de ambos grupos son las mismas (o conocidas). En estudios de defensa perceptual, por ejemplo, esperamos que los sujetos control muestren respuestas medias o normales, mientras que esperamos que los sujetos esperimentales muestren respuestas vigilantes o represivas, recibiendo puntuaciones bajas o altas comparadas con aquellas de los sujetos del grupo control.

En tales estudios, las pruebas estadísticas dedicadas a las diferencias en la tendencia central ocultan, más que revelar, diferencias entre los grupos; nos conducen a la aceptación de la hipótesis nula cuando debería ser rechazada, porque cuando algunos sujetos experimentales muestran respuestas vigilantes y así obtienen puntuaciones de latencia bajos mientras otros muestran respuestas represivas y así obtienen puntuaciones de latencia altos, el promedio de las puntuaciones del grupo experimental puede ser muy cercano al promedio del grupo control (donde cada miembro pudo obtener puntuaciones medianas).

La prueba de Siegel-Tukey está diseñada especialmente para la clase de situaciones descritas. Esta prueba únicamente tiene valor cuando existen razones *a priori* para creer que la condición experimental nos conducirá a puntuaciones extremas en cualquier dirección manteniendo la misma mediana. Así, si σ^2 es la varianza de las variables y si X es el grupo experimental y Y el grupo control, podemos escribir la hipótesis como

$$H_0\colon \sigma_x^2 = \sigma_y^2$$

y

$$H_1\colon \sigma_x^2 > \sigma_y^2$$

La prueba de Siegel-Tukey se centra en el *rango* o extensión de un grupo comparado con otro, y a menudo se denomina *prueba para escalas de diferencias entre dos grupos.* Esto es, si existen m casos en el grupo X y n en el grupo Y, y las $N = m + n$ puntuaciones se ordenan por tamaños de manera ascendente, y si la hipótesis nula (las puntuaciones de X y las puntuaciones de Y provienen de la misma población) es verdadera, entonces deberíamos esperar que las X y Y estuvieran bien mezcladas en la serie ordenada. Esperaríamos según H_0, que algunas puntuaciones extremadamente altas fueran X y algunas fueran Y, y que el rango medio de las puntuaciones incluyera una mezcla de X y Y. Sin embargo, si la hipótesis alterna es verdadera (las puntuaciones X representan las respuestas extremas), entonces esperaríamos que una considerable proporción de puntuaciones X fueran mayores o menores, mientras que relativamente pocas puntuaciones X se encontrarían en la parte media de la serie combinada. Esto es, las puntuaciones Y serían relativamente más compactas, y su rango o variabilidad resultaría relativamente más pequeña que la de las puntuaciones X. La prueba de Siegel-Tukey determina si lo relativamente compacto de las puntuaciones Y respecto de todos los $N = m + n$ puntuaciones nos conduce a rechazar la hipótesis nula de que tanto el grupo X como el grupo Y provienen de la misma distribución.

Método

Para calcular la prueba de Siegel-Tukey, se combinan las puntuaciones de los grupos X y Y y se arreglan en una sola serie ordenada, cuidando de identificar cada puntuación correspondiente a cada grupo. Se asigna un rango a cada puntuación ordenada en la secuencia; *cada rango se asigna alternando los extremos de la secuencia ordenada*. Así, tenemos que en la prueba de Siegel-Tukey, los rangos proceden de las puntuaciones (atípicas) extremas a las centrales (típicas). El lector debe notar que de acuerdo con la lógica aceptada de la prueba, este procedimiento separa los grupos de "puntuaciones extremas" de los grupos de "puntuaciones normales". Por ejemplo, supongamos que se observó un conjunto de puntuaciones X y Y y que $m = 7$ y $n = 6$, y que se ordenaron del menor al mayor:

Grupo:	X	X	Y	X	Y	X	Y	Y	Y	X	Y	X	X
Rango:	1	4	5	8	9	12	13	11	10	7	6	3	2

Entonces calculamos la sumatoria de rangos para los grupos X y Y.

$$W_x = 1 + 4 + 8 + 12 + 7 + 3 + 2 = 37$$

y

$$W_y = 5 + 9 + 13 + 11 + 10 + 6 = 54$$

Si la hipótesis nula de que la dispersión de los dos grupos es la misma resulta verdadera, esperaríamos que la sumatoria de rangos (ajustados para el tamaño de la muestra) fuera aproximadamente la misma. Sin embargo, si la hipótesis alterna de que las puntuaciones X son más variables que las puntuaciones Y es verdadera, esperaríamos una W_x menor y una W_y mayor, reflejando esto que los rangos más pequeños fueron asignados a los extremos de la serie ordenada. El lector notará que ésta es precisamente la lógica de la prueba de Wilcoxon, estudiada en este capítulo. Por tanto, para probar la hipótesis nula, determinamos la probabilidad asociada con la observación de la suma de los rangos tan grande o mayor que la W_y obtenida de nuestra muestra, utilizando la tabla J del Apéndice I. (Alternativamente, podemos calcular la probabilidad de observar una sumatoria de rangos tan pequeña o menor que la W_x obtenida.) Para estos datos, la probabilidad de observar una W_y tan grande o mayor que 54 es $p = 0.051$. Por tanto, si $\alpha = 0.05$, podemos rechazar la hipótesis de que la dispersión o varianza es la misma para los dos grupos.

Ejemplo. En un estudio sobre la discriminación de la duración, Eisler[18] examinó varias formas de funciones exponenciales relacionando duraciones objetivas y subjetivas. Estas funciones se utilizaron para probar el modelo paralelo-reloj para la discriminación de duraciones. Se utilizaron dos grupos de sujetos. La tarea de un grupo abarcaba (entre otras cosas) la estimación de duraciones cortas y el otro grupo, duraciones largas. Se argumentó que aunque ciertos parámetros pueden variar como una función de la condición, el exponente de la función exponencial no sería una función de la duración. Sin embargo, algunos investigadores han argumentado que las diferencias individuales podrían variar como en función de la duración y habría mayor variabilidad en los exponentes asociados con las duraciones más grandes. Hubo ocho sujetos en el grupo de las duraciones grandes (0.9 a 1.2 seg.) y nueve sujetos en el grupo de las duraciones cortas (0.07 a 0.16 seg.).

La hipótesis fue que, para los modelos probados, el exponente de la función exponencial no sería una función de la duración y la variabilidad del exponente no se modificaría.

i. *Hipótesis nula.* H_0: la variabilidad del exponente estimado de la función exponencial en los juicios de duración no se afectará por las duraciones utilizadas. H_1: la variabilidad del exponente de la función exponencial variará de acuerdo con las duraciones utilizadas.
ii. *Prueba estadística.* Puesto que la hipótesis se interesa en la escala de distribuciones de parámetros y se supone que las medianas de las distribuciones son las mismas, es apropiada la prueba de Siegel-Tukey.
iii. *Nivel de significación.* $\alpha = 0.05$, $m = 8$ y $n = 9$.
iv. *Distribución muestral.* La distribución muestral de la prueba de Siegel-Tukey es la misma que la de la prueba de Wilcoxon. Así, la lógica de prueba puede ser aplicada a los rangos asignados ordenados por este procedimiento.
v. *Región de rechazo.* En virtud de que la hipótesis alterna no plantea la dirección de las diferencias, se utiliza una prueba unidireccional. La región de rechazo consiste en todos los valores de las sumatorias de los rangos tan grandes (o mayores) que las observadas en los datos.
vi. *Decisión.* Los valores de los exponentes estimados se presentan en la tabla 5.27, en la cual, los datos de los dos grupos se ordenaron por tamaños de manera ascendente, se les asignó su rango correspondiente y posteriormente, dichos rangos fueron ajustados (como se puede observar en la última columna de la tabla). En la misma tabla se muestran los valores de $W_x = 72$ y de $W_y = 81$. La tabla J del Apéndice I muestra que (en pruebas unidireccionales) la probabilidad de observar un valor de W_x tan pequeño (o más pequeño) como el valor observado de 72, es $p = 0.519$. Por tanto, no podemos rechazar la hipótesis de que la distribución de los exponentes de la función exponencial es la misma en las dos condiciones.

Asignación del orden de los rangos

Aunque asignamos los rangos de orden desde "la parte exterior" de la distribución a la mediana, existen algunos procedimientos alternativos. Para ilustrar el mé-

[18] Eisler, H., "Applicability of the parallel-clock model duration discrimination", en *Perception and Psychophysics*, núm. 29, 1981, págs. 225-233.

Tabla 5.27. Valores de los exponentes para el modelo reloj-paralelo de duración de la discriminación.

Grupo X: duración larga ($m = 8$)
0.62 1.10 0.82 0.68 0.78 0.75 0.76 0.47

Grupo Y: duración corta ($n = 9$)
0.89 0.70 0.80 0.74 0.85 0.67 0.69 0.89 0.77

Datos combinados				
Puntuación	*Grupo*		*Rango*	*Rango ajustado*
0.47	X		1	1
0.62	X		4	4
0.67		Y	5	5
0.68	X		8	8
0.69		Y	9	9
0.70		Y	12	12
0.74		Y	13	13
0.75	X		16	16
0.76	X		17	17
0.77		Y	15	15
0.78	X		14	14
0.80		Y	11	11
0.82	X		10	10
0.85		Y	7	7
0.89		Y	6	4.5
0.89		Y	3	4.5
1.10	X		2	2

$W_x = 72$, $W_y = 81$.

todo, consideremos una situación en la cual existen siete puntuaciones que ya están ordenadas. Los rangos pueden estar asignados así:

$$1\ 4\ 5\ 7\ 6\ 3\ 2$$

Al asignar los rangos alternando los extremos, no podemos tener los mismos rangos en cada lado de la mediana. Sin embargo, el método empleado tiene la ventaja de que la *sumatoria de rangos* para cualquier par adyacente de puntuaciones en un lado de la mediana es igual a la suma de los rangos de las dos puntuaciones que se encuentran a la misma distancia de la mediana en el lado contrario. Así, en el ejemplo anterior, $1 + 4 = 3 + 2$, $4 + 5 = 6 + 3$, etc. Si en lugar de asignar rangos de orden de la puntuación más pequeña a la puntuación mayor, asignamos los rangos de orden de la puntuación mayor a la puntuación más pequeño, entonces los valores de W_x y W_y serán diferentes. Aunque los valores resultantes de W_x y W_y no serían muy diferentes para muestras de tamaño moderado, el investigador debería decidir cuál ordenamiento utilizar *antes* de examinar los datos. Para los datos de la tabla 5.27, si ordenáramos los datos por rangos comenzando por la puntuación mayor, obtendríamos una $W_x = 76$ y una $W_y = 77$, para las cuales su probabilidad (unidireccional) asociada en 0.336. El cambio no afectaría nuestra conclusión, tomada con base en el procedimiento contrario de asignación de rangos.

Debe apuntarse que algunos investigadores asignan el orden de los rangos de la parte interna hacia la parte externa. Esto es, para los datos anteriores, el orden de los rangos pudieran ser los siguientes:

$$7 \quad 4 \quad 3 \quad 1 \quad 2 \quad 5 \quad 6$$

o

$$6 \quad 5 \quad 2 \quad 1 \quad 3 \quad 4 \quad 7$$

Puede utilizarse cualquier método. Sin embargo, en el último método esperaríamos que los rangos extremos fueran los mayores que los rangos de en medio, por lo cual debe ajustarse la prueba de acuerdo con esto.

Medianas conocidas

Si se conocen las medianas de las dos distribuciones, la prueba puede aplicarse restando la mediana de las puntuaciones de cada grupo, antes de ordenar por rangos los datos combinados. El efecto de esto es ajustar las medianas iguales, de tal manera que la prueba pueda aplicarse apropiadamente. Sin embargo, esta corrección es adecuada cuando las medianas de las *poblaciones son conocidas*, y no es apropiado utilizar las medianas de las muestras para ajustar distribuciones similares en tendencia central.

Resumen del procedimiento

Éstos son los pasos que hay que seguir en la aplicación de la prueba de Siegel-Tukey:

1. Determine el orden en que serán asignados los rangos.
2. Después de que hayan sido recolectadas las puntuaciones, ordénelas por rangos en una misma serie, identificando a qué grupo pertenece cada dato. Si las medianas de la población son conocidas y desiguales, reste las medianas a cada una de las puntuaciones (por grupo) antes de ordenarlas en la misma serie. Asigne los rangos de orden a las puntuaciones en la secuencia, de manera que se alternen los extremos de la serie (como se describió anteriormente).
3. Determine los valores de W_x y W_y.
4. Para muestras pequeñas, determine la significación de la W_x observada utilizando la tabla J del Apéndice I. Si el tamaño de la muestra es grande, determine la significación de W_x usando la ecuación (5.11) [o la ecuación (5.13) si hay rangos empatados].

5. Si la probabilidad determinada en el cuarto paso es menor o igual que α, rechace H_0.

Potencia

La potencia de la prueba de Siegel-Tukey es relativamente baja. Cuando se utiliza en datos que tienen una distribución normal, la potencia es 0.61 para N pequeña. Se debe destacar que a menos que se satisfaga el supuesto de medianas iguales, la prueba de Siegel-Tukey no puede ser interpretada ya que un valor significativo puede obtenerse simplemente como resultado de una diferencia en las medianas.

Referencias bibliográficas

El lector encontrará buenos análisis de esta prueba en Siegel y Tukey (1960, 1961), Moses (1963) y Lehmann (1975).

PRUEBA DE RANGOS DE MOSES PARA DIFERENCIAS EN LA ESCALA

Función y racionalización

Como se afirmó en la sección anterior, en las ciencias de la conducta sociales a menudo se tiene interés en evaluar las diferencias en la dispersión de dos grupos. Aunque los investigadores deseen saber acerca de las diferencias en la tendencia central, las diferencias en la escala pueden tener importancia teórica y valor práctico. Por ejemplo, determinar que un grupo es más homogéneo que otro podría ser valioso en cuanto a desarrollar materiales instruccionales especiales para ese grupo. Las diferencias en la heterogeneidad de los grupos podría resultar interesante para el psicólogo social que estudia los factores implicados en el ajuste a nuevos ambientes. La de Siegel-Tukey es una prueba que tiene éxito al comparar diferencias en la escala o variabilidad. Sin embargo, el uso de la prueba requiere que las medianas de los dos grupos sean las mismas o conocidas. Esto es, la prueba de Siegel-Tukey supone que las dos medianas son las mismas o, si son conocidas, que éstas pueden ser restadas de cada puntuación para "ajustar" las medianas iguales. Como muchos lectores sospecharán, hay muchas situaciones en las cuales dichos supuestos no pueden ser justificados. La prueba de rangos de Moses es válida en casos en los cuales las medianas son desconocidas o bien, no puede suponerse que sean iguales. Contraria a la prueba de Siegel-Tukey, la de rangos de Moses supone que las observaciones corresponden a mediciones en, al menos, escala de intervalo.

Las hipótesis puede escribirse como

$$H_0\colon \sigma_x^2 = \sigma_y^2$$

y

$$H_1\colon \sigma_x^2 \neq \sigma_y^2$$

para hipótesis unidireccionales

$$H_1\colon\ \sigma_x^2 > \sigma_y^2$$

si deseamos probar una hipótesis alterna unidireccional de que la variable X es mayor que la variable Y. Por supuesto, la hipótesis alterna podría ser

$$H_1\colon\ \sigma_x^2 < \sigma_y^2$$

Método

Para calcular la prueba de rangos de Moses es necesario dividir las observaciones de los dos grupos en subconjuntos del mismo tamaño. Cada subconjunto debe contener al menos dos observaciones. Si la división es tal que hay observaciones que no pueden colocarse en ningún subconjunto, éstas se descartan del análisis. Es importante dividir los datos en subconjuntos al azar; lo anterior se realiza mejor con una tabla de números aleatorios. Por ejemplo, si hay $m = 25$ observaciones en el conjunto X y $n = 21$ observaciones en el conjunto Y, entonces lo recomendable es utilizar subconjuntos de cinco observaciones que resultarían en $m' = 5$ subconjuntos del grupo X y $n' = 4$ subconjuntos del grupo Y; y se descartaría una observación del grupo Y. O los datos podrían dividirse en subconjuntos de cuatro observaciones, con $m' = 6$ subconjuntos de X y $n' = 5$ subconjuntos de Y, y se descartaría una observación de cada grupo. Por supuesto, pueden emplearse subconjuntos de otros tamaños.

Para cada subconjunto, se calcula la sumatoria de las diferencias al cuadrado de cada dato respecto a la media de cada subconjunto (a cada dato se le resta la media del subconjunto al que pertenece, la diferencia se eleva al cuadrado y todas estas diferencias elevadas al cuadrado se suman para cada subconjunto). El procedimiento es sencillo, pero requiere muchas operaciones. Para denominarlos, utilizaremos un doble subíndice para identificar los subconjuntos individualmente. Primero, k es el número de observación en cada subconjunto, m' es el número de subconjunto de X y n' es el número de subconjunto de Y. Entonces, los datos para el j-ésimo subconjunto de X puede enumerarse como

$$X_{j1}, X_{j2}, \ldots, X_{jk} \qquad j = 1, 2, \ldots, m'$$

y los datos del subconjunto de Y pueden enumerarse como

$$Y_{j1}, Y_{j2}, \ldots, Y_{jk} \qquad j = 1, 2, \ldots, n'$$

Para los subconjuntos de X calculamos un índice de dispersión $D(X_j)$:

$$D(X_j) = \sum_{i=1}^{k} (X_{ji} - \overline{X}_j)^2 \qquad j = 1, 2, \ldots, m' \tag{5.21}$$

donde

$$\overline{X}_j = \frac{\sum_{i=1}^{k} X_{ji}}{k}$$

es la media de las observaciones es el j-ésimo subconjunto de X. De manera similar, para cada uno de los subconjuntos de Y calculamos el índice de dispersión $D(Y_j)$:

$$D(Y_j) = \sum_{i=1}^{k} (Y_{ji} - \overline{Y}_j)^2 \qquad j = 1, 2, \ldots, n' \tag{5.22}$$

donde
$$\overline{Y}_j = \frac{\sum_{i=1}^{k} Y_{ji}}{k}$$

es la media de las observaciones en el j-ésimo subconjunto de Y.

Ahora, si la hipótesis nula de igual variabilidad para los grupos X y Y es verdadera, esperaríamos que los valores de $D(X_j)$ y $D(Y_j)$ estuvieran tan bien "mezclados" en las medidas de dispersión para los subconjuntos, que ambas serían muy similares. Sin embargo, si la hipótesis alterna es verdadera, entonces esperaríamos que los valores de $D(X_j)$ tendieran generalmente a ser más pequeños que las $D(Y_j)$, si los datos de X tienen menos variabilidad que los datos de Y [o los valores de $D(X_j)$ tenderían a ser mayores que las $D(Y_j)$ si los datos de X tienen mayor variabilidad que los datos de Y]. Para probar la hipótesis de igual dispersión, aplicamos la prueba de Wilcoxon a los índices de dispersión calculados para cada uno de los subconjuntos. En la aplicación de esta prueba, los tamaños de la muestra son m' y n'. Esto es, una vez que calculamos las D, se puede aplicar la lógica de la prueba de Wilcoxon. Si rechazamos la hipótesis de D iguales, entonces debemos rechazar la hipótesis de que las variables X y Y tienen la misma dispersión.

Ejemplo. Algunas investigaciones han encontrado que los receptores de la insulina pueden variar como una función de la variación en el metabolismo de la glucosa inducida fisiológica o farmacológicamente. Sin embargo, se desconoce si cambios en los receptores de insulina inducen cambios en el metabolismo de la glucosa.

En un esfuerzo para examinar esta cuestión, algunos investigadores analizaron situaciones en las cuales el metabolismo de la glucosa se pudo medir en función de la modificación de los receptores de insulina.[19] Las personas que padecen de distrofia muscular de Duchenne (DMD) presentan marcados defectos en el nivel de membranas, los cuales se esperaría que resultaran en la modificación de los receptores de la insulina. Sin embargo, tales personas generalmente tienen un metabolismo de los carbohidratos normal. Las investigaciones no demuestran defectos en los receptores de insulina en ausencia de cambios manifiestos en el metabolismo de los carbohidratos.

Para el estudio se seleccionó un grupo de 17 sujetos normales y un grupo de 17 sujetos que padecían DMD. A todos los sujetos se les colocó bajo el mismo régimen dietético. Como parte del estudio, en cada sujeto se midió el "aglutinamiento" de los monocitos por parte de la insulina. Los resultados se presentan en la tabla 5.28. Aunque se esperaban diferencias en el "aglutinamiento", la variabilidad del mismo debería ser diferente en ambos grupos. Esto es, en el grupo de personas normales se esperaba un aglutinamiento más homogéneo que el grupo de sujetos con DMD, y se esperaba una variabilidad de rango más amplio.

[19] De Pirro, R., Lauro, R., Testa, I, Ferreti, G., De Martinis, C. y Dellantonio, R., "Decreased insulin receptors but normal glucose metabolism in Duchenne muscular dystrophy", en *Science*, núm 216, 1982, págs. 311-313.

Tabla 5.28. "Aglutinamiento" de monocitos por insulina.

Sujetos normales	*Sujetos con DMD**
2.50	2.10
2.48	2.00
2.45	1.80
2.32	1.70
2.32	1.60
2.31	1.55
2.28	1.40
2.27	1.40
2.25	1.30
2.22	1.25
2.22	1.10
2.18	1.03
2.16	0.98
2.12	0.86
2.12	0.85
2.05	0.70
1.90	0.65

* Distrofia muscular de Duchenne.

i. *Hipótesis nula.* H_0: los sujetos normales y que padecen DMD muestran igual variación en el aglutinamiento por insulina. H_1: los sujetos con DMD muestran una mayor variabilidad en el aglutinamiento por insulina, que los sujetos normales.

ii. *Prueba estadística.* Puesto que la hipótesis se interesa en las distribuciones de escalas de parámetros, y que se supone que las medianas de las distribuciones no son iguales, y ya que se supone que las distribuciones adyacentes no se distribuyen normalmente, la prueba de rangos de Moses es la apropiada.

iii. *Nivel de significación.* $\alpha = 0.05$, $m = 17$ y $n = 17$.

iv. *Distribución muestral.* La distribución muestral de los estadísticos asociados con la prueba de rangos de Moses es la misma que la de la prueba de Wilcoxon. Por tanto, la lógica de la prueba de Wilcoxon puede aplicarse a los estadísticos derivados.

v. *Región de rechazo.* En virtud de que la hipótesis alterna especifica la dirección de la diferencia, se utiliza una prueba unidireccional. La región de rechazo consiste en todos los valores de las sumatorias de rangos tan grandes o mayores que el valor observado.

vi. *Decisión.* Se optó por utilizar subconjuntos de tamaño $k = 4$. Esto se hizo así por que sólo un dato sería descartado en cada grupo. (Si se hubiera utilizado $k = 3$ o 5, se habrían descartado dos datos de cada grupo.) Utilizando una tabla de números aleatorios, se eliminaron las observaciones número 16 y 15 del grupo de sujetos normales y con DMD, respectivamente. Recurriendo nuevamente a una tabla de números aleatorios, cada grupo se dividió en cuatro subconjuntos. En la tabla 5.29 se muestran las listas de asignaciones a cada subconjunto. Por medio de la ecuación (5.21) se calcularon los índices $D(X_j)$ para los subconjuntos de personas normales, y utilizando la ecuación (5.22) se calcularon los índices $D(Y_j)$ para los subconjuntos de sujetos con DMD. Estos valores se presentan en la tabla 5.29. Lo siguiente es la aplicación de la prueba de Wilcoxon a los ocho índices de disper-

Tabla 5.29. Datos de la tabla 5.28 arreglados en subconjuntos para calcular la prueba de rangos de Moses.

Datos de los sujetos normales arreglados en subconjuntos

Conjunto		*Puntuaciones*			$D(X_i)$
1	2.18	2.31	1.90	2.45	0.1646
2	2.28	2.25	2.12	2.22	0.0145
3	2.22	2.48	2.50	2.30	0.0563
4	2.16	2.12	2.27	2.32	0.0261

Datos de los sujetos DMD arreglados en subconjuntos

Conjunto		*Puntuaciones*			$D(Y_i)$
1	1.55	1.25	1.03	0.70	0.3857
2	2.10	0.98	1.10	0.65	1.1706
3	1.30	2.00	1.40	1.80	0.3275
4	1.40	1.60	0.86	1.70	0.4212

sión. Esto es, $m' = n' = 4$. Para aplicar la prueba, las D deben ser ordenadas de menor a mayor:

Puntuación D:	0.0145	0.0261	0.0563	0.1646	0.3275	0.3857	0.4212	1.1706
Rango:	1	2	3	4	5	6	7	8
Grupo:	X	X	X	X	Y	Y	Y	Y

Con los datos anteriores calculamos la suma de rangos:

$$W_x = 1 + 2 + 3 + 4 = 10$$

y

$$W_y = 5 + 6 + 7 + 8 = 26$$

Puesto que la hipótesis alterna es que los sujetos con DMD (grupo Y) deberían mostrar una mayor variabilidad, la hipótesis alterna es:

$$H_1: \sigma_x^2 < \sigma_y^2$$

Por tanto, debemos rechazar H_0 si la probabilidad asociada con una W_x tan pequeña como 10 (o alternativamente, la probabilidad asociada con una W_y tan grande como 26) es menor que 0.05. En la tabla J del Apéndice I, encontramos que la probabilidad asociada es 0.014, por tanto, debemos rechazar H_0 y concluir que la variabilidad en los sujetos con DMD es mayor que la variabilidad en los sujetos normales.

Empates

Aunque en la aplicación de la prueba de rangos de Moses no hay problema con los empates en los datos originales, sí hay que realizar un ajuste si existen empates en las $D(X_j)$ y en las $D(Y_j)$. Se debe utilizar la corrección usual de la prueba de Wilcoxon (véase la sección correspondiente).

Muestras grandes

Cuando el tamaño de la muestra es grande, se debe emplear la aproximación para muestra grande de la prueba de Wilcoxon (véase la sección correspondiente).

Resumen del procedimiento

Éstos son los pasos que hay que seguir para la aplicación de la prueba de rangos de Moses:

1. Dependiendo del tamaño de la muestra en cada grupo, divida los datos de cada grupo en muestras al azar de tamaño $k \geq 2$, con la ayuda de una tabla de números aleatorios. Descarte cualquier dato extra. Los tamaños de los subconjuntos deben escogerse de tal manera que el número de datos que se descarten sea mínimo, m' será el número de subconjuntos de X, y n' será el número de subconjuntos de Y.
2. Utilice las ecuaciones (5.21) y (5.22) para calcular los índices de dispersión $D(X_j)$ y $D(Y_j)$ para cada subconjunto.
3. Arregle las D en orden y asígneles rangos. Calcule las sumatorias de rangos W_x y W_y.
4. Utilice los tamaños de muestra m' y n' de los subconjuntos y recurra a la tabla J del Apéndice I para determinar la significación de W_x. Si la probabilidad asociada es menor que α, rechace H_0. Si los tamaños de muestra m' y n' son grandes, utilice la aproximación para muestras grandes [ecuación (5.11) o (5.13)].

Potencia-eficacia

La eficacia de la prueba de rangos de Moses es una función del tamaño de los subconjuntos utilizados. La eficacia se incrementa conforme aumenta el tamaño de la muestra. Si la distribución subyacente es normal, la eficacia es 0.61 para subconjuntos de tamaño 4, 0.80 para subconjuntos de tamaño 9 y es asintóticamente 0.95 (cuando los subconjuntos se vuelven infinitamente grandes). Por supuesto, existe una limitante, ya que cuando incremente el tamaño de la muestra decrementa el número de muestras utilizadas en la prueba de Wilcoxon. Debe notarse que la prueba paramétrica F para la igualdad en las varianzas es extremadamente sensible a la violación de los supuestos de normalidad.

Referencias bibliográficas

En Moses (1963) y Hollander y Wolfe (1971) el lector encontrará detalles adicionales de la prueba de rangos de Moses. Acerca de la potencia-eficacia de la prueba, puede consultar el artículo de Moses (la referencia anterior) y Schorak (1969).

ANÁLISIS

En este capítulo hemos presentado nueve pruebas estadísticas que son útiles en la evaluación de la "significación de la diferencia" entre dos muestras independientes. En la selección de alguna de ellas, el investigador puede auxiliarse en este análisis, que tiene la ventaja que describe las pruebas y las contrasta entre sí.

Todas las pruebas no paramétricas para dos muestras independientes evalúan la hipótesis de que las dos muestras provienen de la misma población, sólo que estas pruebas son más o menos sensibles a distintos tipos de diferencias entre las muestras. Por ejemplo, si se desea probar si dos muestras representan poblaciones que difieren en tendencia central, se pueden escoger las siguientes pruebas: la prueba de la mediana (o la prueba exacta de Fisher cuando *N* es pequeña), la prueba de Wilcoxon, la prueba poderosa de rangos ordenados, la prueba de Kolmogorov-Smirnov para dos muestras (unidireccional) y la prueba de las permutaciones. Por otro lado, si ese investigador está interesado en determinar si dos muestras provienen de la misma población y difieren en cualquier aspecto (tendencia central, dispersión, sesgo, etc.), una de las siguientes pruebas es la apropiada: la prueba ji cuadrada, o la prueba de Kolmogorov-Smirnov (bidireccional). Las técnicas restantes, la prueba de Siegel-Tukey y la prueba de rangos de Moses, son aplicables en la evaluación de si un grupo exhibe respuestas extremas en comparación con un grupo independiente.

La elección entre las pruebas que son sensibles a diferencias en la ubicación está determinada por el tipo de medición realizada y por los tamaños de la muestra. La prueba más poderosa es la prueba de las permutaciones. Sin embargo, esta prueba puede utilizarse sólo cuando tenemos la confianza en la naturaleza *numérica* de las mediciones obtenidas y es posible de aplicar sólo cuando las muestras son pequeñas. Con muestras grandes o medidas "débiles", la opción sugerida es la prueba de Wilcoxon, que es casi tan poderosa como la prueba de las permutaciones si la dispersión de los dos grupos es la misma; o la prueba poderosa de rangos ordenados si uno no puede suponer dispersiones (varianzas) iguales para los dos grupos. Si las muestras son muy pequeñas, la prueba de Kolmogorov-Smirnov es un poco más eficaz que la prueba de Wilcoxon. Si las mediciones son tales que sólo dicotomizan las observaciones como "por arriba" o "por abajo" de la mediana, entonces es aplicable la prueba de la mediana. Esta prueba no es tan poderosa como la prueba de Wilcoxon al contrastar diferencias en la ubicación, pero es más apropiada que la prueba de Wilcoxon o la prueba poderosa de rangos ordenados cuando los datos son observaciones que no pueden ser completamente ordenados por rangos. Si el tamaño de las muestras combinadas es muy pequeño, el investigador, cuando aplique la prueba de la mediana, debe hacer el análisis utilizando la prueba exacta de Fisher. Cabe apuntar que la prueba de la mediana puede ser una opción viable aun para datos en escala de intervalo. Por ejemplo, si las observaciones están truncadas que algunos valores del rango total no fueron observados, entonces la prueba t no es apropiada, mientras que la prueba de la mediana es adecuada ya que solamente cuenta aquellas puntuaciones que se encuentran por arriba (o por debajo) de la mediana.

La elección entre las pruebas que son sensibles a todo tipo de diferencias (el segundo grupo mencionado anteriormente) depende de lo poderoso de las mediciones obtenidas, del tamaño de la muestra y del poder relativo de las pruebas disponibles. La prueba ji cuadrada es aplicable a datos que se encuentran en escala nominal o más poderosas. Si se aplica la ji cuadrada y se rechaza la H_0, entonces la tabla de contingencia y los grados de libertad pueden ser divididos en componentes aditivos para determinar exactamente dónde aparecen las diferencias en la tabla. Cuando N es pequeña y los datos se encuentran en una tabla de contingencia de 2×2, se debe utilizar la prueba exacta de Fisher en lugar de la ji cuadrada. En muchos casos la ji cuadrada puede no hacer un uso eficaz de toda la información de los datos. Si las puntuaciones de la población se distribuyen continuamente, podemos escoger la prueba de Kolmogorov-Smirnov (bidireccional) en lugar de la prueba ji cuadrada. De todas las pruebas para cualquier diferencia, la de Kolmogorov-Smirnov es la más poderosa. Si se utiliza con datos que no cubren los supuestos de continuidad, es todavía aplicable, sólo que resulta más conservadora, es decir, el valor obtenido de P en tales casos será un tanto superior de lo que debería ser y, por tanto, la probabilidad de cometer un error de tipo II se incrementará. Si H_0 es rechazada con tales datos, podemos tener absoluta confianza de tal decisión.

Se deben destacar dos puntos acerca del uso de este segundo tipo de pruebas. Primero, si estamos interesados en evaluar la hipótesis alterna de que los grupos difieren en tendencia central, es decir, que una población tiene una mediana mayor que el otro grupo, se debería usar una prueba específicamente diseñada para contrastar diferencias en ubicación: una de las pruebas enumeradas en el primer grupo. Segundo, cuando se rechaza H_0 con base en una prueba que contrasta cualquier tipo de diferencia (una de las pruebas del segundo grupo), se puede afirmar entonces que los dos grupos son diferentes, pero no se puede decir *específicamente de qué manera(s)* son diferentes las poblaciones.

Por último al evaluar diferencias en dispersión o varianzas, la prueba de Siegel-Tukey supone que las medianas para los dos grupos son las mismas (o conocidas). Si las medianas de los dos grupos son diferentes, la prueba adecuada es la de rangos de Moses; sin embargo, el suponer a los datos en escala de intervalo, requiere cálculos adicionales y subdividir cada uno de los grupos en subconjuntos al azar.

Consideradas en conjunto, las pruebas que se reseñaron en este capítulo forman un repertorio útil de procedimientos para analizar diferencias entre dos grupos independientes.

6

El caso de *k* muestras relacionadas

En los primeros capítulos expusimos pruebas estadísticas para evaluar las diferencias entre *a*) una sola muestra y alguna población específica y *b*) dos muestras, ya sea relacionados o independientes. En este capítulo y en el siguiente se presentarán procedimientos para evaluar diferencias entre tres o más grupos. Esto es, las pruebas estadísticas contenidas en estos capítulos evaluarán la hipótesis nula de que k (tres o más) muestras han sido extraídas de la misma población o de poblaciones idénticas. En el presente capítulo se presentan pruebas para evaluar k muestras *relacionadas*, y en el capítulo siguiente, pruebas para evaluar k muestras *independientes*.

En ocasiones, las circunstancias requieren que diseñemos un experimento en donde dos o más muestras o condiciones sean estudiadas simultáneamente. Cuando tres o más muestras o condiciones van a ser comparadas en un experimento, es necesario utilizar una prueba estadística que nos indique si existe una diferencia *global* entre las k muestras o condiciones antes de seleccionar cualquier par de muestras para evaluar la significación de las diferencias entre ellas.

Si deseamos utilizar una prueba estadística para dos muestras, a fin de evaluar las diferencias, por ejemplo, de cinco grupos, debemos calcular, con el objeto de comparar cada par de muestras, 10 pruebas estadísticas. [Cinco muestras tomadas en dos ocasiones $= \binom{5}{2} = 5!/2!3! = 10$.] Un procedimiento como éste, no sólo es tedioso sino que además puede conducirnos a conclusiones superficiales que en su momento pueden resultar inadecuadas. Por ejemplo, deseamos utilizar un nivel de significación de $\alpha = 0.05$. Nuestra hipótesis es que existe una diferencia entre $k = 5$ muestras. Si evaluamos nuestra hipótesis comparando cada una de las cinco muestras con cada una de las muestras restantes, por medio de una prueba de dos muestras (lo que requeriría 10 comparaciones en total), nos proporcionamos 10 oportunidades en lugar de una de rechazar H_0. Al escoger 0.05 como nuestro nivel de significación, tomamos el riesgo de rechazar H_0 erróneamente (cometiendo un error de Tipo I) 5 % de las veces. Pero si realizamos 10 evaluaciones estadísticas de la misma hipótesis, incrementamos la probabilidad a 0.40 de que una prueba estadística para dos muestras encuentre una o más diferencias "significativas" (aún con $\alpha = 0.05$ para cada prueba individual). Es decir, el nivel de significación se convierte de $\alpha = 0.05$ a $\alpha = 0.40$.

En la bibliografía se han reportado casos en los cuales una prueba global de cinco muestras no presentaron diferencias significativas (lo cual condujo a la aceptación de H_0), pero sí lo hicieron pruebas para dos muestras. Tiende a aprovecharse la selección *a posteriori* de las pruebas y, por tanto, no podemos estar seguros de una decisión que involucre k muestras en las cuales el análisis se realice mediante pruebas estadísticas de dos en dos muestras en cada ocasión.

Sólo una prueba global (una prueba de k muestras) es la que nos permite rechazar la hipótesis nula y justifica que utilicemos un procedimiento para evaluar las diferencias entre cualquier par de las k muestras.

La técnica paramétrica para evaluar si varias muestras provienen de poblaciones idénticas es el *análisis de varianza* y estadísticos F asociados. Los supuestos asociados a los estadísticos que subyacen al análisis de varianza son:

1. Las puntuaciones u observaciones son extraídas de manera independiente de poblaciones normalmente distribuidas.
2. Todas las poblaciones tienen la misma varianza.
3. Las medias en las poblaciones normalmente distribuidas son combinaciones lineales de los "efectos", debido a los renglones y las columnas (los efectos son aditivos).

Además, la prueba F requiere de mediciones que se encuentren, al menos, en escala de intervalo.

Si para un investigador estos supuestos no son aplicables, las puntaciones no cubren los requisitos necesarios, o desea evitar hacer suposiciones a fin de incrementar la generalidad de sus descubrimientos, en el presente y en el siguiente capítulo se abordará una de las pruebas estadísticas no paramétricas que sí puede aplicarse de manera apropiada. Además de evitar las suposiciones y los requisitos mencionados, estas pruebas no paramétricas para k muestras tienen la ventaja adicional de posibilitar el análisis de datos de naturaleza categórica u ordinal.

Existen dos diseños básicos para la comparación de k grupos. En el primero, se *igualan* k muestras del mismo tamaño, de acuerdo con cierto criterio o criterios, los cuales pueden afectar los valores de las observaciones. En algunos casos, se logra la igualación comparando los mismos individuos bajo las k condiciones. También cada uno de los N individuos puede ser medido bajo las k condiciones. Para tales diseños se deben utilizar las pruebas estadísticas para k muestras relacionadas (que se presentan en este capítulo). El segundo diseño involucra k muestras *independientes* al azar, no necesariamente del mismo tamaño, y una muestra de cada población. Para este diseño se deben emplear las pruebas estadísticas para k muestras independientes (que se presentarán en el capítulo 7).

La distinción anterior está hecha exactamente en el caso paramétrico. El primer diseño se conoce como el *análisis de varianza bifactorial* o *análisis de varianza* *de medidas repetidas*, y a veces se le llama *diseño de bloques aleatorizados*.[1] El segundo diseño se denomina *análisis de varianza unifactorial*.

La distinción es similar a la que hicimos entre las pruebas para dos muestras relacionadas (capítulo 4) y para dos muestras independientes (capítulo 5).

Las pruebas estadísticas no paramétricas que se presentan en este capítulo son paralelas al análisis de varianza bifactorial o de medidas repetidas. Comenzaremos con una prueba apropiada para datos en medidas categóricas (en escala nominal). La segunda prueba es aplicable a datos que, al menos, se encuentran en escala ordinal. La tercer prueba permite evaluar una hipótesis acerca del ordenamiento de los efectos de variables ordinales. Al final del capítulo compararemos y contrastaremos estas pruebas para k muestras relacionadas y ofreceremos una guía adicional para que el investigador pueda seleccionar la mejor opción de acuerdo con la naturaleza de los datos.

PRUEBA *Q* DE COCHRAN

Función

La prueba de McNemar para dos muestras relacionadas (presentada en el capítulo 4) puede ampliarse para ser utilizada en estudios que involucran más de dos muestras. Como resultado de lo anterior se tiene la *prueba Q de Cochran* para k muestras relacionadas, que provee un método para evaluar si tres o más conjuntos igualados de frecuencias o proporciones difieren significativamente entre ellos mismos. La igualación debe fundamentarse en las características relevantes de los diferentes sujetos o en el hecho de que los mismos sujetos sean utilizados en las diferentes condiciones. La prueba Q de Cochran es particularmente aplicable a datos de tipo categórico (en escala nominal) u observaciones ordinales (o de intervalo) dicotómicas.

Es posible imaginarse una amplia variedad de hipótesis de investigación que pueden analizarse por medio de la prueba Q de Cochran. Por ejemplo, podemos evaluar si varios reactivos de una prueba difieren en dificultad para analizar la información de k reactivos medidos en N individuos. En este diseño, los k grupos se consideran "apareados" o "igualados", debido a que cada persona contesta los k reactivos.

Por otro lado, podemos analizar un solo ítem y comparar las respuestas de N sujetos bajo k condiciones diferentes. Nuevamente, la igualación consiste en tener a los sujetos en cada grupo, sólo que ahora los "grupos" difieren en que cada uno es observado en diferentes condiciones. Esto permite evaluar si existen diferencias en las respuestas de los sujetos en cada una de las k condiciones. Por ejemplo, se puede preguntar a cada uno de los miembros de un panel de votantes a cuál de dos candidatos prefieren en $k = 5$ ocasiones durante la temporada de elecciones (antes de la campaña, en la cúspide de la campaña del candidato A, en la cúspide de

[1] El término de *bloques aleatorizados* se deriva de la experimentación en agronomía, en la cual las parcelas pueden ser utilizadas como unidades experimentales. Un bloque está formado por parcelas adjuntas que se supone son más parecidas que las que están situadas a mayor distancia. Los k tratamientos (por ejemplo, k variedades de fertilizantes o k variedades de semilla) se asignan al azar de manera independiente a cada una de las k parcelas en un bloque.

la campaña del candidato 2B, inmediatamente después de votar e inmediatamente después de ser anunciados los resultados). La prueba Q de Cochran determinaría si estas condiciones tienen un efecto significativo sobre las preferencias de los votantes hacia los candidatos.

De nuevo, debemos comprar las respuestas de un ítem para N conjuntos, con k personas igualadas en cada conjunto. De esta manera, tendremos respuestas de k grupos igualados.

Método

Si los datos de los estudios descritos anteriormente fueran organizados en una tabla de doble entrada con dos renglones y k columnas, sería posible evaluar la hipótesis nula acerca de si la proporción (o frecuencia) de respuestas de una clase particular es la misma en cada columna, excepto para diferencias fortuitas. Cochran (1950) ha demostrado que, si la hipótesis nula es verdadera, es decir, si no hay diferencias en la probabilidad de, por ejemplo, "éxito" bajo cada condición (lo cual es lo mismo que decir que el "éxito" y el "fracaso" se distribuyen al azar en los renglones y las columnas de la tabla de doble entrada), y si el número de renglones no es demasiado pequeño, el estadístico

$$Q = \frac{k(k-1)\sum_{j=1}^{k}(G_j - \overline{G})^2}{k\sum_{i=1}^{N} L_i = \sum_{i=1}^{N} L_i^2} \qquad (6.1)$$

está distribuido aproximadamente como χ^2 con $gl = k - 1$, donde:

G_j = número total de "éxitos" en la j-ésima columna
$\overline{G}$ = media de G_j
L_i = número total de "éxitos" en el i-ésimo renglón

Una ecuación equivalente y derivada con facilidad de la ecuación (6.1), pero que simplifica los cálculos, es

$$Q = \frac{(k-1)\left[k\sum_{j=1}^{k} G_j^2 - \left(\sum_{j=1}^{k} G_j\right)^2\right]}{k\sum_{i=1}^{N} L_i - \sum_{i=1}^{N} L_i^2} \qquad (6.2)$$

Debido a que la distribución muestral de Q es aproximadamente como la distribución de χ^2 con $gl = k - 1$, la probabilidad asociada con la ocurrencia según H_0 de valores tan grandes como una Q observada, puede determinarse con la ayuda de la Tabla C del Apéndice I para un nivel de significación particular y un valor particular de $gl = k - 1$. La implicación es que la proporción (o frecuencia) de éxito difiere significativamente entre varias muestras. Esto es, H_0 puede ser rechazada en ese nivel de significación particular.

Ejemplo. Supóngase que estamos interesados en conocer la influencia de la simpatía del entrevistador sobre las respuestas del padre de familia en una encuesta de opinión. Podemos entrenar a un entrevistador para efectuar tres clases de entrevistas: entrevista 1: mostrando interés, simpatía y entusiasmo; entrevista 2: mostrando formalidad, sobriedad y cortesía, y entrevista 3: mostrando desinterés, brusquedad y formalidad áspera. Al entrevistador se le asignaría visitar tres grupos de 18 casas y se le pediría que utilizara un tipo de entrevista en cada grupo de casas. Así, obtendríamos 18 conjuntos de padres de familia con tres padres de familia igualados (equivalentes en cuanto a variables relevantes) en cada conjunto. Para cada conjunto, los tres miembros serían asignados al azar a las tres condiciones (tipos de entrevista). De esta manera, tendríamos tres muestras igualadas ($k = 3$) con 18 miembros en cada una ($N = 18$). Entonces, podemos evaluar si las diferencias entre los tres estilos de entrevista influenciaron el número de respuestas de "sí" (1) dadas a un ítem particular en los tres grupos igualados. Con la utilización de datos ficticios, una prueba de esta hipótesis sería como sigue:

i. *Hipótesis nula*. H_0: la probabilidad de un "sí" es la misma para los tres tipos de entrevista. H_1: la probabilidad de un "sí" difiere dependiendo del estilo de la entrevista.
ii. *Prueba estadística*. Se escogió la prueba Q de Cochran porque los datos pertenecen a más de dos grupos relacionados ($k = 3$) y son respuestas dicotómicas ("sí" y "no").
iii. *Nivel de significación*. $\alpha = 0.01$ y N es el número de casos en cada uno de los k grupos igualados.
iv. *Distribución muestral*. Cuando la hipótesis nula es verdadera, Q [calculada con las ecuaciones (6.1) o (6.2)] se distribuye aproximadamente como χ^2 con $gl = k - 1$. Esto es, la probabilidad asociada con la ocurrencia según H_0 de cualquier valor tan grande como un valor de Q observado, puede determinarse utilizando la tabla C del Apéndice I.
v. *Región de rechazo*. La región de rechazo consiste en todos los valores de Q que sean tan grandes que la probabilidad asociada con su ocurrencia cuando H_0 es verdadera, sea igual o menor que $\alpha = 0.01$.
vi. *Decisión*. En este ejemplo, representaremos a los "sí" con 1 y a los "no" con 0. Los datos del estudio se muestran en la tabla 6.1. Las puntaciones están arregladas en $N = 18$ renglones y $k = 3$ columnas. Se muestran, además, los valores de L_i (número de "sí" en cada renglón) y de L_i^2. Por ejemplo, en el primer conjunto de padres de familia todos respondieron "no", sin considerar el estilo de la entrevista, tenemos entonces que $L_i = 0 + 0 + 0 = 0$, y $L_i^2 = 0^2 = 0$. En el segundo conjunto, las respuestas a las entrevistas 1 y 2 fueron afirmativas, pero las respuestas a la entrevista 3 fueron negativas, esto es: $L_2 = 1 + 1 + 0 = 2$, y $L_2^2 = 2^2 = 4$.
   Podemos observar que $G_1 = 13$, es el número de respuestas "sí" a la entrevista 1; $G_2 = 13$, es el número de respuestas "sí" a la entrevista 2, y $G_3 = 3$, es el número de respuestas "sí" a la entrevista 3.

El número total de respuestas "sí" en las tres entrevistas es

$$\sum_{j=1}^{3} G_j = 13 + 13 + 3 = 29$$

Observe que $$\sum_{i=1}^{18} L_i = 29$$

Tabla 6.1. Respuestas de "sí" (= 1) y "no" (= 0) de padres de familia ante tres tipos de entrevista.

Grupo	*Entrevista 1*	*Entrevista 2*	*Entrevista 3*	L_i	L_i^2
1	0	0	0	0	0
2	1	1	0	2	4
3	0	1	0	1	1
4	0	0	0	0	0
5	1	0	0	1	1
6	1	1	0	2	4
7	1	1	0	2	4
8	0	1	0	1	1
9	1	0	0	1	1
10	0	0	0	0	0
11	1	1	1	3	9
12	1	1	1	3	9
13	1	1	0	2	4
14	1	1	0	2	4
15	1	1	0	2	4
16	1	1	1	3	9
17	1	1	0	2	4
18	1	1	0	2	4
Total	$G_1 = 13$	$G_2 = 13$	$G_3 = 3$	$\sum_{i=1}^{18} L_i = 29$	$\sum_{i=1}^{18} L_i^2 = 63$

(Los totales por columna y renglón son iguales.) La suma de cuadrados del total por renglón es

$$\sum_{i=1}^{18} L_i^2 = 63$$

la suma de la última columna.

Si sustituimos los valores en la ecuación (6.2), tenemos

$$Q = \frac{(k-1)\left[k\sum_{j=1}^{k} G_j^2 - \left(\sum_{j=1}^{k} G_j\right)^2\right]}{k\sum_{i=1}^{N} L_i - \sum_{i=1}^{N} L_i^2} \tag{6.2}$$

$$= \frac{(3 - 1)\ [3(13^2 + 13^2 + 3)^2 - 29^2]}{(3)(29) - 63}$$

$$= 16.7$$

Al remitirnos a la tabla C del Apéndice I, tenemos que $Q \geq 16.7$ tiene una probabilidad de ocurrencia cuando H_0 es verdadera de $p < 0.001$ con $gl = k - 1 = 3 - 1 = 2$. Esta probabilidad es menor que el nivel de significación $\alpha = 0.01$. Así, tenemos que el valor de Q se encuentra en la región de rechazo y, por tanto, nuestra decisión es rechazar H_0 a favor de H_1. Con base en estos datos, podemos concluir que las probabilidades de obtener una respuesta afirmativa en los diversos estilos de entrevistas, son diferentes.

Debe notarse que Q se distribuye como χ^2 con $gl = k - 1$ si el número de renglones (el tamaño N de la muestra) no es demasiado pequeño (generalmente $N \geq 4$) y si el producto Nk es mayor que 24. Debido a que los renglones consisten en ceros y unos, no afectan el valor de Q y el tamaño "efectivo" de la muestra para aproximarse a la distribución de χ^2 es $N =$ el número de renglones donde no todos son ceros o unos. Para muestras muy pequeñas, la distribución muestral exacta de Q puede construirse a partir de los planteamientos de las permutaciones. Este caso no se analiza porque los cálculos relevantes son especialmente tediosos y la distribución es relativamente aproximada a la χ^2.

Resumen del procedimiento

Los pasos que hay que seguir para la utilización de la prueba Q de Cochran son:

1. A los datos dicotómicos se les asigna la puntuación 1 por cada éxito y 0 por cada fracaso.
2. Los datos se presentan en una tabla de $N \times k$, siendo $N =$ renglones y $k =$ columnas. N es el número de casos en cada uno de los k grupos o condiciones.
3. El valor de Q se determina utilizando la ecuación (6.1) o la ecuación (6.2).
4. La significación del valor observado de Q puede determinarse mediante la tabla C del Apéndice I, dado que Q se distribuye aproximadamente como la χ^2 con $gl = k - 1$. Si la probabilidad asociada con la ocurrencia cuando H_0 es verdadera es tan grande como el valor observado de Q es igual o menor que α, debemos rechazar H_0.

Potencia-eficacia

La noción potencia-eficacia no tiene mucho sentido cuando la Q de Cochran se aplica a datos en escala nominal o dicotómicos, porque las pruebas paramétricas no se aplican a este tipo de datos. Cuando la Q de Cochran se emplea con datos que no son de tipo nominal o dicotómicos, puede proporcionarnos información en exceso. Como ya se mencionó, la distribución de la χ^2 se aproxima exactamente a la distribución de la Q cuando $N \geq 4$ y $Nk > 24$.

Referencias bibliográficas

El lector puede encontrar análisis de la prueba Q de Cochran en Cochran (1950) y Marascuilo y Sweeney (1977). En Patil (1975) se localizan tablas de la distribución muestral exacta para N y k pequeñas.

ANÁLISIS DE VARIANZA BIFACTORIAL POR RANGOS, DE FRIEDMAN

Función

Cuando los datos de k muestras igualadas están al menos en escala ordinal, se puede utilizar el *análisis de varianza de Friedman* para evaluar la hipótesis nula de que las k muestras fueron extraídas de la misma población.

Debido a que las k muestras son igualadas, el número de casos N es el mismo en cada una de las muestras. La igualación puede lograrse estudiando el mismo grupo de sujetos en cada una de las k condiciones. También, el investigador puede obtener N conjuntos, cada uno de k sujetos igualados, y asignar al azar a un sujeto de cada conjunto a la primera condición, otro sujeto a la segunda condición, etc. Por ejemplo, si se desea estudiar las diferencias de aprendizaje bajo cuatro métodos de enseñanza, es necesario obtener conjuntos de $k = 4$ alumnos, cada conjunto formado por niños que han sido igualados respecto a las variables relevantes (edad, aprendizaje previo, inteligencia, estatus socioeconómico, motivación, etc.), y entonces asignar un niño de cada uno de los N conjuntos a cada método de enseñanza: un niño al método de enseñanza A, otro niño al método B, otro al método C y, finalmente, otro al método D.

El análisis de varianza bifactorial de Friedman evalúa la hipótesis nula de que los k grupos igualados o medidas repetidas provienen de la misma población o de poblaciones con la misma mediana. Para especiicar de manera más explícita la hipótesis nula, debemos proponer que θ_j es la mediana poblacional en el j-ésimo grupo o condición. Así mismo, debemos plantear en la hipótesis nula que las medianas son las mismas tanto como H_0: $\theta_1 = \theta_2 = \ldots = \theta_k$. Entonces, la hipótesis alterna es H_1: $\theta_i \neq \theta_j$ en al menos dos condiciones o grupos i y j. Es decir, si la hipótesis alterna es verdadera, al menos un par de condiciones tienen medianas diferentes. Bajo la hipótesis nula, la prueba supone que las variables tienen la misma distribución continua subyacente; así, ésta requiere que las mediciones, se encuentren al menos, en escala ordinal.

Racionalización y método

Para la prueba de Friedman, los datos deben presentarse en una tabla de doble entrada conteniendo N renglones y k columnas. Los renglones representan los sujetos o conjuntos de sujetos igualados, y las columnas, las distintas condiciones. Si lo que se estudia son las puntaciones de los sujetos en las distintas condiciones, entonces cada renglón nos proporciona las puntaciones de cada sujeto en cada una de las k condiciones.

Los datos que emplea esta prueba son rangos. Las puntaciones en cada *renglón* se ordenan por rangos separadamente. Esto es, estudiando k condiciones, los rangos en cualquier renglón varían de 1 a k. La prueba de Friedman determina la probabilidad de que diferentes *columnas* de rangos (muestras) provengan de la misma población, es decir, que las k variables tengan la misma mediana.

Por ejemplo, supóngase que deseamos estudiar las puntaciones de tres grupos en cuatro condiciones ($N = 3$ y $k = 4$). Cada grupo contiene cuatro sujetos igualados y cada uno de éstos será asignado a cada una de las cuatro condiciones. Supongamos que nuestras puntaciones para este estudio son las que se presentan en la tabla 6.2. Para aplicar la prueba de Friedman a estos datos, primero debemos ordenar las puntaciones por rangos en cada renglón. A la puntación menor se le asigna el rango 1, a la siguiente, el rango 2, etc., hasta obtener los datos que se muestran en la tabla 6.3. Observe que los rangos en cada renglón de esta tabla varían de 1 a $k = 4$.

Si la hipótesis nula (todas las muestras —columnas— provienen de la misma población) es verdadera, entonces la distribución de los rangos en cada columna sería cuestión de oportunidad y así, esperaríamos que los rangos 1, 2, 3 y 4 aparecieran en cada columna con, aproximadamente, igual frecuencia. Es decir, si los datos fueran aleatorios, la suma de los rangos en cada columna sería $N(k + 1)/2$. Para los datos de la tabla 6.3, la suma esperada sería $3(4 + 1)2 = 7.5$. Esto indica que para cualquier grupo, es cuestión de oportunidad que bajo cualquier condición ocurran las puntuaciones mayor y menor —este caso ocurriría si realmente las condiciones no fueran diferentes.

Tabla 6.2. Puntuaciones de tres grupos igualados bajo cuatro condiciones.

	Condiciones			
Grupo	*I*	*II*	*III*	*IV*
A	9	4	1	7
B	6	5	2	8
C	9	1	2	6

Tabla 6.3 Rango de tres grupos igualados bajo cuatro condiciones.

	Condiciones			
Grupo	*I*	*II*	*III*	*IV*
A	4	2	1	3
B	3	2	1	4
C	4	1	2	3
R_j	11	5	4	10

Si las puntuaciones de los sujetos fueran independientes de las condiciones, el conjunto de rangos en cada columna representaría una muestra al azar de una distribución discreta rectangular de los rangos 1, 2, 3 y 4, y los rangos totales de las distintas columnas serían, aproximadamente, iguales. Por el contrario, si las puntuaciones de los sujetos fueran dependientes de las condiciones (es decir, si H_0 fuera falsa), entonces los rangos totales variarían de una columna a otra. En vista de que todas las columnas contienen un número igual de casos, un planteamiento equivalente consistiría en que, según H_0, el promedio de los rangos en las distintas columnas sería, aproximadamente, el mismo.

La prueba de Friedman determina si los rangos totales (denominados R_j) para cada condición o variable, difieren significativamente de los valores esperados por oportunidad. Para realizar esta prueba, debemos calcular el valor del estadístico, el cual denotaremos como F_r

$$F_r = \left[\frac{12}{Nk(k+1)} \sum_{j=1}^{k} R_j^2 \right] - 3N(k+1) \qquad (6.3)$$

donde

N = número de renglones (sujetos)
k = número de columnas (variables o condiciones)
R_j = suma de los rangos en la j-ésima columna (suma de los rangos para variable j-ésima)

y

$\sum_{j=1}^{k}$ = la sumatoria de los cuadrados de los rangos de todas las condiciones

Las probabilidades asociadas a varios valores de F_r cuando H_0 es verdadera, han sido tabuladas para varios tamaños de muestras y varios números de variables. La tabla M del Apéndice I nos proporciona las probabilidades asociadas a valores de F_r tan grandes como los valores tabulados para varios valores de N y de k. Si el valor observado de F_r es mayor que el valor registrado en dicha tabla, en el nivel de significación escogido, se debe rechazar H_0 a favor de H_1.

Si el número de variables es mayor que cinco ($k > 5$) o el tamaño de la muestra (N) es mayor que los valores proporcionados por la tabla M del Apéndice, se debe utilizar una aproximación para muestras grandes. Cuando el número de renglones y/o columnas es grande, se puede demostrar que el estadístico F_r que nos proporciona la ecuación (6.3) se distribuye aproximadamente como la χ^2 con $gl = k - 1$. De esta manera, puede utilizarse la tabla C del Apendice I para determinar la significancia de la probabilidad.

Si el valor de F_r [calculado mediante la ecuación (6.3)] es igual o mayor que el valor proporcionado por la tabla M o la tabla C del Apéndice I para un nivel de significación particular, entonces la suma de los rangos (o, equivalentemente, el rango promedio R_j/N) para las distintas columnas difiere significativamente (lo cual indica que el tamaño de las puntuaciones depende de las condiciones en que se obtienen), y por tanto H_0 debe ser rechazada en ese nivel de significancia.

Para ilustrar el cálculo de F_r y el uso de la tabla M del Apéndice I, probaremos la significación de las diferencias de los datos mostrados en la tabla 6.3. Obsérvese que el número de condiciones es $k = 4$ y el número de renglones es $N = 3$. Las sumas R_j de los rangos son 11, 5, 4 y 10, respectivamente. Podemos calcular el valor de F_r para los datos de la tabla 6.3 sustituyendo sus valores en la ecuación (6.3):

$$F_r = \left[\frac{12}{Nk(k+1)} \sum_{j=1}^{k} R_j^2 \right] - 3N(k+1) \qquad (6.3)$$

$$= \frac{12}{(3)(4)(4+1)}(11^2 + 5^2 + 4^2 + 10^2) - (3)(3)(4+1)$$

$$= 7.4$$

Para determinar la probabilidad de la ocurrencia de $F_r \geq 7.4$ cuando H_0 es verdadera, podemos utilizar la tabla M del Apéndice I, la cual nos proporciona los valores críticos observados de F_r para $k = 4$. Así mismo, esta tabla nos muestra que la probabilidad asociada a $F_r \geq 6.5$ cuando $N = 3$ y $k = 4$, es $p \leq 0.05$. Así, para estos datos debemos rechazar la hipótesis nula de que las cuatro muestras fueron extraídas de la misma población con las mismas medianas en el nivel de 0.05 de significación ya que el valor observado de F_r es mayor que el valor registrado en las tablas.

Ejemplo para *N* y *k* grandes. En un estudio acerca del efecto de tres programas de reforzamiento sobre el aprendizaje de discriminación en tres muestras igualadas ($k = 3$) de 18 ratas ($N = 18$),[2] éstas fueron entrenadas bajo tres programas de reforzamiento. La igualación se realizó al utilizar 18 ratas de la misma camada, en grupos de tres. Aunque las 54 ratas recibieron la misma cantidad de reforzamiento, el programa de reforzamiento fue diferente para cada uno de los grupos. Un primer grupo fue entrenado con 100 % de reforzamiento (*RR*); un segundo grupo igualado fue entrenado con reforzamiento parcial en donde cada secuencia de ensayos terminaba con un ensayo no reforzado (*RU*), y el tercer grupo igualado fue entrenado con reforzamiento parcial en donde cada secuencia de ensayo finalizaba con un ensayo reforzado (*UR*).

Después del entrenamiento, el aprendizaje se midió por medio de la velocidad en que las ratas aprendieron una conducta "opuesta" —las ratas que habían sido entrenadas para correr hacia el color blanco, debían correr hacia el color negro—. El mejor aprendizaje inicial debería mostrar una transferencia lenta en la nueva condición. La predicción consistía en que la utilización de los distintos programas de reforzamiento mostraría las diferencias en la capacidad de "transferencia" del aprendizaje.

i. *Hipótesis nula.* H_0: los distintos programas de reforzamiento no tienen efecto diferencial sobre la conducta observada. H_1 los distintos programas de reforzamiento tienen efecto diferencial.
ii. *Prueba estadística.* Se seleccionó el análisis de varianza de Friedman porque el número de errores en la transferencia del aprendizaje probablemente no es una medida de intervalo de la fuerza original del aprendizaje. Aún más, el análisis de varianza paramétrico se excluye porque al evaluar la situación experimental se encontró que uno de los supuestos de la prueba F, probablemente, no fue cumplido.
iii. *Nivel de significación.* $\alpha = 0.05$ y N es el número de ratas en cada uno de los $k = 3$ grupos igualados $= 18$.
iv. *Distribución muestral.* Como se calculó con la ecuación (6.3) y el tamaño de la muestra es grande, F_r se distribuye aproximadamente como χ^2 con $gl = k - 1$. Así, la probabilidad asociada a la ocurrencia según H_0 de un valor tan grande como el valor observado de F_r, puede ser determinada mediante la utilización de la tabla C del Apéndice I.
v. *Región de rechazo.* La región de rechazo consiste en todos los valores de F_r que sean

[2] Grosslight, J. H. y Radlow, R., "Patterning effect of the nonreinforcement-reinforcement sequence in a discrimination situation", en *Journal of Comparative and Physiological Psychology*, núm. 49, 1956, págs. 542-546.

tan grandes que la probabilidad asociada con la ocurrencia cuando H_0 es verdadera, sea menor o igual que $\alpha = 0.05$.

vi. *Decisión*. Se determinó el número de errores cometidos por cada rata en la situación de "transferencia de aprendizaje" y esas puntaciones se ordenaron por rangos para cada uno de los tres grupos de 18 ratas. Estos rangos se muestran en la tabla 6.4.

Obsérvese que la suma de los rangos para los grupo RR es 39.5, para el grupo RU es 42.5 y para el grupo UR es 26.0. Un rango bajo significa un gran número de errores en la transferencia. Podemos calcular el valor de F_r sustituyendo los valores observados en la ecuación (6.3):

$$F_r = \left[\frac{12}{Nk(k+1)} \sum_{j=1}^{k} R_j^2 \right] - 3N(k+1) \tag{6.3}$$

$$= \frac{12}{(18)(3)(3+1)}(39.5^2 + 42.5^2 + 26^2) - (3)(18)(3+1)$$

$$= 8.58$$

Tabla 6.4. Rangos de 18 grupos igualados de la transferencia después del entrenamiento bajo tres condiciones de reforzamiento.

	Tipo de reforzamiento		
Grupo	*RR*	*RU*	*UR*
1	1	3	2
2	2	3	1
3	1	3	2
4	1	2	3
5	3	1	2
6	2	3	1
7	3	2	1
8	1	3	2
9	3	1	2
10	3	1	2
11	2	3	1
12	2	3	1
13	3	2	1
14	2	3	1
15	2.5	2.5	1
16	3	2	1
17	3	2	1
18	2	3	1
R_j	39.5	42.5	26.0

La tabla C del Apéndice I nos indica que $F_r = 8.58$ cuando $gl = k - 1 = 3 - 1 = 2$, es significativo entre los niveles 0.01 y 0.02. Por tanto, ya que $p < 0.02$ es menor que nuestro nivel de significación ($\alpha = 0.05$), la decisión es rechazar H_0. En conclusión, las puntuaciones de las ratas en la transferencia de aprendizaje dependen del programa de reforzamiento en los ensayos de aprendizaje originales.

Empates

Cuando existan empates entre los rangos en cualquier grupo (renglón), el estadístico F_r debe ser corregido para modificar la distribución muestral. La ecuación (6.4) nos proporciona el valor de F_r adecuado cuando ocurren empates. Aunque la ecuación (6.4) puede utilizarse en forma generalizada, es decir, en caso de que haya o no empates, el cálculo se puede volver tedioso.

$$F_r = \frac{12 \sum_{j=1}^{k} R_j^2 - 3N^2k(k+1)^2}{Nk(k+1) + \dfrac{\left(Nk - \sum_{i=1}^{N} \sum_{j=1}^{g_i} t_{i \cdot j}^3\right)}{(k-1)}} \tag{6.4}$$

donde g_i es el número de conjuntos de rangos empatados en el i-ésimo grupo y $t_{i \cdot j}$ es el tamaño del j-ésimo conjunto de rangos empatados en el i-ésimo grupo. Se incluyen los conjuntos de tamaño 1. Como en los casos de otras correcciones para datos empatados, el efecto de rangos empatados consiste en incrementar el tamaño del estadístico F_r de Friedman. Al efectuar la corrección para los empates en el ejemplo anterior, notamos que hay dos rangos empatados en el decimoquinto grupo, hay 52 empates de tamaño 1 y un empate de tamaño 2. Por tanto,

$$\sum_{i=1}^{N} \sum_{j=1}^{g_i} t_{i \cdot j}^3 = 1 + 1 + 1 + \ldots + 1 + 8 + 1 + \ldots + 1 = 60$$

Utilizando la ecuación (6.4) obtenemos $F_r = 8.70$, que es mayor que el valor obtenido (8.58) sin la corrección. Obviamente, debido a que H_0 fue rechazada sin la corrección, también es rechazada con la corrección. Es importante insistir que en este ejemplo el efecto de los empates fue muy pequeño; sin embargo, conforme el número de empates se incremente, el efecto sobre F_r será mayor.

Comparaciones múltiples entre grupos o condiciones

Cuando el valor obtenido de F_r es significativo, este resultado refleja que al menos una de las condiciones difiere con respecto a otra de las condiciones. Pero esto no indica al investigador cuál grupo es el diferente, ni cuántos de los grupos difieren entre sí. Es decir, cuando el valor obtenido de F_r es significativo, evaluamos la hipótesis H_0: $\theta_u = \theta_v$ contra la hipótesis H_1: $\theta_u \neq \theta_v$ para algunas condiciones u y v. Existe un procedimiento sencillo para determinar cuál(es) condición(es) es(son) la(s) que difiere(n). Primero, se determinan las diferencias $|R_u - R_v|$ para todos los pares de grupos o condiciones. Cuando el tamaño de la muestra es grande, estas diferencias se distribuyen aproximadamente de manera normal. Sin embargo, debido a que existe un gran número de diferencias y éstas no son independientes, debe ajustarse apropiadamente el procedimiento de comparación. Supóngase que evaluamos la hipótesis de que no existen diferencias entre las k condiciones o grupos igualados y que la rechazamos en el nivel de significación α. Entonces debemos probar la significación de las diferencias de los pares individuales, utilizando la siguiente desigualdad. Esto es, si

$$|R_u - R_v| \geq z_{\alpha/k(k-1)} \sqrt{\frac{Nk(k+1)}{6}} \qquad (6.5a)$$

o si los datos son expresados en términos de rangos promedios dentro de cada condición, y si

$$|\overline{R}_u - \overline{R}_v| \geq z_{\alpha/k(k-1)} \sqrt{\frac{k(k+1)}{6N}} \qquad (6.5b)$$

entonces podemos rechazar la hipótesis H_0: $\theta_u = \theta_v$, y concluir que $\theta_u \neq \theta_v$. Entonces, si la diferencia entre la suma de rangos (o rangos promedio) excede el valor crítico correspondiente dado por las ecuaciones (6.5*a*) o (6.5*b*), debemos concluir que las dos condiciones son diferentes. El valor $z_{\alpha/k(k-1)}$ es el valor de la abscisa de la distribución normal unitaria donde se encuentra el $\alpha/k(k-1)\%$ de la distribución. Los valores de z pueden obtenerse en la tabla A del Apéndice I.

Debido a que a menudo es necesario obtener los valores basados en probabilidades extremadamente pequeñas, en especial cuando k es grande, se puede utilizar la tabla A_{II} en lugar de la tabla A del Apéndice I. Aquélla es una tabla de la distribución normal estándar que ha sido arreglada de tal forma en que los valores utilizados en las comparaciones múltiples puedan ser obtenidos fácilmente. La tabla está arreglada con base en el número de comparaciones (#*c*) que pueden realizarse. Los valores de la tabla son las probabilidades del extremo superior asociadas a varios valores de α. Cuando hay k grupos, hay $k(k-1)/2$ comparaciones.[3]

Ejemplo. En el ejemplo anterior acerca de las diferencias entre los programas de reforzamiento, la hipótesis nula de que no existían diferencias entre los tres métodos de entrenamiento fue rechazada y concluimos que había diferencias entre los métodos de entrenamiento. Sin embargo, aunque comprobamos la existencia de diferencias, no supimos si había diferencias entre una condición y otra o si las diferencias existieron entre las tres con-

[3] A algunos lectores les puede parecer que existe una pequeña discrepancia entre el subíndice de z, el cual es $\alpha/k(k-1)$, y el número de comparaciones # c, que es $k(k-1)/2$. Nótese que estamos evaluando diferencias absolutas y, por tanto, sólo se utiliza el extremo superior de la distribución tabulada. De aquí que la probabilidad del extremo superior ($\alpha/2$) se divida entre el número de comparaciones $k(k-1)/2$, lo cual resulta en $\alpha/k(k-1)$.

diciones. Para encontrar dónde se encuentran las diferencias, determinaremos las comparaciones múltiples para los tres grupos.

Utilizaremos el mismo nivel de significación del análisis inicial ($\alpha = 0.05$). Primero debemos determinar las diferencias entre las condiciones. Por conveniencia, emplearemos los subíndices *RR*, *RU* y *UR* para referirnos a los tres grupos. Entonces, tenemos que $R_{RR} = 39.5$, $R_{RU} = 42.5$ y $R_{UR} = 26.0$, y las diferencias son:

$$\begin{aligned} |R_{RR} - R_{RU}| &= |39.5 - 42.5| = 3.0 \\ |R_{RR} - R_{UR}| &= |39.5 - 26.0| = 13.5 \\ |R_{RU} - R_{UR}| &= |42.5 - 26.0| = 16.5 \end{aligned}$$

Encontramos la diferencia crítica mediante la ecuación (6.5*a*). Debido a que $\alpha = 0.05$ y $k = 3$, el número de comparaciones $\#c$ es igual a $k(k-1)/2 = (3)(2)/2 = 3$. Recurrimos a la tabla A_{II} del Apéndice I para ver el valor de z, que es $z = 2.394$ [Alternativamente, se puede obtener el valor de z de la tabla A del Apéndice. Para utilizar esta tabla, primero debemos calcular $\alpha/k(k-1) = 0.05/(3)(2) = 0.00833$. En dicha tabla encontraremos (después de la interpolación) que $z_{0.00833} = 2.394$.] Entonces, la diferencia crítica es

$$\begin{aligned} z_{\alpha/k(k-1)} \sqrt{\frac{Nk(k+1)}{6}} &= 2.394 \sqrt{\frac{(18)(3)(3+1)}{6}} \\ &= 2.394\sqrt{36} \\ &= 14.36 \end{aligned}$$

Ya que sólo la tercera diferencia (16.5) excede el valor de la diferencia crítica, concluimos que únicamente la diferencia entre las condiciones *RU* y *UR* es significativa. Nótese que la segunda diferencia, aunque grande, no es de una magnitud suficientemente grande que nos permita concluir que *RR* y *UR* son diferentes.

Comparaciones de grupos o condiciones con un control

A veces un investigador puede pretender hacer una comparación más específica que el conjunto de comparaciones descritas anteriormente. Por ejemplo, supongamos que una condición o un grupo representa la *línea base* contra la que hay que comparar las demás condiciones o grupos. Después de aplicar el análisis de varianza de Friedman y comprobar que es significativo, el investigador puede desear comparar todas las condiciones en contra de una. Por conveniencia, denotaremos a la condición control como condición 1. Entonces, la hipótesis que el investigador debe evaluar es

$$H_0\colon \theta_1 = \theta_u \qquad \text{para } u = 2, 3, \ldots, k$$

contra

$$H_1\colon \theta_1 \neq \theta_u \qquad \text{para algunas } u = 2, 3, \ldots, k$$

El siguiente procedimiento permite al investigador evaluar un conjunto de condiciones contra una condición control.

Al igual que en el procedimiento de comparaciones múltiples descrito en la sección anterior, debemos calcular las diferencias $|R_1 - R_u|$ entre la condición de tratamiento y cada una de las otras condiciones. Cuando el tamaño de las muestras es de moderado a grande, estas diferencias se distribuyen aproximadamente de manera normal. De todas formas, las comparaciones no son independientes y los valores críticos deben obtenerse mediante la tabla A_{III} del Apéndice I. Entonces, podemos evaluar la significación de las diferencias entre una condición de tratamiento y las otras condiciones utilizando la siguiente desigualdad. Esto es, si

$$|R_1 - R_u| \geq q(\alpha, \#c)\sqrt{\frac{Nk(k+1)}{6}} \tag{6.6a}$$

o si los datos son expresados en términos de rangos promedios dentro de cada condición, y si

$$|\bar{R}_1 - \bar{R}_u| \geq q(\alpha, \#c)\sqrt{\frac{k(k+1)}{6N}} \tag{6.6b}$$

entonces podemos rechazar la hipótesis H_0: $\theta_1 = \theta_u$ a favor de H_1: $\theta_1 \neq \theta_u$. Los valores de $q(\alpha, \#c)$ son proporcionados por la tabla A_{III} del Apéndice para los valores seleccionados de α y de $\#c$, donde $\#c = k - 1$, el cual es el número de comparaciones.

Ejemplo. Supongamos que tenemos un conjunto de $N = 12$ sujetos medidos en una condición de línea base y otras cuatro condiciones diferentes; entonces $k = 5$. Tenemos los siguientes valores de $R_1 = 33$, $R_2 = 30$, $R_3 = 43$, $R_4 = 14$ y $R_5 = 60$. Utilizando la ecuación (6.3), el valor de $F_r = 38.47$, el cual es significativo en el nivel de $\alpha = 0.05$.[4] Ahora deseamos evaluar la diferencia entre cada condición y la línea base. Los valores respectivos de $|R_1 - R_u|$ son 3, 10, 19 y 27. Mediante la ecuación (6.6*a*) podemos encontrar los límites para las diferencias. Primero, en la tabla A_{III} del Apéndice I encontramos que $q(\alpha, \#c) = 2.44$ para $\alpha = 0.05$ y $\#c = k - 1 = 4$. Entonces,

$$\begin{aligned} |R_1 - R_u| &\geq q(\alpha, \#c)\sqrt{\frac{Nk(k+1)}{6}} \qquad (6.6a) \\ &\geq 2.44\sqrt{\frac{(12)(5)(5+1)}{6}} \\ &\geq 18.9 \end{aligned}$$

Cualquier diferencia que sea mayor de 18.9 indicará una diferencia significativa entre esa condición y la condición control. Sólo dos de las diferencias exceden ese límite. Por tanto, podemos concluir que las condiciones 4 y 5 son significativamente diferentes de la condición control (1).

[4] Se invita al lector a calcular el valor de F_r en este ejemplo, para asegurar el entendimiento de su cálculo con los datos proporcionados.

Resumen del procedimiento

Éstos son los pasos que hay que seguir para la utilización de el análisis de varianza bifactorial por rangos de Friedman:

1. Presente los puntajes en una tabla de doble entrada con N renglones (sujetos) y k columnas (condiciones o variables).
2. Ordene por rangos los datos en cada renglón, de 1 a k.
3. Determine la sumatoria de los rangos en cada columna (R_j).
4. Calcule el valor de F_r con la ecuación (6.3) si no hay empates, o con la ecuación (6.4) si hay empates en cualquier renglón.
5. El método para determinar la probabilidad de ocurrencia de un valor observado de F_r cuando H_0 es verdadera, depende del tamaño de N y de k:

   a) la tabla M del Apéndice I nos proporciona los valores críticos seleccionados de F_r para N y k pequeñas.
   b) Para N y/o k mayores que los utilizados en la tabla M del Apéndice, la probabilidad asociada puede ser determinada haciendo referencia a la distribución de la χ^2 con $gl = k - 1$ (tabla C del Apéndice I).

6. Si la probabilidad, proporcionada al realizar el paso 5, es igual o menor que α, se rechaza H_0.
7. Si H_0 es rechazada, utilice las comparaciones múltiples [ecuación (6.5)] para determinar cuáles diferencias entre las condiciones son significativas. Si se van a evaluar las diferencias entre las distintas condiciones y una condición control, utilice la ecuación (6.6).

Eficacia relativa

La potencia-eficacia del análisis de varianza bifactorial por rangos de Friedman para datos normalmente distribuidos cuando se comparan con su contraparte normal (la prueba F), es $2/\pi = 0.64$ cuando $k = 2$, y es mayor conforme incrementa k($k = 5$, eficacia $= 0.80$; $k = 10$, eficacia $= 0.87$; $k = 20$, eficacia $= 0.91$). Cuando se comparan muestras con distribución uniforme o exponencial, la eficacia es mayor.

Referencias bibliográficas

Los primeros tratados acerca del análisis de varianza de Friedman se encuentran en Friedman (1937, 1940). Estudios más recientes pueden consultarse en Lehmann (1975) y Randles y Wolfe (1979). El análisis de varianza bifactorial por rangos de Friedman se relaciona funcionalmente con el coeficiente de concordancia de Kendall (otra prueba no paramétrica, la cual se presenta en el capítulo 8).

PRUEBA DE PAGE PARA ALTERNATIVAS ORDENADAS

Función

El análisis de varianza bifactorial por rangos de Friedman evalúa la hipótesis de que k grupos igualados o k medidas repetidas son los mismos, contra la hipótesis alterna de que uno o más grupos son diferentes. En ocasiones, un investigador puede desear mantener una hipótesis alterna más específica. Por ejemplo, en un experimento sobre aprendizaje, cierto investigador pretende evaluar la hipótesis de "no aprendizaje" contra la hipótesis de que los sujetos pueden recordar más en el ensayo 2 que en el ensayo 1, pueden recordar más en el ensayo 3 que en el ensayo 2, etc. En este caso, la hipótesis alterna asociada con el análisis de varianza bifactorial por rangos de Friedman es demasiado general. La *prueba de Page para alternativas ordenadas* descrita en esta sección, evalúa la hipótesis de que los grupos (o las medidas) son los mismos contra la hipótesis alterna de que los grupos (o las medidas) están ordenados en una secuencia específica. Para aclarar las hipótesis nula y alterna de manera más explícita, θ = mediana de la población para el j-ésimo grupo o medida. Entonces, debemos plantear la hipótesis nula de que las medianas son las mismas como

$$H_0\text{: } \theta_1 = \theta_2 = \ldots = \theta_k$$

y la hipótesis alterna debe plantearse de la siguiente manera:

$$H_1\text{: } \theta_1 \leq \theta_2 \leq \ldots \leq \theta_k$$

esto es, las medianas se encuentran ordenadas por magnitud. Si la hipótesis alterna es verdadera, al menos una de las diferencias es una desigualdad estricta ($<$). Es importante notar que para asegurar el uso adecuado de esta prueba, el investigador debe ser capaz de especificar el orden de los grupos (medidas o condiciones) *a priori*.

Para aplicar la prueba de Page, los datos de las k muestras o medidas deben encontrarse al menos en escala ordinal. Suponemos además que hay N casos o conjuntos de observaciones. Como en el análisis de varianza bifactorial por rangos de Friedman, si las k muestras son igualadas, la igualación se logra al obtener N conjuntos de k sujetos igualados y asignar aleatoriamente a cada uno de los sujetos en cada una de las k condiciones.

Racionalización y método

Como ya se dijo, para aplicar la prueba de Page para alternativas ordenadas, el investigador debe especificar primero el ordenamiento, *a priori*, de los grupos. Los datos se colocan en una tabla de dos entradas con N renglones y k columnas. Al igual que en el análisis de varianza bifactorial por rangos de Friedman, los renglones representan los sujetos o conjuntos de sujetos igualados, y las columnas representan las k condiciones (grupos o medidas).

Los datos para la prueba de Page son rangos. Las puntuaciones en cada renglón se ordenan por separado y varían de 1 a k. La hipótesis nula consiste en que el rango promedio en cada una de las columnas es el mismo. La hipótesis alterna es que el rango se incrementa a lo largo de los grupos del 1 al k. Más que utilizar los rangos promedios en los cálculos de la prueba estadística, se utilizan los rangos totales R_j para los j grupos. Para realizar la prueba, se calcula el estadístico L:

$$L = \sum_{j=1}^{k} jR_j = R_1 + 2R_2 + \ldots + kR_k \tag{6.7}$$

donde R_j es la sumatoria de los rangos en las j columnas.

Las probabilidades asociadas con varios valores de L cuando H_0 es verdadera, han sido tabuladas para varios tamaños de N y varios números de variables k. La tabla N del Apéndice I nos proporciona las probabilidades asociadas a valores de L tan grandes como los valores de tablas para varios valores de N y k. Si el valor observado de L es mayor que el valor contenido en las tablas en el nivel de significación escogido, se debe rechazar la hipótesis H_0 a favor de la hipótesis alterna H_1. Por ejemplo, consideremos un experimento en el cual hay $N = 9$ conjuntos de observaciones en $k = 4$ mediciones. Se escogió un nivel de significación de $\alpha = 0.01$. En la tabla N del Apéndice vemos que si $L \geq 246$, debemos rechazar H_0 a favor de H_1.

Muestras grandes

En la tabla N del Apéndice I se presentan los valores críticos de L para $N \leq 20$ con $k = 3$ y $N \leq 12$ con $k = 4, 5, 6, 7, 8, 9$ y 10. Para valores mayores de N o de k, se utiliza una aproximación para muestras grandes en la evaluación de la hipótesis mediante el estadístico L de la prueba de Page. Para valores grandes de N y k, la distribución muestral de L se distribuye de manera aproximadamente normal con

$$\mu_L = \frac{Nk(k+1)^2}{4} \tag{6.8}$$

$$\sigma_L^2 = \frac{Nk^2(k^2-1)^2}{144(k-1)} \tag{6.9}$$

Así, para evaluar la hipótesis H_0 de que las medianas son iguales, contra la hipótesis de que las medianas están ordenadas, debemos calcular el estadístico z_L:

$$z_L = \frac{12L - 3Nk(k+1)^2}{k(k^2-1)} \sqrt{\frac{k-1}{N}} \tag{6.10}$$

Para N grande, el estadístico z_L está distribuido aproximadamente de manera normal, con media igual a cero y desviación estándar igual a uno. La significación de z_L y de L puede ser determinada mediante cualquier tabla de distribución normal estándar (tabla A del Apéndice I). Debido a que las alternativas están ordenadas, la prueba de Page es unidireccional.

Ejemplo para N y k pequeñas. En años recientes se ha incrementado el interés por la capacidad de las personas de percibir patrones táctiles. Se han desarrollado dispositivos que convierten caracteres impresos en patrones táctiles, con el propósito, entre otros, de capacitar a las personas con deficiencias visuales a "leer" textos de manera táctil. Uno de los dispositivos desarrollados, el Optacon, contiene una rejilla de pequeños puntos, cada uno de los cuales puede vibrar independientemente. Las letras del alfabeto producen distintos patrones de puntos que vibran. En investigaciones que involucran la integración temporal de tales patrones táctiles, Craig[5] examinó la cantidad de integración entre los elementos de los patrones vibrotáctiles como una función del tiempo entre el inicio de los elementos individuales. En un estudio subsecuente, el experimentador manipuló los estimulos de inicio asincrónico (EIA) por parte de los patrones y el espaciamiento entre los renglones de puntos vibrantes, los cuales están en contacto con la yema del dedo de los sujetos. La tarea era indicar si había espaciamiento o no en los modelos presentados.

Se evaluaron cuatro sujetos entrenados durante un gran número de ensayos en los cuales el estímulo variaba en la sincronía inicial y el espaciamiento entre los estímulos. A fin de evaluar la precisión de los sujetos para detectar la presencia de un espacio, era necesario determinar qué tan exactos eran los sujetos al reportar la ausencia de un espacio entre los sucesivos estímulos, sólo que también se varió el EIA. En cada sujeto se utilizaron seis diferentes EIA. En la tabla 6.5 se resume la proporción de respuestas correctas para cada sujeto en cada condición. El principal interés del investigador era evaluar la hipótesis de que la exactitud estaba inversamente relacionada con los EIA.

Tabla 6.5. Proporción de respuestas correctas como una función de estímulos de inicio asincrónico (EIA).

	Estímulos de inicio asincrónico (ms)					
Sujeto	204	104	56	30	13	0
A	0.797	0.873	0.888	0.923	0.942	0.956
B	0.794	0.772	0.908	0.982	0.946	0.913
C	0.838	0.801	0.853	0.951	0.883	0.837
D	0.815	0.801	0.747	0.859	0.887	0.902

i. *Hipótesis nula* H_0: los diferentes EIA no tienen efecto en la precisión con que los sujetos reportan el espaciamiento de los puntos en los patrones táctiles. H_1: la exactitud con que los sujetos reportan el espaciamiento de los puntos en los patrones táctiles es inversamente proporcional a los EIA. Esto es, con la asincronía decrementa la proporción de respuestas correctas.

ii. *Prueba estadística*. Se selecciona la prueba de Page para alternativas ordenadas porque el investigador supone un ordenamiento en la exactitud de las respuestas como

[5] Craig, J. C., "Vibratory temporal integration as a function of pattern discriminability", en *Perception & Psychophysics*, núm. 35, 1984, págs. 579-582.

una función de los EIA. Se excluye el análisis de varianza paramétrico ya que los datos muestran una falta de homogeneidad en la varianza y las distribuciones parecen estar sesgadas. Lo anterior indica que los supuestos de la prueba F no se satisfacen. Esto, complementado con una muestra pequeña, nos sugiere que es apropiado utilizar una prueba no paramétrica.

iii. *Nivel de significación* $\alpha = 0.01$, N es el número de sujetos $= 4$, y k es el número de mediciones en cada sujeto $= 6$.

iv. *Distribución muestral.* Los valores de N y k son pequeños, por tanto, la distribución muestral de L calculado con la ecuación (6.7) se encuentra tabulada en la tabla N del Apéndice I.

v. *Región de rechazo.* La región de rechazo consiste en todos los valores de L observados que exceden los valores registrados en las tablas, asociados a los valores apropiados de α, N y k.

vi. *Decisión.* La proporción de respuestas correctas para cada sujeto en cada condición está resumida en la tabla 6.5. Para realizar la prueba de Page, es necesario ordenar por rangos los datos en cada renglón, de 1 a 6. Este ordenamiento se presenta en la tabla 6.6, y en él se calculan las sumatorias de los rangos R_j. Estos valores se sustituyen en la ecuación (6.7):

$$
\begin{aligned}
L &= \sum_{j=1}^{k} jR_j = R_1 + 2R_2 + \ldots + kR_k \qquad (6.7)\\
&= 9 + 2(6) + 3(11) + 4(20) + 5(20) + 6(18)\\
&= 9 + 12 + 33 + 80 + 100 + 108\\
&= 342
\end{aligned}
$$

Tabla 6.6. Rangos ordenados de las proporciones de respuestas correctas para los datos de la tabla 6.5.

	Estímulos de inicio asincrónico (ms)					
Sujeto	204	104	56	30	13	0
A	1	2	3	4	5	6
B	2	1	3	6	5	4
C	3	1	4	6	5	2
D	3	2	1	4	5	6
R_j	9	6	11	20	20	18

En la tabla N del Apéndice I se muestra que el valor crítico de L para $\alpha = 0.01$, $N = 4$ y $k = 6$, es 331. Ya que el valor observado de L (342) es mayor que el valor contenido en las tablas (331) en el nivel de significación seleccionado, el experimentador debe rechazar H_0 y concluir que la exactitud de las respuestas está relacionada inversamente con los EIA. (Debe notarse que H_0 pudo haberse rechazado en el nivel de $\alpha = 0.001$.)

Resumen del procedimiento

Éstos son los pasos que hay que seguir en la aplicación de la prueba de Page para alternativas ordenadas:

1. Arregle los datos en una tabla de doble entrada con N renglones (sujetos) y k columnas (condiciones o variables). El ordenamiento se debe especificar *a priori*.
2. Ordene por rangos los datos en cada renglón de 1 a k.
3. Determine las sumatorias de los rangos (R_j) en cada columna.
4. Calcule el valor de L con la ecuación (6.7).
5. El método para determinar la probabilidad asociada a L cuando H_0 es verdadera, depende del tamaño de N y de k:

   a) La tabla N del Apéndice I nos proporciona los valores críticos de L seleccionados para $N \leq 20$ cuando $k = 3$, y $N \leq 12$ cuando $4 \leq k \leq 10$.

   b) Si el número de observaciones y/o variables excluye el uso de la tabla N, se puede utilizar una tabla de distribución normal. El valor de z_L puede ser calculado utilizando la ecuación (6.10) y la tabla A del Apéndice I puede utilizarse para determinar si z_L y, por tanto, L, se encuentran en la región de rechazo. Ya que H_1 especifica una hipótesis de orden alterno, la prueba es de una cola. Si H_0 es rechazada, se puede utilizar el procedimiento de comparaciones múltiples descrito en la sección anterior. Sin embargo, cuando se hacen comparaciones, debe tenerse en cuenta que las pruebas son unidireccionales y, por tanto, los valores z deben ser ajustados.

Eficacia relativa

La eficacia de la prueba de Page para alternativas ordenadas, cuando se compara con su distribución normal alternativa (la prueba t), es la misma que la del análisis de varianza bifactorial por rangos de Friedman (véase la página 216). Sin embargo, comparada con la prueba de Friedman, la prueba de Page es más potente en términos de poder detectar alternativas ordenadas.

Referencias bibliográficas

La prueba de Page fue propuesta por Page (1963). Esta prueba relaciona con el coeficiente de correlación de rangos ordenados de Spearman (véase el capítulo 8). La potencia de la prueba ha sido analizada por Hollander (1967).

ANÁLISIS

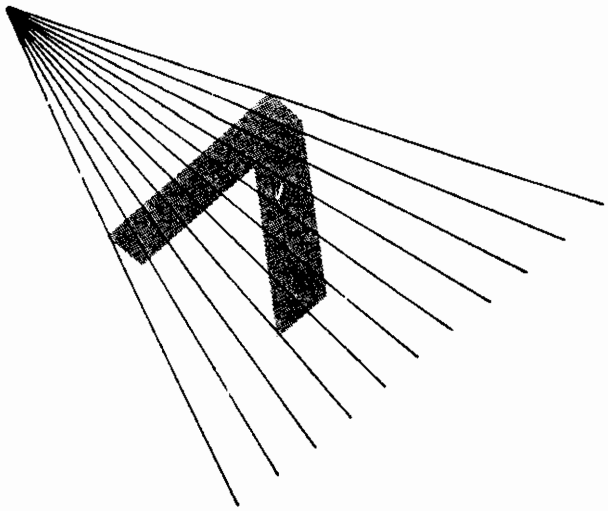
En este capítulo se presentaron tres pruebas estadísticas no paramétricas para evaluar H_0 en el caso de k muestras relacionadas o grupos igualados. La primera, la prueba Q de Cochran, es aplicable cuando las mediciones de la variable en estudio son categóricas (es decir, están en escala ordinal o en escala ordinal dicotomizada). Esta prueba capacita al investigador para determinar la probabilidad de que las k muestras relacionadas pudieran haber sido extraídas de la misma población, respecto a la proporción o frecuencia de "éxitos" en las distintas muestras o condiciones. Esto es, representa una evaluación global de si las k muestras exhiben diferencias significativas en la frecuencia de "éxitos".

La segunda prueba no paramétrica presentada, el análisis de varianza bifactorial por rangos de Friedman, es adecuada cuando las mediciones de las variables están al menos en escala ordinal. Esta prueba evalúa la probabilidad de que las k muestras relacionadas provengan de la misma población respecto a los rangos promedios. Esto es, constituye una evaluación global de si los valores de los datos varían como una función de las condiciones bajo las cuales fueron observados.

La prueba de Friedman debe utilizarse preferentemente a la prueba Q de Cochran cuando las características de los datos sean las adecuadas. Una ventaja de la prueba de Friedman es que tiene tablas de probabilidades exactas para muestras muy pequeñas, mientras que la prueba Q de Cochran está descartada cuando N (el número de renglones o conjuntos de observaciones) es muy pequeño.

Cuando utilizamos la prueba de Friedman y la hipótesis H_0 se rechaza pueden utilizarse las comparaciones múltiples para determinar cuáles condiciones difieren entre sí. Si el investigador tiene una hipótesis más precisa acerca de las diferencias entre una condición (por ejemplo una condición control) y otras condiciones, la prueba de Friedman ofrece la posibilidad de realizar tales comparaciones específicas.

La última prueba, la prueba de Page para alternativas ordenadas, al igual que la prueba de Friedman, supone que los datos se encuentran al menos en una escala ordinal. Sin embargo, para el análisis de varianza bifactorial por rangos de Friedman, la hipótesis alterna es que los grupos o las mediciones son diferentes. En contraste, la hipótesis alterna para la prueba de Page consiste en que los grupos están ordenados *a priori* con respecto a sus medianas. Debido a que la hipótesis alterna es más precisa, debe optarse por la prueba de Page cuando se tiene la disyuntiva de elegir entre una y otra, en una investigación en particular. Finalmente, es importante señalar que la hipótesis alterna especificada por la prueba de Page, con frecuencia se encuentra en los estudios experimentales de las ciencias sociales y de la conducta.



El caso de *k* muestras independientes

En el análisis de los datos de investigación, el investigador necesita decidir si varias muestras independientes pueden considerarse provenientes de la misma población. La hipótesis de investigación es que las k poblaciones son diferentes y la hipótesis estadística que va a ser evaluada es H_0: pob 1 $=$ pob 2 $= \ldots =$ pob k. El investigador extrae una muestra de cada población. Los valores de la muestra casi siempre de alguna manera varían; por tanto, el problema es determinar si las diferencias observadas en las muestras significan realmente diferencias en las poblaciones respectivas o si son meramente el tipo de diferencias que se pueden esperar entre las distintas muestras al azar de la misma población.

En este capítulo se presentarán procedimientos para evaluar la significación entre tres o más muestras o grupos independientes. Es decir, técnicas estadísticas para evaluar la hipótesis nula de que las k muestras independientes se han extraído de la misma población o de k poblaciones idénticas.

En la introducción al capítulo 6 intentamos distinguir entre dos tipos de pruebas para k muestras. El primer tipo de pruebas se utiliza para analizar datos de k muestras *igualadas* u observaciones repetidas de una sola muestra, y en ese mismo capítulo se presentaron pruebas estadísticas no paramétricas para esta clase de datos. La segunda de las pruebas para k muestras es apropiada para analizar datos de k muestras *independientes*. Estas últimas pruebas serán presentadas en este capítulo.

La prueba paramétrica usual para evaluar si varias muestras independientes provienen de la misma población, es el análisis de varianza unifactorial o prueba F. Los supuestos asociados con el modelo estadístico que subyace a la prueba F son que las observaciones fueron extraídas independientemente de poblaciones normalmente distribuidas y que todas ellas tienen la misma varianza. Para una interpretación verdaderamente significativa de los resultados de la prueba F, el requisito de las mediciones de la variable debe estar, al menos, en una escala de intervalo.

Si un investigador encuentra que tales supuestos son inapropiados para los datos derivados del problema de investigación, puede utilizarse una de las pruebas estadísticas no paramétricas para k presentadas en este capítulo. La elección de una prueba en particular depende de la naturaleza de los datos y de las suposiciones que debe hacer el investigador. Algunas de las pruebas descritas en este capítulo pueden aplicarse con datos que son necesariamente categorías (datos en escala nominal) y algunas otras lo hacen con rangos (datos en escala ordinal).

Presentaremos cuatro pruebas no paramétricas para el caso de k muestras independientes y concluiremos el capítulo con una análisis del uso comparativo de estas pruebas.

PRUEBA JI CUADRADA PARA *k* MUESTRAS INDEPENDIENTES

Función

Cuando los datos experimentales consisten en frecuencias en categorías discretas (nominales o categóricas y, en ocasiones, ordinales), puede utilizarse la prueba ji cuadrada para evaluar la significación de las diferencias entre k grupos independientes. La prueba ji cuadrada para k muestras o grupos independientes es una sencilla extensión de la prueba ji cuadrada para dos muestras independientes, presentada en el capítulo 5. En general, la prueba es similar tanto para dos como para k muestras o grupos independientes.

Método

El método de calcular el estadístico de la prueba ji cuadrada para muestras independientes se describirá brevemente, junto con un ejemplo de su aplicación. El lector encontrará una explicación más amplia de la prueba en el capítulo 5.

Para aplicar la prueba ji cuadrada, primero debemos presentar los datos (frecuencias) en una tabla $r \times k$ donde los datos en cada columna corresponden a las frecuencias en cada una de las r categorías de respuesta para cada uno de los k grupos o muestras diferentes. La hipótesis nula es que las k muestras de frecuencias provienen de la misma población o de poblaciones idénticas. Esta hipótesis, de que las k poblaciones no difieren entre sí, puede evaluarse aplicando las ecuaciones (5.2) o (5.2*a*):

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{5.2}$$

o

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{n_{ij}^2}{E_{ij}} - N \tag{5.2a}$$

donde

n_{ij} = número de casos que se categorizaron (observados) en el i-ésimo renglón de la j-ésima columna
E_{ij} = número de casos esperados en el i-ésimo renglón de la j-ésima columna cuando H_0 es verdadera

y la doble sumatoria se realiza sobre el total de renglones y columnas de la tabla (es decir, en la sumatoria intervienen todas y cada una de las celdillas). El lector recordará (cap. 5) que los valores esperados se determinan calculando $E_{ij} = R_iC_j/N$ para cada una de las celdillas de la tabla. Los valores de X^2 obtenidos al utilizar las ecuaciones (5.2) o (5.2*a*) se distribuyen asintóticamente (conforme N es más grande) como ji cuadrada con $gl = (r - 1)(k - 1)$, donde r es el número de renglones y k es el número de columnas (o grupos independientes) en la tabla de contingencia. Así, la probabilidad asociada con la ocurrencia de valores tan grandes como una X^2 observada nos la proporciona la tabla C del Apéndice I. Si un valor observado de X^2 es igual o mayor que el valor proporcionado por la tabla C del Apéndice para un nivel de significación particular y para $gl = (r - 1)(k - 1)$, entonces puede rechazarse H_0 en ese nivel de significación. Como se verá en los ejemplos y en la sección "cuándo utilizar la prueba ji cuadrada", es importante que los valores esperados (las E_{ij}) no sean demasiado pequeños, a fin de que pueda realizarse una interpretación adecuada del estadístico.

Ejemplo. En un macroproyecto para determinar la eficacia de varios tratamientos a pacientes externos para depresión clínica, dos investigadores dieron un tratamiento de cuatro posibles (psicoterapia, terapia conductual, terapia con fármacos o terapia de relajación) durante 10 semanas a 178 pacientes moderadamente deprimidos.[1] Los investigadores seleccionaron cuidadosamente a los pacientes para asegurarse que cada uno de ellos cumplía con los criterios de selección del estudio. Estos criterios incluían las puntuaciones dentro del rango de moderado y más allá en pruebas psicométricas para la depresión. Después de asignar al azar a los pacientes a una de las cuatro condiciones mencionadas anteriormente, cada uno de ellos fue tratado por un terapeuta. Éstos eran psicólogos, médicos o psiquiatras que fueron seleccionados de acuerdo con el prestigio en el uso de la técnica que iban a aplicar en el estudio.

Después del periodo de 10 semanas de tratamiento, los pacientes contestaron un cuestionario que incluía el Inventario de Beck para la Depresión (IBD), que es un instrumento común para medir depresión. Las puntuaciones del inventario se clasificaron en tres categorías posibles: normal (puntuaciones ≤ 7), benigno ($7 <$ puntuaciones < 23) y moderado a severo (puntuaciones ≥ 23). (Una puntuación de 23 o mayor en el IBD, era un criterio de selección para participar en el estudio.)

i. *Hipótesis nula*. H_0: la proporción de sujetos en cada una de las puntuaciones-categorías del IBD es la misma en cada uno de los diferentes tratamientos. H_1: la proporción de sujetos en cada una de las puntuaciones-categorías del IBD es diferente de acuerdo con los diferentes tratamientos.
ii. *Prueba estadística*. Puesto que los grupos del estudio son independientes y más de dos, es apropiado utilizar una prueba estadística para k muestras independientes.

[1] McLean, P. D. y Hakstian, A. R., "Clinical depression: Comparative efficacy of outpatient treatments", en *Journal of Consulting and Clinical Psychology*, núm. 47, 1979, págs. 818-836.

En virtud de que los datos consisten en categorías discretas, la prueba ji cuadrada es adecuada.

iii. *Nivel de significación.* $\alpha = 0.05$ y N es el número de sujetos que participaron en el estudio $= 178$.

iv. *Distribución muestral.* Según la hipótesis nula, cuando X^2 se calcula mediante las ecuaciones (5.2) o (5.2*a*), se distribuye aproximadamente como ji cuadrada con $gl = (r - 1)(k - 1)$. Cuando H_0 es verdadera, la probabilidad asociada con la ocurrencia de valores tan grandes o mayores que una X^2 observada, se encuentra en la tabla C del Apéndice I.

v. *Región de rechazo.* La región de rechazo consiste en todos los valores de X^2 que sean tan grandes, que la probabilidad asociada con su ocurrencia cuando H_0 es verdadera, sea menor o igual que $\alpha = 0.05$.

vi. *Decisión.* En la tabla 7.1 se resume la frecuencia de las puntuaciones dentro de cada categoría para cada grupo (tratamiento). Además, muestra las frecuencias esperadas (marcadas con un asterisco) en cada una de las categorias del IBD si H_0 fuera verdadera (si no existieran diferencias entre los distintos tratamientos, estos valores se determinan de los totales marginales de acuerdo con el método descrito en la sección "Prueba ji cuadrada para dos muestras independientes", del capítulo 5). Por ejemplo, mientras 11 de los sujetos que recibieron psicoterapia tuvieron puntuaciones de 7 o menores, si H_0 fuera verdadera esperaríamos $(56 \times 44)/178 = 13.84$ sujetos con puntuaciones de 7 o menores. Mientras 21 de los sujetos que recibieron terapia conductual tuvieron puntuaciones de 7 o menores, si H_0 fuera verdadera esperaríamos $(56 \times 42)/178 = 13.21$ sujetos con puntuaciones de 7 o menores. Si de los 42 sujetos que recibieron terapia conductual, tres tuvieron puntuaciones de 23 o mayores, si H_0 fuera verdadera esperaríamos $(34 \times 42)/178 = 8.02$ sujetos con puntuaciones de 23 o mayores.

El tamaño de X^2 refleja la magnitud de la discrepancia entre los valores observados y los esperados en cada una de las celdillas. Podemos calcular X^2 para los valores de la tabla 7.1 por medio de la ecuación (5.2*a*):

$$x^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{n_{ij}^2}{E_{ij}} - N \tag{5.2a}$$

$$\begin{aligned}
&= \frac{13^2}{8.40} + \frac{8^2}{8.21} + \frac{10^2}{9.36} + \frac{3^2}{8.02} + \frac{20^2}{21.75} + \frac{23^2}{21.26} + \frac{27^2}{24.22} + \frac{18^2}{20.76} \\
&\qquad + \frac{11^2}{13.84} + \frac{12^2}{13.53} + \frac{12^2}{15.42} + \frac{21^2}{13.21} - 178 \\
&= 20.12 + 7.80 + 10.68 + 1.12 + 18.39 + 24.88 + 30.10 + 15.61 \\
&\qquad + 8.84 + 10.64 + 9.34 + 33.38 - 178 \\
&= 12.80
\end{aligned}$$

Observamos que para los datos de la tabla 7.1, $X^2 = 12.80$ con

$$gl = (r - 1)(k - 1) = (3 - 1)(4 - 1) = 6$$

La tabla C del Apéndice I revela que el valor de X^2 es significativo más allá del nivel $\alpha = 0.05$. (El valor crítico para $\alpha = 0.05$ y $gl = 6$ es 12.59.) Por tanto, nuestra decisión es rechazar H_0. Concluimos que las puntuaciones en el IBD postratamiento varían en función del tratamiento empleado.

Tabla 7.1. Frecuencia de respuesta al nivel de tratamiento.

Rango de puntuaciones en el IBD (postratamiento)	*Psicoterapia*		*Terapia de relajación*		*Farmacoterapia*		*Terapia conductual*		Total
Moderado a severo (puntuación ≥ 23)	13	**8.40*	8	**8.21*	10	**9.36*	3	**8.02*	34
Benigno ($7 <$ puntuaciones < 23)	20	**21.75*	23	**21.26*	27	**24.22*	18	**20.76*	88
Normal (puntuación ≤ 7)	11	**13.84*	12	**13.53*	12	**15.42*	21	**13.21*	56
Total	44		43		49		42		178

* Frecuencia esperada.

Partición de los grados de libertad en tablas de contingencia $r \times k$

Si al analizar una tabla de contingencia $r \times k$ rechazamos H_0, el investigador puede concluir, con toda seguridad, que los k grupos difieren en la variable medida. Sin embargo, aunque se puede concluir que los k grupos son diferentes, el resultado de la ji cuadrada en sí mismo no dice al investigador dónde se encuentra la diferencia. Esto es, una X^2 significativa sólo nos dice *que en algún lado* de la tabla las frecuencias observadas no son simplemente desviaciones por oportunidad de las frecuencias teóricas o esperadas. A la mayoría de los investigadores les gustaría saber en qué parte de la tabla de contingencia las discrepancias son más importantes. El procedimiento de partición descrito en esta sección capacita al investigador para realizar un análisis adicional de una tabla de contingencia para la que la X^2 es significativa. En el capítulo 5 se describieron procedimientos para dividir tablas de contingencia de $r \times 2$. Los procedimientos para dividir una tabla de $r \times k$ son similares.

Para dividir una tabla de contingencia, se debe construir una serie de subtablas de 2×2 (una por cada grado de libertad). Por conveniencia para la exposición, comenzamos la partición en la esquina superior izquierda de la tabla; las particiones sucesivas se construyen combinando los renglones y columnas de manera apropiada. Puesto que la medición de la variable se encuentra en escala nominal y los k grupos pueden ser detallados en cualquier orden sin cambiar la X^2 del conjunto global, la tabla puede ser arreglada *a priori* de tal modo que tengan sentido en el contexto del problema sujeto a investigación.

Para ilustrar el método, enumeramos las particiones para una tabla de contingencia 3×3. La primera partición consiste en las cuatro frecuencias de la esquina superior izquierda de la tabla:

$$
\begin{array}{c|c}
n_{11} & n_{12} \\
\hline
n_{21} & n_{22}
\end{array}
$$

(1)

La segunda partición se forma al "replegar las columnas de la primera partición de 2 $\times$ 2 para formar el primer renglón de la segunda partición:

$$
\begin{array}{c|c}
n_{11} & n_{12} \\
+ & + \\
n_{21} & n_{22} \\
\hline
n_{31} & n_{32}
\end{array}
$$

(2)

Las particiones restantes para la tabla de 3 $\times$ 3 son las siguientes:

$$
\begin{array}{c|c}
n_{11} + n_{12} & n_{13} \\
\hline
n_{21} + n_{22} & n_{23}
\end{array}
\qquad\qquad
\begin{array}{c|c}
n_{11} + n_{12} & n_{13} \\
+ & + \\
n_{21} + n_{22} & n_{23} \\
\hline
n_{31} + n_{32} & n_{33}
\end{array}
$$

(3) (4)

Aunque el arreglo puede parecer "voluminoso", el sistema es realmente sencillo. La celdilla inferior derecha de una partición asociada con la ij-ésima celdilla consiste de una sola frecuencia (n_{ij}), la celdilla superior izquierda es la sumatoria de todas las frecuencias "de arriba" y a "la izquierda" de la ij-ésima celdilla. La frecuencia inferior izquierda es la sumatoria de las frecuencias a la izquierda de la ij-ésima celdilla, y la celdilla superior derecha es la sumatoria de las frecuencias que están por encima de la ij-ésima celdilla.

Como se vio en el capítulo 5, cada partición se evalúa utilizando la prueba estadística ji cuadrada. Sin embargo, no es apropiado utilizar la fórmula "usual" para la X^2 ya que las frecuencias esperadas deben ajustarse para cada tabla de 2 $\times$ 2 para reflejar la tabla entera (y la población) y no solamente la subtabla representada en la partición.

El valor de la X^2 dividida para la t-ésima partición nos lo proporciona la siguiente ecuación:

$$
X_t^2 = \frac{N\left[C_j\left(R_i \sum_{h=1}^{i-1} \sum_{m=1}^{j-1} n_{hm} \quad \sum_{h=1}^{i-1} R_h \sum_{m=1}^{j-1} n_{im}\right) - \sum_{m=1}^{j-1} C_m \left(R_i \sum_{h=1}^{i-1} n_{hj} - n_{ij} \sum_{h=1}^{i-1} R_h\right)\right]^2}{C_j R_i \sum_{m=1}^{j-1} C_m \sum_{m=1}^{j} C_m \sum_{h=1}^{i-1} R_h \sum_{h=1}^{i} R_h} \tag{7.1}
$$

donde $t = i + (r - 1)(j - 2) - 1$. Vale decir, X_t^2 es la prueba para la t-ésima partición asociada con la ij-ésima celdilla. Cada uno de los estadísticos X_t^2 se distribuye como ji cuadrada con $gl = 1$. Los valores de X_t^2 de todas las particiones se suman para obtener el valor de X^2 de la tabla completa. El cálculo de X_t^2 con la ecuación (7.1) es sencillo, aunque parezca muy complicado. Por tanto, puede utilizarse el programa para computadora del Apéndice para realizar los cálculos. (Este programa se emplea también en la aplicación de los procedimientos de partición descritos en el capítulo 5.)

Ejemplo. En el ejercicio anterior, referente a la eficacia de varios tratamientos para la depresión clínica de pacientes externos, encontramos que las puntuaciones postratamiento en el IBD fueron diferentes de acuerdo con las distintas condiciones. ($X^2 = 12.80$ con $gl = 6$ para los datos de la tabla 7.1.) Aunque el investigador puede concluir, con toda confianza, que existen diferencias evidentes entre los tratamientos, es deseable determinar si la eficacia diferencial de los tratamientos varió entre todos los tratamientos o si ésta se concentró en uno o dos de ellos.

Para determinar dónde se encuentran en la tabla 7.1 las diferencias en los efectos de los tratamientos, dicha tabla fue dividida. Las particiones correspondientes se muestran en la tabla 7.2. Al final de la tabla se presentan las X^2 divididas. En las primeras dos particiones se comparan los resultados de la psicoterapia con los resultados de la terapia de relajación. En la primera partición se comparan las puntuaciones en el IBD de severo a moderado con el benigno, para estos dos grupos. Con la ecuación (7.1) (o con el programa para computadora) el valor resultante de $X_1^2 = 1.62$. X_1^2 se distribuye como ji cuadrada con $gl = 1$, el cual, claramente, no es significativo. La siguiente partición consiste en combinar las puntuaciones de severo a moderado y benigno y su comparación con los sujetos que obtuvieron una puntuación de normal en el IBD. Para ésta, el valor de $X_2^2 = 0.09$; este valor tampoco es significativo. Ahora, podemos concluir que no existen diferencias entre las puntuaciones en el IBD en los tratamientos de psicoterapia, terapia de relajación y farmacoterapia.

Estos dos grupos de tratamiento fueron combinados y comparados con el grupo de farmacoterapia en términos de las puntuaciones en el IBD postratamiento. Los resultados de estas particiones fueron $X_3^2 = 0.42$ y $X_4^2 = 0.06$. Dichos resultados permiten concluir al investigador que no existen diferencias entre los tratamientos de psicoterapia, terapia de relajación y farmacoterapia.

Lo que resta es comparar la terapia conductual con las otras tres terapias. Los primeros tres grupos se mezclan y la distribución de puntuaciones combinadas en el IBD se comparan con las puntuaciones del grupo de terapia conductual. Las particiones relevantes son la (5) y (6) de la tabla 7.2. Los valores resultantes de X^2 son 1.84 y 8.76, respectivamente. Así, tenemos que X_6^2 es la única X^2 significativa asociada con las particiones. Por tanto, sería apropiado que el investigador concluya que para este estudio, la psicoterapia, la terapia de relajación y la farmacoterapia produjeron resultados similares. Sin embargo, la terapia conductual es diferente de las otras tres terapias. El revisar con cuidado la tabla 7.1 y los valores de X^2 de partición nos muestra el *locus* de la diferencia: significativamente más sujetos en el grupo de terapia conductual obtuvieron puntuaciones de normal en el IBD.

Se advierte que la secuencia que se siga en el procedimiento de partición puede tener un efecto sobre las X^2 resultantes. Si la tabla de contingencia original se divide en un orden diferente, es decir, si los renglones, columnas o ambos son rearreglados, los valores de las particiones, con mucha certeza, variarán. Para un adecuado uso e interpretación de el análisis de partición, es importante que el investigador sea capaz de especificar un orden de partición *a priori* que tenga sentido en el contexto particular del problema de investigación.

Tabla 7.2. Particiones derivadas de la tabla de contingencia 7.1.

IBD	*Psicoterapia*	*Terapia de relajación*
Moderado a severo	13	8
Benigno	20	23

(1)

IBD	*Psicoterapia*	*Terapia de relajación*
Moderado a severo + benigno	33	31
Normal	11	12

(2)

IBD	*Psicoterapia + terapia de relajación*	*Farmaco-terapia*
Moderado a severo	21	10
Benigno	43	27

(3)

IBD	*Psicoterapia + terapia de relajación*	*Farmaco-terapia*
Moderado a severo + benigno	64	37
Normal	23	12

(4)

IBD	*Psicoterapia + terapia de relajación + farmaco-terapia*	*Terapia conductual*
Moderado a severo	31	3
Benigno	70	18

(5)

IBD	*Psicoterapia + terapia de relajación + farmaco-terapia*	*Terapia conductual*
Moderado a severo + benigno	101	21
Normal	35	21

(6)

Resumen de la X^2 dividida

Partición	X^2
1	1.62
2	0.09
3	0.42
4	0.06
5	1.84
6	8.76
Total de conjunto	12.80

Si se desea construir la partición *a posteriori*, se pueden utilizar los procedimientos descritos en este capítulo. Sin embargo, el valor crítico para la significación de cada partición debe ser cambiado. La región de probabilidad debe cambiarse de α a α/p, donde p = número de particiones. Así, tenemos que, si algún investigador quiso construir las seis particiones *a posteriori* en el ejemplo anterior, los valores críticos de X^2 en los que tuvo que basarse son $\alpha/p = 0.05/6 = 0.0083$.

Análisis de residuos

Cuando el valor obtenido de X^2 para una tabla de contingencia $r \times k$ es significativo, el método de partición ayuda al investigador a determinar dónde se encuentran las diferencias en la tabla. El método descrito anteriormente a menudo basta para evidenciar las diferencias. Sin embargo, para algunas tablas, después de aplicar al método de partición, el investigador puede desear (aún) analizar los datos de alguna manera adicional que le permita entender mejor dónde se encuentran las diferencias. Es posible complementar el método de partición analizando los residuos (las discrepancias entre los valores esperados y los observados), a fin de determinar cuáles son mayores que los que se podrían esperar por oportunidad. El residuo e_{ij}, para la ij-ésima celdilla en una tabla de $r \times k$ es el siguiente:

$$e_{ij} = \frac{n_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$

donde $E_{ij} = R_i C_j / N$. La varianza de este residuo puede estimarse mediante

$$v_{ij} = \frac{1 - R_i/N}{1 - C_j/N} = \frac{N - R_i}{N - C_j}$$

El residuo *ajustado* o residuo *estandarizado* para la ij-ésima celdilla se calcula mediante

$$d_{ij} = \frac{n_{ij} - E_{ij}}{\sqrt{E_{ij}}} \sqrt{\frac{N - C_j}{N - R_i}} \qquad (7.2)$$

Conforme la muestra de tamaño N incrementa su tamaño, la d_{ij} tiene una distribución aproximadamente normal con media igual a cero y varianza igual a uno. Así, tenemos que la sigificación de d_{ij} puede determinarse mediante la tabla A del Apéndice I. El investigador que analice los residuos debe ser cuidadoso de que las d_{ij} no sean independientes; por tanto, debe ser precavido en la interpretación de los resultados. Un procedimiento prudente consiste en combinar el análisis de los residuos con el análisis de las particiones.

Para el ejemplo anterior, el residuo ajustado d_{11} para la celdilla superior izquierda se calcula mediante la ecuación (7.2), de la siguiente manera:

$$\begin{aligned} d_{11} &= \frac{13 - 8.4}{\sqrt{8.4}} \sqrt{\frac{178 - 44}{178 - 34}} \\ &= 1.53 \end{aligned}$$

Las d_{ij} restantes del primer renglón son $-$ 0.07, 0.20, $-$ 1.72. Las d_{ij} para el segundo y tercer renglones son $-$ 0.46, 0.46, 0.67, $-$ 0.75, $-$ 0.80, $-$ 0.44, $-$ 0.89 y 2.26, respectivamente. La única diferencia significativa en el nivel $\alpha = 0.05$ (bidireccional) es la de la celdilla inferior derecha (d_{34}) de la tabla 7.1. Este resultado agrega fuerza adicional al argumento de que la diferencia encontrada entre la terapia conductual y las otras tres terapias ha producido el efecto encontrado en la tabla: el uso de la terapia conductual trae como resultado, significativamente más puntuaciones en el IBD normal que el empleo de las otras terapias.

Resumen del procedimiento

Éstos son los pasos que hay que seguir para la aplicación de la prueba ji cuadrada para k muestras o grupos independientes:

1. Arregle las frecuencias observadas en una tabla de contingencia $r \times k$, utilizando las k columnas para las muestras o grupos.
2. Determine la frecuencia esperada según H_0 para cada celdilla encontrando el producto de los totales marginales comunes a la celdilla y dividiendo este producto entre N. Vale decir, encuentre las frecuencias esperadas $E_{ij} = R_1C_j/N$. (N es la sumatoria de cada uno de los totales marginales y representa el número total de observaciones *independientes*. Cuando se "inflan" los valores de N debido a múltiples observaciones de cada sujeto, esto invalida la prueba.) Si las frecuencias esperadas son pequeñas, combine categorías tal como se muestra en la siguiente sección.
3. Calcule X^2 utilizando las ecuaciones (5.2) o (5.2*a*). Determine los grados de libertad $gl = (r - 1)(k - 1)$.
4. Determine la significación de la X^2 observada recurriendo a la tabla C del Apéndice I. Si la probabilidad para el valor observado de X^2 con un valor particular de grados de libertad (el que corresponda) es menor o igual que α, rechace la hipótesis nula en favor de H_1.
5. Si se rechaza H_0, el valor global de X^2 puede ser dividido, utilizando la ecuación (7.1) para determinar dónde se encuentran (en la tabla de contingencia original) las diferencias en la variable medida en los distintos grupos. Cada uno de los valores de X^2 de partición se distribuyen como ji cuadrada con $gl = 1$, conforme incrementa el tamaño de N. Después de dividir la tabla, pueden analizarse los residuos (diferencias entre los valores esperados y los valores observados) mediante la ecuación (7.2).

Cuándo utilizar la prueba ji cuadrada

Para aplicación adecuada de la prueba ji cuadrada se requiere que las frecuencias esperadas (las E_{ij}) en cada celdilla no sean demasiado pequeñas. Cuando se viola este requisito, los resultados de la prueba no pueden ser interpretados porque la distribución de X^2 no está bien aproximada a la distribución de la ji cuadrada de la tabla C del Apéndice I. Cochran (1954) recomienda que en pruebas ji cuadrada donde los grados de libertad son mayores que uno (esto es, que r o k sean mayores que dos), el número de celdillas que presenten frecuencias esperadas menores que cinco, no debe ser mayor del 20 % y ninguna celdilla debe presentar frecuencias esperadas menores que uno.[2]

Si los requisitos no son cubiertos por los datos de acuerdo con la forma en que se recolectaron y no fue posible contar con una muestra grande, el investigador debe combinar categorías para incrementar las E_{ij} en varias celdillas. Sólo después de combinar categorías y después de que menos del 20 % de las celdillas tengan frecuencias esperadas menores que cinco, y ninguna celdilla tenga una frecuencia esperada menor que uno, el investigador puede interpretar útilmente los resultados de la prueba ji cuadrada. La combinación de las categorías debe hacerse juiciosamente. Esto es, los resultados de la prueba estadística pueden no ser interpretables si las categorías han sido combinadas caprichosamente. Las categorías que son combinadas deben tener alguna propiedad en común o identificación mutua, por si la interpretación del resultado de la prueba, después de haber combinado los renglones o columnas, tiene cierta utilidad. El investigador debe ser precavido contra la necesidad de combinar categorías si se utiliza una muestra suficientemente grande.

La prueba ji cuadrada es insensible a los efectos de orden, Así, cuando se ordenan las categorías de respuesta o los grupos (o ambos), la prueba ji cuadrada puede no ser la mejor. Cochran (1954) ha presentado métodos que fortalecen la prueba ji cuadrada común cuando H_0 es verdadera contra alternativas específicas.

Finalmente, debe notarse que la prueba ji cuadrada es aplicada a frecuencias. Así, es importante utilizar las ecuaciones (5.2) o (5.2*a*) en las frecuencias de los datos. No es correcto emplear porcentajes o cualquier otra transformación de los datos al aplicar la prueba.

Potencia

Generalmente no existe una opción clara a la prueba de la ji cuadrada cuando ésta se utiliza en datos nominales y, por tanto, no es posible calcular un valor exacto de su potencia. Sin embargo, Cochran (1952) ha mostrado que el límite de la potencia de la distribución de la prueba ji cuadrada tiende a uno conforme se incrementa el tamaño de N.

Referencias bibliográficas

Para explicaciones adicionales de la prueba ji cuadrada, el lector puede consultar Cochran (1952, 1954), Delucchi (1983), Everitt (1977), Lewis y Burke (1949) y

[2] El lector notará que el modo empírico con el cual se resta atención a las frecuencias esperadas pequeñas, parece ser algo arbitrario. Esto es así por que las autoridades difieren en "qué tan cerca" debe estar la distribución muestral de X^2 respecto a la distribución de la ji cuadrada para que ésta sea considerada suficientemente buena. Esto hace parecer que el mayor número de renglones y columnas en la tabla de contingencia, el menor número de valores esperados pequeños, puede hacer que la aproximación sea buena. (Un investigador comprobó que para una tabla de 50 celdillas y *todos* los valores esperados menores que uno, la aproximación era muy buena. Nosotros esperamos no tener que utilizar una tabla como ésa en nuestra práctica cotidiana.)

McNemar (1969). Los procedimientos de partición se examinan en detalle en Castellan (1965). Lienert y Netter (1987) describen ajustes a los procedimientos de partición cuando el orden no se determina *a priori*. Shaffer (1973) explica una alternativa a la partición. El método para el análisis de los residuos ha sido estudiado por Haberman (1973).

EXTENSIÓN DE LA PRUEBA DE LA MEDIANA

Función

La *extensión de la prueba de la mediana* determina si k grupos independientes (no necesariamente del mismo tamaño) se han extraído de la misma población o de poblaciones distintas, pero con medianas iguales. Esta prueba es útil y apropiada cuando la variable en estudio ha sido medida en, al menos, escala ordinal. Es particularmente apropiada cuando, por alguna razón, no ha sido posible observar el valor exacto de las puntuaciones extremas, es decir, cuando algunos de los datos observados se encuentran por arriba del punto de corte.

Método

Para aplicar la extensión de la prueba de la mediana, primero debemos determinar el valor de la mediana para las puntuaciones de las k muestras combinadas, es decir, encontramos la mediana *común* para todas las puntuaciones en los k grupos. Después, debemos reemplazar cada puntuación por un signo de más (+) si la puntuación es mayor que la mediana o por un signo de menos (−) si éste es más pequeño que la mediana. (Puede suceder que una o más puntuaciones coincidan con el valor de la mediana; entonces, se deben dicotomizar las puntuaciones, es decir, se asigna un más (+) si la puntuación es mayor que la mediana o un menos (−) si la puntuación es igual o menor que la mediana.)

Podemos presentar los conjuntos de puntuaciones resultantes en una tabla de contingencia de $2 \times k$, con los números en el cuerpo de la tabla representando las frecuencias de signos de más (+) (puntuaciones por arriba de la mediana) y signos de menos (−) (puntuaciones por abajo de la mediana) en cada uno de los k grupos. Un ejemplo de lo anterior es la siguiente tabla:

	Grupos			
	1	2	···	k
Observaciones por arriba de la mediana	n_{11}	n_{12}	···	n_{1k}
Observaciones por abajo de la mediana	n_{21}	n_{22}	···	n_{2k}

Para probar la hipótesis nula de que las k muestras provienen de la misma población con respecto a las medianas, debemos calcular el valor del estadístico X^2 utilizando las ecuaciones (5.2) o (5.2*a*):

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{5.2}$$

o

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{n_{ij}^2}{E_{ij}} - N \tag{5.2a}$$

donde

n_{ij} = número de casos observados que fueron categorizados en el i-ésimo renglón de la j-ésima columna
E_{ij} = número de casos esperados en el i-ésimo renglón de la j-ésima columna cuando H_0 es verdadera

y la doble sumatoria es sobre todos los renglones y columnas de la tabla (es decir, sumatoria de todas las celdillas). Los valores de X^2 resultantes de la aplicación de la ecuación (5.2) se distribuyen (para N grande) como ji cuadrada con $gl = (r - 1)(k - 1)$, donde r es el número de renglones y k es el número de columnas (grupos) de la tabla de contingencia. Para la prueba de la mediana $r = 2$, tenemos que

$$gl = (r - 1 (k - 1) = (2 - 1)(k - 1) = k - 1$$

Cuando H_0 es verdadera, la probabilidad asociada con la ocurrencia de valores tan grandes como una X^2 observada, se proporciona en la tabla C del Apéndice I. Si la X^2 observada es igual o mayor que valor de la tabla C para el nivel de significación previamente determinado y para el valor observado de $gl = k - 1$, entonces podemos rechazar H_0 en ese mismo nivel de significación.

Si es posible dicotomizar las puntuaciones exactamente en la mediana, entonces cada E_{ij} es la mitad del total marginal para su columna. Cuando las puntuaciones son dicotomizadas como aquellas que son mayores e iguales y menores que la mediana, el método para encontrar las frecuencias esperadas es el que se presentó en la sección "Prueba ji cuadrada para dos muestras independientes" del capítulo 5.

Una vez que las puntuaciones se han categorizado como signos de más (+) y signos de menos (−) respecto a la mediana, y las frecuencias resultantes se han arreglado en una tabla de contingencia de $2 \times k$, el procedimiento de cálculo para esta prueba es exactamente el mismo que para la ji cuadrada para k muestras independientes, que se desarrolla en el primer tema de este capítulo. Esto se ilustra con el ejemplo siguiente.

Ejemplo. Supongamos que un investigador en educación desea estudiar la influencia del nivel escolar de las madres sobre el interés en sus hijos en edad escolar. Como medida del nivel de escolaridad, el investigador utiliza el grado escolar más alto que la madre haya completado, y como medida del interés en sus hijos se toma como índice el número de visitas voluntarias que cada madre realiza durante el año escolar, es decir, obras de teatro de los hijos, reuniones y conferencias para los padres de familia de parte de los maestros y funcionarios de la escuela, etc. El investigador obtiene una muestra al azar del 10 % de los 440 niños inscritos en la escuela, de esta lista se obtienen los nombres de las 44 madres que pertenecen a la muestra. La hipótesis es que el número de visitas a la escuela por parte de una madre varía de acuerdo con el número de años escolares que esa madre completó.

i. *Hipótesis nula.* H_0: no existe diferencia en la frecuencia de las visitas a la escuela entre las madres con diferente nivel de escolaridad; esto es, la frecuencia en las visitas es independiente de la cantidad de años escolares que las madres hayan completado. H_1: la frecuencia de visitas a la escuela por parte de las madres difiere dependiendo de su nivel de escolaridad.
ii. *Prueba estadística.* Puesto que los grupos de madres con distintos niveles educativos son independientes y ya que se utilizan varios grupos, es recomendable usar una prueba para k muestras o grupos independientes. En virtud de que el número de años escolares estudiados constituye una medida en escala ordinal del nivel educativo y ya que el número de visitas a la escuela es una medida ordenada (ordinal) del interés en el desempeño escolar de los niños, es apropiada la extensión de la prueba de la mediana para evaluar la hipótesis concerniente a las diferencias entre las medianas de cada grupo.
iii. *Nivel de significación.* $\alpha = 0.05$ y N es el número de madres pertenecientes a la muestra $= 44$.
iv. *Distribución muestral.* Según la hipótesis nula, el estadístico X^2 calculado mediante la ecuación (5.2) se distribuye aproximadamente como ji cuadrada con $gl = k - 1$ cuando $r = 2$. (En la prueba de la mediana, el número de renglones r en la tabla de contingencia asociada es siempre dos.) La probabilidad asociada con la ocurrencia de valores tan grandes como un X^2 observada cuando H_0 es verdadera nos la proporciona la tabla C del apéndice I.
v. *Región de rechazo.* La región de rechazo consiste en todos los valores de X^2 que sean tan grandes que la probabilidad asociada con su ocurrencia cuando H_0 es verdadera, sea menor o igual que $\alpha = 0.05$.
vi. *Decisión.* En este ejemplo (ficticio), el investigador recolectó los datos que se presentan en la tabla 7.3. La mediana combinada para estos 44 datos es 2.5. Vale decir, la mitad de las madres visitaron la escuela de sus hijos dos o menos veces durante el año escolar, y la otra mitad visitó la escuela tres veces o más. Estos datos se dividen por la mediana combinada para obtener los datos que se presentan en la tabla 7.4, la cual nos muestra el número de madres en cada nivel educativo que están por encima o por debajo de la mediana del número de visitas a la escuela. Por ejemplo, del número de madres cuyo nivel educativo se limitó al nivel elemental (octavo grado), cinco de ellas visitaron la escuela tres o más veces durante el año, y cinco visitaron la escuela dos veces o menos. De aquellas madres que asistieron a algunos años de licenciatura, tres visitaron la escuela tres o más veces y una visitó la escuela dos veces o menos.

En la parte inferior de cada renglón de la tabla 7.4 se presenta el número esperado de visitas de cada grupo, de acuerdo con el supuesto de que H_0 es verdadera. Obsérvese que en las puntuaciones dicotomizadas exactamente por la mediana, la frecuencia esperada es precisamente la mitad de la sumatoria de las frecuencias para la columna donde está ubicada la celdilla. El investigador observó que la mitad de las frecuencias esperadas de la tabla de contingencia son menores que cinco. La distribución muestral del estadístico X^2 no se aproxi-

Tabla 7.3. Número de visitas a la escuela de acuerdo con el nivel educativo de las madres (datos ficticios).

Nivel escolar de la madre					
Escuela elemental (8o. grado)	*Décimo grado*	*Preparatoria (12o. grado)*	*Algunos semestres de licenciatura*	*Licenciatura*	*Posgrado*
4	2	2	9	2	2
3	4	0	4	4	6
0	1	4	2	5	
7	6	3	3	2	
1	3	8			
2	0	0			
0	2	5			
3	5	2			
5	1	1			
1	2	7			
	1	6			
		5			
		1			

Tabla 7.4. Número de visitas a la escuela de acuerdo con el grado educativo de las madres (datos ficticios).

	Nivel escolar de la madre						
	Escuela elemental (8o. grado)	*Décimo grado*	*Preparatoria (12o. grado)*	*Algunos semestres de licenciatura*	*Licenciatura*	*Posgrado*	*Total*
Número de madres cuyas visitas fueron más frecuentes que la mediana	5 *5*	4 *5.5*	7 *6.5*	3 *2*	2 *2*	1 *1*	22
Número de madres cuyas visitas fueron menos frecuentes que la mediana	5 *5*	7 *5.5*	6 *6.5*	1 *2*	2 *2*	1 *1*	22
Total	10	11	13	4	4	2	44

ma muy bien a la distribución de la ji cuadrada cuando más del 20 % de las celdillas tienen frecuencias esperadas menores que cinco. (Véase la explicación acerca de cuándo utilizar la prueba ji cuadrada, en la sección correspondiente de este capítulo.) Por tanto, el investigador decidió combinar categorías a fin de tener frecuencias esperadas suficientemente grandes. Puesto que las categorías con frecuencias esperadas pequeñas abarcan a mujeres con distintos niveles educativos, el investigador decidió mezclar tres de estas categorías, el investigador decidió mezclar tres de estas categorías en una sola: Licenciatura (uno o más años, incluido posgrado).[3, 4] Haciendo esto, se obtuvieron los valores que se muestran en la tabla 7.5. Obsérvese que en dicha tabla, todas las frecuencias esperadas son mayores que cinco.

Podemos calcular el valor del estadístico X^2 sustituyendo los valores de los datos de la tabla 7.5 en la ecuación (5.2).

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{5.2}$$

$$= \frac{(5-5)^2}{5} + \frac{(4-5.5)^2}{5.5} + \frac{(7-6.5)^2}{6.5} + \frac{(6-5)^2}{5}$$

$$+ \frac{(5-5)^2}{5} + \frac{(7-5.5)^2}{5.5} + \frac{(6-6.5)^2}{6.5} + \frac{(4-5)^2}{5}$$

$$= 0 + 0.409 + 0.0385 + 0.2 + 0 + 0.409 + 0.0385 + 0.2$$

$$= 1.295$$

Mediante el cálculo anterior determinamos que $X^2 = 1.295$ y $gl = k - 1 = 4 - 1 = 3$. La tabla C del Apéndice nos revela que un valor de $X^2 \geq 1.295$ para $gl = 3$ tiene una probabilidad de ocurrencia entre 0.80 y 0.70 cuando H_0 es verdadera. Puesto que esta probabilidad es mayor que nuestro nivel de significación $\alpha = 0.05$, nuestra decisión es que con base en estos datos (ficticios) no podemos rechazar la hipótesis nula de que el interés en la educación de sus niños (medido en términos de las visitas escolares hechas por las madres) es independiente del nivel escolar de las madres.

[3] Cabe destacar que para este ejemplo en particular, el valor esperado *a priori* en cada celdilla de la tabla original es $44/12 = 3.67 < 5$. Hay varias maneras de resolver el problema de los valores esperados pequeños. Pudiera haberse seleccionado una muestra de mayor tamaño, el número de categorías de nivel educativo pudieron haber sido menos, o combinarse las categorías después de recabar los datos. El investigador seleccionó este último. Utilizar una muestra de mayor tamaño no solamente es más costoso, sino que tampoco asegura que los valores esperados serán suficientemente grandes. La utilización *a priori* de menos categorías no sólo sacrifica información, sino que tampoco asegura que los valores esperados serán suficientemente grandes. La elección realizada por el investigador parece ser la mejor.

[4] La elección particular de combinar los grupos tiene la ventaja adicional de hacer que las frecuencias esperadas en cada grupo sean muy similares. El poder de la ji cuadrada es mayor cuando las frecuencias esperadas en cada celdilla son iguales.

Tabla 7.5. Número de visitas a la escuela de acuerdo con el grado educacional de las madres (datos ficticios).

	Nivel escolar de la madre				
	Escuela elemental (8o. grado)	*Décimo grado*	*Preparatoria (12o. grado)*	*Licenciatura (uno o más años)*	*Total*
Número de madres cuyas visitas fueron más frecuentes que la mediana	5 *5*	4 *5.5*	7 *6.5*	6 *5*	22
Número de madres cuyas visitas fueron menos frecuentes que la mediana	5 *5*	7 *5.5*	6 *6.5*	4 *5*	22
Total	10	11	13	10	44

Resumen del procedimiento

Éstos son los pasos que hay que seguir durante la aplicación de la extensión de la prueba de la mediana:

1. Determine la mediana común de las puntuaciones de los k grupos.
2. Asigne signos de más (+) a todas las puntuaciones que se encuentren por encima de la mediana y signos de menos (−) a todos aquellas que se encuentren por debajo de la mediana, por ello, cada uno de los grupos de puntuaciones se dividió en dos categorías (las anteriores). Presente las frecuencias resultantes en una tabla de contingencia de $2 \times k$.
3. Utilizando los datos de la tabla que se ha formado, calcule el valor de X^2 mediante la ecuación (5.2) o (5.2a). Determine los grados de libertad $gl = k - 1$.
4. Determine la significación del valor observado de X^2 utilizando la tabla C del Apéndice I. Si la probabilidad asociada dada por los valores tan grandes como la X^2 observada es igual o menor que α, rechace H_0 en favor de la H_1.

Como hemos mencionado, la extensión a la prueba de la mediana es, de hecho, una prueba ji cuadrada para k muestras o grupos independientes. Si existen varios grupos, el investigador puede desear dividir la tabla de contingencia para determinar dónde se encuentran las diferencias. Para información concerniente a las condiciones en las cuales la prueba se emplea adecuadamente, y acerca del poder de la prueba, el lector puede recurrir a la exposición del tema en la sección anterior de este capítulo. En la siguiente sección examinaremos una prueba que es una alternativa más poderosa que se utiliza cuando los datos pueden ordenarse completamente de acuerdo con las mediciones de la variable.

Referencias bibliográficas

El lector encontrará análisis relevantes en las referencias que se detallan al final de la sección anterior.

ANÁLISIS DE VARIANZA UNIFACTORIAL POR RANGOS, DE KRUSKAL-WALLIS

Función

El análisis de varianza unifactorial por rangos, de Kruskal-Wallis, es una prueba extremadamente útil para decidir si k muestras independientes provienen de diferentes poblaciones. Los valores de la muestra invariablemente difieren de alguna manera, y la pregunta es si las diferencias entre las muestras significan diferencias genuinas en la población o si sólo representan la clase de variaciones que pueden esperarse en muestras que se obtienen al azar de la misma población. La técnica de Kruskal-Wallis prueba la hipótesis nula de que las k muestras provienen de la misma población o de poblaciones idénticas con la misma mediana. Para especificar explícitamente las hipótesis nula y alterna, θ_j debe ser la mediana de la población para el j-ésimo grupo o muestra. Entonces podemos escribir la hipótesis nula de que las medianas son las mismas como H_0: $\theta_1 = \theta_2 = \ldots \theta_k$; y la hipótesis alterna puede ser escrita como H_1: $\theta_i \neq \theta_j$ para algunos grupos i y j. Esto es, si la hipótesis alterna es verdadera, al menos un par de grupos tienen medianas diferentes. Según la hipótesis nula, la prueba supone que las variables en estudio tienen la misma distribución subyacente; además, requiere que las mediciones de la variable se encuentre, al menos, en escala ordinal.

Racionalización y método

En la aplicación del análisis de varianza unifactorial por rangos, de Kruskal-Wallis, los datos se presentan en una tabla de doble entrada donde cada columna representa cada grupo o muestra sucesiva. Así, los datos deben arreglarse de la siguiente manera:

Grupo			
1	2	$\cdots$	k
X_{11}	X_{12}	$\cdots$	X_{1k}
X_{21}	X_{22}	$\cdots$	X_{2k}
$\vdots$			$\vdots$
$X_{n_1 1}$	$\vdots$	$\cdots$	$X_{n_k k}$
	$X_{n_2 2}$		

donde X_{ij} es el dato para la i-ésima observación en el j-ésimo grupo y n_{ij} es el número de observaciones en el j-ésimo grupo.

En el cálculo de la prueba de Kruskal-Wallis, cada una de las N observaciones se reemplaza por un rango (el que le corresponda). Esto es, todas las puntuaciones de todas las k muestras se combinan en una sola serie y se ordenan por rangos. La puntuación más pequeña se reemplaza por el rango uno, la puntuación que le sigue en tamaño (ascendente) se reemplaza por el rango dos, y la puntuación mayor se reemplaza por el rango N, donde N es el número total de observaciones independientes de las k muestras.

Cuando lo anterior se hubo realizado, se debe encontrar la suma de rangos en cada muestra (columna). A partir de estas sumatorias de rangos, podemos calcular los rangos promedio para cada muestra. Ahora, si las muestras provienen de la misma o de idénticas poblaciones, los rangos promedio deberían ser (aproximadamente) los mismos, mientras que las muestras provienen de poblaciones con medianas diferentes, los rangos promedio deberán ser distintos. La prueba de Kruskal-Wallis evalúa la diferencia entre los rangos promedios para determinar si son lo suficientemente dispares, de tal suerte que no sea probable que las muestras hayan sido extraídas de la misma población.

Presentaremos dos formas para la prueba Kruskal-Wallis y los términos necesarios para calcular el estadístico Kruskal-Wallis (KW):

$$KW = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \, (\bar{R}_j - \bar{R})^2$$

o

$$KW = \left[\frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \, \bar{R}_j^2 \right] - 3(N+1) \qquad (7.3)$$

donde

k = número de muestras o grupos
n_j = número de casos en la j-ésima muestra
N = número de casos en la muestra combinada (la suma de n)

R_j = sumatoria de los rangos en la j-ésima muestra o grupo
$\overline{R}_j$ = promedio de los rangos en la j-ésima muestra o grupo
$\overline{R}$ = $(N + 1)/2$ = promedio de los rangos en la muestra combinada (la gran media)

y la sumatoria se realiza en todas las k muestras.

Si las k muestras realmente se extrajeron de la misma población o de poblaciones idénticas, esto es, si H_0 es verdadera, entonces puede calcularse la distribución muestral del estadístico KW, y tabularse la probabilidad de observar diferentes valores de KW. Sin embargo, cuando hay más de $k = 3$ grupos y cuando el número de observaciones en cada grupo es mayor que cinco la distribución muestral de KW está bien aproximada a la distribución de la ji cuadrada con $gl = k - 1$. La aproximación se mejora cuando tanto el número de grupos, k, como el número de observaciones en cada grupo, n, se incrementan. Así, tenemos que cuando hay más de cinco casos en los distintos grupos, esto es, cuando todas las $n_j > 5$, y cuando H_0 es verdadera, la probabilidad asociada con valores tan grandes como un KW observado puede determinarse haciendo referencia a la tabla C del Apéndice I. Si el valor del KW observado es igual o mayor que el valor de tabla de X^2 proporcionado por tabla C en el nivel de significación previamente determinado y para los grados de libertad $gl = k - 1$, entonces podemos rechazar H_0 en ese nivel de significación.

Cuando $k = 3$ y el número de casos en cada una de las tres muestras es igual o menor que cinco, las probabilidades asociadas con cada KW puede obtenerse de la tabla O del Apéndice I. Ésta proporciona significaciones seleccionadas de valores de KW para n_1, n_2 y n_3 pequeñas, esto es, para muestras de tamaño mayor que cinco. Estas probabilidades son aquellas asociadas con la ocurrencia de valores tan grandes o mayores que un valor de tabla de KW cuando H_0 es verdadera.

Ejemplo para muestras pequeñas. En un estudio experimental de toma de decisiones, los investigadores han dedicado esfuerzo teórico y empírico para entender tareas de decisión que son aprendidas en una manera prudente. En una serie de estudios donde se les pidió a los sujetos que aprendieran la relación de dos señales respecto a un resultado probabilístico, una tarea requería que los sujetos aprendieran relaciones funcionales del tipo $X + Y + c = Z$, en las que X y Y estaban relacionadas probabilísticamente con el criterio Z y c era una constante arbitraria. Los sujetos aprendieron la tarea fácilmente cuando se proporcionaban las señales X y Y. Sin embargo, las primeras investigaciones sugirieron que si la tarea era dividida en dos partes, es decir, aprender la relación entre *una* señal y el resultado y entonces aprender la relación de *ambas* señales con el resultado, los sujetos tendrían considerable dificultad en aprender la tarea compuesta. En un estudio,[5] una señal era un predictor válido (pero imperfecto) del resultado, mientras que la otra señal, Y, no estaba relacionada con el resultado y no tenía utilidad a menos que la señal X fuera presentada al mismo tiempo. Para evaluar la habilidad de la gente en hacer predicciones en este tipo de tarea y aprender tareas de inferencias más complejas, los sujetos fueron divididos en tres grupos (a uno de los grupos se le presentaban ambas señales; a otro grupo se le presentaba sólo la señal válida en primer lugar y al tercer grupo sólo se le presentaban señales irrelevantes en primer lugar). Para este ejemplo, los datos consisten en las predicciones de los sujetos en el estadio

[5] Castellan, N. J. Jr. y Jenkins, R., *Deprivation conditions in multiple-cue probability learning* (manuscrito inédito).

final del experimento en el cual se presentaban ambas señales. El índice de ejecución es un estadístico que proporciona la exactitud del juicio del sujeto. La hipótesis de investigación fue que los tres grupos diferirían en su ejecución terminal en la tarea de predicción.

i. *Hipótesis nula.* H_0: no existen diferencias en la mediana de la ejecución de los sujetos en las tres tareas de predicción. H_1: los grupos difieren respecto a la ejecución en las tareas de predicción.

ii. *Prueba estadística.* En virtud de que se van a estudiar tres grupos, es necesario utilizar una prueba para k muestras independientes. Adicionalmente, puesto que el índice de exactitud en los juicios está medido en una escala ordinal, es apropiado utilizar la prueba de Kruskal-Wallis.

iii. *Nivel de significación*: $\alpha = 0.05$, N es el número total de sujetos en el estudio $= 12$, n_1 es el número de sujetos en la condición de primero señales irrelevantes $= 3$, n_2 es el número de sujetos en la condición de primero señales válidas $= 4$ y n_3 es el número de sujetos que aprendieron utilizando ambas señales $= 5$.

iv. *Distribución muestral.* Para $k = 3$ y n_j pequeña, la distribución muestral de KW es proporcionada por la tabla O del Apéndice I.

v. *Región de rechazo.* La región de rechazo consiste en todos los valores de KW que son tan grandes que la probabilidad asociada con su ocurrencia, cuando H_0 es verdadera, es igual o menor que $\alpha = 0.05$.

vi. *Decisión.* El índice de la exactitud en el juicio para cada sujeto en cada condición del estudio se presenta en la tabla 7.6. Si ordenamos por rangos los 12 datos de manera ascendente, obtenemos los rangos que se muestran en la tabla 7.7. Estos rangos se suman en cada uno de los tres grupos para obtener $R_1 = 17$, $R_2 = 21$ y $R_3 = 40$, como se muestra en la tabla 7.7. En ésta se pueden encontrar los rangos promedio por cada grupo: 5.67, 5.25 y 8.00, respectivamente.

Tabla 7.6. Índices de la exactitud de juicios para sujetos que aprenden la relación $X + Y + c$.

Entrenamiento		
Primero señal irrelevante	*Primero señal válida*	*Ambas señales*
0.994	0.795	0.940
0.872	0.884	0.979
0.349	0.816	0.949
	0.981	0.890
		0.978

Tabla 7.7. Índices de la exactitud de juicios para sujetos que aprenden la relación $X + Y + c$ (ordenados por rangos).

	Entrenamiento		
	Primero señal irrelevante	*Primero señal válida*	*Ambas señales*
	12	2	7
	4	5	10
	1	3	8
		11	6
			9
R_j	17	21	40
$\overline{R}_j$	5.67	5.25	8

Ahora, con estos datos podemos calcular el valor de *KW* mediante la ecuación (7.3):

$$KW = \left[\frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \bar{R}_j^2\right] - 3(N+1) \tag{7.3}$$

$$= \frac{12}{12(12+1)} [3(5.67)^2 + 4(5.25)^2 + 5(8.00)^2] - 3(12+1)$$

$$= 1.51$$

La tabla O del Apéndice I nos muestra que cuando las n_j son 3, 4 y 5, $KW \geq 1.51$ tiene la probabilidad de ocurrencia según la hipótesis nula de no diferencia entre los grupos mayor que 0.10. Así, tenemos que con estos datos no podemos rechazar H_0.[6]

Observaciones empatadas

Cuando ocurren empates entre dos o más puntaciones (al margen del grupo), a cada puntuación se le asigna el promedio de los rangos en los que ocurrieron los empates.

Puesto que la varianza de la distribución muestral de *KW* es influida por los empates, se puede desear realizar la corrección de los cálculos teniendo en cuenta los empates. Para realizar la corrección del efecto de los empates, *KW* debe calcularse mediante la ecuación (7.3) y dividirlo entre

$$1 - \frac{\sum_{i=1}^{g} (t_i^3 - t_i)}{N^3 - N} \tag{7.4}$$

donde

g = número de grupos de rangos empatados
t_i = número de rangos empatados en el i-ésimo grupo
N = número total de observaciones entre todas las muestras

[6] El lector debe ser cuidadoso respecto a no encontrar diferencias significativas en este ejemplo (y en otros). No rechazar H_0 *no implica* que deba ser aceptada y que, por lo tanto, no existen diferencias entre los grupos. Cuando los tamaños de las muestras son pequeños, sólo se detectan por nuestros procedimientos estadísticos, diferencias relativamente grandes que nos conducirían al rechazo de H_0. Esto es así por que cuando el tamaño de la muestra es pequeño y H_0 es, de hecho, verdadera, la probabilidad de grandes variaciones en los resultados es además muy grande. Como consecuencia, es difícil distinguir entre resultados que meramente reflejan desviaciones por oportunidad (cuando H_0 es verdadera) y diferencias reales (cuando H_1 es verdadera). Si H_0 no es rechazada, entonces puede no haber diferencias entre los grupos: o el tamaño de las muestras es demasiado pequeño y/o la variabilidad de la muestra es demasiado grande y/o las diferencias reales son tan pequeñas que no pueden ser detectadas. Antes de *aceptar* H_0 en tales casos, el investigador debe buscar evidencia con la cual corroborar u obtener datos adicionales. Como nota final, esta precaución no implica que no debamos tener confianza en las diferencias entre grupos, si fuimos capaces de rechazar H_0 en un nivel de significación dado Este argumento se aplica con igual fuerza tanto a las pruebas paramétricas como a las no paramétricas.

Como siempre, la magnitud del factor de corrección depende del tamaño de los grupos con empates, es decir, los valores de t_i tanto como el porcentaje de observaciones implicadas. Este tema se examinó en el capítulo 5.

La expresión general para la corrección por empates de KW es

$$KW = \frac{\left[\dfrac{12}{N(N+1)} \displaystyle\sum_{j=1}^{k} n_j \bar{R}_j^2\right] - 3(N+1)}{1 - \left[\displaystyle\sum_{i=1}^{g} (t_i^3 - t_i)\right] \Big/ (N^3 - N)} \tag{7.5}$$

El efecto de corregir los empates es incrementar el valor de KW y así volver el resultado más significativo que si no se hubiera realizado la corrección. Por tanto, si se rechazó H_0 sin haber realizado la corrección [es decir, utilizando la ecuación (7.3) para calcular KW], se podrá ser igualmente capaz de rechazar H_0 en un nivel de significación más riguroso si se efectúa la corrección para los empates.

Frecuentemente, el efecto de la corrección es despreciable. Si menos del 25 % de observaciones presentan empates, la probabilidad asociada con una KW calculada sin corrección para empates, es decir, haber utilizado la ecuación (7.3), raramente es cambiada en más del 20 % cuando se realiza la corrección, esto es, si KW se calcula mediante la ecuación (7.5).

En el ejemplo siguiente existen varias observaciones empatadas, y el valor de KW se calcula con corrección y sin ella para ilustrar el efecto que los empates tienen en este caso.

Ejemplo para muestras grandes. Como parte de la investigación descrita en el ejemplo para muestras pequeñas de esta sección, otro experimentó se centró en el aprendizaje de relaciones funcionales de la forma $aX + bY = Z$. Como en el experimento descrito anteriormente, la señal X era parcialmente válida, la señal Y era irrelevante y a y b eran constantes. Si un sujeto resolvía el problema, en cada ensayo podía ocurrir una respuesta correcta. Los sujetos fueron asignados a una de tres condiciones (las mismas condiciones del ejemplo anterior, para muestras pequeñas). Los sujetos aprendieron primero con ambos predictores o con alguno de los dos, tanto señales válidas parcialmente como señales irrelevantes, y entonces se realizaban las predicciones utilizando ambos predictores para el balance de la sesión experimental. La hipótesis experimental fue que el modo de recibir el entrenamiento inicial no tendría efecto sobre la ejecución final en las tareas de predicción.

i. *Hipótesis nula*. H_0: no existen diferencias entre los tres grupos en los niveles finales de exactitud en la tarea de predicción. H_1: los grupos difieren en los niveles finales de exactitud en la tarea de predicción.

ii. *Prueba estadística*. Puesto que los tres grupos son independientes, es apropiado utilizar una prueba estadística para k muestras independientes. En virtud de que el índice de exactitud en los juicios es continuo y en una escala ordinal, se satisfacen los supuestos de la prueba de Kruskal-Wallis.

iii. *Nivel de significación*. $\alpha = 0.05$ y N es el número total de sujetos en el experimento $= 18$.

iv. *Distribución muestral*. Ya que el tamaño de la muestra es mayor que 5, la distribución muestral de KW se aproxima a la distribución de ji cuadrada con $gl = k - 1$. Así, la probabilidad asociada con la ocurrencia cuando H_0 es verdadera, de valores tan grandes como un KW observado, puede determinarse utilizando la tabla C del Apéndice I.

v. *Región de rechazo*. La región de rechazo consiste en todos los valores de KW que sean tan grandes que la probabilidad asociada con su ocurrencia cuando H_0 es verdadera y cuando $gl = k - 1 = 2$, sea igual o menor que $\alpha = 0.05$.

vi. *Decisión*. Los valores finales de la exactitud en el juicio se muestran en la tabla 7.8 para cada sujeto en cada condición. Si los $N = 18$ sujetos son ordenados por rangos del menor al mayor, obtenemos los rangos que se muestran en la tabla 7.9. Obsérvese que los datos se han ordenado por rangos en una sola serie, tal como lo requiere esta prueba. La puntuación menor es 0.21 y se le asigna el rango uno. Hay un triple empate en la puntuación 0.80; así, el rango mayor es 17 $[(16 + 17 + 18)/3 = 17]$. En la tabla 7.9 se muestran, además, las sumatorias de rangos (las R_j) para cada grupo, así como el rango promedio de cada grupo ($\overline{R}_j$). Con los datos de la tabla 7.9 podemos calcular el valor de KW, sin la corrección para empates, mediante la ecuación (7.3):

$$KW = \left[\frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \overline{R}_j^2\right] - 3(N+1) \tag{7.3}$$

$$= \frac{12}{18(18+1)} [6(4.17)^2 + 6(10.83)^2 + 6(13.50)^2] - 3(18+1)$$

$$= 66.72 - 57$$

$$= 9.72$$

La tabla C del Apéndice I nos indica que un $KW \geq 5.99$ con $gl = 3 - 1 = 2$ tiene una probabilidad de ocurrencia cuando H_0 de $p < 0.05$. Así, puesto que el valor observado de KW (9.72) es mayor que 5.99, la hipótesis de no diferencia en

Tabla 7.8. Índices de exactitud de juicios para sujetos que aprenden la relación $aX + bY$.

Entrenamiento		
Primero señal irrelevante	*Primero señal válida*	*Ambas señales*
0.44	0.70	0.80
0.44	0.77	0.76
0.54	0.48	0.34
0.32	0.64	0.80
0.21	0.71	0.73
0.28	0.75	0.80

Tabla 7.9. Índices de exactitud de juicios para sujetos que aprenden la relación $aX + bY$ (ordenados por rangos).

	Entrenamiento		
	Primero señal irrelevante	*Primero señal válida*	*Ambas señales*
	5.5	10	17
	5.5	15	14
	8	7	4
	3	9	17
	1	11	12
	2	13	17
R_j	25	65	81
$\overline{R}_j$	4.17	10.83	13.50

la exactitud de los juicios es rechazada, y concluimos que existen diferencias entre los grupos entrenados de manera distinta. Efectivamente, la revisión de la tabla 7.9 nos muestra que las ejecuciones son peores en la condición de primero señal irrelevante, que en los otros dos grupos.

Corrección para los empates

Para realizar la correción de los empates, debemos determinar cuántos grupos de empates hay y cuantas puntuaciones se encuentran empatadas en cada grupo. Para estos datos, hay dos grupos de puntuaciones empatadas (dos puntuaciones se encuentran empatados en 0.44 —con un rango de 5.5— y tres se encuentran empatadas en 0.80 —con rango 17—). Así, en la aplicación de la corrección de acuerdo con la ecuación (7.4), tenemos $g = 2$ que el número de grupos con empates, $t_1 = 2$ es el número de puntuaciones empatadas en el primer grupo y $t_2 = 2$ es el número de puntuaciones empatadas en el segundo grupo. Así, la corrección es

$$1 - \frac{\sum_{i=1}^{g}(t_i^3 - t_i)}{N^3 - N} = 1 - \frac{(2^3 - 2) + (3^3 - 3)}{18^3 - 18} \tag{7.4}$$

$$= 1 - \frac{(8 - 2) + (27 - 3)}{5832 - 18}$$

$$= 1 - 0.005$$

$$= 0.995$$

Cuando la corrección se aplica al valor de *KW* encontrado anteriormente, el valor corregido es $KW = 8.72/0.995 = 8.76$. Por supuesto, ya que rechazamos H_0 con el primer valor obtenido, también será rechazada con el valor corregido. Debe notarse que aún con cinco de las 18 observaciones implicadas en empates, la corrección produjo un cambio muy pequeño en *KW*.

Comparaciones múltiples entre tratamientos

Cuando el valor obtenido de *KW* es significativo, indica que al menos uno de los grupos es diferente de al menos otro de los grupos. Esto no dice al investigador qué grupos son diferentes, ni le indica cuántos de los grupos son diferentes de cada uno. Lo que se necesita es un procedimiento que nos posibilite determinar cuáles grupos son diferentes. Esto es, vamos a probar la hipótesis H_0: $\theta_u = \theta_v$ en contra de la hipótesis H_1: $\theta_u \neq \theta_v$ para algunos grupos u y v. Existe un procedimiento sencillo para determinar cuáles pares de grupos son diferentes. Empezamos determinando las diferencias $|\bar{R}_u - \bar{R}_v|$ para todos los pares de grupos. Cuando el tamaño de la muestra es grande, estas diferencias se distribuyen aproximadamente de manera normal. Sin embargo, ya que hay una cantidad muy grande de diferencias y que las diferencias no son independientes, el procedimiento de comparación debe ajustarse apropiadamente. Supongamos que la hipótesis de no diferencia entre los k fue probada y rechazada en el nivel α de significación. Podemos probar la significación de los pares individuales de diferencias utilizando la siguiente desigualdad. Si

$$|\overline{R}_u - \overline{R}_v| \geq z_{\alpha/k(k-1)} \sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_u} + \frac{1}{n_v}\right)} \qquad (7.6)$$

podemos rechazar la hipótesis H_0: $\theta_u = \theta_v$ y concluir que $\theta_u \neq \theta_v$. El valor de $z_{\alpha/k(k-1)}$ es el valor de abcisa de la distribución normal sobre el cual se ubica el $\alpha/k(k - 1)$ % de la distribución. Los valores de z pueden obtenerse de la tabla A del Apéndice I.

Puesto que a menudo es necesario obtener valores basados en probabilidades extremadamente pequeñas, especialmente cuando k es grande, puede utilizarse la tabla A_{II} del Apéndice I, en lugar de la tabla A. Ésta es una tabla de la distribución normal estándar (estandarizada) que ha sido construida de tal forma que los valores usados en comparaciones múltiples pueden obtenerse fácilmente. La tabla está construida con base en el número de comparaciones que pueden realizarse. Los valores de tabla son los valores de z asociados con varios valores de α. La entrada de los renglones (#c) son el número de comparaciones. Cuando hay k grupos, hay $k(k - 1)/2$ comparaciones posibles.

Ejemplo. En el ejemplo para muestras grandes en esta sección, rechazamos H_0 y concluimos que las medianas no eran iguales. Puesto que hay $k = 3$ grupos, hay $3(3 - 1)/2 = 3$ comparaciones posibles. Si tomamos las diferencias entre los rangos promedio, tenemos

$$|\overline{R}_1 - \overline{R}_2| = |4.17 - 10.83| = 6.66$$

$$|\overline{R}_1 - \overline{R}_3| = |4.17 - 13.50| = 9.33$$

$$|\overline{R}_2 - \overline{R}_3| = |10.83 - 13.50| = 2.67$$

Para encontrar cuál de estas diferencias es significativa, podemos aplicar la prueba de comparaciones múltiples descrita en esta sección. Es necesario encontrar el valor crítico de z. Ya que escogimos $\alpha = 0.05$ en el análisis original, aquí se debe utilizar el mismo nivel, y en virtud de que el número de comparaciones es $\#c = k(k - 1)/2 = 3(3 - 1)/2 = 3$, encontramos el valor crítico de z de la tabla A_{II} del Apéndice I: el valor de $z = 2.394$. (Éste es el mismo valor que encontraríamos si hubiéramos utilizado la tabla A: $z_{\alpha/k(k-1)} = z_{0.05/3(3-1)} = z_{0.0083} \approx 2.39$.) La diferencia crítica se encuentra mediante la ecuación (7.6):

$$\begin{aligned} z_{\alpha/k(k-1)} \sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_u} + \frac{1}{n_v}\right)} \qquad (7.6) \\ = 2.394 \sqrt{\frac{18(18+1)}{12}\left(\frac{1}{6} + \frac{1}{6}\right)} \\ = 2.394\sqrt{9.5} \\ = 7.38 \end{aligned}$$

Puesto que sólo la diferencia entre los grupos 1 y 3 (primero señal irrelevante contra ambas señales) es mayor que el valor crítico 7.38, sólo esa comparación fue significativa y podemos concluir que estas medianas son diferentes.

Debe notarse, con sumo cuidado, que en la aplicación de la ecuación (7.6) a las comparaciones múltiples en el ejemplo anterior, se calculó solamente una diferencia crítica. Esto fue posible porque cada uno de los k grupos eran del mismo tamaño. Al tener las muestras tamaños desiguales, cada una de las diferencias observadas tendrían que ser comparadas contra las distintas diferencias críticas.

Comparación de tratamientos contra control

A veces un investigador incluye un grupo control o estándar como uno de los k grupos. Un ejemplo sería cuando se desea evaluar los efectos sobre la conducta de varias drogas. Aunque el mayor interés puede ser si los grupos difieren en la variable medida (seleccionada), el interés principal puede ser si existen diferencias entre la conducta bajo administración de *cualquier* droga y la conducta cuando no se administra droga (o bien, se aplica un placebo). En este caso, el investigador podría aplicar el análisis de varianza unifactorial por rangos de Kruskal-Wallis, si se satisfacen los supuestos para su uso. Sin embargo, si H_0 es rechazada, el investigador puede tener interés en demostrar si cualquiera de los grupos a los que se administró droga difiere del grupo control. Esto es, si θ_c es la mediana del grupo control, y θ_u es la mediana del u-ésimo grupo, al investigador le gustaría evaluar H_0: $\theta_c = \theta_u$ contra H_1: $\theta c \neq \theta_u$, (o tal vez H_0: $\theta_c > \theta_u$). Ya que no estamos interesados en comparar todos los grupos, el método para comparaciones múltiples dado por la ecuación (7.6) debe ajustarse para tener en consideración el pequeño número de comparaciones. Cuando existen k grupos en la prueba global, habrá $k - 1$ comparaciones con el grupo control; así $\#c = k - 1$. Las relaciones apropiadas para las comparaciones múltiples en este caso son las siguientes:

Para probar H_1: $\theta_c \neq \theta_u$,

$$|\overline{R}_c - \overline{R}_u| \geq z_{\alpha/2(k-1)} \sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_c} + \frac{1}{n_u}\right)} \tag{7.7}$$

Para probar H_1: $\theta_c > \theta_u$,

$$\overline{R}_c - \overline{R}_u > z_{\alpha/(k-1)} \sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_c} + \frac{1}{n_u}\right)} \tag{7.8}$$

Los valores críticos de z se encuentran utilizando la tabla A o la tabla A_{II} del Apéndice I con $\#c = k - 1$. [*Nota*: Si los tamaños de las muestras son iguales, se obtiene una mejor aproximación cuando los valores de z de las ecuaciones (7.7) y (7.8) son reemplazados por $q(\alpha, \#c)$ y se utiliza la tabla A_{III} del Apéndice.]

Resumen del procedimiento

Éstos son los pasos que hay que seguir para la aplicación de la prueba de análisis de varianza unifactorial por rangos, de Kruskal-Wallis:

1. Ordene por rangos las observaciones de los k grupos en una sola serie, asignando los rangos de uno a N. (A las observaciones empatadas se les asigna el valor promedio de los rangos empatados.)
2. Determine los valores de R_j (la sumatoria de los rangos) y $\bar{R}_j$ (los rangos promedios) para cada uno de los k grupos de rangos.
3. Si una gran proporción de las observaciones se encuentran empatadas, calcule el valor de KW mediante la ecuación (7.5); en caso contrario, emplee la ecuación (7.3).
4. El método para determinar la significación de un valor observado de KW depende del número de grupos (k) y del tamaño de los grupos (n_j):

   a) Si $k = 3$ y si n_1, n_2, y $n_3 \leq 5$, debe utilizarse la tabla O del Apéndice I para determinar, en el supuesto de que H_0 es verdadera, la probabilidad asociada de un KW tan grande como el observado.
   b) En otros casos, la significación de un valor tan grande como el valor observado de KW puede ser evaluado mediante la tabla C del Apéndice I, con $gl = k - 1$.

5. Si la probabilidad asociada con el valor observado de KW es menor o igual que el nivel de significación (α) previamente elegido, rechace H_0 en favor de H_1.
6. Si H_0 es rechazada, debe utilizarse el método de comparaciones múltiples [ecuación (7.6)] para determinar cuáles diferencias son significativas. Si la prueba implica comparaciones entre tratamientos y grupo control, debe utilizarse el método de comparación dado por las ecuaciones (7.7) y (7.8).

Potencia-eficacia

Comparada con la prueba paramétrica más poderosa, la prueba F, y en condiciones donde se cubren los supuestos asociados con el modelo estadístico del análisis de varianza paramétrico, la prueba de Kruskal-Wallis tiene una eficacia asintótica de $3/\pi = 95.5$ %.

La prueba de Kruskal-Wallis es más eficaz que la extensión de la prueba de la mediana porque utiliza más de la información que contienen las observaciones, convirtiendo las puntuaciones en rangos, más que simplemente dicotomizarlas como "por arriba" o "por debajo" de la mediana.

Referencias bibliográficas

El lector encontrará explicaciones de la prueba del análisis de varianza unifactorial por rangos, de Kruskal-Wallis, en Kruskal y Wallis (1952) y en Kruskal (1952). Otros análisis útiles se hallan en Lehman (1975) y Hettmansperger (1984).

PRUEBA DE JONCKHEERE PARA NIVELES ORDENADOS DE LA VARIABLE

Función

La prueba del análisis de varianza unifactorial por rangos, de Kruskal-Wallis, prueba la hipótesis de que k grupos o muestras independientes son los mismos, en contra de la hipótesis alterna: de que uno o más de esos grupos difieren de los otros. Sin embargo, en algunas situaciones experimentales, el investigador puede desear mantener una hipótesis alterna más específica. Por ejemplo, en un experimento sobre el efecto de distintas dosis de droga en la ejecución de tareas de aprendizaje, el investigador puede desear evaluar la hipótesis de "no diferencia", en contra de la hipótesis alterna de que el incremento en la dosis resultará en el "deterioro" de la ejecución. En este caso la hipótesis alterna asociada al análisis de varianza unifactorial por rangos de Kruskal-Wallis, aunque válida, es demasiado general. La prueba de Jonckheere para niveles ordenados de la variable se presenta en esta sección, prueba la hipótesis de que las muestras (o los grupos) se encuentran ordenadas en una secuencia específica *a priori*. Para especificar las hipótesis nula y alterna más explícitamente, θ_j será la mediana de la población para la j-ésima muestra (o grupo). Entonces, podemos plantear la hipótesis nula de que las medianas son las mismas como H_0: $\theta_1 = \theta_2 = \ldots = \theta_k$ y la hipótesis alterna puede plantearse como H_1: $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_k$, esto es, las medianas se encuentran ordenadas por magnitud. Si la hipótesis alterna es verdadera, al menos unas de las diferencias es estrictamente desigual ($<$). Es importante notar que a fin de asegurar el uso adecuado de la prueba, el investigador debe ser capaz de especificar el orden de los grupos o medidas *a priori*. Esto equivale a decir que uno no puede ver las k medianas y especificar la hipótesis alterna. El orden debe ser especificado *antes* de recabar los datos.

Para aplicar la prueba de Jonckheere, los datos de las k muestras o grupos independientes deben encontrarse en, al menos, escala ordinal, y según la hipótesis nula es de suponer que cada una de las muestras proviene de la misma población. Nosotros supondremos que existen N casos u observaciones de los que hay n_j datos en cada j-ésimo grupo.

Racionalización y método

Para aplicar la prueba de Jonckheere para niveles ordenados de la variable el investigador primero debe especificar *a priori* el orden de los grupos. Después, los datos se presentan en una tabla de doble entrada donde cada columna representa sucesivamente a cada grupo o muestra, arreglados de acuerdo con la hipótesis del orden de las medianas. Esto es, los grupos se ordenan; el primero de ellos es el grupo que se "conjetura" que tiene la mediana más pequeña y el grupo k será el grupo que se "conjetura" que tiene la mediana mayor. Así, los datos se presentan de la siguiente manera:

Grupo			
1	2	...	k
X_{11}	X_{12}	...	X_{1k}
X_{21}	X_{22}	...	X_{2k}
$\vdots$			$\vdots$
$X_{n_1 1}$	$\vdots$	...	$X_{n_k k}$
	$X_{n_2 2}$		

La prueba de Jonckheere requiere el recuento del número de veces que una observación en el i-ésimo grupo o muestra es precedida por una observación en el j-ésimo grupo o muestra. Aunque el procedimiento de recuento parece ser más bien tedioso, realmente es muy sencillo si se aplican sistemáticamente algunos procedimientos computacionales.

Primero, debemos definir el estadístico, a veces denominado recuento de Mann-Whitney:

$$U_{ij} = \sum_{h=1}^{n_i} \#(X_{hi}, j) \tag{7.9}$$

donde $\#(X_{hi}, j)$ es el número de veces que el dato X_{hi} precede (es más pequeño que) al dato de las muestras j, donde $i < j$. El estadístico J de la prueba de Jonckheere es, entonces, el número total de este recuento:

$$J = \sum_{i<j}^{k} U_{ij} = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} U_{ij} \tag{7.10}$$

La distribución muestral de J ha sido tabulada para muestras pequeñas y se proporciona en la tabla P del Apéndice I. Las entradas de la tabla proporcionan las probabilidades asociadas a los valores de J tan grandes o mayores que los valores de tabla para distintos valores de J, las n_j y la probabilidad de α. El lector notará que la tabla está compuesta por dos partes diferentes. En la primera parte se presenta la distribución de J para $k = 3$ y n_j menores que 9, y en la segunda, se presenta la distribución de J para $k = 4, 5, 6$ y n_j mayores que 6. Si el valor observado de J es mayor que el valor de tabla para el nivel de significación seleccionado, entonces debe rechazarse la hipótesis nula H_0 en favor de la hipótesis alterna H_1. Por ejemplo, consideremos un caso en el que hay $k = 3$ grupos y $n_1 = 3$, $n_2 = 4$ y $n_3 = 4$. Seleccionamos un nivel de significación de $\alpha = 0.05$. El valor calculado del estadístico de Jonckheere fue $J = 26$. La tabla P de Apéndice I nos muestra que la probabilidad de observar un valor de $J \geq 26$ es mayor que 0.10; por tanto, no podemos rechazar la hipótesis nula H_0 de que las medianas para los tres grupos son iguales.

Conforme el tamaño de la muestra es mayor, la distribución muestral se aproxima a la normal, con media

$$\mu_J = \frac{N^2 - \sum_{j=1}^{k} n_j^2}{4} \tag{7.11}$$

y varianza

$$\sigma_J^2 = \frac{1}{72}\left[N^2(2N + 3) - \sum_{j=1}^{k} n_j^2(2n_j + 3) \right] \tag{7.12}$$

Así, el estadístico

$$J^* = \frac{J - \mu_J}{\sigma_J} \tag{7.13}$$

se distribuye aproximadamente de manera normal con media igual a cero y desviación estándar igual a uno. Por tanto, puede utilizarse la tabla A del Apéndice I para evaluar la hipótesis acerca de J^* y dado esto, J. Por supuesto, ya que las alternativas se encuentran ordenadas, la prueba se considera unidireccional.

Ejemplo. Cuando se mezclan sacarosa y cloruro de sodio, hay un mutuo "enmascaramiento" en los juicios que se mencionan acerca de lo dulce o salado de las mezclas. Existen varios factores que afectan la cantidad de enmascaramiento. Uno es el tipo de compuesto enmascarador utilizado (por ejemplo, quinina) y otro es la concentración del compuesto utilizado. Un tercer factor es la proporción relativa del estímulo de prueba respecto al estímulo neutral o base. Los experimentos que implican juicios de sabor son a menudo tareas psicofísicas que requieren mucho ensayos, algunos con el estímulo relevante (ensayos-señal) y el resto con el estímulo y la mezcla (ensayos-señal más ruido). En un experimento diseñado para evaluar el efecto de la proporción relativa de estímulos puros y mezclados sobre juicio de sabor, Kroeze[7] varió la proporción relativa de estímulos puros y mezclados. En otras áreas de la conducta se han registrado tales efectos de las frecuencias, que además han sido compatibles con una explicación derivada de la teoría del nivel de adaptación de Helson.[8] En una serie de cuatro muestras independientes, la intensidad física (= concentración) del cloruro de sodio se mantuvo constante en 0.32 mol/1; en el estímulo prueba, la concentración de sacarosa fue mantenida constante en 0.32 mol/1. En todos los grupos, la frecuencia relativa [NaC1/(NaC1 + sacarosa)] se varió en los ensayos de prueba. Los juicios individuales de salinidad para las distintas proporciones se presentan en la tabla 7.10. Kroeze supuso que los juicios de salinidad se incrementarían conforme se decrementara la proporción de NaC1 en los ensayos de prueba.

i. *Hipótesis nula.* H_0: las mezclas relativas de ensayos NaC1 y ensayos NaC1 + sacarosa no tienen efecto sobre los juicios de salinidad. H_1: los juicios de los sujetos

[7] Kroeze, J. H., "The influence of relative frequencies of pure and mixed stimuli on mixture suppression in taste", en *Perception & Psychophysics*, núm. 31, 1982, págs. 276-278.
[8] Helson, H., *Adaptation-level theory*, Harper & Row, Nueva York, 1964.

Tabla 7.10. Juicios individuales sobre la salinidad de un estímulo compuesto como función del porcentaje de la salinidad del NaCl puro.

Porcentaje del NaCl puro			
80	50	17	10
8.82	13.53	19.23	73.51
11.27	28.42	67.83	85.25
15.78	48.11	73.68	85.82
17.39	48.64	75.22	88.88
24.99	51.40	77.71	90.33
39.05	59.91	83.67	118.11
47.54	67.98	86.83	
48.85	79.13	93.25	
71.66	103.05		
72.77			
90.38			
103.13			

Nota. Cada columna representa una muestra de observaciones separada e independiente. Los datos se han ordenado de manera ascendente, dentro de cada grupo, para facilitar el cálculo de U_{ij}. Si se utiliza alguna rutina para computadora, no es necesario ordenar los datos en cada muestra.

acerca de la salinidad se relacionan inversamente con la proporción de ensayos de prueba NaCl del experimento.

ii. *Prueba estadística.* Puesto que el investigador supone un ordenamiento en los juicios acerca de la salinidad, es apropiada una prueba para alternativas ordenadas.

iii. *Nivel de significación.* $\alpha = 0.05$. El número de sujetos es $n_1 = 12$, $n_2 = 9$, $n_3 = 8$ y $n_4 = 6$ en los cuatro grupos.

iv. *Distribución muestral.* En virtud de que los tamaños de las muestras son diferentes y el número de grupos es mayor de tres, se utilizará la distribución muestral para muestras grandes de la prueba de Jonckheere; esto es, se calculará el estadístico J^* definido en la ecuación (7.13) y la significación de su probabilidad será determinada mediante la tabla A del Apéndice I.

v. *Región de rechazo.* La región de rechazo consiste en todos los valores de J^* que sean mayores que 1.645, el valor de la distribución normal estándar asociada con $\alpha = 0.05$.

vi. *Decisión.* Utilizando los datos de la tabla 7.10, se calcularon los valores de los estadísticos U_{ij}; éstos se presentan en la tabla 7.11. Por ejemplo, consideremos el dato 47.54 del grupo 1. Éste se encuentra antes que siete datos del grupo 2 (48.11, 48.64, 51.40, 59.91, 79.13, 67.98, 103.05), siete datos del grupo 3 (67.83, 73.68, 75.22, 77.71, 83.67, 86.83, 93.25), y los seis datos del grupo 4. Las U_{ij} son las columnas de los recuentos precedentes en la tabla 7.11. Así, el valor del estadístico de la prueba de Jonckheere para estos datos es el siguiente:

$$J = 66 + 73 + 62 + 52 + 48 + 36$$

$$= 337$$

Tabla 7.11. Número de # (X_{hi}, j) para los datos de la tabla 7.10.

				Porcentaje de NaCl puro						
				80			50		17	10
i	*1*	*1*	*1*		*2*	*2*		*3*		
j	*2*	*3*	*4*		*3*	*4*		*4*		
	9	8	6	8.82	8	6	13.58	6	19.23	73.51
	9	8	6	11.27	7	6	28.42	6	67.83	85.25
	8	8	6	15.78	7	6	48.11	5	73.68	85.82
	8	8	6	17.39	7	6	48.64	5	75.22	88.88
	8	7	6	24.99	7	6	51.40	5	77.71	90.33
	7	7	6	39.05	7	6	59.91	5	83.67	118.11
	7	7	6	47.54	6	6	67.98	3	86.83	
	5	7	6	48.85	3	5	79.13	1	93.25	
	2	6	6	71.66	0	1	103.05			
	2	6	6	72.77						
	1	1	1	90.38						
	0	0	1	103.13						
U_{ij}	66	73	62		52	48		36		

Es necesario calcular la media y la desviación estándar del estadístico J, mediante las ecuaciones (7.11) y (7.12):

$$\mu_J = \frac{N^2 - \sum_{j=1}^{k} n_j^2}{4} \tag{7.11}$$

$$= \frac{35^2 - 12^2 - 9^2 - 8^2 - 6^2}{4}$$

$$= 225$$

y $$\sigma_J^2 = \frac{1}{72}\left[N^2(2N + 3) - \sum_{j=1}^{k} n_j^2 (2n_j + 3)\right] \tag{7.12}$$

$$= 1/72\,\{35^2(70 + 3) - [12^2(24 + 3) + 9^2(18 + 3) +$$

$$8^2(16 + 3) + 6^2(12 + 3)]\}$$

$$= 1\,140$$

Con estos valores podemos calcular J^*

$$J^* = \frac{J - \mu_J}{\sigma_J} \tag{7.13}$$

$$= \frac{337 - 225}{33.76}$$

$$= 3.32$$

En virtud de que el valor observado de J^* es mayor que el valor crítico de 1.645, podemos rechazar la hipótesis de que las medianas de los cuatro grupos son iguales, y concluir que éstas se incrementan en magnitud. (Observemos que rechazar H_0 implica que al menos una diferencia entre medianas sucesivas es significativa.)

Observaciones empatadas

Cuando ocurren empates entre dos o más puntuaciones en el momento de contar las puntuaciones precedentes [ecuación (7.9)], el recuento debe incrementarse en 0.5 (1/2, más que en 1) por cada empate. Como en la prueba de Kruskal-Wallis, la varianza de J [ecuación (7.12)] puede verse afectada por los empates, pero a menos que el número de empates sea grande o que haya muchos datos empatados en la misma puntuación, el efecto en la distribución muestral de J^* es, practicamente, despreciable.

Resumen del procedimiento

Éstos son los pasos que hay que seguir en la aplicación de la prueba de Jonckheere para niveles ordenados de la variable.

1. Presente las puntuaciones en una tabla de doble entrada en la que las k columnas representan las muestras o grupos arreglados en un orden *a priori* va desde la mediana hipotéticamente más pequeña a la mediana hipotéticamente mayor.
2. Calcule el recuento de las puntuaciones precedentes y el recuento de Mann-Whitney (la U_{ij}) utilizando la ecuación (7.9).
3. Determine el estadístico J de la prueba de Jonckheere, que es la sumatoria de los recuentos determinada en el paso 2.
4. El método para determinar la significación del valor observado de J depende del número de grupos (k) y del tamaño de los grupos o muestras (n_j):

   a) Si $k = 3$ y n_1, n_2 y $n_3 \leq 8$, puede utilizarse la tabla P del Apéndice I para determinar, en el supuesto de que H_0 es verdadera, la probabilidad asociada de J que sea tan grande como el valor observado.

   b) Si $k = 4$, 5, o 6, y las muestras (n_j) son del mismo tamaño y menos de siete, puede utilizarse la tablas P del Apéndice I para determinar, en el supuesto de que H_0 es verdadera, la probabilidad asociada de J que es tan grande como el valor observado.

c) Si el número de grupos o el número de observaciones en un grupo es demasiado grande para utilizar la tabla P del Apéndice I, el estadístico J^* puede calcularse mediante la ecuación (7.13), y la probabilidad asociada con su valor se determina mediante la tabla A del Apéndice I.

Si el valor de J (o J^*) es lo suficientemente grande como para rechazar H_0, el investigador puede aplicar las técnicas de comparaciones múltiples descritas en la sección anterior. Sin embargo, en este caso las comparaciones son unidireccionales y el valor de z debe ajustarse.

5. Para facilitar el cálculo de J y J^*, el lector puede utilizar un programa para computadora como el que se presenta en el Apéndice II. En el ejemplo anterior, los cálculos se efectuaron con mucho detalle. Estos resultados pueden revisarse utilizando el programa muestra que incluye los datos como un ejemplo.

Potencia-eficacia

La eficacia asintótica de la prueba de Jonckheere es $3/\pi = 95.5$ % cuando se la compara con una prueba t (F) apropiada para alternativas ordenadas. Así, cuando se compara con la prueba paramétrica apropiada para datos normalmente distribuidos, la eficacia de la prueba de Jonckheere es la misma que la de la prueba de Kruskal-Wallis.

Referencias bibliográficas

La prueba de Jonckheere para niveles ordenados de la variable se explica en Jonckheere (1954), Lehmann (1975) y Terpstra (1952). Pueden encontrarse análisis relativos a la eficacia de la prueba de Jonckheere en Puri (1975). Potter y Strum (1981) han examinado la rapidez con que se incrementa el poder de la prueba de Jonckheere. En Lehmann (1975) pueden encontrarse fórmulas para la corrección de la varianza.

ANÁLISIS

En este capítulo se presentaron cuatro pruebas estadísticas no paramétricas para analizar datos de k muestras o grupos independientes. La primera de éstas, la prueba ji cuadrada para k muestras independientes, es útil cuando los datos corresponden a frecuencias y cuando las mediciones de las variables en estudio se encuentran en escala nominal o categórica. La prueba ji cuadrada es conveniente cuando los datos son categorías discretas en una escala ordinal; sin embargo, algunos de los otros métodos examinados en este capítulo pueden resultar más apropiados en tales casos. La prueba ji cuadrada evalúa si las proporciones o frecuencias en las distintas categorías son independientes de la condición (muestra o grupo) en la que fueron observadas, es decir, prueba la hipótesis nula de que las k muestras provienen de la misma población o de poblaciones idénticas con respecto a la proporción de observaciones en las distintas categorías.

La segunda prueba presentada, la extensión de la prueba de la mediana, requiere que las mediciones de la variable en estudio se encuentren en, al menos, escala ordinal para una adecuada interpretación de los resultados del análisis. Esta técnica prueba si los k grupos o muestras independientes se han extraído de poblaciones que presentan medianas idénticas.

El análisis de varianza unifactorial por rangos, de Kruskal-Wallis, la tercera prueba examinada, requiere mediciones de la variable, al menos, en escala ordinal. Esta técnica prueba la hipótesis de que las k muestras o grupos independientes han sido extraídos de la misma población o de poblaciones idénticas con la misma distribución continua de respuestas (pero desconocida).

La cuarta prueba presentada, la de Jonckheere para niveles ordenados de la variable, requiere que las mediciones de la variable se encuentren en escala ordinal. Este procedimiento prueba la hipótesis de que las k muestras o grupos independientes pudieran haber sido extraídos de la misma población o de poblaciones idénticas con la misma distribución continua (pero desconocida), en contra de la hipótesis alterna de que las medianas de las distribuciones se encuentran ordenadas por magnitud de acuerdo con alguna hipótesis *a priori*.

No tenemos opción entre estas pruebas si los datos corresponden a frecuencias más que a puntuaciones, es decir, si tenemos una enumeración de datos o si las mediciones se encuentran en escala nominal o categórica. La ji cuadrada para k muestras independientes es la mejor prueba de las que se presentaron en este capítulo, para esta clase de datos.

La extensión de la prueba de la mediana, la prueba de Kruskal-Wallis y la de Jonckheere pueden aplicarse a la misma clase de datos; es decir, estas pruebas pueden realizar las mismas suposiciones según H_0: que las variables provienen de poblaciones que tienen distribuciones continuas idénticas. Cuando los datos son tales que puede aplicarse la extensión de la prueba de la mediana y la prueba de Kruskal-Wallis, esta última puede ser más eficaz ya que utiliza más información disponible en las observaciones. Ésta convierte las puntuaciones en rangos, mientras que la extensión de la prueba de la mediana convierte las puntuaciones en signos de más $(+)$ o de menos $(-)$, dependiendo si los datos se encuentran por arriba o por debajo de la mediana. Así, la prueba de Kruskal-Wallis preserva la magnitud de los datos más completamente que la prueba de la mediana. Por esta razón, generalmente es más sensible a las diferencias entre las k muestras o grupos. Sin embargo, como se anotó en el análisis de la prueba de la mediana en el capítulo 5, hay situaciones que implican datos ordenados para los que la prueba de la mediana es la única opción. Esto podría ocurrir cuando los valores de la variable medida son demasiado extremos para ser codificados y ordenados por rangos con precisión. En tales casos, puede aplicarse la prueba de la mediana, pero los rangos no pueden determinarse a fin de aplicar la prueba de Kruskal-Wallis o la de Jonckheere.

En caso de existir un ordenamiento *a priori* de las medianas de los grupos de la población, la prueba de Jonhckheere es más poderosa que la prueba de Kruskal-Wallis. Esto es así porque la hipótesis que será evaluada es más específica que la de la prueba de Kruskal-Wallis.

Una característica de las cuatro pruebas presentadas en este capítulo es que si la prueba estadística es significativa, nos permite concluir que existen diferencias entre los k grupos. Sin embargo, ninguna de estas pruebas nos dirá dónde se encuentran las diferencias. Afortunadamente, existen procedimientos que auxilian al investigador a localizar las diferencias. Para las pruebas ji cuadrada y la extensión de la prueba de la mediana, los grados de libertad pueden dividirse para ayudar a localizar las diferencias; adicionalmente, el análisis de los residuos puede proveer detalles adicionales acerca de dónde se encuentran las diferencias más significativas en la tabla. En el caso de la prueba de Kruskal-Wallis y la de Jonckheere, podemos utilizar técnicas de comparaciones múltiples que nos ayuden a determinar dónde se encuentran las diferencias más significativas. Aunque las técnicas de partición y de comparaciones múltiples son instrumentos poderosos que aíslan los efectos, el investigador debe tener cautela al aplicar estos procedimientos sólo a datos para los que la prueba es significativa.

Hay algunas otras pruebas no paramétricas para diferenciar entre k muestras o grupos independientes. Una de ellas es la de deslizamiento para k muestra (Mosteller, 1948; Mosteller y Tukey, 1950). Chacko (1963) y Puri (1965) han propuesto pruebas que son generalizaciones de la prueba de Jonckheere. Tal vez la prueba más poderosa es la de la *sombrilla* (Mack y Wolfe, 1981), que prueba la hipótesis de las medianas, de acuerdo con algunos supuestos *a priori* primero incrementan hasta un máximo y después decrementan. Esto es, prueba la hipótesis H_0: $\theta_1 = \theta_2 = \ldots = \theta_k$ contra la hipótesis H_1: $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_h \geq \ldots \geq \theta_{k-1} \geq \theta_k$ para algunos grupos h predeterminados. Esta prueba es relativamente sencilla, como si se aplicaran dos pruebas de Jonckheere: una para un conjunto de desigualdades y otra para el otro conjunto de desigualdades.

8

Medidas de asociación y sus pruebas de significación

En el proceso de investigación en las ciencias conductuales, frecuentemente deseamos conocer si dos series de puntuaciones están relacionadas y, si es así, el grado de su relación. Establecer que existe una correlación entre dos variables podría ser el objetivo de una investigación, como en algunos estudios de dinámica de la personalidad, percepción, semejanza entre grupos, etc. O bien establecer una correlación podría ser sólo un paso en un estudio que tiene otra finalidad, como cuando usamos medidas de correlación para probar la confiabilidad de nuestras observaciones.

Este capítulo está dedicado a la presentación de medidas de correlación no paramétricas y las pruebas estadísticas que determinan la probabilidad asociada con la ocurrencia de una correlación tan grande como la que se ha observado en la muestra, según la hipótesis nula de que las variables son independientes o no están relacionadas en la población. Es decir, además de presentar medidas de asociación, presentaremos pruebas estadísticas que determinan la "significación" de la asociación observada. El problema de medir el *grado* de asociación entre dos series de puntuaciones es más general que el de probar la *existencia* de algún grado de asociación en alguna población. Naturalmente, es de cierto interés ser capaz de establecer el grado de asociación entre dos series de puntuaciones obtenidas de un grupo dado de sujetos. Pero quiza sea de mayor relevancia determinar si alguna asociación observada en una *muestra* de puntuaciones indica que las variables en estudio están o no asociadas de la *población* de la cual se extrajo la muestra. La correlación observada por sí misma representa un estimador del grado de asociación. Las pruebas de significación de ese coeficiente determina en el nivel establecido de confianza, la probabilidad de que muestras aleatorias de una población en la que no existiera asociación alguna entre las variables, resultara en una correlación tan grande (o más grande) que la obtenida.

En el caso paramétrico, la medida usual de correlación es el coeficiente de correlación producto-momento r de Pearson. Este estadístico requiere variables que estén medidas en al menos una escala de intervalos iguales, para una adecuada interpretación del estadístico. Si deseamos probar la significación de un valor observado o r, debemos no sólo encontrar la medida requerida, sino también suponer que las observaciones se muestren de una distribución normal bivariada. Más aún, el coeficiente de correlación producto-momento de Pearson mide el grado en el cual existe una relación de función lineal entre las variables.

Si, para un conjunto determinado de datos la suposición asociada con el coeficiente de correlación producto-momento r de Pearson no es sostenible o no realista, entonces se debe usar uno de los coeficientes de correlación y las pruebas estadísticas asociadas no paramétricas presentadas en este capítulo. Las medidas no paramétricas de correlación están disponibles para datos tanto categóricos como ordenados. Las pruebas no hacen o bien, hacen pocas suposiciones acerca de la distribución de la población de la cual se extrajeron las puntuaciones. Algunas suponen que las variables tienen una continuidad subyacente, mientras que otras no hacen ni aún esta suposición. Algunas prueban relaciones monotónicas entre las variables (pero no necesariamente lineales), mientras que otras miden asociaciones de cualquier tipo. Más aún, el investigador encontrará que, especialmente con muestras pequeñas, el cómputo de las medidas de asociación y las pruebas de significación no paramétricas no es más difícil y frecuentemenrte es más fácil que el cómputo de la r de Pearson.

Los usos y las limitaciones de cada índice se examinarán al ser presentada la medida. Al final del capítulo se presenta un análisis comparativo de los méritos y usos de las diferentes medidas.

EL COEFICIENTE *C* DE CRAMÉR

Función

El *coeficiente C de Cramér* es una medida del grado de asociación o relación entre dos series de atributos o variables. Se usa únicamente cuando tenemos sólo información categórica (escala nominal) acerca de uno o de ambos conjuntos de atributos o variables. Esto es, puede emplearse cuando la información acerca de los atributos consiste en una serie no ordenada de categorías.

Para usar el coeficiente de Cramér, no es necesario suponer continuidad subyacente para las diferentes categorías usadas al medir uno o ambos conjuntos de atributos. De hecho, no necesitamos siquiera ser capaces de ordenar las categorías en alguna forma particular. El coeficiente de Cramér, al ser calculado de una tabla de contingencia, proporciona los mismos valores sin considerar cómo fueron ordenadas las categorías en las filas y columnas.

Método

Empecemos suponiendo que tenemos datos en dos series de variables categóricas no ordenadas. Por conveniencia, denotaremos estas variables como A y B.

Para calcular el coeficiente de Cramér entre puntuaciones de dos conjuntos de variables categóricas A, con categorías $A_1, A_2, \ldots, A_k$, y B, con categorías $B_1, B_2, \ldots B_r$, arreglamos las frecuencias dentro de la siguiente tabla de contingencia:

	A_1	A_2	$\cdots$	A_k	*Total*
B_1	n_{11}	n_{12}	$\cdots$	n_{1k}	R_1
B_2	n_{21}	n_{22}	$\cdots$	n_{2k}	R_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
B_r	n_{r1}	n_{r2}	$\cdots$	n_{rk}	R_r
Total	C_1	C_2	$\cdots$	C_k	N

Los datos pueden consistir en cualquier número de categorías. Es decir, se puede calcular un coeficiente de Cramér para datos de una tabla de 2×2, una tabla de 2×5, una de 4×4, una de 3×7, o cualquier tabla de $r \times k$.

En tal tabla, podemos tener frecuencias esperadas para cada celda (los E_{ij}), al determinar qué frecuencias se esperaría que ocurrieran si no existiera asociación entre las dos variables; esto es, las frecuencias esperadas en cada celda si las variables fueran independientes o no relacionadas. Mientras mayor sea la discrepancia entre esos valores esperados y los valores observados, más alto es el grado de asociación entre las dos variables y, por tanto, más grande el valor del coeficiente de Cramér.

El grado de asociación entre dos conjuntos de atributos al medirse por medio del coeficiente de Cramér, aunque sean o no ordenables e independientemente de la naturaleza de la variable (puede ser continua o discreta) y de la distribución subyacente del atributo (la distribución poblacional puede ser normal o de cualquier forma), puede encontrarse de una tabla de contingencia de frecuencias de observaciones por

$$C = \sqrt{\frac{X^2}{N(L-1)}} \tag{8.1}$$

donde

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{5.2}$$

o

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{n_{ij}^2}{E_{ij}} - N \tag{5.2a}$$

se calcula por el método presentado anteriormente en el capítulo 5 y L es el *mínimo* del número de filas o columnas en la tabla de contingencia. En otras palabras, con el propósito de calcular C, primero se calcula el valor de X^2 mediante la ecuación (5.2) y después se sustituye ese valor dentro de la ecuación (8.1) para obtener C. Debe notarse que, como la correlación producto-momento de Pearson, el coeficiente de Cramér tiene un valor máximo de uno y C será igual a cero cuando las variables o los atributos sean independientes. A diferencia de la correlación producto-momento de Pearson, el coeficiente de Cramér no puede ser negativo. Esto es de esperar, ya que el estadístico mide la relación entre variables categóricas que no poseen ningún orden inherente.

Ejemplo.[1] Como parte de un estudio acerca del proceso por el cual son modificados los estándares de contabilidad con información financiera, Hussein[2] desarrolló un cuestionario que fue enviado a los miembros del consejo consultivo del Financial Accounting Standards Board (FASB) y a miembros de varios comités que se especializan en estándares de contabilidad financiera en organizaciones patrocinadoras del FASB. El FASB es la organización por medio de la cual deben aprobarse los cambios en los estándares y procedimientos de contabilidad. El estudio, cuyos detalles no son relevantes en este ejemplo, se diseñó para evaluar los factores informacionales, económicos, organizacionales y cognoscitivos implicados en el proceso de establecer los estándares. En la investigación, es poco usual que para clasificar las respuestas se envíen por correo cuestionarios demasiado grandes. Sin embargo, para que la investigación de los diferentes grupos fuera significativa, la clasificación de las respuestas de varias organizaciones debería ser similar. Si esto no es así, entonces las respuestas (o las no respuestas) de un grupo pueden resultar en un punto de vista sesgado acerca del proceso total.

Para determinar si la clasificación inicial de respuestas estaba asociada con la organización, esto es, variaba entre las diferentes organizaciones patrocinadoras, se analizaron los datos concernientes a la clasificación de respuestas. Seis organizaciones o grupos recibieron los cuestionarios ($k = 6$) y hubo tres posibles disposiciones para cada cuestionario: reci-

[1] Para probar la significación de una medida de asociación, seguimos los mismos seis pasos que hemos señalado a lo largo de este libro para todas las otras pruebas estadísticas. Dichos pasos son los siguientes: 1. La hipótesis nula H_0 es que las dos variables no están relacionadas o son independientes en la población, mientras que H_1 es que están relacionadas o asociadas en la población. 2. La prueba estadística es la prueba de significación que resulta apropiada para la medida de asociación seleccionada. 3. El nivel de significación se especifica con anterioridad y puede ser cualquier probabilidad pequeña, por ejemplo, $\alpha = 0.01$, etc., mientras que N es el número de casos para los que existen puntuaciones en ambas variables. 4. La distribución muestral es la distribución teórica del estadístico usado para probar H_0, las probabilidades exactas o los valores críticos que son proporcionados en las tablas que se usan para probar la significación del estadístico. 5. La región de rechazo consiste en todos los valores de la medida de asociación que son tan extremos, que la probabilidad asociada con su ocurrencia según H_0 es menor o igual que una (y se usa una región de rechazo unidireccional cuando el signo de la asociación es predicho en H_1). 6. La decisión consiste en determinar el valor observado de la medida de asociación y después determinar la probabilidad, bajo la suposición de que H_0 es cierta, de tal valor extremo; si y sólo si esa probabilidad es igual o menor que α, la decisión es rechazar H_0 en favor de H_1.

Debido a que los mismos, o similares, conjuntos de datos se usan repetidamente como material ilustrativo en las explicaciones de las diferentes medidas de asociación, con el propósito de ilustrar las diferencias o similitudes entre estas medidas, la repetición constante de los seis pasos de la inferencia estadística en los ejemplos nos conduciría a una redundancia innecesaria. Por tanto, hemos decidido no incluir estos seis pasos en la presentación de los ejemplos en este capítulo. Mencionamos aquí que ellos podrían muy bien haber sido incluidos para señalar al lector que el procedimiento de toma de decisiones usado para probar la significación de una medida de asociación, es idéntico al procedimiento que se emplea en otra clase de pruebas estadísticas.

[2] Hussein, M. E., "The innovative process in financial accounting standards", en *Accounting, Organizations, and Society*, núm. 6, 1981, págs. 27-37.

Tabla 8.1. Respuestas al cuestionario.

Disposición al cuestionario	*Organización*						
	AAA	*AICPA*	*FAF*	*FASB*	*FEI*	*NAA*	*Total*
Completado	8 *7.49*	8 *7.15*	3 *6.46*	11 *10.89*	17 *11.91*	2 *5.10*	49
Rechazado	2 *3.51*	5 *3.35*	1 *3.04*	2 *5.11*	0 *5.59*	13 *2.40*	23
No respondido	12 *11.00*	8 *10.50*	15 *9.50*	19 *16.00*	18 *17.50*	0 *7.50*	73
Total	22	21	19	32	35	15	144

bido y completado, rechazado y no respondido ($r = 3$). Estos datos están resumidos en la tabla 8.1.

Para calcular el coeficiente C de Cramér es necesario calcular primero el estadístico ji cuadrada X^2. Un primer paso para calcular X^2 es encontrar los valores esperados, los E_{ij}, para cada celda de la tabla. Éstos se proporcionan en cursivas en la tabla 8.1. Al usar la ecuación (5.2), encontramos el valor del estadístico ji cuadrada:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{5.2}$$

$$= \frac{(8 - 7.49)^2}{7.49} + \frac{(8 - 7.15)^2}{7.15} + \ldots + \frac{(0 - 7.50)^2}{7.50}$$

$$= 75.25$$

En seguida, usamos la ecuación (8.1) para calcular C

$$C = \sqrt{\frac{X^2}{N(L - 1)}} \tag{8.1}$$

$$= \sqrt{\frac{75.25}{144(3 - 1)}}$$

$$= \sqrt{0.2613}$$

$$= 0.51$$

Así, encontramos que existe un grado moderado de asociación entre la disposición de la respuesta al cuestionario y la organización a la cual pertenece el receptor.

Prueba de la significación del coeficiente de Cramér

Las puntuaciones u observaciones con las que tratamos en la investigación, pertenecen frecuentemente a individuos en los que estamos interesados debido a que constituyen una muestra aleatoria de una población de interés. Cuando observamos una correlación entre dos conjuntos de atributos en la muestra, en ocasiones deseamos determinar que es plausible concluir que están asociados en la población representada por la muestra.

Si un grupo de sujetos constituye una muestra aleatoria de alguna población, podemos determinar si la asociación que existe entre dos conjuntos de puntuaciones de la muestra indica que existe una asociación en la población al probar la "significación" de la asociación. Al probar la significación de una medida de asociación, estamos probando la hipótesis nula de que no existe correlación en la población, esto es, que los valores observados de la medida de asociación en la muestra pudieran haber surgido al azar en una muestra aleatoria de una población en la que las dos variables son independientes, es decir, no correlacionadas. La hipótesis alterna es que las variables no son independientes.

Para probar la hipótesis nula, primero averiguamos la distribución muestral nula del estadístico (en este caso la medida de asociación), con la suposición de que H_0 es cierta. Después, usamos una prueba estadística asociada para determinar si el valor observado de ese estadístico puede razonablemente considerarse que tiene lugar según H_0, con referencia a algún nivel predeterminado de significación. Si la probabilidad asociada con la ocurrencia según H_0 de un valor tan grande como el valor observado del estadístico es igual o menor que nuestro nivel predeterminado de significación, esto es, si $p \leq \alpha$, entonces podemos rechazar H_0 y concluir que la asociación observada en dicha muestra no es el resultado de una desviación al azar de la independencia en la población, sino que representa una relación genuina entre las variables en la población. Sin embargo, si la prueba estadística revela que es probable que nuestro valor observado puede haberse originado según H_0, esto es, $p > \alpha$, entonces nuestro datos no nos permiten concluir que existe una relación entre las variables en la población de la cual se extrajo la muestra; esto es, no podemos concluir que las variables no son independientes en la población. Este método de probar hipótesis debe ahora ser completamente familiar al lector. En el capítulo 1 se proporciona una explicación más completa del método y los ejemplos de su uso se encuentran a lo largo de todo el libro.

Ahora el lector debe conocer que la significación del coeficiente de correlación producto-momento r de Pearson, puede ser probada por medio del método descrito anteriormente. Más adelante en este capítulo se descubrirá que la significación de las diferentes medidas de asociación no paramétricas se prueban exactamente por el mismo método. Sin embargo, el coeficiente de Cramér como tal es un caso especial. Una razón de que no nos referimos a la distribución muestral de C para probar la significación de una C observada, es que las complejidades matemáticas de tal procedimiento son considerables. Sin embargo, una razón mejor es que, en el curso del cálculo del valor de C, calculamos un estadístico que proporciona por sí mismo una indicación simple y adecuada de la significación de C. Este valor es, naturalmente, X^2 que se distribuye como χ^2 cuando el tamaño de la muestra es grande. Podemos probar si una C observada difiere significativamente de cero simplemente al determinar la significación del estadístico X^2 para la tabla de contingencia asociada, debido a que C es una función lineal de X^2. Ya que sabemos que la distribución muestral de X^2, conocemos la de C^2 y, por tanto, la de C.

Para cualquier tabla de contingencia $r \times k$, podemos determinar la significación del grado de asociación (la significación de C) averiguando la probabilidad asociada con la ocurrencia, cuando H_0 es cierta, de valores tan grandes a los valores observados de X^2, con $gl = (r - 1)(k - 1)$. Si esa probabilidad es igual o menor que α, la hipótesis nula puede ser rechazada en ese nivel de significación. En la tabla C del Apéndice I se proporciona la probabilidad asociada con la ocurrencia según H_0 de valores tan grandes como una X^2 observada. Si la X^2 para el estadístico de la muestra es significativo, entonces podemos concluir que en la población la asociación entre las dos series de atributos no es cero, esto es, que los atributos o las variables no son independientes.

Ejemplo. Hemos mostrado en el último ejemplo que la relación entre miembros de organización y la disposición de respuestas al cuestionario es 0.51 al ser medida por el coeficiente C de Cramér. En el curso del cálculo de C, determinamos que $X^2 = 75.25$. Ahora bien, si consideramos a los individuos a quienes los cuestionarios fueron enviados como una muestra aleatoria de la población de individuos responsables del proceso de establecimiento de estándares, esto es, una población de personas que reúnen el criterio de selección del estudio, podemos probar si ser miembro de la organización está asociado con la disposición a responder al probar la significación de $X^2 = 75.25$. Con referencia a la tabla C del Apéndice I, podemos determinar que $X^2 \geq 75.25$ con $gl = (r - 1)(k - 1) = (3 - 1)(6 - 1) = 10$, tiene una probabilidad de ocurrencia cuando H_0 es cierta, menor que 0.001. Así, podemos rechazar H_0 en el nivel de significación $\alpha = 0.001$ y concluir que la disposición a responder el cuestionario varía entre las diferentes organizaciones encuestadas. Esto es, podemos concluir que ya que C es significativamente diferente de cero, la asociación en la población es mayor que cero.

Resumen del procedimiento

Éstos son los pasos que hay que seguir para el uso del coeficiente de Cramér:

1. Arregle las frecuencias observadas en unta tabla de contingencia $r \times k$ como la tabla 8.1; donde r es el número de categorías en las cuales se clasifica una variable y k es el número de categorías en las cuales se clasifica la otra variable.
2. Determine la frecuencia esperada según H_0 para cada celda, al multiplicar los datos totales marginales comunes a esa celda y dividiendo después este producto por el número total de casos N. Es decir, para cada celda en la tabla de contingencia calcule $E_{ij} = R_iC_i/N$. Si más de cerca del 20 % de las celdas tienen frecuencias esperadas menores que cinco o si cualquiera de las celdas tiene una frecuencia esperada menor que uno combine las categorías (ya sea filas o columnas) para incrementar las frecuencias esperadas que sean deficientes (véase el capítulo 5).
3. Mediante la ecuación (5.2) o la (5.2a), calcule el valor de X^2 para los datos.
4. Use este valor de X^2 para calcular el valor de C mediante la ecuación (8.1).
5. Para probar si el valor observado de C indica que existe una asociación sig-

nificativa entre las dos variables en la población muestreada, determine la probabilidad asociada según H_0 de un valor tan grande como el observado X^2 con $gl = (r - 1)(k - 1)$, consultando la tabla C del Apéndice I. Si esa probabilidad es igual o menor que, rechace H_0 es favor de H_1.

Limitaciones del coeficiente de Cramér

La amplia aplicabilidad y relativa facilidad de cálculo de C puede hacer parecer que ésta es una medida ideal de asociación. Aunque resulta extremadamente útil, hay algunas limitaciones o deficiencias del estadístico con las cuales el investigador debe estar familiarizado.

En general, es deseable que un índice de correlación muestre al menos las siguientes características: 1. cuando las variables sean independientes y exista una carencia completa de asociación entre las variables, el valor del índice debe ser cero, y 2. cuando las variables muestren completa dependencia una de la otra, esto es, cuando estén perfectamente correlacionadas, el estadístico debe ser igual a la unidad o uno. El coeficiente de Cramér tiene la primera característica: es igual a cero cuando no existe asociación entre las variables en la muestra. Naturalmente, cuando no existe asociación entre las variables en la población, por lo general observaremos un valor de C en la muestra más grande (pero no significativamente más grande) que cero. Sin embargo, cuando es igual a la unidad, pudiera no ser una correlación "perfecta" entre las variables. Ésta es la primera limitación de C.

Cuando $C = 1$, esto indica que las variables están perfectamente correlacionadas cuando la tabla de contingencia asociada es cuadrada, esto es, cuando $r = k$. En ese caso, cada fila y cada columna tendrán solo una celda *única* en la cual existan frecuencias diferentes de cero. Sin embargo, si la tabla de contingencia no es cuadrada, es aún posible que C sea igual a la unidad. Sin embargo, en este caso existe asociación perfecta entre las variables en solamente *una dirección*. Para entender esta situación, supóngase que $r < k$. Entonces, si $C = 1$, habrá solo una entrada diferente de cero en cada *columna*, pero debe haber algunas *filas* con más de una entrada diferente de cero. (Realmente, habrá $r - k$ celdas "extra" con frecuencias diferentes de cero.) Así, en esta situación, existe una perfecta asociación de la variable *columna* a la variable fila, pero *no existe* una perfecta asociación de la variable *fila* a la variable columna. La relación contraria se sostiene cuando $C = 1$ y $r > k$. Se puede considerar $C = 1$ para una tabla de contingencia no cuadrada como representando una relación perfecta "asimétrica": es perfecta en una dirección, pero no en la otra.

Una segunda limitación de C es que los datos deben ser fáciles de usar con el estadístico X^2, con el propósito de que su significación pueda ser interpretada apropiadamente. El lector debe recordar que la significación de la prueba de independencia ji cuadrada supone que los valores esperados son grandes. En la práctica, la regla común concerniente a los valores esperados es que la prueba puede aplicarse apropiadamente sólo si menos del 20 % de las celdas en la tabla de contingencia tienen frecuencias esperadas menores que cinco y ninguna celda tiene una frecuencia esperada menor que uno.

Una tercera limitación de C es que no resulta directamente comparable con cualquier otra medida de correlación, por ejemplo, la r de Pearson (excepto cuando la tabla de contingencia es 2 $\times$ 2, la r_s de Spearman o la T de Kendall). Estas medidas se aplican a variables ordenadas, mientras que el coeficiente de Cramér es apropiado para usarse con variables categóricas (escala nominal). Aunque por lo general el coeficiente de Cramér no es apropiado para usarse con variables ordenadas, puede emplearse para evaluar el grado de asociación no monotónica entre dos variables ordenadas.

Finalmente, los lectores acostumbrados a pensar en la r^2 (la correlación producto-momento de Pearson al cuadrado) como una proporción de la varianza explicada por la relación entre dos variables, deben ser precavidos contra tal interpretación de C y C^2. Aunque podemos interpretar valores mayores de C como indicadores de un grado de relación más grande que los indicados por valores menores, las diferencias en la magnitud no tienen interpretación directa.

A pesar de estas limitaciones, el coeficiente de Cramér es una medida de asociación extremadamente útil debido a su amplia aplicabilidad. Dicho coeficiente no hace suposiciones acerca de la forma de las distribuciones poblacionales de donde provienen las variables que están siendo evaluadas, y no requiere continuidad subyacente en las variables, sino sólo mediciones categóricas de las mismas. Debido a esta libertad en las suposiciones, C puede usarse frecuentemente para indicar el grado de relación entre dos conjuntos de variables a las cuales ninguna otra medida de asociación que presentaremos es aplicable.

Otra *ventaja* del coeficiente de Cramér es que permite al investigador comparar tablas de contingencia de diferentes tamaños y, lo más importante, tablas basadas en diferentes tamaños de muestra. Aunque es estadístico X^2 no mide la independencia de dos variables, es sensible al tamaño de la muestra. El coeficiente de Cramér hace que las comparaciones de las relaciones obtenidas en diferentes tablas de contingencia resulten más fáciles.

Potencia

Debido a su naturaleza y sus limitaciones, no podemos esperar que el coeficiente de Cramér sea muy potente para detectar una relación en la población. Sin embargo, su facilidad de cálculo y su completa libertad de suposiciones restrictivas hace recomendable su uso si otras medidas de correlación son inaplicables. Debido a que C es una función del estadístico ji cuadrada X^2, su limitado poder de distribución, como el de X^2, tiende a 1 al aumentar el tamaño de N (Cochran, 1952).

Referencias bibliográficas

Para otros detalles relativos al coeficiente de Cramér, se recomienda al lector consultar Kendall (1975) y McNemar (1969).

COEFICIENTE PHI PARA TABLAS 2 × 2: r_ϕ

Función

El *coeficiente phi* r_ϕ es una evaluación de la asociación o relación entre dos conjuntos de atributos medidos en una escala nominal, cada uno de los cuales puede tomar sólo dos valores. De hecho, es idéntico en valor al coeficiente de Cramér presentado en la sección anterior.[3] Se supondrá que el lector ya ha leído esa sección; así, la presentación aquí será breve.

Método

Para calcular el coeficiente phi, es conveniente arreglar los datos en una tabla de 2 × 2. Ya que los datos son dicotómicos, supondremos que los datos son codificados como cero y uno para cada variable, aunque puede ser usada cualquier asignación de valor binario.

	Variable X		
Variable Y	*0*	*1*	*Total*
1	A	B	$A + B$
0	C	D	$C + D$
Total	$A + C$	$B + D$	N

Ya que esta tabla de contingencia es mucho más simple que la tabla de contingencia general descrita en la sección anterior, hemos reemplazado las frecuencias de la celda n_{ij} con A, B, C y D. El coeficiente phi para una tabla 2 × 2 es definido como

$$r_\phi = \frac{|AD - BC|}{\sqrt{(A + B)(C + D)(A + C)(B + D)}} \tag{8.2}$$

cuyo rango puede ser desde cero hasta uno. El coeficiente phi está relacionado con el estadístico X^2 que se usa para probar la independiencia de variables categóricas (medidas nominalmente). De aquí que la significación del coeficiente phi pueda probarse al usar el estadístico X^2 presentado en el capítulo 5:

[3] En algunas otras referencias, el coeficiente phi r_ϕ se define para *todas* las tablas de contingencias. Aquí se examina en el contexto de tablas de 2 × 2 sólo debido a la superioridad del coeficiente C de Cramér para otras tablas. Una desventaja de r_ϕ como índice de asociación para tablas más grandes, es que no es igual a la unidad cuando existe una asociación perfecta en tablas de frecuencia no cuadradas (véase la secc. anterior).

$$X^2 = \frac{N(|AD - BC| - N/2)^2}{(A + B)(C + D)(A + C)(B + D)} \tag{5.3}$$

el cual, como hemos visto en esa sección, está distribuido como χ^2 con $gl = 1$. Este estadístico prueba la hipótesis H_0 de que el coeficiente phi en la población de la cual las variables fueron muestreadas, es cero (esto es, que las variables son independientes), contra la hipótesis H_1 de que las variables están relacionadas.

Se advierte que si el tamaño de la muestra es pequeño, la significación de r_ϕ puede probarse mediante la prueba exacta de Fisher (véase el capítulo 5).

Ejemplo. En un experimento que implicaba los efectos de la conducta trasmitida por los medios sobre las preferencias individuales, se diseñó un experimento en el cual una audiencia podía mostrar aprobación (por medio de aplausos) de la presentación de un orador en un grupo grande de discusión.[4] El tema de la discusión era si los miembros de partidos políticos radicales deberían rehusar o no empleos públicos. Había dos oradores, uno en favor de cada postura. Había dos condiciones: en una, la audiencia mostraba fuerte aprobación de un argumento (en favor) y en la otra, la audiencia mostraba fuerte aprobación del otro argumento (en contra). A los sujetos que vieron el debate y las reacciones de la audiencia se les pidió que indicaran su propia preferencia por uno de los dos oradores. Los investigadores suponían que la aprobación de la audiencia debería afectar la preferencia de sus sujetos; específicamente, el orador aplaudido debería ser más favorecido y el no aplaudido, menos favorecido. Los sujetos calificaron su propia posición sobre la cuestión antes y después del debate. Los datos consistieron en el cambio en esas calificaciones. La magnitud del cambio fue ignorado, y sólo la *dirección* del cambio fue codificada. Los datos del experimento están resumidos en la tabla 8.2.

Los investigadores deseaban determinar la fuerza de la relación entre la conducta de la audiencia y el cambio en la preferencia de los observadores. Ya que los datos son dicotómicos y sólo categóricos, el coeficiente phi es el índice apropiado. Al usar los datos de la tabla 8.2, el valor de r_ϕ puede ser determinado con la ecuación (8.2):

$$r_\phi = \frac{|AD - BC|}{\sqrt{(A + B)(C + D)(A + C)(B + D)}} \tag{8.2}$$

$$= \frac{|(21)(14) - (37)(26)|}{\sqrt{(21 + 37)(26 + 14)(21 + 26)(37 + 14)}} = 0.28$$

Así, existe una moderada relación entre los cambios en la preferencia de los sujetos y la aprobación de la audiencia. Para determinar si esta relación es significativa, se usa la prueba X^2 para tabla de contingencia de 2×2 [ecuación (5.3)]:

$$X^2 = \frac{N(|AD - BC| - N/2)^2}{(A + B)(C + D)(A + C)(B + D)} \tag{5.3}$$

$$= \frac{98\,[|(21)(14) - (37)(26)| - 98/2]^2}{(21 + 37)(26 + 14)(21 + 26)(37 + 14)} = 6.75$$

[4] Stocker-Kreichgauer, G. y von Rosenstiel, L., "Attitude change as a function of the observation of vicarious reinforcement and friendliness/hostility in a debate", en B. Brandstatter, J. H. Davis y G. Stocker-Kreichgauer (eds.), *Group decision making*, Academic Press, Nueva York, 1982, págs. 241-255.

Tabla 8.2. Número de gente que cambia su preferencia hacia el orador (en favor o en contra).

Audiencia que apoya al orador	*Cambio en la preferencia hacia*		
	En favor	*En contra*	*Total*
En contra	21	37	58
En favor	26	14	40
Total	47	51	98

Ya que el estadístico X^2 está distribuido como χ^2 con $gl = 1$, podemos determinar su significación y, de aquí, la significación de r_ϕ al consultar la tabla C del Apéndice I. En esta tabla se muestra que $X^2 \geq 6.75$ con $gl = 1$ tiene una probabilidad de ocurrencia cuando H_0 es cierta menor que 0.01. Así, podemos rechazar H_0 en el nivel de significación $\alpha = 0.01$ y concluir que la reacción de la audiencia tuvo un efecto sobre la preferencia hacia los oradores (y sus argumentos) en los debates y que la relación entre los cambios en la preferencia y la aprobación de la audiencia es diferente de cero.

Resumen del procedimiento

Éstos son los pasos al usar el coeficiente phi:

1. Arregle las frecuencias observadas en una tabla de contingencia 2×2.
2. Use las frecuencias en la tabla 2×2 para calcular el coeficiente r_ϕ phi mediante la ecuación (8.2).
3. Para probar si el valor observado de r_ϕ indica que existe una asociación significativa entre las dos variables en la población muestreada, determine el estadístico asociado ji cuadrada X^2 usando la ecuación (5.3). Después, determine la probabilidad según H_0 de obtener un valor tan grande como el observado X^2 con $gl = 1$, consultando la Tabla C del Apéndice I. Si esa probabilidad es igual o menor que α, rechace H_0 en favor de H_1.

Potencia-eficacia

Ya que la prueba para el coeficiente phi es similar a la prueba del coeficiente de Cramér (ambos están basados en la distribución χ^2), el lector puede referirse a la explicación de la potencia en la sección anterior. Sin embargo, el lector debe estar consciente de que, si las variables están ordenadas, el sacrificio de información para formar la tabla 2×2 y calcular el coeficiente phi es muy grande. Para variables ordenadas. El investigador debe usar uno de los métodos presentados en secciones subsecuentes de este capítulo.

Referencias bibliográficas

Se recomiendan al lector las referencias de la sección previa y de la sección "Prueba ji cuadrada para dos muestras independientes", del capítulo 5.

COEFICIENTE DE CORRELACIÓN r_s DE SPEARMAN DE RANGOS ORDENADOS

Función

De todos los estadísticos basados en rangos, *el coeficiente de correlación r_s de Spearman de rangos ordenados*, fue el primero en desarrollarse y quizá sea el mejor conocido actualmente. Es una medida de asociación entre dos variables que requiere que ambas estén medidas en al menos una escala ordinal, de manera tal que los objetos o individuos en estudio puedan ser colocados en rangos en dos series ordenadas.

Racionalización

Supóngase que N individuos son ordenados en rangos en cada una de dos variables. Por ejemplo, podríamos arreglar un grupo de estudiantes en orden de sus puntuaciones en una prueba de admisión a un colegio y de nuevo en orden de su puntuación promedio al final del año escolar. Si los *rangos* de los estudiantes en la prueba de admisión se denotan como $X_1, X_2, \ldots, X_N$ y los *rangos* de la puntuación promedio son representados por $Y_1, Y_2, \ldots, Y_N$, podemos usar una medida de correlación de rangos ordenados para determinar la relación de los X y los Y.

Podemos ver que la correlación entre los rangos del examen de admisión y los rangos de la puntuación promedio deberían ser perfectos si y sólo si $X_i = Y_i$ para todos los i, esto es, si cada persona tiene el mismo rango en ambas variables. Por tanto, podría parecer lógico usar las variadas diferencias

$$d_i = X_i - Y_i$$

como una indicación de la disparidad entre los dos conjuntos de rangos. Supóngase que Mary McCord recibió la puntuación más alta en el examen de admisión, pero se colocó en el quinto lugar en su clase en la puntuación promedio. Su d debería ser $1 - 5 = -4$. John Stanislowski, por otra parte, se colocó en décimo lugar en el examen de admisión, pero encabeza la clase en la puntuación promedio; para él, $d = 10 - 1 = 9$. La magnitud de estas diferentes d nos da una idea de qué tan cercana es la relación entre las puntuaciones del examen de admisión y el logro académico. Si la relación entre los dos conjuntos de rangos fuera perfecta, cada d debería ser cero. Mientras más grandes sean las d_i, menos perfecta es la asociación entre las dos variables.

Al computar un coeficiente de correlación podría haber problemas o inconvenientes para usar directamente las d_i. Una dificultad es que las d_i negativas podrían cancelar a las positivas cuando tratamos de determinar la magnitud total de la discrepancia entre los rangos, aun pensando que es la *magnitud*, más que el signo de la discrepancia, la que representa un índice de la disparidad de los rangos. Sin embargo, si se emplea d_i^2 en lugar de d_i, esta dificultad se elimina. Queda claro que mientras más grandes sean las diferentes d_i, más grande será el valor de Σd_i^2, que es la suma del cuadrado de las diferencias para N pares de datos.

La derivación de la fórmula computada para r_s es bastante sencilla. Se hace simplificando la fórmula del coeficiente de correlación producto-momento r de Pearson cuando los datos están compuestos por rangos. Proporcionaremos dos expresiones alternativas para r_s. Una de estas formas alternativas es útil en el cálculo del coeficiente y la otra se usará posteriormente cuando encontremos necesario corregir el coeficiente, cuando se presenten puntuaciones empatadas en los datos. Si $x = X - \overline{X}$, donde $\overline{X}$ es la media de las *puntuaciones* en la variable, X, y si $y = Y - \overline{Y}$, donde $\overline{Y}$ es la media de las *puntuaciones* en la variable Y, entonces una expresión general para el coeficiente de correlación producto-momento de Pearson es

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} \tag{8.3}$$

en el cual las sumas están sobre los valores de N en la muestra.[5] Ahora cuando las X y las Y son rangos, $r = r$. Conociendo que los datos que están en rangos, podemos simplificar la ecuación (8.3) para obtener la expresión siguiente para el coeficiente de correlación de Spearman de rangos ordenados:

$$r_s = \frac{\Sigma x^2 + \Sigma y^2 - \Sigma d^2}{2\sqrt{\Sigma x^2 \Sigma y^2}} \tag{8.4}$$

y

$$r_s = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N^3 - N} \tag{8.5}$$

Recuérdese que $d_i = X_i - Y_i$, la diferencia en rangos en las dos variables. La simplificación de la ecuación (8.4) a la forma proporcionada en la ecuación (8.5) es posible notando que cuando los datos están en rangos y no existen empates en los datos, $\Sigma x^2 = \Sigma y^2 = (N^3 - N)/12$. Debe notarse que si existen empates, el uso de la ecuación (8.3) o de la (8.4) proporcionará el valor correcto de r_s; más tarde se proporcionará una correccción para los empates para la ecuación (8.5).

Método

Para computar r_s, haga una lista de los N sujetos u observaciones. En seguida, a cada sujeto asignese el rango para la variable X y el rango para la variable Y, asig-

[5] En esta sección usaremos la forma abreviada del operador de suma Σ, en el que omitiremos el índice para la sumatoria así como el subíndice para la variable indexada. El contexto debería esclarecer sobre qué variables y rangos se toma la suma. En este caso, la sumatoria es sobre todas las N variables.

nándose el rango de 1 a la X más pequeña y el rango de N a la X más grande, etc. Determine en seguida los valores de d_i, que es la diferencia entre los rangos y Y para la iésima observación. Obtenga el cuadrado de cada d_i, y después sume todos los valores de d_i^2 para obtener Σd_i^2. Luego, coloque este valor y el valor de N (el número de observaciones o sujetos) directamente en la ecuación (8.5).

Ejemplo. Como parte de un estudio sobre el efecto de la presión del grupo sobre el conformismo individual en una situación que implica riesgo monetario, dos investigadores[6] administraron la escala F, una medida de autoritarismo y una escala diseñada para medir estatus de lucha social[7] a 12 estudiantes. Se deseaba tener información acerca de la correlación entre las puntuaciones de autoritarismo y aquellas de estatus de lucha social (dicho estatus fue indicado por el acuerdo en juicios tales como "La gente no debería casarse con personas de un nivel social más bajo", "Asistir al hipódromo es mejor que asistir a un juego de beisbol", "Vale la pena elaborar nuestro árbol genealógico"). En la tabla 8.3 se proporciona cada una de las puntuaciones de los 12 estudiantes en las dos escalas.

Tabla 8.3. Puntuaciones de autoritarismo y estatus de lucha social.

	Puntuaciones	
Sujeto	*Autoritarismo*	*Estatus de lucha social*
A	82	42
B	98	46
C	87	39
D	40	37
E	116	65
F	113	88
G	111	86
H	83	56
I	85	62
J	126	92
K	106	54
L	117	81

Para calcular la correlación de Spearman de rangos ordenados entre estos dos conjuntos de puntuaciones, es necesario ordenar los rangos de las dos series. Los rangos de las puntuaciones proporcionadas en la tabla 8.3 se muestran en la tabla 8.4, que también presenta los diferentes valores de d_i y d_i^2. Así, por ejemplo, en la tabla 8.4 se muestra que el estudiante (sujeto *J*) que exhibió el mayor autoritarismo (en la escala *F*), también exhibió el estatus de

[6] Siegel, S. y Fagan, J., "The Asch effect under conditions of risk" (estudio inédito). Los datos que se presentan aquí corresponden a un estudio piloto.

[7] Siegel, A. E. y Siegel, S., "An experimental test of some hypotheses in reference group theory", (estudio inédito).

lucha social más extremo y, por tanto, se le asignó un rango de 12 en ambas variables. El lector observará que ningún rango de los sujetos en una variable, estuvo más de tres rangos distantes del rango en la otra variable, esto es, la d_i más grande es de tres.

Tabla 8.4. Rangos de autoritarismo y estatus de lucha social.

	Rangos			
Sujeto	*Autoritarismo*	*Estatus de lucha social*	d_i	d_i^2
A	2	3	−1	1
B	6	4	2	4
C	5	2	3	9
D	1	1	0	0
E	10	8	2	4
F	9	11	−2	4
G	8	10	−2	4
H	3	6	−3	9
I	4	7	−3	9
J	12	12	0	0
K	7	5	2	4
L	11	9	2	4
			$\Sigma d_i^2 = 52$	

A partir de los datos mostrados en la tabla 8.4, podemos computar el valor de r aplicando la ecuación (8.5) a estos datos:

$$r_s = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N^3 - N} \tag{8.5}$$

$$= 1 - \frac{6(52)}{(12)^3 - 12} = 0.82$$

Observamos que para estos 12 estudiantes, la correlación entre el autoritarismo y el estatus social es $r_s = 0.82$.

Observaciones empatadas

Ocasionalmente, dos o más sujetos reciben la misma puntuación en la misma variable. Cuando ocurren puntuaciones empatadas, a cada una de ellas se le asigna el promedio de los rangos que habrían sido asignados si no hubieran ocurrido los empates, lo cual es nuestro procedimiento usual para asignar rangos a observaciones empatadas.

Si la proporción de las observaciones empatadas no es grande, su efecto sobre r_s es insignificante y la ecuación (8.5) puede aún usarse para su cálculo. Sin embargo, si la proporción de empates es grande, entonces debe incorporarse un factor de corrección en el cálculo de r_s.

El efecto de los rangos empatados en la variable X es reducir la suma de cuadrados (Σx^2) abajo del valor de $(N^3 - N)/12$, esto es, cuando existen empates,

$$\Sigma x^2 < \frac{N^3 - N}{12}$$

Por tanto, es necesario corregir la suma de cuadrados, teniendo en cuenta los empates (los rangos empatados no surten efecto en la media o Σx lo cual es siempre $= 0$). El factor de corrección es

$$T_x = \sum_{i=1}^{g} (t_i^3 - t_i) \tag{8.6}$$

donde g es el número de grupos de diferentes rangos empatados y r es el número de rangos empatados en el i-ésimo grupo. Cuando la suma de cuadrados es corregida por empates, se convierte en

$$\Sigma x^2 = \frac{N^3 - N - T_x}{12}$$

Los empates que ocurren en la variable Y requieren corrección de la misma manera, y el factor de corrección se denota T_y. Cuando está presente un número considerable de empates, para calcular r_s puede suponerse una de las siguientes ecuaciones:

$$r_s = \frac{\Sigma x^2 + \Sigma y^2 - \Sigma d^2}{2\sqrt{\Sigma x^2 \Sigma y^2}} \tag{8.4}$$

o

$$r_s = \frac{(N^3 - N) - 6\Sigma d^2 - (T_x + T_y)/2}{\sqrt{(N^3 - N)^2 - (T_x + T_y)(N^3 - N) + T_x T_y}} \tag{8.7}$$

Ejemplo con empates. En el estudio citado en el ejemplo anterior, cada estudiante se observó individualmente en la situación de presión de grupo desarrollada por Asch.[8] En esta situación, a un grupo de sujetos se les pedía individualmente que eligieran cuál de un conjunto de líneas alternativas era de la misma longitud que una línea estándar. Todos menos uno de los sujetos eran confederados del investigador, y en ciertos ensayos ellos elegían desanimadamente una línea incorrecta. El sujeto ingenuo, que estaba sentado de tal manera que era la última persona a la que se le pedía mencionar su juicio, tenía la opción de pararse sólo al seleccionar la línea correcta (lo cual era inequívoco para la gente en las si-

[8] Asch, S. E., *Social psychology*, Prentice-Hall, Nueva York, 1952, págs. 451-476.

tuaciones donde ninguna presión de grupo contradictorio está implícita) o "responder" a la presión del grupo estableciendo que la línea incorrecta era la igual.

La modificación que Siegel y Fagan introdujeron en este experimento fue acordar pagar a cada sujeto 50 centavos por cada juicio correcto y penalizarlo con 50 centavos por cada juicio incorrecto. Se le dieron dos dólares a cada sujeto al comienzo del experimento, y ellos entendieron que podían guardar todo el dinero que poseyeran al final de la sesión. Como ya sabían los sujetos ingenuos, este acuerdo se había hecho con todos los miembros del grupo que emitía los juicios. Cada sujeto ingenuo participó en 12 igualaciones "decisivas". Cada sujeto ingenuo podía "responder" tanto como 12 veces.

Como parte del estudio, los investigadores querían saber si responder en esta situación estaba correlacionado con el estatus de lucha social, que se midió con la escala descrita en el ejemplo previo. Esto fue determinado al calcular la correlación de Spearman de rangos ordenados entre las puntuaciones de cada uno de los 12 sujetos ingenuos sobre la escala de estatus de lucha social y el número de veces que cada uno de ellos respondía hacia la presión grupal. Los datos de estas dos variables se presentan en la tabla 8.5. Obsérvese que dos de los sujetos ingenuos no respondieron del todo (sujetos A y B), mientras que sólo un sujeto (sujeto L) respondió en cada ensayo decisivo. Los rangos para las puntuaciones originales enumeradas en la tabla 8.5 están dadas en columnas separadas en esa tabla. Obsérvese que para estos datos existen tres grupos de observaciones empatadas en la variable X (número de respuestas). Cuando existen empates, el rango asignado es el *promedio de los rangos* que habrían sido asignados si los valores hubieran diferido ligeramente.[9] Dos sujetos empataron en cero; a ambos se les dio el rango de 1.5. Dos empataron en uno; a ambos se les dio el rango de 3.5. Dos empataron en ocho; a ambos se les dieron rangos de 10.5.

Tabla 8.5. Puntuaciones originales y rangos sobre el rendirse y el estatus de lucha social.

	Número de rendiciones		*Estatus de lucha social*			
Sujeto	*Datos*	*Rango*	*Datos*	*Rango*	d_i	d_i^2
A	0	1.5	42	3	− 1.5	2.25
B	0	1.5	46	4	− 2.5	6.25
C	1	3.5	39	2	1.5	2.25
D	1	3.5	37	1	2.5	6.25
E	3	5	65	8	− 3.0	9.00
F	4	6	88	11	− 5.0	25.00
G	5	7	86	10	− 3.0	9.00
H	6	8	56	6	2.0	4.00
I	7	9	62	7	2.0	4.00
J	8	10.5	92	12	− 1.5	2.25
K	8	10.5	54	5	− 5.5	30.25
L	12	12	81	9	3.0	9.00
					$\Sigma d_i^2 =$	109.50

[9] En esta sección se supone que el lector sabe cómo ejecutar los rangos de los datos cuando existen empates en las observaciones. El procedimiento para puntear rangos empatados se examina con detalle en el capítulo 4.

Debido a la proporción relativamente grande de observaciones empatadas en la variable X, para calcular el valor de r_s se debe usar la ecuación (8.7). Para usar esa ecuación, debemos primero determinar los valores de Σx^2 y Σy^2 corregidos para empates, esto es, debemos encontrar T_x y T_y.

Ahora con $g = 3$ grupos de observaciones ligadas en la variable X, donde $t_i = 2$ en cada conjunto, tenemos

$$T_x = (2^3 - 2) + (2^3 - 2) + (2^3 - 2)$$
$$= 18$$

y

$$\Sigma x^2 = \frac{N^3 - N - T_x}{12}$$
$$= \frac{12^3 - 12 - 18}{12}$$
$$= 141.5$$

Esto es, la corrección para empates, $\Sigma x^2 = 141.5$. Encontramos Σy^2 por un método comparable. Sin embargo, ya que no hay empates en las puntuaciones de Y (las puntuaciones de lucha social) en estos datos, $T_y = 0$ y

$$\Sigma y^2 = \frac{N^3 - N - T_y}{12}$$
$$= \frac{12^3 - 12 - 0}{12}$$
$$= 143$$

Así, corregida para empates, $\Sigma x^2 = 141.5$ y $\Sigma y^2 = 143$. De la adición mostrada en la tabla 8.5, sabemos que $\Sigma d^2 = 109.5$. Sustituyendo estos valores en la ecuación (8.7), tenemos

$$r_s = \frac{(N^3 - N) - 6\Sigma d^2 - (T_x + T_y)/2}{\sqrt{(N^3 - N)^2 - (T_x + T_y)(N^3 - N) + T_xT_y}} \tag{8.7}$$
$$= \frac{1\,716 - 6(109.5) - 18/2}{\sqrt{1\,716^2 - (18)(1\,716) + 0}}$$
$$= \frac{1\,050}{1\,706.976}$$
$$= 0.615$$

Corrigiendo para empates, la correlación entre la cantidad de respuestas y el grado de estatus social de lucha es $r_s = 0.615$. Si hubiéramos calculado r_s con la ecuación (8.5), esto es, si no hubiéramos corregido para empates, habríamos encontrado $r_s = 0.617$. Esto ilustra el efecto relativamente pequeño de los empates sobre el valor del coeficiente de correlación de Spearman de rangos ordenados, cuando existen pocos grupos de empates o el número de empates dentro de un grupo de empates es pequeño. Nótese, sin embargo, que el efecto de los empates en los rangos es inflar el valor de la correlación r_s (no corregida). Por esta razón, debe usarse la corrección cuando existe una gran proporción de empates ya sea en una o en ambas variables X y Y o bien, si el número de empates en un grupo de empates es grande.

Prueba de la significación de r_s

Si los sujetos cuyas puntuaciones se emplean para calcular r_s fueron elegidos aleatoriamente de una población, podermos usar esas puntuaciones para determinar si las dos variables están asociadas en la población. Esto es, podemos probar la hipótesis nula de que las dos variables en estudio *no* están asociadas (esto es, son independientes) en la población y el valor observado de r_s difiere de cero sólo al azar. Así, podemos probar la hipótesis H_0: no existe asociación entre X y Y, contra la hipótesis H_1: existe asociación entre X y Y (una prueba bidireccional) o H_1: existe una asociación positiva (o negativa) entre X y Y (una prueba unidireccional). Puede notarse que no hemos especificado las dos hipótesis como H_0: $\rho_s = 0$ contra H_1: $\rho_s \neq 0$ debido a que no es probable el caso en el cual las variables estén normalmente distribuidas, $\rho_s = 0$ no necesariamente significa que las variables son independientes, mientras que si son independientes, entonces $\rho_s = 0$. Como resultado, debemos tener cuidado al interpretar la significación de r_s.

Muestras pequeñas

Supóngase que la hipótesis nula es cierta. Esto es, supóngase que no existe relación en la población entre las variables X y Y. Ahora bien, si una muestra de puntuaciones X y Y se toma aleatoriamente de una población, para una ordenación de rangos dada de las puntuaciones Y, cualquier ordenación de rango de las puntuaciones X es tan probable como cualquier otra ordenación de rangos de las puntuaciones X; y para cualquier orden dado en las puntuaciones X, todos los órdenes posibles de las puntuaciones Y son igualmente probables. Para los N sujetos, existen $N!$ rangos posibles de puntuaciones X que pueden ocurrir en asociación con cualesquiera rangos dados de puntuaciones Y. Ya que éstos son igualmente probables, la probabilidad de ocurrencia de cualquier rango particular de puntuaciones X con un rango dado de puntuaciones Y es $1/N!$

Para cada uno de los posibles rangos de Y existirá un valor asociado de r_s. Cuando H_0 es cierta, la probabilidad de ocurrencia de cualquier r_s particular es así proporcional al número de permutaciones que dan lugar a ese valor.

Al usar la ecuación (8.5), la fórmula computacional de r_s, encontramos que para $N = 2$, solo dos valores de r_s son posibles: $+1$ y -1. Cada uno de estos tiene una probabilidad de ocurrencia según H_0 de 1/2.

Para $N = 3$, los posibles valores de r_s son -1, $-1/2$, $+1/2$ y $+1$. Cuando H_0 es cierta, las probabilidades respectivas son 1/6, 1/3, 1/3 y 1/6.

En la tabla Q del Apéndice I se proporcionan los valores críticos de r_s que se han obtenido por un método similar de generar todos los rangos posibles. Para N desde 4 hasta 50, en la tabla se proporcionan los valores críticos de la correlación de rangos ordenados r_s de Spearman según H_0 para varios valores de α entre 0.25 y 0.0005. La tabla es unidireccional, esto es, las probabilidades establecidas se aplican cuando el valor observado de r_s está en la dirección predicha, ya sea positiva o negativa. Si un valor observado de r_s es igual o excede un valor particular de la tabla, ese valor observado es significativo (para una prueba unidireccional) al nivel indicado. Para una prueba bidireccional en la cual la hipótesis alterna H_1 es que las dos variables están relacionadas, pero no hace suposiciones acerca de la dirección de la relación entre ellas, las probabilidades en la tabla Q del Apéndice I se duplican. Por conveniencia, las probabilidades bidireccionales no están anotadas en la tabla.

Ejemplo. Ya hemos encontrado que para $N = 12$, la correlación de Spearman de rangos ordenados entre el autoritarismo y el estatus de lucha social es $r_s = 0.82$. En la tabla Q del Apéndice I se muestra que un valor tan grande como éste es significativo en el nivel $P < 0.001$ (prueba unidireccional). Así podemos rechazar H_0 en el nivel $\alpha = 0.001$ y concluir que, en la población de estudiantes de la cual se obtuvo la muestra, el autoritarismo y el estatus de lucha social no son independientes.

También hemos visto que la relación entre el estatus de lucha social y la cantidad de respuesta a la presión grupal es $r_s = 0.62$ en nuestro grupo de 12 sujetos. Al consultar la tabla Q del Apéndice I, podemos determinar que $r_s \geq 0.62$ tiene una probabilidad de ocurrencia, cuando H_0 es cierta, entre $p = 0.025$ y $p = 0.01$ (prueba unidireccional). Así, podríamos concluir, en el nivel $\alpha = 0.025$, que estas dos variables no son independientes en la población de la cual se extrajo la muestra.

Muestras grandes

Cuando N es más grande que cerca de 20 o 25, la significación de una r_s obtenida según la hipótesis nula también puede ser probada mediante el estadístico

$$z = r_s\sqrt{N - 1} \qquad (8.8)^{10}$$

Para N grande, el valor definido por la ecuación (8.8) está distribuido de manera aproximada normal con media cero y desviación estándar uno. Así, la probabilidad asociada cuando H_0 es cierta para cualquier valor tan extremo como un valor observado de r_s, puede determinarse al calcular la z asociada con ese valor usando la ecuación (8.8), y después determinando la significación de la z al consultar la tabla A del Apéndice I. Aunque la prueba de grandes muestras puede emplearse cuando N es tan pequeña como 20, el uso de la tabla Q del Apéndice I es preferible para $N \leq 50$.

[10] Algunos estudiosos recomiendan el estadístico ligeramente mejor

$$t = r_s\sqrt{\frac{N - 2}{1 - r_s^2}}$$

que está distribuido aproximadamente como la t de Student con $gl = N - 2$ (Tabla B del apéndice I). Debido a la disponibilidad de la tabla Q del Apéndice I que tabula las probabilidades exactas unidireccionales de la distribución muestral de r_s para $N \leq 50$, hemos optado por la expresión más simple de la ecuación (8.8). En la práctica real y con N más grandes, la ventaja de t sobre z es pequeña.

Ejemplo. Ya hemos determinado que la relación entre el estatus de lucha social y la cantidad de respuestas a la presión grupal es $r_s = 0.62$ para $N = 12$. Aunque N es pequeña, usaremos la aproximación de muestras grandes como un ejemplo para probar la significación de esta r_s:

$$\begin{aligned} z &= r_s\sqrt{N-1} \qquad (8.8) \\ &= 0.62\sqrt{12-1} \\ &= 2.05 \end{aligned}$$

En la tabla A del Apéndice I se muestra que una z tan grande como 2.05 es significativa en el nivel 0.05, pero no en el nivel 0.01 para una prueba unidireccional. Éste es esencialmente el mismo resultado que obtuvimos al usar la tabla Q del Apéndice I. En ese caso podemos rechazar H_0 en $\alpha = 0.025$, concluyendo que el estatus de lucha social y la cantidad de respuestas están asociados en la población de la cual fueron muestreados los 12 estudiantes.

Resumen del procedimiento

Éstos son los pasos que hay que seguir en el uso del coeficiente de correlación de Spearman de rangos ordenados:

1. Asigne rangos a los sujetos (observaciones) en la variable X desde 1 hasta N. Asigne rangos a las observaciones de la variable Y desde 1 hasta N. Para X empates (o Y), asigne a cada una el valor promedio de los rangos asociados.
2. Haga una lista de los N sujetos. Coloque el rango de cada sujeto en la variable X y la variable Y.
3. Determine el valor de d_i para cada sujeto sustrayendo el rango de Y del correspondiente rango de X. Eleve al cuadrado este valor para determinar d_i^2. Sume los d_i^2 para los N casos para determinar Σd_i^2.
4. Si la proporción de empates en cualquiera de las observaciones de X o Y es grande, use la ecuación (8.7) para calcular r_s.[11] En otros casos, use la ecuación (8.5).
5. Si los sujetos constituyen una muestra aleatoria de alguna población, se puede probar si el valor observado de r_s indica una asociación entre las variables X y Y en la población. Las hipótesis son H_0: no existe asociación entre X y Y, y H_1: existe una asociación entre X y Y. El método para hacer esto depende del tamaño de la muestra N:

   a) Para N desde 4 hasta 50, los valores críticos de r_s entre los niveles de significación (unidireccionales) 0.25 y 0.0005, están proporcionados en la tabla Q del Apéndice I. Para una prueba bidireccional, las probabilidades de significación correspondiente se duplican.

   b) Para $N > 50$, la probabilidad asociada con un valor tan grande como el

[11] La ecuación (8.3) podría usarse para calcular r_s si existen o no empates, pero su uso puede resultar más problemático. Sin embargo, muchas calculadoras facilitan el cálculo (correcto) de r_s usando la ecuación (8.3) si existen o no empates. La elección de la fórmula se deja al lector.

valor observado de r_s, puede ser aproximado al calcular la z asociada con ese valor usando la ecuación (8.8) y determinando después la significación de ese valor de z al consultar la tabla A del Apéndice I.

6. Si el valor de r_s (o de z) excede el valor crítico, rechace H_0 en favor de H_1.

Eficacia relativa

La eficacia del coeficiente de correlación de Spearman de rangos ordenados, cuando es comparado con las correlaciones paramétricas más poderosas —el coeficiente de correlación producto-momento r, de Pearson—, es cercana al 91 %. Es decir, cuando r_s se usa con una muestra para probar la existencia de asociación en una población para la cual se encuentran las suposiciones y requisitos que subyacen a la r_s de Pearson, esto es, cuando la población tiene una distribución normal bivariada, entonces r_s es 91 % tan eficaz como r para rechazar H_0. Si existe una correlación entre X y Y en esta población, con 100 casos r_s revelará esa correlación con la misma significación con que r lo logra con 91 casos.

Referencias bibliográficas

Para otros detalles acerca del coeficiente de correlación de Spearman de rangos ordenados, el lector puede consultar McNemar (1969) o Gibbons (1985).

COEFICIENTE DE CORRELACIÓN *T* DE KENDALL DE RANGOS ORDENADOS

Función

El *coeficiente de correlación T de Kendall de rangos ordenados* es adecuado como una medida de correlación con la misma clase de datos para los cuales es útil r_s.[12] Esto es, si al menos se han logrado medidas ordinales de ambas variables X y Y, tal que a cada sujeto pueda serle asignado un rango tanto en X como en Y, entonces T_{xy} (o simplemente T, si el contexto es claro), proporcionará una medida del grado de asociación o correlación entre los dos conjuntos de rangos. La distribución muestral de T según la hipótesis nula de independencia es conocida y, por tanto, T como r_s puede ser usada en pruebas de significación.

Una ventaja de T sobre r_s es que T puede ser generalizada a un coeficiente de correlación parcial. Este coeficiente parcial se presenta en la siguiente sección. El coeficiente T también es particularmente adecuado para evaluar el acuerdo entre jueces múltiples, que se examinarán posteriormente.

[12] Algunos autores se refieren al coeficiente examinado en esta sección como τ (tau) de Kendall. Sin embargo, distinguiremos entre T, un estadístico basado en una muestra, y τ, el parámetro de la población.

Racionalización

Supóngase que para poner rango a cuatro objetos preguntamos al juez X y al juez Y. Por ejemplo, podríamos pedirles que pusieran el rango a cuatro ensayos en orden de calidad de estilo de exposición. Representamos los cuatro ensayos como *a*, *b*, *c* y *d*. Los rangos obtenidos fueron:

Ensayo:	*a*	*b*	*c*	*d*
Juez X:	3	4	2	1
Juez Y:	3	1	4	2

Si rearreglamos ei orden de los ensayos de tal modo que los rangos del juez X aparezcan en el orden natural[13] (esto es, 1, 2, . . . , N), tenemos

Ensayo:	*d*	*c*	*a*	*b*
Juez X:	1	2	3	4
Juez Y:	2	4	3	1

Ahora podemos determinar el grado de correspondencia entre los jueces X y Y. Si los rangos del juez X están en su orden natural, procedemos a determinar cuántos pares de rangos en el conjunto del juez Y están en su orden correcto (natural), respecto a aquellos del juez X. Contamos el número de acuerdos en el ordenamiento y el número de desacuerdos en el ordenamiento de rangos observado.

Considérense primero todos los posibles *pares* de rangos en los cuales el rango del juez Y es 2 (el primer rango en este conjunto) y el otro miembro es un rango "posterior" (a la derecha). El primer par (2 — 4) tiene el orden correcto –2 precede a 4. Ya que el orden es "natural", asignamos una puntuación de $+1$ a este par. Los rangos 2 y 3 constituyen el segundo par (2 — 3). Este par está también en el orden correcto, de modo que se le asigna así mismo una puntuación de $+1$. Ahora el tercer par (2 — 1) consiste en los rangos 2 y 1. Estos rangos no están en orden natural –2 precede a 1. Por tanto, asignamos a este par una puntuación de -1. Para todos los pares que incluyen el rango 2, el total de puntuaciones es el siguiente:

$$(+1) + (+1) + (-1) = +1$$

Ahora consideramos todos los posibles pares de rangos que incluyen el rango 4 (que es el segundo rango de la izquierda en el conjunto del juez Y). Los pares son (4 — 3) y (4 — 1); ya que ambos pares no están en el orden natural, se les asigna una puntuación de -1 a cada uno. El total de estas puntuaciones es el siguiente:

$$(-1) + (-1) = -2$$

[13] Por *orden natural* queremos decir el orden en que pueden ser colocados los valores observados de la variable. Debe notarse que la colocación de una variable en el orden natural es necesaria sólo para hacer más fácil el estadístico de rangos ordenados. Más aún, no importa qué variable sea colocada en orden –el investigador puede colocar cualquiera en orden natural–, el valor del estadístico de rangos ordenados resultante no es afectado.

Cuando consideramos el rango 3 y los rangos subsecuentes, existe sólo el par (3 — 1). Los dos miembros de este par están en el orden incorrecto; por tanto, este par recibe una puntuación de -1.

El total de las puntuaciones que hemos asignado es

$$(+1) + (-2) + (-1) = -2$$

Esta suma es el número de acuerdos en el ordenamiento entre los rangos menos el número de desacuerdos en el ordenamiento entre los rangos.

Ahora bien, ¿cuál es el *máximo posible* total que pudiéramos haber obtenido de las puntuaciones asignadas a todos los pares en los rangos del juez Y? El máximo posible total pudiera haber ocurrido si los rangos de los jueces X y Y hubieran acordado perfectamente; para entonces, cuando los rangos del juez X fueron arreglados en su orden natural, cada par de rangos del juez Y pudieran también haber estado en el orden correcto y, por tanto, cada par habría podido recibir una puntuación de $+1$. El máximo posible total, aquel que podría haber ocurrido en el caso de perfecto acuerdo entre X y Y, sería la combinación de cuatro objetos tomados de dos en dos en el tiempo, o

$$\binom{4}{2} = 6$$

lo cual es el número de pares diferentes que pueden hacerse de cuatro objetos.

El grado de relación entre los dos conjuntos de rangos es indicado por la razón del total real de $+1$ y -1 por el máximo posible total, que es el número de pares posibles. El coeficiente de correlación por orden de rangos de Kendall es la razón:

$$T = \frac{\#\text{ de acuerdos} - \#\text{ de desacuerdos}}{\text{número total de pares}} = \frac{-2}{6} = -0.33$$

Esto es, $T = -0.33$ es una medida del acuerdo entre los rangos asignados a los ensayos por el juez X y aquellos asignados por el juez Y.

Se puede pensar que T es una función del número mínimo de inversiones o intercambios entre rangos cercanos que se requieren para transformar un rango en otro. Esto es, T es una clase de coeficiente de desorden.

Método

Hemos visto que

$$T = \frac{\#\text{ de acuerdos} - \#\text{ de desacuerdos}}{\text{número total de pares}}$$

En general, el máximo posible total será $\binom{N}{2}$, que puede ser expresado como $N(N - 1)/2$. Esta última expresión puede ser el denominador de la fórmula para T. Para el numerador, denotemos la suma observada de puntuaciones $+1$ (acuerdos) y puntuaciones -1 (desacuerdos) para todos los pares como S. Entonces

$$T = \frac{2S}{N(N-1)} \tag{8.9}$$

donde N es el número de objetos o individuos colocados en rangos tanto para X como para Y.

Como veremos, el cálculo de S puede abreviarse considerablemente mediante el método descrito cuando examinamos la lógica de la medida.

Cuando los rangos del juez X estaban en el orden natural, los rangos correspondientes del juez Y estaban en este orden:

Juez Y: 2 4 3 1

Podemos determinar S empezando con el primer número de la izquierda y contando el número de rangos a su derecha que son *más grandes*: éstos son los acuerdos en el orden. Sustraemos de éste el número de rangos a su derecha, que son *más pequeños*: éstos son los desacuerdos en el orden. Si hacemos lo mismo para todos los rangos y después sumamos los resultados, obtenemos S. Este procedimiento se delinea enseguida:

Juez Y:	2	4	3	1	Total
	2→	+	+	−	+1
		4→	−	−	−2
			3→	−	−1
				1→	0
			Gran total	=	−2

Así, el número total de acuerdos en el ordenamiento menos el número de acuerdos en el ordenamiento es $S = -2$. Conociendo S, podemos usar la ecuación (8.9) para calcular el valor de T para los rangos asignados por los dos juicios:

$$T = \frac{2S}{N(N-1)} \tag{8.9}$$

$$= \frac{2(-2)}{(4)(4-1)}$$

$$= -0.33$$

Ejemplo. En la sección anterior calculamos la r_s de Spearman para las 12 puntuaciones de los estudiantes sobre autoritarismo y estatus de lucha social. Las puntuaciones de los 12 estudiantes están presentadas en la tabla 8.3 y los rangos de estas puntuaciones se muestran en la tabla 8.4. Podemos calcular el valor de la T de Kendall para estos mismos datos.

Los datos conjuntos de rangos que van a ser correlacionados (mostrados en la tabla 8.4) son los siguientes:

Sujetos:	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*
Rangos del estatus de lucha social:	3	4	2	1	8	11	10	6	7	12	5	9
Rangos de autoritarismo:	2	6	5	1	10	9	8	3	4	12	7	11

Para computar T, rearreglamos el orden de los sujetos de tal modo que los rangos en el estatus de lucha social estén en el orden natural:

Sujetos:	*D*	*C*	*A*	*B*	*K*	*H*	*I*	*E*	*L*	*G*	*F*	*J*	
Rangos del estatus de lucha social:	1	2	3	4	5	6	7	8	9	10	11	12	
Rangos de autoritarismo:	1	5	2	6	7	3	4	10	11	8	9	12	Total
	1→	+	+	+	+	+	+	+	+	+	+	+	+11
		5→	−	+	+	−	−	+	+	+	+	+	+ 4
			2→	+	+	+	+	+	+	+	+	+	+ 9
				6→	+	−	−	+	+	+	+	+	+ 4
					7→	−	−	+	+	+	+	+	+ 3
						3→	+	+	+	+	+	+	+ 6
							4→	+	+	+	+	+	+ 5
								10→	+	−	−	+	0
									11→	−	−	+	− 1
										8→	+	+	+ 2
											9→	+	+ 1
												12→	0
										Gran total =			+ 44

Una vez arreglados los rangos en la variable X en su orden natural, determinamos el valor de S para el correspondiente orden de rangos en la variable Y:

$$\begin{aligned} S &= (11 - 0) + (7 - 3) + (9 - 0) + (6 - 2) + (5 - 2) \\ &\quad + (6 - 0) + (5 - 0) + (2 - 2) + (1 - 2) + (2 - 0) \\ &\quad + (1 - 0) \\ &= 44 \end{aligned}$$

El rango de autoritarismo más lejano hacia la izquierda es uno. Este rango tiene 11 rangos que son más grandes hacia su derecha y cero rangos que son más pequeños, de manera que su contribución para S es $11 - 0 = 11$. El siguiente rango es 5. Tiene siete rangos a su derecha que son más grandes y tres a su derecha que son más pequeños, tal que su contribución para S es $(7 - 3) = 4$. Procediendo en esta forma, obtenemos los diferentes valores mostrados arriba, los cuales se suman para dar por resultado $S = 44$. Nótese que las sumas individuales están proporcionadas en la última columna. Sabiendo que $S = 44$ y $N = 12$, podemos usar la ecuación (8.9) para calcular T:

$$T = \frac{2S}{N(N-1)} \tag{8.9}$$

$$= \frac{2(44)}{(12)(12-1)} = 0.67$$

El valor $T = 0.67$ representa el grado de relación entre el autoritarismo y el estatus de lucha social que muestran los 12 estudiantes.

Observaciones empatadas

Cuando dos o más observaciones están empatadas ya sea en las variables X o Y, utilizamos nuestro procedimiento usual en colocar rangos a las puntuaciones empatadas: se les da a las observaciones ligadas el promedio de los rangos que deberían haber recibido si no hubiera habido empates.

El efecto de los empates es cambiar el denominador de nuestra ecuación para T. En el caso de empates, T se convierte en

$$T = \frac{2S}{\sqrt{N(N-1) - T_x}\ \sqrt{N(N-1) - T_y}} \tag{8.10}$$

donde

$T_x = \Sigma t(t - 1)$, siendo t el número de observaciones empatadas en cada grupo de empates en la variable X

$T_y = \Sigma t(t - 1)$, siendo t el número de observaciones empatadas en cada grupo de empates en la variable Y

La determinación de los valores de t se examinó en la sección anterior. [El lector debe notar que T_x y T_y son diferentes del estadístico similar definido por la ecuación (8.6).] Los cálculos que se requieren para la ecuación (8.10) se ilustran en el siguiente ejemplo.

Ejemplo con empates. Repetiremos un ejemplo que fue presentado en el estudio de la r_s de Spearman. Correlacionamos las puntuaciones de 12 sujetos en una escala que medía el estatus de lucha social con el número de veces de cada respuesta dada a la presión de grupo, al juzgar la longitud de líneas. Los datos para este estudio se presentan en la tabla 8.5 y los rangos correspondientes están en esa misma tabla.

Los primeros dos conjuntos de rangos que van a ser correlacionados (primero presentados en la tabla 8.5) son los siguientes:

Sujetos:	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*
Rangos del estatus de lucha social:	3	4	2	1	8	11	10	6	7	12	5	9
Rangos de responder:	1.5	1.5	3.5	3.5	5	6	7	8	9	10.5	10.5	12

Como es usual, primero rearreglamos el orden de los sujetos, de manera tal que los rangos en la variable X ocurran en un orden natural:

Sujetos:	*D*	*C*	*A*	*B*	*K*	*H*	*I*	*E*	*L*	*G*	*F*	*J*	
Rangos del estatus de lucha social:	1	2	3	4	5	6	7	8	9	10	11	12	
Rangos de responder:	3.5	3.5	1.5	1.5	10.5	8	9	5	12	7	6	10.5	Total
	3.5→	0	−	−	+	+	+	+	+	+	+	+	6
		3.5→	−	−	+	+	+	+	+	+	+	+	6
			1.5→	0	+	+	+	+	+	+	+	+	8
				1.5→	+	+	+	+	+	+	+	+	8
					10.5→	−	−	−	+	−	−	0	−4
						8→	+	−	+	−	−	+	0
							9→	−	+	−	−	+	−1
								5→	+	+	+	+	4
									12→	−	−	−	−3
										7→	−	+	0
											6→	+	1
												10.5→	0
											Gran total =		25

Podemos entonces computar el valor de S en el modo usual:

$$\begin{aligned} S &= (8-2)+(8-2)+(8-0)+(8-0)+(1-5) \\ &\quad + (3-3)+(2-3)+(4-0)+(0-3)+(1-1) \\ &\quad + (1-0) \\ &= 25 \end{aligned}$$

Puede notarse que cuando existen observaciones empatadas, los rangos estarán empatados y ningún rango en un par de comparación precede al otro, de modo que un valor de 0 es asignado en el cálculo de S.

Una vez que hemos determinado que $S = 25$, determinemos ahora los valores de T_x y T_y. No existen empates entre las puntuaciones del estatus de lucha social, esto es, en los rangos de X y, por tanto, $T_x = 0$.

En la variable Y (responder) existen tres conjunto de rangos empatados. Dos sujetos están empatados en el rango 1.5, dos están empatados en 3.5 y dos están empatados en 10.5. En cada uno de estos casos, el número de observaciones ligadas $T = 2$. Por tanto, T_y puede ser calculado:

$$\begin{aligned} T_y &= \Sigma t(t-1) \\ &= 2(2-1) + 2(2-1) + 2(2-1) \\ &= 6 \end{aligned}$$

Con $T_x = 0$, $T_y = 6$, $S = 25$ y $N = 12$, podemos determinar el valor de T usando la ecuación (8.10):

$$T = \frac{2S}{\sqrt{N(N-1) - T_x}\ \sqrt{N(N-1) - T_y}} \qquad (8.10)$$

$$T = \frac{2(25)}{\sqrt{(12)(12-1) - 0}\ \sqrt{(12)(12-1) - 6}}$$

$$= 0.39$$

Si no hubiéramos corregido el coeficiente anterior para empates, esto es, si hubiéramos usado la ecuación (8.9) para calcular T, habríamos encontrado $T = 0.38$. Obsérvese que el efecto de corrección para empates es relativamente pequeño, a menos que la proporción de rangos empatados sea grande o el número de empates en un grupo sea grande.

Comparación de T y r_s

En dos casos hemos calculado tanto T como r_s para los mismos datos. El lector habrá notado que los valores numéricos de T y r_s no son idénticos cuando ambos son calculados de los mismos pares de rangos. Para la relación entre autoritarismo y estatus de lucha social, $r_s = 0.82$, mientras que $T = 0.67$. Para la relación entre estatus de lucha social y número de respuestas a la presión de grupo, $r_s = 0.62$ y $T = 0.39$.

Estos ejemplos ilustran el hecho de que T y r_s tienen diferentes escalas subyacentes y numéricamente no son comparables uno con el otro. Esto es, si medimos el grado de correlación entre las variables A y B usando r_s, y después hacemos lo mismo para A y C usando T, no podemos decir si A está más cercanamente relacionada a B o a C, debido a que hemos usado medidas de correlación no comparables. Sin embargo, puede advertirse que existe una relación entre las dos medidas, que está expresada de mejor manera en la siguiente desigualdad:

$$-1 \leq 3T - 2r_s \leq 1$$

Existen también diferencias en la interpretación de las dos medidas. El coeficiente de correlación de rangos ordenados de r_s de Spearman es el mismo que el coeficiente de correlación producto-momento de Pearson, calculado entre variables cuyos valores consisten en rangos. Por otra parte, el coeficiente de correlación de rangos ordenados de Kendall tiene una interpretación diferente. Ésta es la *diferencia* entre la probabilidad de que, en los datos observados, X y Y estén en el mismo orden y la probabilidad de que los datos de X y Y estén en un orden diferente. T_{xy} es la diferencia en las frecuencias relativas en la muestra y τ_{xy} es la diferencia entre las probabilidades en la población.

Sin embargo, ambos coeficientes utilizan la misma cantidad de información en los datos y, por tanto, ambos tienen la misma sensibilidad para detectar la existencia de asociación en la población. Esto es, las distribuciones muestrales de T y r_s son tales que para un conjunto determinado de datos, ambos conducirán al rechazo de la hipótesis nula (de que las variables no están relacionadas en la población), en el mismo nivel de significación. Sin embargo, debemos recordar que las medidas son diferentes y miden asociación en modos diferentes. Esto se clarificará después de analizar la prueba de significación de τ.

Prueba de significación de T

Si una muestra aleatoria se extrae de alguna población en la cual X y Y no están relacionadas y se les ponen rangos a los miembros de la muestra en X y Y, entonces para cualquier orden dado de los rangos de X, todos los posibles órdenes para los rangos de Y son igualmente probables. Esto es, para un orden dado de rangos X, cualquiera de los posibles órdenes de rangos Y es tan probable de ocurrir como cualquier otro orden posible de los rangos de Y. Supóngase que ordenamos los rangos de X en orden natural, esto es, 1, 2, 3, . . . , N. Para ese orden de los rangos de X, todos los $N!$ posibles de órdenes de los rangos de Y son igualmente probables según

H_0. Por tanto, cualquier orden particular de los rangos de Y tiene una probabilidad de ocurrencia, cuando H_0 es cierta, de $1/N!$

Para cada uno de los $N!$ posibles rangos de Y, existirá un valor asociado de T. Estos posibles valores de T variarán desde $+1$ hasta -1 y pueden ser obtenidos en una distribución de frecuencias. Por ejemplo, para $N = 4$ existen $4! = 24$ posibles arreglos de los rangos de Y, y cada uno de ellos tiene un valor asociado de T. Su frecuencia de ocurrencia cuando X y Y son independientes, se muestra en la tabla 8.6. Podemos computar similares tablas de probabilidades para otros valores de N, pero, naturalmente, al aumentar el valor de N este método se vuelve cada vez más tedioso.

Tabla 8.6. Probabilidad de T según H_0 para $N = 4$.

Valor de T	*Frecuencia de ocurrencia según H_o*	*Probabilidad de ocurrencia según H_o*
-1.0	1	$\frac{1}{24}$
-0.67	3	$\frac{3}{24}$
-0.33	5	$\frac{5}{24}$
0	6	$\frac{6}{24}$
0.33	5	$\frac{5}{24}$
0.67	3	$\frac{3}{24}$
1.0	1	$\frac{1}{24}$

Afortunadamente, para $N > 10$, la distribución muestral de T se aproxima a la distribución normal. Por tanto, para N grande, podemos usar la tabla de la distribución normal (tabla A del Apéndice) para determinar la probabilidad asociada con la ocurrencia de cualquier valor tan extremo como un valor observado de T cuando H_0 es cierta.

Sin embargo, cuando N es 10 o menos, la tabla R del Apéndice I puede emplearse para determinar la probabilidad exacta asociada con la ocurrencia (unidireccional) según H_0, de cualquier valor tan extremo como un T observado. Para tales muestras pequeñas, la significación de una relación observada entre dos muestras de rangos puede ser determinada simplemente encontrando el valor de T y después consultando la tabla R, para determinar la probabilidad (unidireccional) asociada con ese valor. Si la p tabulada $\leq \alpha$, H_0 puede rechazarse. Por ejemplo, supongamos que $N = 8$ y $T = 0.357$. La tabla R_I del Apéndice muestra que $T, \geq 0.357$ para $N = 8$ tiene una probabilidad de ocurrencia, según H_0, de $p = 0.138$.

Cuando el tamaño de la muestra está entre 11 y 30, se puede utilizar la tabla R_{II} del Apéndice. En dicha tabla se proporcionan los valores críticos de la T de Kendall para niveles de significación seleccionados. Cuando N es mayor que 10, T se distribuye aproximadamente en forma normal con

$$\text{Media} = \mu_T = 0$$

y

$$\text{Varianza} = \sigma_T^2 = \frac{2(2N + 5)}{9N(N - 1)}$$

Esto es,

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{3T\sqrt{N(N - 1)}}{\sqrt{2(2N + 5)}} \tag{8.11}$$

está aproximadamente distribuida de manera normal con media cero y varianza uno. Así, la probabilidad asociada cuando H_0 es cierta, con la ocurrencia de cualquier valor observado T tan extremo, puede determinarse al calcular los valores de z definidos por la ecuación (8.11) y después determinando la significación de esa z consultando la tabla A del Apéndice I.

Ejemplo para N > 10. Ya hemos determinado que entre 12 estudiantes, la correlación entre el autoritarismo y el estatus de lucha social es $T = 0.67$. Si consideramos que estos 12 estudiantes son una muestra aleatoria de alguna población, podemos probar la hipótesis de que estas dos variables son independientes en esa población, consultando la tabla R_{II} del Apéndice I. Los datos de esta tabla indican que la probabilidad de obtener un valor de muestra de $T \geqslant 0.67$ cuando H_0 es cierta es menor que 0.005.

Ya que $N > 10$, podemos también usar la aproximación normal para la distribución muestral de T usando la ecuación (8.11):

$$z = \frac{3T\sqrt{N(N - 1)}}{\sqrt{2(2N + 5)}} \tag{8.11}$$

$$z = \frac{(3)(0.67)\sqrt{(12)(12 - 1)}}{\sqrt{2[2(12) + 5]}}$$

$$= 3.03$$

Al consultar la tabla A del Apéndice I, vemos que $z \geqslant 3.03$ tiene una probabilidad de ocurrencia, cuando H_0 es cierta, de $p = 0.0012$. Así, podemos rechazar H_0 al nivel de significación $\alpha = 0.0012$ y concluir que las dos variables no son independientes en la población de la cual se extrajo esta muestra. Esto, naturalmente, es consistente con el resultado obtenido al usar la tabla R_{II} del Apéndice I.

Ya hemos mencionado que T y r_s tienen capacidades similares para rechazar H_0. Esto es, aún pensando que T y r_s son numéricamente diferentes para el mismo conjunto de datos, sus distribuciones muestrales nulas son tales, que con los mismos datos H_0 puede ser rechazada en aproximadamente el mismo nivel de significación por medio de las pruebas asociadas a ambas medidas. Sin embargo, en el caso de no nulidad (cuando H_1 es verdadera), ellas son sensibles a diferentes aspectos de la dependencia entre las variables.

En el presente caso, $T = 0.67$. Asociado con este valor $z = 3.03$, que nos permite rechazar H_0 en $\alpha = 0.0012$. Cuando el coeficiente de correlación por orden de rangos, de Spearman, se calculó para los mismos datos, encontramos $r_s = 0.82$. Cuando aplicamos a ese valor la prueba de significación para r_s [ecuación (8.8)], encontramos que $z = 2.72$. En la tabla A del Apéndice I se muestra que $z \geqslant 2.72$ tiene una probabilidad de ocurrencia, cuando H_0 es cierta, de ligeramente más que 0.003. Así, T y r_s para el mismo conjunto de datos tienen pruebas de significación que rechazan a H_0 en esencialmente el mismo nivel de significación.

Resumen del procedimiento

Éstos son los pasos que hay que seguir en el uso del coeficiente de correlación por orden de rangos T, de Kendall:

1. Ordene las observaciones en la variable X desde 1 hasta N. Ordene las observaciones en la variable Y desde 1 hasta N.
2. Arregle la lista de N sujetos de manera tal que los rangos de los sujetos en la variable X estén en su orden natural; esto es, $1, 2, 3, \ldots, N$.
3. Observe los rangos de Y en el orden en que ocurrieron cuando los rangos X están en el orden antural. Determine los valores de S, el número de acuerdos en el orden menos el número de desacuerdos en el orden, para los órdenes observados en los rangos de Y.
4. Si no hay empates entre las observaciones X o Y, use la ecuación (8.9) para calcular el valor de T. Si existen empates, use la ecuación (8.10).
5. Si los N sujetos constituyen una muestra aleatoria de alguna población, se puede probar la hipótesis de que las variables son independientes en esa población. El método para hacer esto depende del tamaño de N:

   a) Para $N \leqslant 10$, en la tabla R del Apéndice I se proporcionan las probabilidades asociadas (unidireccionales) de un valor tan grande como una T observada.

   b) Para $N > 10$, pero menor que 30, en la tabla R del Apéndice se proporcionan las probabilidades asociadas (unidireccionales) de un valor tan grande como una T observada.

   c) Para $N > 30$ (o para valores de significación intermedios para $10 < N \leqslant 30$), calcule el valor de z asociado con T usando la ecuación (8.11). Se puede entonces usar la tabla A del Apéndice I para determinar la probabilidad asociada de un valor tan grande como la z observada y, de aquí, de T.

6. Si la probabilidad resultante por el método apropiado es igual o menor que α, H_0 puede ser rechazada en favor de H_1.

Eficacia

La r_s de Spearman y la T de Kendall son similares en su capacidad para rechazar H_0, puesto que hacen un uso semejante de la información en los datos. Cuando se usan con datos para los cuales el coeficiente de correlación producto-momento r de Pearson es apropiadamente aplicable, tanto T como r_s tienen una eficacia del 91 %. Esto es, T es una prueba de independencia de dos variables en una población normal bivariada con una muestra de 100 casos, aproximadamente tan sensible como la r de Pearson con 91 casos (Moran, 1951).

Referencias bibliográficas

El lector encontrará otros detalles útiles de la τ de Kendall en Kendall (1975) y Everitt (1977).

COEFICIENTE DE CORRELACIÓN PARCIAL $T_{xy.z}$ DE *KENDALL DE RANGOS ORDENADOS*

Función

Cuando se observa correlación entre dos variables, existe siempre la posibilidad de que la correlación se deba a la asociación entre cada una de las dos variables y una tercera variable. Por ejemplo, entre un grupo de niños de escuela elemental de diversas edades, se puede encontrar una alta correlación entre el tamaño del vocabulario y la estatura. Esta correlación no puede reflejar alguna relación genuina o directa entre estas dos variables, sino ser el resultado del hecho de que tanto el tamaño del vocabulario como la estatura estén asociadas con una tercera variable: la edad.

Estadísticamente, este problema puede ser atacado por métodos de *correlación parcial*. En la correlación parcial, se eliminan los efectos de variación en una tercera variable sobre la relación entre las variable X y Y. En otras palabras, se encuentra la correlación entre X y Y manteniéndose constante la tercera variable Z.

Al diseñar un experimento, se tiene la alternativa de introducir controles experimentales para eliminar la influencia de una tercera variable o bien, de usar métodos estadísticos para eliminar su influencia. Por ejemplo, se puede desear estudiar la relación entre la habilidad de memorización y la habilidad para resolver cierta clase de problemas. Ambas habilidades pueden estar relacionadas con la inteligencia; por tanto, para determinar la relación directa de una con la otra, debe ser controlada la influencia de las diferencias en la inteligencia. Para efectuar control *experimental*, debemos elegir sujetos con igual inteligencia. Pero si los controles experimentales no son factibles, entonces se deben aplicar controles *estadísticos*. Por medio de la técnica de correlación parcial, podemos mantener constante el efecto de la inteligencia sobre la relación entre la habilidad de memorización y la habilidad para resolver problemas y, por tanto, determinar la extensión de la relación directa o no contaminada entre estas dos habilidades.

En esta sección presentaremos un método de control estadístico que puede usarse con la correlación τ de rangos ordenados de Kendall. Para usar este método no paramétrico de correlación parcial, debemos tener datos que estén medidos en al menos una escala ordinal. No se necesita hacer suposiciones acerca de la forma de la distribución de puntuaciones en la población.

Racionalización

Supóngase que obtenemos rangos de cuatro sujetos en tres variables X, Y y Z. Queremos determinar la correlación entre X y Y cuando Z esta parcializada, esto es, se mantiene constante. Los rangos son los siguientes:

Sujetos:	*a*	*b*	*c*	*d*
Rango en Z:	1	2	3	4
Rango en X:	3	1	2	4
Rango en Y:	2	1	3	4

Ahora bien, si consideramos los posibles pares de rangos en cualquier variable, sabemos que para estos sujetos existen $\binom{4}{2}$ pares posibles: cuatro objetos tomados de dos en dos. Habiendo arreglado los rangos de Z en orden natural, examinamos cada par posible en los rangos de X, los rangos de Y y los rangos de Z. Asignamos una $+$ para cada uno de aquellos pares en los cuales la variable con el rango inferior precede a la variable con el rango superior, y un $-$ a cada par en el cual la variable con el rango mayor precede al inferior:

	Par					
Rango	(*a*, *b*)	(*a*, *c*)	(*a*, *d*)	(*b*, *c*)	(*b*, *d*)	(*c*, *d*)
Z	+	+	+	+	+	+
X	−	−	+	+	+	+
Y	−	+	+	+	+	+

Primero, nótese que la variable Z está en orden natural; todos sus pares precedentes son codificados como $+$. En seguida, nótese que para la variable X, la puntuación para el par (a, b) es codificado como $-$, debido a que los rangos para a y b, 3 y 1, respectivamente, ocurren en el orden "equivocado": la variable con el rango más alto precede al más bajo. Para la variable X, la puntuación para el par (a, c) también es codificada como $-$, debido a que el rango de a, que es 3, es más alto que el rango de c, que es 2. Para la variable Y, el par (a, c) recibe $+$ debido a que el rango de a, que es 2, es más bajo que el rango de c, que es 3.

Podemos resumir la información que hemos obtenido colocándola en una tabla de 2×2, la tabla 8.7. Considérese primero los tres signos bajo (a, b). Para ese conjunto de rangos, tanto a X como a Y se les asignó $-$, mientras que a Z se le asignó +. Así, podemos decir que tanto X como Y están en "desacuerdo" con Z. Resumimos esta información colocando el par (a, b) en la celda D de la tabla 8.7. Considérese en seguida el par (a, c). Aquí Y tiene signo de acuerdo con el signo de Z, pero el signo de X está en desacuerdo con el signo de Z. Por tanto, el par (a, c) es asignado a la celda C de la tabla 8.7. En cada uno de los pares restantes, tanto los signos de Y como los de X están de acuerdo con el signo de Z; así, estos cuatro pares se colocan en la celda A de la tabla 8.7.

Tabla 8.7. Puntuaciones ordenadas de X y Y comparadas con el orden de Z.

Y par / *X par*	*Signo en acuerdo con el signo de Z*	*Signo en desacuerdo con el signo de Z*	*Total*
Signo en acuerdo con el signo de Z	A 4	B 0	4
Signo en desacuerdo con el signo de Z	C 1	D 1	2
Total	5	1	6

En general, para tres conjuntos de rangos de N objetos, podemos usar el método ilustrado anteriormente para construir la clase de ordenamiento cuyo modelo es la tabla 8.8. El *coeficiente de correlación parcial* $T_{xy.z}$ *de Kendall de rangos ordenados* (léase la correlación entre X y Y manteniendo Z constante), se define como

$$T_{xy.z} = \frac{AD - BC}{\sqrt{(A + B)(C + D)(A + C)(B + D)}} \tag{8.12}$$

Tabla 8.8. Forma de colocar los datos para calcular $T_{xy.z}$ por medio de la ecuación (8.12).

Y par / *X par*	*Signo en acuerdo con el signo de Z*	*Signo en desacuerdo con el signo de Z*	*Total*
Signo en acuerdo con el signo de Z	A	B	$A+B$
Signo en desacuerdo con el signo de Z	C	D	$C+D$
Total	$A + C$	$B + D$	$\binom{N}{2}$

En el caso de los cuatro objetos que hemos considerado, es decir, en el caso de los datos mostrado en la tabla 8.7:

$$T_{xy.z} = \frac{(4)(1) - (0)(1)}{\sqrt{(4)(2)(5)(1)}} = 0.63$$

Así, la correlación entre X y Y si se mantiene constante el efecto de Z, se expresa por $T_{xy.z} = 0.63$. Si hubiéramos calculado la correlación entre X y Y sin considerar el efecto de Z, habríamos encontrado $T_{xy} = 0.67$. Esto sugiere que las relaciones entre X y Z y entre Y y Z están influyendo sólo ligeramente en la relación observada entre X y Y. Sin embargo, esta clase de inferencia debe hacerse con reservas, a menos que existan antecedentes relevantes *a priori* para esperar que cualquier efecto sea observado. El lector notará que la ecuación (8.12) es similar al coeficiente phi r_ϕ presentado anteriormente en este capítulo. Esta semejanza sugiere que $T_{xy.z}$ mide la extensión en la cual X y Y acuerdan, *independientemente* de su relación con Z.

Método

Aunque el método mostrado para calcular $T_{xy.z}$ es útil para revelar la naturaleza del coeficiente de correlación parcial, al incrementarse N este método se vuelve más tedioso debido al rápido incremento en el valor de $\binom{N}{2}$, el número de pares de las N observaciones. Afortunadamente, se ha desarrollado una forma de cálculo sencilla para $T_{xy.z}$.

Kendall ha mostrado que

$$T_{xy.z} = \frac{T_{xy} - T_{xz}T_{yz}}{\sqrt{(1 - T_{xz}^2)(1 - T_{yz}^2)}} \qquad (8.13)^{14}$$

La ecuación (8.13) es más fácil de calcular que la ecuación (8.12). Para usarla, se deben encontrar primero las correlaciones (las T entre X y Y, X y Z, Y y Z). Una vez obtenidos estos valores, podemos usar la ecuación (8.13) para encontrar $T_{xy.z}$.

Para los rangos X, Y y Z hemos considerado $T_{xy} = 0.67$, $T_{xz} = 0.33$ y $T_{yz} = 0.67$. Insertando estos valores en la ecuación (8.13), tenemos

$$T_{xy.z} = \frac{0.67 - (0.33)(0.67)}{\sqrt{[1 - (0.33)^2][1 - (0.67)^2]}}$$
$$= 0.63$$

Al usar la ecuación (8.13), llegamos al mismo valor para $T_{xy.z}$ que ya habíamos encontrado mediante la ecuación (8.12).

Ejemplo. Ya hemos visto que en los datos recopilados por Siegel y Fagan, la correlación entre las puntuaciones sobre el autoritarismo y las puntuaciones sobre el estatus de lucha social es $T = 0.67$. Sin embargo, también hemos observado que existe una correlación entre el estatus de lucha social y la cantidad de conformidad (responder) a la presión de grupo $-T = 0.39$. Lo que puede sorprendernos es si la primera correlación simplemente repre-

[14] Esta fórmula es directamente comparable a la que se usa para encontrar la correlación parcial producto-momento. Sin embargo, Kendall (1975) y otros han notado que esta similitud en la forma debería ser construida como coincidente.

senta la operación de una tercera variable, a saber, conformidad a la presión del grupo. Esto es, puede ser que la necesidad de los sujetos para conformarse, afecte sus respuestas tanto en la escala de autoritarismo como en la escala de estatus de lucha social y, por tanto, la correlación entre las puntuaciones en estas dos escalas pueda deberse a una asociación entre cada una de estas variables y la necesidad de conformarse. Podemos verificar si esto es cierto al calcular la correlación parcial entre el autoritarismo y el estatus de lucha social, parcializando el efecto de la necesidad de conformarse, indicado por la cantidad de responder en la situación de Asch.

Las puntuaciones para los 12 sujetos en cada una de las tres variables se muestran en las tablas 8.3 y 8.5. Los tres conjuntos de rangos se combinan en la tabla 8.9. Obsérvese que la variable en la cual deseamos parcializar el efecto —la conformidad— es la variable *Z*.

Tabla 8.9. Rangos de estatus de lucha social, autoritarismo y conformidad.

Sujeto	*Estatus de lucha social X*	*Autoritarismo Y*	*Conformidad (rendirse) Z*
A	3	2	1.5
B	4	6	1.5
C	2	5	3.5
D	1	1	3.5
E	8	10	5
F	11	9	6
G	10	8	7
H	6	3	8
I	7	4	9
J	12	12	10.5
K	5	7	10.5
L	9	11	12

Ya hemos determinado que la correlación entre el estatus de lucha social (la variable *X*) y el autoritarismo (la variable *Y*) es $T_{xy} = 0.67$. También hemos determinado que la correlación entre el estatus de lucha social y la conformidad es $T_{xz} = 0.39$ (este valor está corregido para empates). De los datos presentados en la tabla 8.9, podemos rápidamente determinar, al usar la ecuación (8.10), que la correlación entre el autoritarismo y la conformidad es $T_{yz} = 0.36$ (este valor está corregido para empates). Con esta información, podemos determinar el valor de $T_{xy.z}$ al usar la ecuación (8.13):

$$T_{xy.z} = \frac{T_{xy} - T_{xz}\,T_{yz}}{\sqrt{(1 - T_{xz}^2)(1 - T_{yz}^2)}} \tag{8.13}$$

$$= \frac{0.67 - (0.39)(0.36)}{\sqrt{[1 - (0.39)^2][1 - (0.36)^2]}}$$

$$= 0.62$$

Ya hemos determinado que cuando la conformidad está parcializada o controlada estadísticamente, la correlación entre el estatus de lucha social y el autoritarismo es $T_{xy.z} = 0.62$. Ya que este valor no es mucho más pequeño que $T_{xy} = 0.67$, podemos concluir que la relación entre el estatus de lucha social y el autoritarismo (medidos por estas escalas), es relativamente independiente de la influencia de la conformidad (medida en términos de la cantidad de respuestas a la presión del grupo).

Prueba de significación para $T_{xy.z}$

Si una muestra aleatoria se extrae de alguna población en la cual X y Y no están relacionadas cuando la variable Z está controlada, entonces todos los posibles ordenamientos de rangos son igualmente probables. Sin embargo, a diferencia del coeficiente de correlación T_{xy} de Kendall de rangos ordenados, en el cual para cada ordenamiento de X existen $N!$ posibles ordenamientos de Y, el número de posibles ordenamientos que deben considerarse cuando calculamos la distribución del coeficiente de correlación parcial de Kendall, de rangos ordenados es $(N!)^2$. De acuerdo con la suposición de que cada uno de los ordenamientos es igualmente posible cuando no existe relación entre las variables, se puede calcular la distribución de $T_{xy.z}$. Debido a que los cálculos son extremadamente laboriosos aun para muestras pequeñas, debemos recurrir a las tablas de la distribución muestral. En la tabla S del Apéndice I se proporcionan los valores críticos para $T_{xy.z}$ para todos los $N < 20$ y para valores seleccionados de N mayores de 20.

La tabla S del Apéndice I puede usarse para determinar la probabilidad exacta asociada con la ocurrencia (unidireccional) según H_0, de cualquier valor tan extremo como una $T_{xy.z}$ observada. Al probar hipótesis acerca de la correlación $\tau_{xy.z}$ de Kendall de rangos ordenados, la hipótesis nula H_0 es: $\tau_{xy.z} = 0$, o "X y Y son independientes para una Z fija".

La hipótesis alterna puede ser que $\tau_{xy.z} > 0$ (prueba unidireccional) o, más comúnmente, la hipótesis alterna H_1 es: $\tau_{xy.z} \neq 0$, o "X y Y no son independientes para una Z fija", que es una prueba bidireccional.

Por ejemplo, supongamos que hemos escogido $\alpha = 0.05$ y que $N = 11$, y calculamos que la correlación parcial de Kendall de rangos ordenados es $T_{xy.z} = 0.48$. Deseamos probar la hipótesis de que X y Y son independientes para una Z fija (o, de manera equivalente, cuando mantenemos la Z constante), contra la hipótesis de que X y Y no son independientes para una Z fija. Entrando en la tabla S con $N = 11$ y $\alpha/2 = 0.25$ (debido a que queremos una prueba bidireccional), encontramos que el valor crítico es 0.453. Ya que el valor observado de $T_{xy.z} = 0.48$ excede el valor crítico 0.453, podemos rechazar en el nivel $\alpha = 0.05$ de significación, la hipótesis de que X y Y son independientes para valores fijos de la variable Z.

Para grandes valores de N, la distribución de $T_{xy.z}$ es complicada, pero se aproxima a la distribución normal. Una aproximación de la varianza es la siguiente:

$$\sigma^2_{T_{xy.z}} = \frac{2(2N + 5)}{9N(N - 1)} \tag{8.14}$$

que es la misma que la varianza de T_{xy} proporcionada en la sección anterior. Por tanto, cuando N es grande, podemos probar la hipótesis H_0: $\tau_{xy.z} = 0$ al calcular

$$z = \frac{3T_{xy.z}\sqrt{N(N-1)}}{\sqrt{2(2N+5)}} \tag{8.15}$$

que está distribuida de manera aproximadamente normal con media cero y desviación estándar uno. Así, la probabilidad asociada con la ocurrencia de un valor tan extremo como un valor observado de $T_{xy.z}$ cuando H_0 es cierta, puede determinarse usando la ecuación (8.15) y consultando la tabla A del Apéndice I para encontrar la significación de esa z.

Ejemplo. En el experimento conducido por Siegel y Fagan, la correlación entre el estatus de lucha social y la conformidad fue $T_{xz} = 0.39$. Sin embargo, cada una de estas variables está correlacionada con las puntuaciones de autoritarismo ($T_{xy} = 0.67$ y $T_{zy} = 0.36$, respectivamente). Quisiéramos saber si esta correlación está mediada por la relación conjunta de cada variable con el autoritarismo. Esto es, para niveles fijos de autoritarismo, ¿son independientes el estatus de lucha social y la conformidad? Para determinar esto, necesitamos calcular la correlación parcial entre el estatus de lucha social y la conformidad cuando el autoritarismo se mantiene constante, la cual puede encontrarse mediante la ecuación (8.13):

$$\begin{aligned} T_{xz.y} &= \frac{T_{xz} - T_{xy}T_{zy}}{\sqrt{(1-T_{xy}^2)(1-T_{zy}^2)}} \qquad (8.13)\\ &= \frac{0.39 - (0.67)(0.36)}{\sqrt{(1-0.67^2)(1-0.36^2)}}\\ &= 0.21 \end{aligned}$$

Para probar la hipótesis acerca de la independencia condicional de X y Z, podemos consultar la tabla S del Apéndice I para determinar la probabilidad de obtener un valor de $T_{xz.y} \geq 0.21$, cuando los rangos de las variables son independientes. Entrando en esa tabla para $N = 12$, encontramos que $0.20 \leq p \leq 0.40$. Por tanto, no podemos rechazar la hipótesis de que el estatus de lucha social y la conformidad son independientes para niveles fijos de autoritarismo.

Debe notarse que para evaluar la hipótesis nula, el que las dos variables sean independientes para niveles fijos de una tercera variable, está basada en la suposición de que todos los rangos de las tres variables son igualmente probables. En algunas aplicaciones, puede ser apropiado probar la misma hipótesis, pero no suponer que todos los rangos debidos a la tercera variable son igualmente probables; esto es, se puede no desear suponer que τ_{xz} y τ_{yz} sean cero. Todo parece indicar que la prueba de significación proporcionada aquí es relativamente poderosa en tales casos.

Una precaución acerca de los coeficientes de correlación parcial

El lector debe estar consciente de que los coeficientes de correlación parcial deben calcularse e interpretarse con sumo cuidado. Si un investigador quiere analizar el efecto que una variable tiene sobre la relación entre otras dos variables

—ya sea mostrar que la dependencia observada entre las dos variables está mediada por una tercera variable ($\tau_{xy.z} \approx 0$) o bien, que una tercera variable tiene poco efecto sobre la relación entre las dos variables ($\tau_{xy.z} \approx \tau_{xy}$)—, la racionalización para analizar el efecto de una tercera variable debe estar basada sobre algunas nociones *a priori* acerca de las relaciones que debieran obtenerse. Existe un considerable riesgo implícito en la estrategia de simplemente calcular todas las posibles correlaciones parciales y probar su significación, debido a que al incrementarse el número de variables, la posibilidad de obtener diferencias espurias se incrementa a causa del gran número de pruebas realizadas.

Resumen del procedimiento

Éstos son los pasos en el uso del coeficiente de correlación parcial $T_{xy.z}$, de Kendall de rangos ordenados:

1. Sean X y Y las dos variables cuya relación se desea determinar, y sea Z la variable cuyo efecto sobre X y Y se va a parcializar o mantener constante.
2. Póngase rangos a las observaciones sobre la variable X desde 1 hasta N. Hágase lo mismo para las observaciones sobre las variables de Y y Z.
3. Ya sea con la ecuación (8.9) (si no existen rangos empatados o con la ecuación (8.10) (si existen empates), determine los valores observados de T_{xy}, T_{xz}, T_{yx}.
4. Usando esos valores, calcule el valor de $T_{xy.z}$ usando la ecuación (8.13).
5. Para probar la significación de $T_{xy.z}$, esto es, para probar la hipótesis de que las variables X y Y son independientes para niveles fijos de la variable Z, el valor obtenido de $T_{xy.z}$ se compara con los valores críticos del estadístico proporcionado en la tabla S del Apéndice I. Para valores grandes de N, la significación de $T_{xy.z}$ puede determinarse al calcular la z mediante la ecuación (8.15) y encontrar la probabilidad asociada de un valor tan grande como la z observada y, de aquí, la correlación parcial $T_{xy.z}$ de rangos ordenados. Al probar la hipótesis de que las dos variables son independientes dados niveles fijos de una tercera variable, la hipótesis alterna es generalmente que las dos variables *no son* independientes; en ese caso, la prueba de significación es bidireccional.

Eficacia

Poco se sabe acerca de la eficacia de las pruebas basadas en el coeficiente de correlación parcial de Kendall de rangos ordenados. Aunque se sabe que la prueba de H_0: $\tau_{xy.z} = 0$, supone que los rangos en las tres variables son igualmente probables, las pruebas parecen ser relativamente poderosas con respecto a las violaciones de estas suposiciones que conciernen a τ_{xz} y τ_{yz}.

Referencias bibliográficas

El lector encontrará otros detalles de este estadístico en Kendall (1975) y en Morán (1951). Para explicaciones que conciernen a las pruebas de significación del coeficiente de correlación parcial de Kendall de rangos ordenados, se le recomienda consultar Johnson (1979), Maghsoodloo (1975) y Pallos (1981).

COEFICIENTE DE CONCORDANCIA *W* DE KENDALL

Función

En las secciones previas de este capítulo hemos estado interesados en medidas de la correlación entre *dos* conjuntos de rangos de N objetos o individuos. En esta sección y en la siguiente, consideraremos dos medidas de la relación entre *varios* rangos de N objetos o individuos.

Cuando tenemos k conjuntos de rangos, podemos determinar la asociación entre ellos usando el *coeficiente de concordancia W de Kendall*. Mientras que la r_s de Spearman y la T de Kendall expresan el grado de asociación entre dos variables medidas en, o transformadas a rangos, la W expresa el grado de asociación entre k variables, esto es, la asociación entre k conjuntos de rangos. Tal medida puede ser particularmente útil para estudios de interjuicio o confiabilidad interprueba y también tiene aplicaciones en estudios de agrupaciones de variables.

Racionalización

Como una solución al problema de averiguar los acuerdos totales entre k conjuntos de rangos, parecería razonable encontrar los coeficientes de correlación por orden de rangos de Spearman (los r_s) o de Kendall (los T) entre todos los pares posibles de rangos, y después calcular el promedio de estos coeficientes para determinar la asociación total. Si usamos tal procedimiento, podríamos necesitar $\binom{k}{2}$ coeficientes de correlación de rangos ordenados. A menos que k fuera muy pequeño, tal procedimiento sería extremadamente tedioso.

El cálculo de W es mucho más simple; más aún, se convierte en una relación lineal para el promedio de la r_s tomada sobre todos los grupos. Si denotamos al valor promedio de los coeficientes de correlación de rangos ordenados entre los $\binom{k}{2}$ posibles pares de rangos como ave(r_s), entonces se puede mostrar que

$$\text{ave}(r_s) = \frac{kW - 1}{k - 1} \tag{8.16}$$

Otra aproximación podría ser imaginar cómo lucirían nuestros datos si no existiera acuerdo entre los diferentes conjuntos de rangos, y después imaginar cómo lucirían si hubiera perfecto acuerdo entre los diferentes conjuntos de rangos. El coeficiente de concordancia podría entonces ser un índice de la divergencia del acuerdo real mostrado en los datos del máximo posible o del perfecto acuerdo. En términos generales, W es precisamente tal coeficiente.

Supongamos que a tres compañías ejecutivas se les pide entrevistar seis solicitantes de empleo y se les ponen rangos para ordenar el juicio de adecuación para un trabajo abierto. Tres conjuntos independientes de rangos son proporcionados por los ejecutivos X, Y y Z a los solicitantes, desde a hasta f como se muestra en la tabla 8.10. En las últimas dos filas de la tabla 8.10 se proporcionan las sumas de los rangos (rotulada R_i) y el promedio de los rangos ($\overline{R}_i$) asignadas a cada solicitante.

Tabla 8.10. Rangos asignados a seis solicitantes de empleo por tres ejecutivos de una compañía (datos ficticios).

	Solicitantes					
Evaluador	*a*	*b*	*c*	*d*	*e*	*f*
Ejecutivo X	1	6	3	2	5	4
Ejecutivo Y	1	5	6	4	2	3
Ejecutivo Z	6	3	2	5	4	1
R_i	8	14	11	11	11	8
$\overline{R}_i$	2.67	4.67	3.67	3.67	3.67	2.67

Ahora bien, si los tres ejecutivos ($k = 3$) hubieran estado en *perfecto* acuerdo acerca de los solicitantes, esto es, si les hubieran dado los rangos a los seis solicitantes en el mismo orden, entonces, un solicitante habría recibido tres rangos de 1 y la suma correspondiente de los rangos R_i sería $1 + 1 + 1 = 3 = k$. El solicitante a quien todos los ejecutivos le hubieran asignado el segundo lugar tendría

$$R_i = 2 + 2 + 2 = 6 = 2k$$

El solicitante de la calificación más baja entre los seis tendría

$$R_i = 6 + 6 + 6 = 18 = 6k = Nk$$

De hecho, con perfecto acuerdo entre los ejecutivos, las diferentes sumas de rangos R_i serían 3, 6, 9, 12, 15, 18, aunque no necesariamente en ese orden. En general, cuando existe un perfecto acuerdo entre los k conjuntos de rangos, obtenemos, para la R_i, la serie $k, 2k, 3k, \ldots, Nk$ y el promedio de los rangos sería $1, 2, 3, \ldots, N$.

Por otra parte, si hubiera habido acuerdo aleatorio entre los tres ejecutivos, entonces las diferentes R_i serían aproximadamente iguales.

De este ejemplo, puede quedar claro que el grado de acuerdo entre los k juicios es reflejado por el grado de variación entre las N sumas de rangos. W, el coeficiente de concordancia, es una función de ese grado de varianza.

Método

Para calcular W, se arreglan primero los datos en una tabla $k \times N$ con cada fila representando los rangos asignados por un juez particular a los N objetos. En seguida, encontramos la suma de los rangos R_i en cada columna de la tabla y dividimos cada una de ellas por k para encontrar el promedio de los rangos $\overline{R}_i$. Entonces, sumamos los $\overline{R}_i$ y dividimos el total por k para obtener la gran media de los $\overline{R}_i$. Cada uno de los $\overline{R}_i$ puede entonces expresarse como una desviación de la gran media. Hemos argumentado anteriormente que mientras más grandes sean estas desviaciones, más grande será el grado de asociación entre los k conjuntos de rangos. Así, se encuentra la suma de cuadrados de estas desviaciones. Sabiendo estos valores, podemos calcular el valor de W:

$$W = \frac{\sum_{i=1}^{N} (\overline{R}_i - \overline{R})^2}{N(N^2 - 1)/12} \tag{8.17a}$$

donde

k = número de conjuntos de rangos, por ejemplo, el número de juicios

N = número de objetos (o individuos) a quienes se les están asignando rangos

$\overline{R}_i$ = promedio de los rangos asignados al i-ésimo objeto o sujeto

$\overline{R}$ = promedio (o gran media) de los rangos asignados a través de todos los objetos o sujetos

$N(N^2 - 1)/12$ = suma máxima posible de las desviaciones cuadradas, esto es, el numerador que ocurriría si hubiera perfecto acuerdo entre los k rangos, y el promedio de los rangos fuera 1, 2, . . . , N

Para los datos mostrados en la tabla 8.10, los rangos totales son 8, 14, 11, 11, 11 y 8, y el promedio de los rangos es 2.67, 4.67, 3.67, 3.67, 3.67 y 2.67, respectivamente. La gran media de estos promedios es 3.5.

Para obtener el numerador de W en la ecuación (8.17a), elevamos las desviaciones de cada rango promedio $\overline{R}_i$ del valor medio y después sumamos esos cuadrados:

$$\begin{aligned}\sum_{i=1}^{N} (\overline{R}_i - \overline{R})^2 &= (2.67 - 3.5)^2 + (4.67 - 3.5)^2 + (3.67 - 3.5)^2 \\ &\quad + (3.67 - 3.5)^2 + (3.67 - 3.5)^2 + (2.67 - 3.5)^2 \\ &= 2.833\end{aligned}$$

Una vez obtenido el numerador, encontramos el valor de W para los datos de la tabla 8.10, usando la ecuación (8.17a):

$$W = \frac{2.833}{6(6^2 - 1)/12}$$

$$= 0.16$$

$W = 0.16$ expresa el grado de acuerdo entre los tres ejecutivos al poner rangos a los seis solicitantes de empleo.

Aunque la ecuación (8.17*a*) muestra la racionalización "intuitiva" para el estadístico W, se puede usar una fórmula un poco más simple. Ya que se conocen los valores de los datos con anterioridad cuando están en forma de rangos, el valor $\overline{R}$, la gran media de todos los rangos, se conoce con anterioridad también. Ya que la suma de N rangos es $N(N + 1)/12$, la media es, por tanto, $(N + 1)/2$. Usando este valor, puede simplificarse la ecuación (8.17*a*):

$$W = \frac{12\Sigma\overline{R}_i^2 - 3N(N + 1)^2}{N(N^2 - 1)} \qquad (8.17b)$$

o podemos simplemente adelantar usando los rangos totales R_i en lugar de la media de los rangos $\overline{R}_i$:

$$W = \frac{12\Sigma R_i^2 - 3k^2N(N + 1)^2}{k^2N(N^2 - 1)} \qquad (8.17c)$$

donde ΣR_i^2 es la suma de las sumas cuadradas de los rangos para cada uno de los N objetos o individuos a quienes se les están poniendo rangos. Para los datos de la tabla 8.10,

$$\Sigma R_i^2 = 8^2 + 14^2 + 11^2 + 11^2 + 11^2 + 8^2$$

$$= 687$$

Al usar este valor y sustituirlo en la ecuación (8.17*c*), encontramos

$$W = \frac{12(687) - 3(3^2)(6)(6 + 1)^2}{3^2(6)(6^2 - 1)}$$

$$= 0.16$$

Naturalmente, este valor es el mismo que el obtenido por la expresión equivalente, ecuación (8.17*a*). La elección de la fórmula se deja al usuario. La ecuación (8.17*c*) es más fácil para cálculos rápidos. Muchas calculadoras pueden determinar directamente la suma de desviaciones cuadradas, de manera tal que la ecuación (8.17*a*) puede resultar apropiada en ese caso.

Para los mismos datos, podríamos haber encontrado ave(r_s) por cualquiera de los dos métodos. Un modo podría ser encontrar las tres correlaciones por orden de rangos $r_{s_{xy}}$, $r_{s_{xz}}$ y $r_{s_{yz}}$. En seguida se promediarían estos tres valores. Para los datos de la tabla 8.10, $r_{s_{xy}} = 0.31$, $r_{s_{xz}} = -0.54$ y $r_{s_{yz}} = -0.54$. El promedio de estos valores es

$$\text{ave}(r_s) = \frac{0.31 + (-0.54) + (-0.54)}{3}$$

$$= -0.26$$

Otro modo de encontrar ave(r_s) sería usar la ecuación (8.16):

$$\text{ave}(r_s) = \frac{kW - 1}{k - 1} \tag{8.16}$$

$$= \frac{3(0.16) - 1}{3 - 1} = -0.26$$

Ambos métodos resultan en el mismo valor, ave$(r_s) = -0.26$. Como se muestra arriba, este valor es una función lineal del valor de W.

Una diferencia entre usar W y ave(r_s) para expresar el acuerdo entre k rangos, es que ave(r_s) puede tomar valores entre $-1/(k-1)$ y $+1$, mientras que W varía entre 0 y $+1$, sin considerar el número de conjuntos de rangos. La razón de que W no pueda ser negativa es que cuando más de dos conjuntos de rangos están implicados, los rangos no pueden estar en completo desacuerdo. Por ejemplo, si el juicio X y el juicio Y están en desacuerdo, el juicio X está también en desacuerdo con el juicio Z, entonces los juicios Y y Z deben concordar. Esto es, cuando están implicados más de dos juicios, el acuerdo y el desacuerdo no son simétricamente opuestos. Un grupo de k jueces pueden todos estar de acuerdo, pero no pueden estar completamente en desacuerdo. Por tanto, W debe ser cero o positivo. También, como se notó en la racionalización de W, el numerador es un índice de la variabilidad de los rangos. Cuando no hay consenso entre los evaluadores, la variabilidad de los rangos será cero, esto es, el promedio de los rangos será el mismo para todos los objetos a quienes se les va a poner rangos.

Ya que el rango de ave(r_s) depende del número de evaluadores, el límite inferior de $-1/(k-1)$ no es directamente comparable a través de los conjuntos de datos. En el ejemplo anterior, los primeros dos evaluadores (X y Y) no acordaron ($r_s = -1$), el evaluador Z también estuvo en perfecto desacuerdo con el evaluador X ($r = -1$) y, por necesidad, Y y Z deben estar de acuerdo ($r_s = 1$). En este caso, ave$(r_s) = -\frac{1}{3}$. El mínimo posible ave(r_s) para $k = 3$ evaluadores es $-\frac{1}{2}$.

El lector debe notar que W produce una relación lineal para r_s, pero parece no producir una relación ordenada para la T de Kendall. Esto revela una de las ventajas que r_s tiene sobre T; sin embargo, como veremos en la siguiente sección, existe un índice de concordancia correspondiente para T.

Observaciones empatadas

Cuando ocurren observaciones empatadas, se asigna a cada una de las observaciones el promedio de los rangos que se les habría asignado si no hubieran ocurrido empates, lo cual es nuestro procedimiento usual en las puntuaciones de rangos empatados.

El efecto de rangos empatados es reducir el valor de W encontrado mediante la ecuación (8.17) (y cualquiera de sus formas). Si la proporción de rangos empatados es pequeña, el efecto es insignificante, y así la ecuación (8.17) puede aún emplearse. Sin embargo, si la proporción de empates es grande o el investigador quiere una estimación más precisa, se debe usar una corrección. Esta corrección resultará en un incremento ligero en el valor de W comparado con el valor que se habría obtenido si no se hubiera hecho corrección alguna. El factor de corrección es el mismo que el usado con el coeficiente de correlación r_s de Spearman de rangos ordenados:

$$T_j = \sum_{i=1}^{g_j} (t_i^3 - t_i)$$

donde t_i es el número de rangos empatados en el iésimo grupo de empates, y g_j es el número de grupos de empates en el j-ésimo conjunto de rangos. Así, T_j es el factor de corrección requerido para el j-ésimo conjunto de rangos.

Con la corrección incorporada para empates, la fórmula para el coeficiente de concordancia de Kendall es la siguiente:

$$W = \frac{12\Sigma\overline{R}_i^2 - 3N(N + 1)^2}{N(N^2 - 1) - (\Sigma T_j)/k} \tag{8.18a}$$

o

$$W = \frac{12\Sigma R_i^2 - 3k^2N(N + 1)^2}{k^2N(N^2 - 1) - k\Sigma T_j} \tag{8.18b}$$

donde ΣT_j nos dirige a la suma de los valores de T_j para todos los k conjuntos de rangos.

Ejemplo. Un grupo profesional y académico, The Society for Cross-Cultural Research (SCCR) —Sociedad para investigación transcultural—, decidió realizar una investigación de sus miembros, concerniente a la elección de sitios para sus reuniones anuales.[15] Para evaluar el interés de la sociedad, se pidió a una muestra de los miembros que evaluara y le pusiera rangos a aquellas características que podían usarse para describir los factores que afectan la atención potencial en las reuniones de la sociedad. Estos factores incluyen características tales como aire acondicionado, clima y contenido del programa.

Además de obtener el rango promedio asignado a cada uno de los factores que afectan la atención en las reuniones, es deseable saber si los evaluadores pueden juzgar haber alcanzado consenso. Un modo de medir el consenso es determinar el grado de acuerdo entre los evaluadores en sus juicios. El coeficiente de concordancia de Kendall es una medida que proporcionaría tal índice. Los rangos asignados a cada uno de los $N = 8$ factores o atributos para cada uno de los $k = 22$ sujetos, se proporcionan en la tabla 8.11. Un rango de uno significa que la característica podría ser importante en decidir la asistencia a la reunión anual, y se asignó un rango de ocho al aspecto menos importante.

Para calcular el coeficiente de concordancia, es necesario calcular primero la suma de rangos para cada uno de los reactivos a los que los sujetos les colocaron rangos. (Los datos no

[15] Starr, B. J. "A report from the SCCR Secretary-Treasurer", en *SCCR Newsletter*, Society for Cross-Cultural Research, otoño de 1982, págs. 3 y 4.

Tabla 8.11. Calificaciones a los factores que afectan la decisión de atender a un encuentro profesional.

	Factores							
Evaluador	*Aire acondicionado*	*Clima*	*Tiempo de encuentro*	*Gente*	*Programa*	*Anuncio*	*Presente*	*Interés declinado*
1	2	7	3	5	4	6	1	8
2	6	5	7	3	4	2	1	8
3	1	6	4	5	2	7	3	8
4	5	6	7	1	2	4	3	8
5	1	8	6	5	2	4	3	7
6	2	7	5	1	3	6	4	8
7	2	7	1	4	3	6	5	8
8	1	4	7	2	3	6	5	8
9	1	7	3	6	2	4	5	8
10	1	6	7	3	2	4	5	8
11	4	5	1	3	2	7	6	8
12	1	4	6	7	2	5	3	8
13	1	5	2	3	4	6	7	8
14	1	6	5	2	3	4	7	8
15	1	7	2	4.5	3	4.5	6	8
16	1	6	5	2.5	2.5	7	4	8
17	1	7	6	4	3	5	2	8
18	3	7	5	6	1	4	2	8
19	1	6	2	4	5	7	3	8
20	1	6	5	3	4	7	2	8
21	1	7	6	2	3	5	4	8
22	1.5	8	1.5	4.5	3	6	4.5	7
R_i	39.5	137	96.5	80.5	62.5	116.5	85.5	174

habían sido recabados como rangos, por lo que sería necesario primero transformar los datos registrados en rangos.) Las sumas de los rangos se proporcionan en la parte inferior de la tabla 8.11. La suma de los cuadrados de los rangos es la siguiente:

$$\Sigma R_i^2 = 39.5^2 + 137^2 + 96.5^2 + 80.5^2 + 62.5^2 + 116.5^2$$
$$+ 85.5^2 + 174^2$$
$$= 91\ 186.5$$

Cabe destacar que es posible aquí una verificación de los cálculos, ya que R debe ser igual a $kN(N + 1)/2$. Puesto que la suma observada es 792, y $22(8)(9)/2 = 792$, hemos verificado en forma parcial los cálculos.

En seguida observamos que los sujetos 15, 16 y 22 tienen empates en sus rangos. Por tanto, es necesario encontrar los términos de corrección (las T_j) para calcular el valor de W corregida para empates. Para el sujeto 15, existe un grupo de empates de tamaño dos; de aquí $g_{15} = 1$ y $t_1 = 2$; así,

$$T_{15} = 2^2 - 2 = 6$$

De manera similar, ya que el sujeto 16 tiene un grupo de empates de tamaño dos, también $T_{16} = 6$. Sin embargo, el sujeto 22 tiene dos grupos de empates, así que

$$\begin{aligned} T_{22} &= (2^3 - 2) + (2^3 - 2) \\ &= 12 \end{aligned}$$

Con estos resultados, y ya que $N = 8$ y $k = 22$, podemos encontrar el valor de W usando la ecuación (8.18*b*):

$$W = \frac{12\Sigma R_i^2 - 3k^2N(N + 1)^2}{k^2N(N^2 - 1) - k\Sigma T_i} \tag{8.18b}$$

$$\begin{aligned} W &= \frac{12(91\,186.5) - 3(22^2)(8)(8 + 1)^2}{22^2(8)(8^2 - 1) - 22(6 + 6 + 12)} \\ &= \frac{153\,342}{243\,408} \\ &= 0.630 \end{aligned}$$

Así, podemos concluir que existe un buen acuerdo entre los sujetos en los rangos de los factores importantes en su decisión de asistir a las reuniones de la sociedad. También podemos concluir que el costo del aire acondicionado y el contenido del programa son juicios que resultan los más importantes (en ese orden), y declinar el interés en el área del trabajo transcultural y el clima, son juzgados los factores menos importantes que determinan la asistencia a las reuniones anuales.

Se advirtió anteriormente que W está relacionada con el coeficiente de correlación de Spearman de rangos ordenados. Si hubiéramos calculado el valor de r_s para cada uno de los $\binom{22}{2} = 22(21)/2 = 231$ pares de sujetos, tendríamos también un índice de acuerdo si promediamos los valores. Sin embargo, en lugar de calcular todos esos pares, podemos usar la ecuación (8.16):

$$\text{ave}(r_s) = \frac{kW - 1}{k - 1} \tag{8.16}$$

$$\begin{aligned} &= \frac{22(0.630) - 1}{22 - 1} \\ &= 0.61 \end{aligned}$$

Así, el promedio de acuerdo intersujetos concerniente a los factores que afectan la atención en los encuentros, es 0.61.

Finalmente, cabe destacar que tendríamos que no considerar los empates en el cálculo de W; esto es, si hubiéramos usado la ecuación (8.17) en lugar de la ecuación (8.18), habríamos encontrado $W = 0.6286$, que es un poco más pequeño que el valor obtenido con la corrección. El efecto de los empates es pequeño en este caso, debido a que el número de grupos de rangos empatados es pequeño, y cada grupo de empates contiene no más de dos empates.

Prueba de la significación de W

Al igual que con las otras técnicas estadísticas no paramétricas presentadas en este libro, el método para probar la significación del coeficiente de concordancia de Kendall, depende del tamaño de la muestra: en este caso, el número de objetos a los que se les asignan rangos.

MUESTRAS PEQUEÑAS

Podemos probar la significación de cualquier valor observado de W al determinar la probabilidad asociada con la ocurrencia, cuando H_0 es cierta, de un valor tan grande como el valor observado. Si obtenemos la distribución muestral de W para todas las permutaciones en los N rangos en todos los posibles modos entre los k rangos, tendremos $(N!)^k$ conjuntos de posibles rangos. Con éstos podemos probar la hipótesis nula de que los k conjuntos de rangos son independientes, al tomar de esta distribución la probabilidad asociada con la ocurrencia según H_0, de un valor tan grande como una W observada.

Por medio de este método, la distribución de W según H_0 (la suposición de que los rangos son independientes), ya se ha trabajado y han sido tabulados ciertos valores críticos. En la tabla T del Apéndice I se proporcionan los valores críticos de W para los valores de significación $\alpha = 0.05$ y $\alpha = 0.01$. Esta tabla es aplicable para k desde 3 hasta 20 y para N desde 3 hasta 7. Si un valor observado de W es más grande o igual al mostrado en la tabla T para un nivel particular de significación, entonces H_0 puede ser rechazada en ese nivel de significación. Debemos recordar que como un índice de significación, $0 \leqslant W \leqslant 1$, tal que solo son apropiadas pruebas concernientes a W unidireccional.

Por ejemplo, vemos que cuando $k = 3$ ejecutivos ficticios, asignan rangos a $N = 6$ solicitantes de empleo, su acuerdo fue $W = 0.16$. La tabla T del Apéndice I revela que el valor de W no es significativo en el nivel $\alpha = 0.05$. Para que la concordancia hubiera sido significativa en el nivel $\alpha = 0.05$, la W observada tendría que haber sido 0.660 o más grande.

MUESTRAS GRANDES

Cuando N es mayor que 7, no se puede usar la tabla T del Apéndice I para determinar la significación de una W observada. Sin embargo, la cantidad

$$X^2 = k(N - 1)W \tag{8.19}$$

está aproximadamente distribuida como la ji cuadrada con $N - 1$ grados de libertad. Esto es, la probabilidad asociada cuando H_0 es cierta, con la ocurrencia de un valor tan grande como una W observada, puede determinarse al encontrar X^2 usando la ecuación (8.19) y después determinando la probabilidad asociada con un valor tan grande de X^2 al consultar la tabla C del Apéndice I.

Si el valor de X^2 calculado de la ecuación (8.19) iguala o excede al mostrado en la tabla C del Apéndice I para un nivel particular de significación y un valor particular de $gl = N - 1$, entonces la hipótesis nula H_0 de que los k rangos no están relacionados (o son independientes), puede rechazarse en ese nivel de significación.

Ejemplo.[16] En el estudio de los factores que afectan la decisión de asistir a las reuniones de la Society for Cross-Cultural Research, $k = 22$ sujetos evaluaron $N = 8$ factores y encontramos que $W = 0.630$. Podemos determinar la significación de este concordancia aplicando la ecuación (8.19):

$$X^2 = k(N - 1)W \tag{8.19}$$

$$= 22(8 - 1)(0.630)$$

$$= 97.02$$

Consultando la tabla C del Apéndice I, encontramos que $X^2 \geq 97.02$ con

$$gl = N - 1 = 8 - 1 = 7$$

tiene una probabilidad de ocurrencia según H_0 de $p < 0.001$. Podemos concluir una considerable confianza que el acuerdo entre los 22 sujetos es más alto de lo que habría sido si se hubieran elegido los rangos al azar o de manera independiente. La muy baja probabilidad asociada según H_0 con el valor observado de W, nos capacita para rechazar la hipótesis nula de que las evaluaciones de los sujetos no están relacionadas una con la otra y concluir que existe un buen consenso entre los miembros, referente a los factores que afectan las decisiones de asistir a las reuniones de la sociedad.

Resumen del procedimiento

Éstos son los pasos en el uso de W, el coeficiente de concordancia de Kendall:

1. Sea N el número de entidades u objetos a los que se les van a asignar rangos y sea k el número de jueces que van a asignar los rangos. Coloque los rangos observados en una tabla $k \times N$.
2. Para cada objeto, determine R_i, la suma de los rangos asignados a ese objeto, por cada uno de los k jueces.
3. Determine los valores cuadrados de cada una de las sumas (R_i^2).
4. Si no existen empates o la proporción de rangos empatados es pequeña, calcule el valor de W mediante una de las formas de la ecuación (8.17). Si la proporción de empates entre los N rangos es grande, use la ecuación (8.18) para determinar el valor de W.
5. El método para determinar si el valor observado de W es significativamente diferente de cero, depende del tamaño de N, el número de objetos a los que se les asignó rangos:

   a) Si $N \leq 7$, en la tabla T del Apéndice I se proporcionan los valores críticos de W para los niveles de significación $\alpha = 0.05$ y $\alpha = 0.01$.

[16] Estos datos también forman parte del estudio SCCR señalado anteriormente.

b) Si $N > 7$, se puede usar la ecuación (8.19) para calcular un valor de W^2 que esté aproximadamente distribuido como ji cuadrada, y cuya significación para $gl = N - 1$ puede probarse consultando la tabla C del Apéndice I.

6. Si W es mayor que el valor crítico encontrado al usar la tabla C o la tabla T del Apéndice I, rechace H_0 y concluya que los rangos no son independientes.

Interpretación de *W*

Un valor alto o significativo de W puede interpretarse como un reflejo de que los k observadores o jueces están aplicando esencialmente los mismos estándares al poner rangos a los N objetos en estudio. Con frecuencia, su ordenamiento de intereses puede utilizarse como un "estándar", especialmente cuando no existe un criterio externo relevante para ordenar los objetos.

Cabe destacar que un valor alto o significativo de W *no* indica que los ordenamientos observados sean *correctos*. De hecho, pueden ser todos incorrectos respecto a algún criterio externo. Por ejemplo, los 22 sujetos en el ejemplo acordaron bien al juzgar qué factores fueron importantes para determinar la atención en las reuniones anuales de la sociedad; sin embargo, sólo el tiempo puede decir si sus juicios fueron acertados. Es posible que una variedad de jueces puedan concordar al ordenar objetos debido a que todos ellos emplean el criterio "incorrecto". En este caso, una W alta o significativa simplemente muestra que todos más o menos concordaron en el uso de un criterio "incorrecto". Para decirlo de otra manera: un alto grado de acuerdo acerca de un ordenamiento, no necesariamente significa que el orden acordado sobre éste sea el "objetivo". En las ciencias conductuales, con frecuencia se cree que ordenamientos "objetivos" y ordenamientos "consensuales" son sinónimos.

Kendall sugiere que el mejor estimador de una colocación de rangos "verdadera" de los N objetos, es el proporcionado cuando W es significativa, por el orden de las diferentes sumas de rangos R_i o, de manera equivalente, el promedio de los rangos $\overline{R}_i$. Si se acepta el criterio de que los diferentes jueces han concordado en los rangos de las N entidades (evidenciado por la magnitud y significación de W), entonces el mejor estimador de los rangos "verdaderos" de esos estimados, es el que proporciona el orden de las sumas (o promedios) de los rangos. Este "mejor estimador" está asociado, en cierto sentido, con un estimador de mínimos cuadrados. Así, en el ejemplo del empleo citado anteriormente, nuestro mejor estimador sería que ya sea el solicitante a o el f (véase la tabla 8.10), deberían ajustarse para el trabajo abierto, para que en cada uno de sus casos la suma de rangos sean iguales $-R_1 = R_6 = 8-$ el valor observado más bajo. Y nuestro mejor estimador sería que de los ocho factores que afectan la asistencia en las reuniones de la SCCR, el aire acondicionado es el factor más importante y la falta de interés, el factor menos importante.

Finalmente, se debe notar que el coeficiente de concordancia W de Kendall está estrechamente relacionado con el estadístico F_r de Friedman examinado en el capítulo 6. El lector cuidadoso advertirá que al estudiar el análisis de varianza bifactorial de Friedman, el modelo se describió como un conjunto de k medidas en cada uno de los N sujetos. En nuestro análisis de W, describimos el modelo implicando un conjunto de k jueces que asignaban rangos a cada uno de N objetos. Los dos estadísticos están linealmente relacionados pero, en nuestra presentación, N y k son intercambiados entre los dos estadísticos.

Eficacia

No existe análogo paramétrico directo de W interpretado como un índice de acuerdo entre un conjunto de k rangos. Sin embargo, como una prueba de igualdad de N rangos, podemos apelar a su relación con el análisis de varianza bifactorial de Friedman. En ese caso, cuando están satisfechas las suposiciones del análisis de varianza, la eficacia de W es baja cuando $N = 2(2/\pi = 0.64)$, pero incrementa a 0.80 cuando $N = 5$ y a $3/\pi = 0.955$ cuando N es grande. Así, la eficacia de la prueba se incrementa al aumentar el número de objetos a los que se les asigna rangos.

Referencias bibliográficas

Contenidas en Friedman (1940) y Kendall (1975) se encuentran explicaciones acerca del coeficiente de concordancia de Kendall. Otros análisis recientes pueden consultarse en Gibbons (1985).

COEFICIENTE DE ACUERDO *u* DE KENDALL DE RANGOS PARA COMPARACIONES APAREADAS

Al examinar el coeficiente de concordancia W de Kendall, fue descrito como un índice de semejanza de ordenamientos de rangos producido por cada uno de k jueces. En esta sección estudiaremos una medida similar, W_T, que está basada en el *coeficiente de acuerdo* u *de Kendall*. Algunas veces, en lugar de pedir a un grupo de jueces que asigne rangos a un conjunto de objetos, podemos presentarles pares de los objetos y solicitar a cada juez que indique una preferencia por uno de los dos objetos. Una tarea en la que pedimos a sujetos que indiquen preferencia por uno de un par de objetos, se llama *comparaciones apareadas*.

En el método de comparaciones apareadas, las preferencias entre los conjuntos de objetos pueden ser inconsistentes. Esto es, si existen tres objetos para ser comparados, por decir A, B y C, el sujeto puede preferir A a B, B a C, pero preferir C a A. Pedir a los sujetos que asignen rangos a los objetos sería imposible, ya que cuando a los objetos se les asignan rangos, las preferencias por pares deben ser consistentes.[17] Aunque podemos tratar de evitar preferencias inconsistentes en un estudio de investigación en particular, debe notarse que aquéllas pueden ocurrir más frecuentemente de lo que se supone. Considérese el siguiente ejemplo: pe-

[17] Las comparaciones apareadas que son consistentes son también *transitivas*. Véase al análisis de las escalas ordinales en el capítulo 2.

dimos a un niño que ponga rangos a un grupo de compañeros estudiantes desde aquel con el que le gustaría jugar más hasta aquel con el que le gustaría jugar menos. Tal tarea es difícil debido a que estamos pidiendo al niño que asigne rangos a un grupo desde el primero hasta el último, lo cual no es una conducta "natural"; más aún, puede no ser posible debido a que las preferencias quizá no sean transitivas. Sin embargo, si presentáramos al niño los nombres de dos compañeros, sería posible, y ciertamente más natural, indicar una preferencia por una persona en cada par.

Cuando se recaban datos por el método de comparaciones apareadas, es posible calcular el grado de acuerdo entre los individuos en sus preferencias. En esta sección examinaremos un coeficiente de acuerdo u adecuado para datos de comparaciones apareadas. Además, veremos que este coeficiente está relacionado con el promedio del coeficiente de correlación T, de Kendall de rangos ordenados, cuando los datos están en rangos.

Racionalización y método

Para calcular el coeficiente de acuerdo, necesitamos observar sólo las preferencias de cada individuo y después agregar éstas en un índice simple. Supongamos que a una persona se le pide que indique preferencias para $N = 4$ objetos. Para hacer esto, podríamos necesitar presentarle $\binom{4}{2} = (4)(3)/2 = 6$ pares al sujeto, el que indicaría una preferencia por uno de los miembros de cada par. Cada par puede ser denotado (a, b), y la persona que expresa una preferencia como $a > b$ o $b > a$. (Léase $>$ como "es preferido a".) Así, para los seis pares presentados, supongamos que las preferencias han sido las siguientes:

Par	*Preferencia*
(a, b)	a
(a, c)	a
(a, d)	d
(b, c)	b
(b, d)	d
(c, d)	d

Estas preferencias pueden resumirse en una matriz la cual consiste en una tabla que resume el número de veces que cada objeto es preferido a cualquier otro objeto (o bien, se le asigna rango antes). La tabla contiene una entrada para cada par en el que la variable *fila* es preferida a la variable *columna*. La matriz de preferencia para las preferencias dadas anteriormente se muestra en la página siguiente.

Si existen varios jueces o sujetos ejecutando los rangos, entonces sus preferencias se combinan dentro de una matriz de preferencia. Para ilustrar el cálculo, usa-

Matriz de preferencia

	a	*b*	*c*	*d*
a	—	1	1	—
b	—	—	1	—
c	—	—	—	—
d	1	1	1	—

remos el ejemplo de los rangos de la sección anterior. Estos datos están resumidos nuevamente en la porción superior de la tabla 8.12. En seguida, transformamos estos rangos en la tabla de preferencias proporcionada en la parte inferior de la tabla 8.12. Cabe destacar que si hubiera habido acuerdo completo entre los tres ejecutivos, exactamente 15 celdas de la tabla tendrían entradas, y cada entrada sería igual a tres. [En general, si existe completo acuerdo entre k jueces haciendo comparaciones apareadas entre N objetos, entonces $N(N-1)/2$ celdas tendrían frecuencias iguales a k. Las restantes $N(N-1)/2$ celdas contendrían cero.] Kendall ha propuesto un coeficiente de acuerdo entre los jueces, que es

$$u = \frac{2\sum_{i=1}^{N}\sum_{j=1}^{N}\binom{a_{ij}}{2}}{\binom{k}{2}\binom{N}{2}} - 1 \qquad (8.20a)$$

donde a_{ij} es el número de veces que el objeto asociado con la fila i es preferido al objeto asociado con la fila j. Aunque el cálculo de u incluye operaciones laboriosas y tediosas, pueden hacerse cercanamente directos. Si manipulamos las expresiones combinatorias y simplificamos, la ecuación (8.20*a*) puede reescribirse en la siguiente forma:

$$u = \frac{4\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}(a_{ij}-1)}{k(k-1)N(N-1)} - 1 \qquad (8.20b)$$

De nuevo, notando algunas relaciones (principalmente entre las celdas en la mitad superior y la mitad inferior de la matriz), podemos simplificar aún más la fórmula:

$$u = \frac{8(\Sigma a_{ij}^2 - k\Sigma a_{ij})}{k(k-1)N(N-1)} + 1 \qquad (8.20c)$$

donde la sumatoria es tomada sobre las a_{ij} por debajo o por arriba de la diagonal. Si existen menores entradas diferentes de cero (o entradas más pequeñas) en cada lado de la diagonal, ese lado puede ser elegido por conveniencia cuando se aplica la ecuación (8.20*c*) para el cálculo del coeficiente de acuerdo.

Tabla 8.12. Rangos asignados a los seis solicitantes de empleo por tres ejecutivos de una compañía (datos ficticios).

	Solicitante					
Evaluador	*a*	*b*	*c*	*d*	*e*	*f*
Ejecutivo *X*	1	6	3	2	5	4
Ejecutivo *Y*	1	5	6	4	2	3
Ejecutivo *Z*	6	3	2	5	4	1

Matriz de preferencia

	a	*b*	*c*	*d*	*e*	*f*
a	–	2	2	2	2	2
b	1	–	1	1	1	0
c	1	2	–	1	2	1
d	1	2	2	–	1	1
e	1	2	1	2	–	1
f	1	3	2	2	2	–

Para la matriz de preferencias proporcionada en la tabla 8.12, tenemos las siguientes sumas para las a_{ij} por abajo de la diagonal:

$$\Sigma a_{ij} = 1 + 1 + 1 + 1 + 1 + 2 + 2 + 2 + 3 + 2 + 1 + 2 + 2 + 2 + 2 = 25$$

y $$\Sigma a_{ij}^2 = 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 2^2 + 2^2 + 2^2 + 3^2 + 2^2 + 1^2 + 2^2 + 2^2 + 2^2 + 2^2 = 47$$

Con estos valores, calculamos u usando la ecuación (8.20*c*):

$$u = \frac{8(\Sigma a_{ij}^2 - k\Sigma a_{ij})}{k(k-1)N(N-1)} + 1 \qquad (8.20c)$$

$$= \frac{8[47 - 3(25)]}{(3)(2)(6)(5)} + 1$$

$$= \frac{8(-28)}{(6)(30)} + 1$$

$$= -0.244$$

El lector puede verificar que éste es el mismo valor que obtuvimos cuando aplicamos la ecuación (8.20*c*) a las entradas por arriba de la diagonal o bien, usando las ecuaciones (8.20*a*) o (8.20*b*).

Un aspecto útil de este coeficiente es que si las comparaciones apareadas para cada sujeto son consistentes, esto es, puede hacerse un rango de los N objetos, entonces *u es igual al promedio de T*. Alternativamente, si calculamos la correlación de Kendall de rangos ordenados para cada par de jueces, entonces el promedio de todas las T sería igual a u. En el ejemplo de los ejecutivos que evalúan a los solicitantes de empleo, $T_{xy} = 0.20$, $T_{yz} = -0.467$ y $T_{xz} = -0.467$; $\text{ave}(T) = (0.20 - 0.467 - 0.467)/3 = -0.244$, que es el valor obtenido al usar la ecuación (8.20).

Como notamos en el análisis del coeficiente de concordancia W de Kendall, ese índice fue una función del promedio del coeficiente de correlación de Spearman de rangos ordenados. Como el estadístico $\text{ave}(r_s)$, u será igual a uno cuando exista completo acuerdo entre los jueces. Sin embargo, aunque cada valor de T puede tener un rango desde -1 a $+1$, el promedio de T no puede alcanzar un mínimo de -1. Esto es, debido a que cuando existen más de dos conjuntos de rangos, no pueden estar todos ellos en desacuerdo (o rangos ordenados "a la inversa") respecto cada uno de ellos al otro. De hecho, el valor mínimo de u es $-1/(k-1)$ cuando k es par y $-1/k$ cuando k es impar. Para tener un índice de acuerdo similar al coeficiente de concordancia de Kendall, podemos definir W_T como

$$W_T = \frac{(k-1)u + 1}{k} \quad \text{si } k \text{ es par} \tag{8.21a}$$

y

$$W_T = \frac{ku + 1}{k + 1} \quad \text{si } k \text{ es impar} \tag{8.21b}$$

Así, como W, W_T puede variar desde cero hasta uno. Para el ejemplo de los tres ejecutivos

$$W_T = \frac{3(-0.244) + 1}{3 + 1}$$

$$= 0.067$$

lo cual indica que existe poco acuerdo entre los ejecutivos. Como esperaríamos, el valor es congruente con el valor del coeficiente de concordancia de Kendall calculado anteriormente en este capítulo, donde encontramos que $W = 0.16$.

Ejemplo. La teoría de decisiones multiatributos se aplica al proceso de toma de decisiones por la gente, en un esfuerzo para desarrollar modelos de toma de decisiones que ayuden no sólo a los psicólogos a entender dicho proceso, sino que también sirva para mejorarlo cuando las decisiones se toman con incertidumbre. En los modelos de la teoría multiatributos, el proceso de decisión se considera un modelo lineal; esto es, se supone que las decisiones pueden ser modeladas como una suma ponderada de las variables o factores implícitos en la decisión. Los pesos aplicados a los factores son generalmente pesos de "importancia" basados sobre los juicios subjetivos de los individuos acerca de la importancia de cada factor en la toma de decisiones. En un estudio diseñado para evaluar la aplicabilidad de los modelos de utilidad multiatributos para decisiones concernientes al uso del suelo,[18] se pidió a los sujetos que evaluaran la importancia de cinco factores generales que describen el efecto de políticas particulares de uso del suelo. Los factores identificados en el estudio fueron los siguientes:

1. Uso múltiple, por ejemplo, localización, acceso y tipo de actividades posibles en el sitio.
2. Belleza, recreación y estilo de vida; por ejemplo, apoyo potencial de la población, deportes al aire libre y áreas escénicas.
3. Fuentes de productividad; por ejemplo, el potencial para gasolina y gas, productos forestales, agricultura y minería.
4. Potencial de entrada para el gobierno; por ejemplo, entradas reales y costos de mantenimiento.
5. Condiciones económicas; por ejemplo, efecto en la base local de taxis, empleo y efecto sobre grupos vulnerables.

A fin de determinar la importancia de cada factor para cada persona en el estudio, se dio a los sujetos cada uno de los cinco factores en pares y se les pidió que indicaran cuál de los dos consideraban más importantes para evaluar las decisiones concernientes al uso del suelo. Ya que había cinco factores, cada sujeto juzgó 10 pares de factores. Debido al modo en que se hicieron las comparaciones apareadas, fue posible para un sujeto juzgar ambos miembros de un par como iguales en importancia.

Aunque se espera que en cualquier evaluación de factores concerniente al uso del suelo la gente expresará una amplia variedad de opiniones, es deseable determinar el grado de consenso entre la gente concerniente a los factores que afectan el uso del suelo. Ya que los datos en este estudio son comparaciones apareadas, el coeficiente de acuerdo de Kendall es un estadístico apropiado para estimar el acuerdo entre los evaluadores.

En una condición en el estudio, $k = 10$ sujetos o jueces hicieron comparaciones apareadas entre los $N = 5$ factores. Sus elecciones están resumidas en la matriz de preferencias que se proporciona en la tabla 8.13. Esta matriz de preferencia fue formada agregando las matrices de preferencia para cada uno de los 10 sujetos. (Debe notarse que cuando un sujeto es indiferente a los elementos de un par, se registra una cantidad de un medio en cada una de las celdas correspondientes en la matriz.) El coeficiente de acuerdo fue entonces calculado:

$$u = \frac{8(\Sigma a_{ij}^2 - k\Sigma a_{ij})}{k(k-1)N(N-1)} + 1 \tag{8.20c}$$

$$= \frac{8[(7^2 + 5.5^2 + \ldots + 9^2) - 10(7 + 5.5 + \ldots + 9)]}{(10)(10-1)(5)(5-1)} + 1$$

$$= \frac{8[(308) - 10(48)]}{(10)(9)(5)(4)} + 1$$

$$= 0.236$$

[18] Sawyer, T. A. y Castellan, N. J. Jr., "Preferences among predictions and the correlation between predicted and observed judgments", 1983.

Tabla 8.13. Matriz de preferencia para 10 sujetos en el estudio de uso del suelo.

	Uso múltiple	*Belleza*	*Recursos*	*Ingresos*	*Condiciones económicas*
Uso múltiple	–	3	$4\frac{1}{2}$	$7\frac{1}{2}$	$2\frac{1}{2}$
Belleza	7	–	8	10	7
Recursos	$5\frac{1}{2}$	2	–	6	$2\frac{1}{2}$
Ingresos	$2\frac{1}{2}$	0	4	–	1
Condiciones económicas	$7\frac{1}{2}$	3	$7\frac{1}{2}$	9	–

Así, vemos que existe un moderno acuerdo entre los sujetos en sus preferencias para los factores. En la siguiente sección determinaremos si este grado de acuerdo representa una desviación considerable del acuerdo aleatorio entre los jueces.

Aunque puede parecer deseable calcular W_T para estas evaluaciones, debemos recordar que sería apropiado sólo si las evaluaciones estuvieran en rangos. Ya que las evaluaciones fueron comparaciones apareadas, el índice de concordancia W_T no se calcula para estos datos.

Prueba de la significación de *u*

El estadístico u puede ser concebido como un estimador de un parámetro de la población υ, que representa el grado real de acuerdo en la población. En este caso, la población consiste en los objetos a los que se les asignaron rangos. A diferencia de muchos otros estadísticos examinados en este libro, al probar hipótesis concernientes al coeficiente de acuerdo, existen dos casos que debemos considerar, ya que la distribución muestral de u depende de si los datos son comparaciones apareadas o rangos. Veamos cada una de ellas. Cabe destacar desde el principio que con el propósito de probar apropiadamente hipótesis acerca de υ, el investigador debe conocer la naturaleza de los datos en los que el coeficiente de acuerdo sea calculado.

Prueba de la significancia cuando los datos son comparaciones apareadas

Cuando los datos usados para calcular el coeficiente de acuerdo son comparaciones apareadas, podemos probar la hipótesis nula H_0: $\upsilon = 0$ contra la hipótesis alterna H_1: $\upsilon \neq 0$. Esto es, la hipótesis nula es que no existe acuerdo entre los evaluadores, y la hipótesis alterna es que el grado de acuerdo es mayor del que se esperaría que tuvieran las comparaciones apareadas si se hubieran hecho al azar. Si el número de jueces o evaluadores es pequeño ($k \leqslant 6$) y el número de variables o factores a los que se les asignaron rangos es también pequeño ($N \leqslant 8$), entonces puede usarse la tabla U del Apéndice I para probar las hipótesis concernientes al acuerdo. Para cada valor de k y N, la tabla presenta los valores posibles de $u \geqslant 0$ junto con la probabilidad de obtener un valor de u igual o mayor que el valor tabulado. Supongamos que $k = 4$ jueces evaluaron a un grupo de $N = 6$ objetos por el método de comparaciones apareadas. Supongamos posteriormente que el valor observado de u fue 0.333. Al consultar la tabla U, vemos que la probabilidad de observar un valor de $u \geqslant 0.333$ tiene una probabilidad de ocurrencia de 0.0037, habiendo los jueces distribuido sus preferencias al azar. En este caso, sería apropiado concluir que existe un acuerdo significativo entre los jueces. Por conveniencia, también se incluyen en la tabla U los valores de S que corresponden a las sumatorias de la ecuación (8.20*c*):

$$S = \Sigma a_{ij}^2 - k\Sigma a_{ij}$$

En algunos casos, puede ser más conveniente determinar la significación de u usando S en lugar de u.

Para otros valores de k y N podemos usar una aproximación de grandes muestras para la distribución muestral. En este caso la prueba estadística es

$$\begin{aligned} X^2 &= \binom{N}{2}[1 + u(k-1)] \\ &= \frac{N(N-1)[1 + u(k-1)]}{2} \end{aligned} \tag{8.22}$$

la cual está asintóticamente distribuida como χ^2 con $\binom{N}{2} = N(N-1)/2$ grados de libertad. La prueba está estrechamente relacionada con la prueba ji cuadrada de la bondad de ajuste, que se examinó en el capítulo 3.

Ejemplo. En el ejemplo de toma de decisiones proporcionado anteriormente en esta sección, el valor del coeficiente de acuerdo de Kendall fue $u = 0.236$. Para probar la hipótesis de que existe acuerdo entre los $k = 10$ sujetos al evaluar los $N = 5$ factores que afectan el uso del suelo, no es posible usar la tabla U del Apéndice I debido a que la tabla está limitada a $k \leqslant 6$. Por tanto, probaremos la hipótesis H_0: $\upsilon = 0$ usando la ecuación (8.22):

$$X^2 = \frac{N(N-1)[1 + u(k-1)]}{2} \tag{8.22}$$

$$= \frac{(5)(5-1)[1 + (0.236)(10-1)]}{2}$$

$$= 10(1 + 2.124)$$

$$= 31.24$$

que está asintóticamente distribuida como χ^2 con $\binom{N}{2} = 5(5 - 1)/2 = 10$ grados de libertad. La tabla C del Apéndice I muestra que podemos rechazar la hipótesis nula H_0: $\upsilon = 0$ en el nivel $\alpha = 0.001$ y concluir que existe un fuerte acuerdo entre los sujetos en sus evaluaciones acerca de la importancia de los factores que afectan el uso del suelo.

PRUEBA DE LA SIGNIFICACIÓN CUANDO LOS DATOS ESTÁN EN RANGOS

Cuando los datos usados para calcular el coeficiente de acuerdo están basados en rangos, la prueba de significación puede ser escrita en términos de $\bar{\tau}$, que es el valor de la población para el promedio τ. Entonces la hipótesis nula es H_0: $\bar{\tau} = 0$; la hipótesis alterna es que $\bar{\tau} \neq 0$. (De manera equivalente, podemos considerar la hipótesis de que el valor de la población $W_\tau = 0$ contra la hipótesis de que el valor de la población $W_\tau \neq 0$.) La prueba de significación es

$$X^2 = \frac{6(2N + 5)\binom{N}{2}\binom{k}{2}}{(k - 2)(2N^2 + 6N + 7)}|u| + f$$

$$= \frac{3(2N + 5)N(N - 1)k(k - 1)}{2(k - 2)(2N^2 + 6N + 7)}|u| + f \tag{8.23}$$

que está distribuida aproximadamente como χ^2 con f grados de libertad:

$$f = \frac{2(2N + 5)^3\binom{N}{2}\binom{k}{2}}{(k - 2)^2 (2N^2 + 6N + 7)^2}$$

$$= \frac{(2N + 5)^3 N(N - 1)k(k - 1)}{2(k - 2)^2 (2N^2 + 6N + 7)^2} \tag{8.24}$$

Debe notarse que, en general, los grados de libertad determinados con la ecuación (8.24) no serán un entero. Para el uso apropiado de la aproximación, es suficiente reducir f al entero siguiente más pequeño cuando se entra en una tabla de la distribución χ^2 tal como la tabla C del Apéndice I.

En el ejemplo de los tres ejecutivos examinado anteriormente (véase la tabla 8.12), encontramos que $u = -0.244$. Para probar la hipótesis H_0: $\bar{\tau} = 0$, primero usamos la ecuación (8.24) para encontrar f, los grados de libertad:

$$f = \frac{(2N + 5)^3 N(N - 1)k(k - 1)}{2(k - 2)^2 (2N^2 + 6N + 7)^2} \tag{8.24}$$

$$= \frac{[(2)(6) + 5]^3 (6)(6 - 1)(3)(3 - 1)}{2(3 - 2)^2 [(2)(6^2) + (6)(6) + 7]^2}$$

$$= \frac{(17^3)\ (6)(5)(3)}{115^2} = 33.43$$

En seguida, usamos la ecuación (8.23) para encontrar el valor de X^2:

$$X^2 = \frac{3(2N + 5)N(N - 1)k(k - 1)}{2(k - 2)(2N^2 + 6N + 7)}|u| + f \tag{8.23}$$

$$= \frac{3[(2)(6) + 5](6)(6 - 1)(3)(3 - 1)}{2(3 - 2)[(2)(6^2) + (6)(6) + 7]}| - 0.244| + f$$

$$= \frac{(3)(17)(6)(5)(3)(2)}{2[(2)(36) + (6)(6) + 7}| - 0.244| + f$$

$$= \frac{(9180)| - 0.244|}{230} + 33.43$$

$$= 43.17$$

Al consultar la tabla C del Apéndice I con $f = 33$ grados de libertad, se indica que no podemos rechazar la hipótesis de que los rangos de los ejecutivos acerca de los solicitantes no están relacionados (o son independientes), en el nivel $\alpha = 0.05$. Este resultado es consistente con lo visto en la sección anterior.

Resumen del procedimiento

Éstos son los pasos en la determinación de u, el coeficiente de acuerdo de Kendall:

1. Sea N el número de entidades u objetos a ser evaluados (ya sea por rangos o por comparaciones apareadas), y sea k el número de jueces que asignan las evaluaciones. Coloque los datos en una matriz de preferencias $N \times N$ como la descrita en esta sección. Si existen rangos ligados, agregue $\frac{1}{2}$ a cada celda ij y ji en la que ocurra cada liga. Denote la frecuencia total en la ij-ésima celda como a_{ij}.
2. Con el uso de las frecuencias ya sea por arriba o por debajo de la diagonal (cualquiera de las dos es conveniente), calcule Σa_{ij}^2 y Σa_{ij} y determine el valor de u mediante la ecuación (8.20*c*).
3. El método para determinar si el valor observado de u es significativamente diferente de cero, depende de si los datos se obtuvieron por comparaciones apareadas o por rangos:

   a) Si los datos fueron obtenidos por el método de comparaciones apareadas, la tabla U del Apéndice I proporciona las probabilidades de u para $k \leq 6$ y $N \leq 8$. Si la magnitud de k o de N excluye el uso de la tabla U, puede usarse la ecuación (8.22) para calcular un valor de X^2 que esté aproximadamente distribuido como χ^2, y cuya significación $gl = N(N - 1)/2$ pueda determinarse mediante la tabla C del Apéndice I.

b) Si los datos se obtuvieron por el método de rangos, puede usarse la ecuación (8.23) para calcular un valor de X^2 que esté aproximadamente distribuido como χ^2 con grados de libertad proporcionados por la ecuación (8.24). La significación de u puede obtenerse usando la tabla C del Apéndice I. [Si los grados de libertad obtenidos con la ecuación (8.24) no son enteros, reduzca el valor al siguiente entero más bajo antes de entrar en la tabla C.]

4. Si la probabilidad obtenida con la tabla U o con la tabla C del Apéndice I es menor o igual a la probabilidad predeterminada α, rechace H_0 y concluya que las evaluaciones (comparaciones apareadas o rangos) no son independientes.

Correlación T_c entre varios jueces y un rango criterio

Una ventaja del coeficiente de acuerdo de Kendall u sobre el uso de W, el coeficiente de concordancia de Kendall, es que es el *promedio* de la correlación por orden de rangos de Kendall entre varios jueces. Otra ventaja es que se generaliza directamente a la correlación entre varios jueces y un rango criterio. Supongamos que hubo varios entrenadores clínicos a quienes se les pidió que asignaran rangos a un grupo de pacientes en orden de la severidad de su trastorno patológico. El coeficiente de correlación r_s de Spearman de rangos ordenados y el coeficiente de correlación T de Kendall de rangos ordenados proporcionan un índice de la relación entre dos jueces, y el coeficiente de concordancia W de Kendall y el coeficiente de acuerdo u de Kendall proporcionan una indicación del acuerdo (o concordancia) *entre* los jueces; sin embargo, estas medidas no indican qué tan cercanamente acuerdan los rangos con algún criterio especificado. En esta sección delineamos un procedimiento para calcular T_c, la correlación entre k conjuntos de rangos y un rango criterio. Cabe destacar desde el principio que T_c es el *promedio* de los coeficientes de correlación de Kendall de rangos ordenados entre cada juez y el rango criterio. Sin embargo, encontraremos que existe un modo relativamente simple de calcular la correlación T_c, y podemos también ejecutar una prueba de significación de T_c.

CÁLCULO DE T_c

El primer paso para calcular T_c es determinar el rango criterio para N objetos. Úsese este rango para construir una matriz de preferencias en la que los objetos (variables) estén enumerados en el orden criterio. En seguida, para cada uno de los k jueces, introduzca los rangos en la matriz de preferencias usando el método delineado al principio de esta sección. Despues, denote la suma de las frecuencias arriba de la diagonal como $\Sigma^+ a_{ij}$ y aquéllas debajo de la diagonal como $\Sigma^- a_{ij}$, y así podemos calcular T_c, la correlación con un rango criterio:

$$T_c = \frac{2(\Sigma^+ a_{ij} - \Sigma^- a_{ij})}{kN(N-1)} \tag{8.25}$$

De manera alternativa y con frecuencia más conveniente, las fórmulas de cálculo para T_c son las siguientes:

$$T_c = \frac{4\Sigma^+ a_{ij}}{kN(N-1)} - 1 \tag{8.25a}$$

y

$$T_c = 1 - \frac{4\Sigma^- a_{ij}}{kN(N-1)} \tag{8.25b}$$

Debe notarse que $\Sigma^+ a_{ij}$ es el número de acuerdos en los rangos con el criterio tomado a través de los jueces. De manera similar, $\Sigma^- a_{ij}$ es el número de desacuerdos en el orden entre los rangos.

Prueba de significación de T_c

Las probabilidades para la distribución muestral de T_c se proporcionan en la tabla V del Apéndice I para $k = 2$ y 3 y $2 \leq N \leq 5$. Para otros valores, la distribución muestral de T_c es aproximadamente normal. Por tanto, para probar la hipótesis H_0: $\tau_c = 0$ contra la hipótesis alterna H_1: $\tau_c > 0$, podemos usar el estadístico

$$z = \left[T_c \pm \frac{2}{kN(N-1)}\right] \frac{3\sqrt{kN(N-1)}}{\sqrt{2(2N+5)}} \tag{8.26}$$

que se distribuye de manera aproximadamente normal con media cero y desviación estándar uno. Puede usarse la tabla A del Apéndice I para estimar las probabilidades asociadas con los valores de T_c. Para calcular z, se sustrae $2/kN(N-1)$ en el numerador si $T_c > 0$; en otro caso, la cantidad se suma (lo cual sería el caso si hubiéramos probado la hipótesis H_1: $\tau_c < 0$).

Ejemplo.[19] Supongamos que $k = 5$ jueces han asignado rangos a $N = 5$ objetos y deseamos determinar la correlación entre los rangos de los jueces y un rango criterio. Por conveniencia, el rango criterio de los objetos sigue el orden de su código de rotulamiento, esto es, *A*, *B*, *C*, *D*, *E*. Los rangos asignados por los jueces a los objetos están proporcionados en la tabla 8.14. El rango criterio se usa para rotular las filas y columnas de la matriz de preferencias en la porción inferior de la tabla 8.14. Usando los rangos, los datos se resumen entonces en la tabla de preferencias. Para estos datos, encontramos que $\Sigma^+ a_{ij} = 37$ y $\Sigma^- a_{ij} = 13$.

[19] Estos datos corresponden a un ejemplo que proporcionan Stilson y Campbell (1962).

Tabla 8.14. Datos de los rangos para el cálculo de T_c la correlación entre varios rangos y el rango criterio.*

	Pacientes				
Juez	*A*	*B*	*C*	*D*	*E*
I	1	2	3	4	5
II	2	1	4	3	5
III	4	1	3	2	5
IV	1	3	5	2	4
V	1	4	3	5	2

Matriz de preferencia

	A	*B*	*C*	*D*	*E*
A	–	3	4	4	5
B	2	–	4	4	4
C	1	1	–	2	3
D	1	1	3	–	4
E	0	1	2	1	–

* Los objetos (pacientes) se enumeran en el orden de criterio de los rangos.

Para calcular T_c, usaremos la ecuación (8.25*a*):

$$T_c = \frac{4\Sigma^+ a_{ij}}{kN(N-1)} - 1 \tag{8.25a}$$

$$= \frac{(4)(37)}{(5)(5)(4)} - 1 \quad = 0.48$$

El lector puede verificar que se habrían obtenido los mismos valores si hubiéramos usado la ecuación (8.25) o la (8.25*b*).

Para probar la hipótesis de que el acuerdo observado entre los rangos de los sujetos y el criterio excede lo que se esperaría si los rangos se hubieran asignado al azar, usamos la ecuación (8.26) para probar la hipótesis de que el valor de la población $\tau_c = 0$ contra la hipótesis de que el valor de la población $\tau_c > 0$:

$$z = \left[T_c \pm \frac{2}{kN(N-1)}\right] \frac{3\sqrt{kN(N-1)}}{\sqrt{2(2N+5)}} \tag{8.26}$$

$$= (0.48 - 0.02)\ \frac{3\sqrt{(5)(5)(5-1)}}{\sqrt{2[(2)(5)+5]}} \quad = 2.52$$

La tabla A del Apéndice I revela que la probabilidad de obtener un valor de $z \geq 2.52$ es 0.0059 (unidireccional). Por tanto, podemos concluir con un alto grado de confianza que los evaluadores de un grupo muestran fuerte acuerdo con el rango criterio.

Referencias bibliográficas

El coeficiente de acuerdo es examinado por Kendall (1975), quien también derivó la distribución muestral de u cuando los datos están basados en comparaciones apareadas. La distribución muestral de u cuando los datos están basados en rangos también se presenta en la monografía de Kendall; se puede encontrar un análisis útil en Ehrenberg (1952). Para información adicional de u y sobre la correlación entre un conjunto de rangos y un rango criterio T_c, el lector puede consultar a Hays (1960) y Stilson y Campbell (1962). Hays también examina un índice apropiado para evaluar el acuerdo entre varios grupos de jueces. En aquella época, era poco lo que se conocía acerca del poder de los diferentes índices que se estudian en esta sección. En Feigin y Cohen (1978) se presentan otras aproximaciones al análisis de datos derivados de comparaciones apareadas y rangos.

DATOS EN ESCALAS NOMINALES Y EL ESTADÍSTICO KAPPA K

En las dos secciones previas estudiamos dos medidas de acuerdo entre un conjunto de k jueces que han asignado rangos o comparado N objetos (entidades o sujetos). Esas medidas, ave(r_s), el coeficiente de concordancia W de Kendall, el coeficiente de acuerdo u de Kendall y sus correspondientes medidas de concordancia W_T, suponen que a los objetos se les pueden asignar rangos o, en el caso del coeficiente de acuerdo de Kendall, que se pueden hacer comparaciones apareadas entre los objetos. En algunas situaciones, los objetos pueden no estar ordenados sino simplemente asignados a categorías que pueden no tener ningún orden inherente a ellas. Un ejemplo sería el caso de un grupo de k psicólogos que desean asignar a cada miembro de un grupo de N pacientes o clientes, a uno de los m diagnósticos o categorías de tratamiento. Las categorías de tratamiento son simplemente clasificaciones nominales. Supongamos que cada uno de los evaluadores categoriza a cada paciente independientemente de los otros pacientes o los otros evaluadores. Dada esta situación, sería posible para un evaluador determinado asignar a cada paciente a la misma categoría o distribuir a los pacientes a través de las categorías. Lo que el investigador desearía conocer acerca de las asignaciones es si los evaluadores concuerdan uno con el otro acerca de la pertenencia a la categoría de cada paciente. En un extremo, los evaluadores podrían tener completo acuerdo entre cada uno de ellos, y en el otro extremo, sus asignaciones pudieran no mostrar acuerdo y parecer aleatorias. (Debe notarse que aún si los evaluadores asignan aleatoriamente los pacientes a las categorías, existiría algún pequeño acuerdo entre sus propias asignaciones al azar, especialmente si el número de evaluadores k excede el número de categorías m.)

El *estadístico kappa* que se examina en esta sección describe una de un número de medidas de acuerdo que se han propuesto para variables categóricas. Estas medidas son todas similares; algunas de ellas están especializadas en evaluar el acuerdo entre sólo dos evaluadores o un simple evaluador evaluando pares de objetos. Nuestra elección es un estadístico conceptualmente similar a nuestras primeras medidas de acuerdo y se puede aplicar a las asignaciones hechas por un número arbitrario de evaluadores. Las referencias dirigirán al lector a alguna de las demás medidas.

Racionalización y método

Considérese un grupo de N objetos o sujetos, cada uno de los cuales va a ser asignado a una de m categorías. Se supone que estas categorías son nominales. Cada uno de un grupo de k evaluadores asigna cada objeto a una categoría. Los datos de las asignaciones pueden ser colocados en una tabla de $N \times m$:

	Categoría						
Objeto	*1*	*2*	$\cdots$	j	$\cdots$	m	
1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1m}	S_1
2	n_{21}						S_2
$\vdots$				$\vdots$			$\vdots$
i	n_{i1}		$\cdots$	n_{ij}	$\cdots$	n_{im}	S_i
$\vdots$				$\vdots$			$\vdots$
N	n_{N1}		$\cdots$	n_{Nj}	$\cdots$	n_{Nm}	S_N
	C_1	C_2	$\cdots$	C_j	$\cdots$	C_m	

donde n_{ij} es el número de evaluadores que asignan el i-ésimo objeto a la j-ésima categoría. Ya que cada evaluador clasifica cada objeto, la suma de las frecuencias en cada fila es igual a k. Sin embargo, el número de veces que un objeto es asignado a una categoría particular, varía de categoría a categoría. Sea C_j el número de veces que un objeto es asignado a la j-ésima categoría, lo cual es simplemente la columna sumada de frecuencias:

$$C_j = \sum_{i=1}^{N} n_{ij}$$

Ahora bien, si los evaluadores están en completo acuerdo concerniente a sus asignaciones, una frecuencia en cada fila sería igual a k y las otras frecuencias serían iguales a cero. Si no hay consenso entre los evaluadores, las asignaciones serían aleatorias y las frecuencias en cada fila serían proporcionales a los totales de la columna. Naturalmente, si los evaluadores fueran a hacer asignaciones aleatorias, esperaríamos que ocurriera algún acuerdo puramente por azar.

El coeficiente de acuerdo kappa es la razón de la proporción de veces que los evaluadores están de acuerdo (corregida para acuerdo aleatorio), a la proporción máxima de veces que los evaluadores podrían concordar (corregida para acuerdo aleatorio):[20]

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{8.27}$$

donde $P(A)$ es la proporción de veces que los k evaluadores concuerdan y $P(E)$ es la proporción de veces que esperaríamos que los k evaluadores concordaran por azar. Si existe completo acuerdo entre los evaluadores, entonces $K = 1$; mientras que si no existe acuerdo entre los evaluadores (diferentes del acuerdo que se esperaría que ocurriera por azar), entonces $K = 0$.

Para encontrar $P(E)$ notamos que la proporción de objetos asignados a la j-ésima categoría es $P_j = C_j/Nk$. Si los evaluadores hacen sus asignaciones al azar, la proporción esperada de acuerdo para cada categoría sería P_j^2, y el acuerdo total esperado a través de todas las categorías sería

$$P(E) = \sum_{j=1}^{m} p_j^2 \tag{8.28}$$

La extensión del acuerdo entre los evaluadores concerniente al i-ésimo sujeto es la proporción del número de pares para los cuales existe acuerdo, a los posibles pares de asignaciones. Para el i-ésimo sujeto esto es

$$S_i = \frac{\sum_{j=1}^{m} \binom{n_{ij}}{2}}{\binom{k}{2}} = \frac{1}{k(k-1)} \sum_{j=1}^{m} n_{ij}(n_{ij} - 1)$$

Para obtener la proporción total de acuerdo, encontramos el acuerdo de estas proporciones a través de todos los objetos evaluados:

$$P(A) = \frac{1}{N} \sum_{i=1}^{N} S_i = \left[\frac{1}{Nk(k-1)} \sum_{i=1}^{N} \sum_{j=1}^{m} n_{ij}^2 \right] - \frac{1}{k-1} \tag{8.29}$$

Los valores de $P(E)$ y $P(A)$ van entonces a combinarse usando la ecuación (8.27) para encontrar el estadístico kappa K.

Ejemplo.[21] *Los investigadores de la conducta animal han observado que el pez macho varilla cambia de color durante el ciclo de anidamiento y de cortejo. Cuando se coloca en un medio adecuado, el varilla macho establece territorios, construye nidos y exhibe conductas de cortejo y agresión cuando son introducidos peces como estímulos dentro del medio.*

[20] En muchos libros e informes de investigación es común denotar el estadístico kappa usando la letra griega κ. Así mismo, muchos de los diferentes estadísticos "similares a kappa" también se denotan con κ. En este libro, usaremos κ para denotar el parámetro que es estimado por el estadístico kappa K.

[21] Rowland, W. J., "The relationships among nuptial coloration, aggression, and courtship of male three-spined sticklebacks", en *Gasterosteus aculeatus. Canadian Journal of Zoology*, núm. 62, 1984, págs.

Para analizar la relación entre el color y otras conductas durante el estudio expenmental, fue necesario codificar al pez en términos de su coloración. Ya que el pez debe ser observado desde fuera de su ambiente y debido a las variaciones en las condiciones observacionales, $k = 4$ observadores entrenados evaluaron la coloración de cada pez. Las coloraciones fueron divididas en $m = 5$ categorías. La primera categoría era para aquellos peces con mínimo desarrollo del color y la última categoría representaba el desarrollo de color y coloración máximos; las otras tres categorías abarcaban diferentes grados de coloración. En este estudio, se observó un grupo de $N = 29$ peces. Los datos se resumen en la tabla 8.15. Nótese que los evaluadores estuvieron en completo acuerdo acerca de la coloración del pez 1 y que dividieron sus evaluaciones del pez 2. Un examen de las filas de la tabla muestra que hubo completo acuerdo para algunos peces, pero bajo acuerdo acerca de otros.

Para evaluar el consenso total entre los evaluadores, se calculará el coeficiente de acuerdo kappa K. Primero, encontramos C_j, el número de veces que un pez fue asignado a la j-ésima categoría. Sumamos las frecuencias en cada columna para obtener los valores proporcionados en la segunda a la última fila en la tabla. Cada una de éstas se divide por $Nk = (29)(4) = 116$ para obtener p_j, la proporción de observaciones asignadas a la categoría j. Encontramos que $p_1 = C_1/Nk = \dfrac{42}{116} = 0.362$, etc. Estos valores se proporcionan en la última fila de la tabla. A partir de tales valores podemos determinar el valor de $P(E)$, la proporción de acuerdo que esperaríamos por azar:

$$P(E) = \sum_{j=1}^{m} p_j^2 \tag{8.28}$$

$$= 0.362^2 + 0.026^2 + 0.319^2 + 0.069^2 + 0.224^2 = 0.2884$$

En seguida debemos encontrar $P(A)$, la proporción de veces que los evaluadores concordaron. Un modo es determinar el valor de S para cada pez y después promediar estos valores. El otro modo es proceder a $P(A)$ directamente usando el lado derecho de la ecuación (8.29). Ilustraremos ambos métodos. Los valores de S_i se proporcionan en la tabla, de manera que el lector pueda entender su cálculo:

$$S_1 = \frac{1}{k(k-1)} \sum_{j=1}^{m} n_{1j}(n_{1j} - 1)$$

$$= \frac{1}{(4)(3)} [0 + 0 + 0 + 0 + (4)(3)]$$

$$= \frac{12}{12} = 1$$

$$S_2 = \frac{1}{(4)(3)} [(2)(1) + 0 + (2)(1) + 0 + 0]$$

$$= \frac{4}{12} = 0.333$$

999-1004. Aunque la coloración cambia con el tiempo (un proceso continuo), las coloraciones son distintas. Un observador entrenado para identificar la coloración sería completamente inconsciente de los aspectos secuenciales. Por tanto, es apropiado un índice de acuerdo categórico.

Pez	Categoría de coloración					S_i
	1	*2*	*3*	*4*	*5*	
1	–	–	–	–	4	$\frac{12}{12} = 1.$
2	2	–	2	–	–	$\frac{4}{12} = 0.333$
3	–	–	–	–	4	$\frac{12}{12} = 1.$
4	2	–	2	–	–	$\frac{4}{12} = 0.333$
5	–	–	–	1	3	$\frac{6}{12} = 0.50$
6	1	1	2	–	–	$\frac{2}{12} = 0.167$
7	3	–	1	–	–	$\frac{6}{12} = 0.50$
8	3	–	1	–	–	$\frac{6}{12} = 0.50$
9	–	–	2	2	–	$\frac{4}{12} = 0.333$
10	3	–	1	–	–	$\frac{6}{12} = 0.50$
11	–	–	–	–	4	$\frac{12}{12} = 1.$
12	4	–	–	–	–	$\frac{12}{12} = 1.$
13	4	–	–	–	–	$\frac{12}{12} = 1.$
14	4	–	–	–	–	$\frac{12}{12} = 1.$
15	–	–	3	1	–	$\frac{6}{12} = 0.50$
16	1	–	2	1	–	$\frac{2}{12} = 0.333$
17	–	–	–	2	2	$\frac{4}{12} = 0.333$
18	–	–	–	–	4	$\frac{12}{12} = 1.$
19	–	–	3	–	1	$\frac{6}{12} = 0.50$
20	–	1	3	–	–	$\frac{6}{12} = 0.50$
21	–	–	1	–	3	$\frac{6}{12} = 0.50$
22	–	–	3	1	–	$\frac{6}{12} = 0.50$
23	4	–	–	–	–	$\frac{12}{12} = 1.$
24	4	–	–	–	–	$\frac{12}{12} = 1.$
25	2	–	2	–	–	$\frac{4}{12} = 0.333$
26	1	–	3	–	–	$\frac{6}{12} = 0.50$
27	2	–	2	–	–	$\frac{4}{12} = 0.333$
28	2	–	2	–	–	$\frac{4}{12} = 0.333$
29	–	1	2	–	1	$\frac{2}{12} = 0.167$
C_j	42	3	37	8	26	
P_j	0.362	0.026	0.319	0.069	0.224	

* Las entradas de las celdas son el número de evaluadores que concuerdan en esa categoría. Una celda vacía indica que la categoría particular no fue escogida por ningún evaluador para ese pez.

El lector notará que el valor de S_i es una medida de acuerdo para el i-ésimo pez. Entonces, usando estos valores, encontramos $P(A)$:

$$P(A) = \frac{1}{N} \sum_{i=1}^{N} S_i \tag{8.29}$$

$$= \frac{1 + 0.333 + 1 + 0.333 + 0.50 + \ldots + 0.333 + 0.167}{29}$$

$$= 0.5804$$

De manera alternativa, podríamos haber eludido por completo el cálculo de S_i, sumando los cuadrados de las frecuencias de las celdas:

$$P(A) = \left[\frac{1}{Nk(k-1)} \sum_{i=1}^{N} \sum_{j=1}^{m} n_{ij}^2 \right] - \frac{1}{k-1} \tag{8.29}$$

$$= \frac{1}{(29)(4)(3)} (4^2 + 2^2 + 2^2 + 4^2 \ldots + 1^2 + 2^2 + 1^2) - \frac{1}{4-1}$$

$$= \frac{318}{348} - \frac{1}{3}$$

$$= 0.5804$$

Podemos usar estos valores de $P(E)$ y $P(A)$ para encontrar K:

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{8.27}$$

$$= \frac{0.580 - 0.288}{1 - 0.288}$$

$$= 0.41$$

Así, concluimos que existe un moderado acuerdo entre los evaluadores. Si este valor representa una diferencia significativa de cero, se examinará en la siguiente sección.

Prueba de significación de K

Después de determinar el valor del estadístico kappa K, generalmente se desaría determinar si el valor observado fue más grande que el valor que se esperaría por azar. Nótese que aunque sustraemos un término de la proporción de acuerdos en las evaluaciones para corregir el acuerdo por azar, tal corrección sustrae solo el *acuerdo esperado* debido al azar. Naturalmente, el acuerdo al azar no será una constante, sino que variará alrededor de algún valor central o esperado. Aunque la distribución muestral de K es complicada para N pequeña, se ha encontrado que para N grande, K se distribuye de manera aproximadamente normal con media cero y varianza

$$\text{var}(K) \approx \frac{2}{Nk(k-1)} \frac{P(E) - (2k-3)[P(E)]^2 + 2(k-2)\,\Sigma p_j^3}{[1-P(E)]^2} \tag{8.30}$$

Por tanto, podemos usar el estadístico

$$z = \frac{K}{\sqrt{\text{var}(K)}} \tag{8.31}$$

para probar la hipótesis H_0: $\kappa = 0$ contra la hipótesis H_1: $\kappa > 0$.

Ejemplo. Para las evaluaciones de coloración proporcionadas en el ejemplo anterior, se encontró que $K = 0.41$. Para probar H_0: $\kappa = 0$ contra H_1: $\kappa > 0$, debemos encontrar la varianza de K. Se escoge el nivel de significación $\alpha = 0.01$. Recuérdese que $N = 29$ (objetos evaluados), $m = 5$ (categorías por evaluar) $k = 4$ (evaluadores) y $P(E) = 0.288$. La única información que se requiere es Σp_j^3. Usando los valores de P_j proporcionados en la tabla 8.15, tenemos

$$\Sigma p_j^3 = 0.362^3 + 0.026^3 + 0.319^3 + 0.069^3 + 0.224^3 = 0.092$$

Entonces,

$$\begin{aligned}\text{var}(K) &\approx \frac{2}{Nk(k-1)} \frac{P(E) - (2k-3)[P(E)]^2 + 2(k-2)\,\Sigma p_j^3}{[1-P(E)]^2} \qquad (8.30)\\ &= \frac{2}{(29)(4)(3)} \frac{0.288 - [(2)(4) - 3](0.288^2) + (2)(4-2)(0.092)}{(1-0.288)^2}\\ &= \frac{2}{348}\left(\frac{0.2413}{0.5069}\right)\\ &= 0.002736\end{aligned}$$

Usando este valor para var(K), podemos encontrar z:

$$\begin{aligned}z &= \frac{K}{\sqrt{\text{var}(K)}} \qquad (8.31)\\ &= \frac{0.41}{\sqrt{0.002736}}\\ &= 7.84\end{aligned}$$

Este valor excede el nivel de significación $\alpha = 0.01$ (cuando $z = 2.32$). Por tanto, el investigador puede concluir que los observadores exhiben acuerdo significativo sobre sus evaluaciones.

Resumen del procedimiento

Éstos son los pasos en la determinación del estadístico kappa K, el coeficiente de acuerdo para datos en escalas nominales:

1. Sea N el número de objetos (sujetos o entidades) que van a ser evaluadas, sea m el número de categorías dentro de las cuales los objetos van a ser asignados, y sea k el número de evaluadores o jueces que producen las evaluaciones. Para cada objeto, cuente el número de veces que los evaluadores lo asignan a cada categoría. Coloque estas frecuencias en una tabla de evaluación de $N \times m$ como la descrita al principio de esta sección. Note que las frecuencias en cada fila de la tabla sumarán k, el número de evaluadores.
2. Para cada categoría j encuentre el número de veces que cualquier objeto es asignado a esa categoría. Este número es C_j. En seguida, encuentre p_j, la proporción de evaluaciones asignadas a la j-ésima categoría. Después, usando la ecuación (8.28) encuentre $P(E)$, la proporción esperada de acuerdo entre los evaluadores que hubieran evaluado los objetos al azar.
3. Luego, mediante la ecuación (8.29), encuentre $P(A)$, la proporción promedio de acuerdo.
4. Para encontrar K, el coeficiente de acuerdo, use los valores calculados de $P(E)$ y $P(A)$ en la ecuación (8.27).
5. Finalmente, para probar la hipótesis H_0: $\kappa = 0$ contra H_1: $\kappa > 0$, encuentre la varianza de K mediante la ecuación (8.30) y encuentre el correspondiente valor de z mediante la ecuación (8.31). Si el valor obtenido de z excede el valor crítico apropiado de z en la tabla A del Apéndice I, rechace H_0.

Una nota de varias versiones del estadístico kappa K

Como fue notado anteriormente, hay varios estadísticos que se han propuesto para medir acuerdo para datos en escalas nominales. Éstos se denotan en muchas referencias K (kappa), sin considerar la forma del estadístico. Estos estadísticos se derivan de los argumentos básicos de Scott (1955) y Cohen (1960) para medidas de acuerdo en escalas nominales.[22] Ésta es la forma desarrrollada por Cohen (para el acuerdo entre dos evaluadores o para N pares de evaluadores) que ha motivado muchas generalizaciones. La forma del estadístico kappa que se proporciona en esta sección es una generalización del estadístico de Cohen para k evaluadores, que se debe a Fleiss (1971). Sin embargo, en virtud de algunos argumentos impuestos concernientes al significado de acuerdo "al azar", cuando $k = 2$, nuestro estadístico K de kappa es el mismo que el anterior índice propuesto por Scott. La suposición de Scott y Fleiss es que las P_j son las mismas para todos los evaluadores, esto es, la probabilidad de que un objeto sea asignado a una categoría particular no varía a través de los evaluadores. Aunque algunos investigadores pueden disentir con este punto de vista, según la hipótesis nula de no acuerdo, los evaluadores serían

[22] Se han propuesto estadísticos similares a kappa para otros propósitos. Hammond, Householder y Castellan (1970) describen una medida de dispersión (variabilidad) para datos categóricos, que es una función del estadístico kappa descrito en esta sección.

incapaces de distinguir un objeto de otro. Fleiss argumentó que "tal inhabilidad implica que los evaluadores aplican las evaluaciones totales de asignaciones, (p_j), a todos y cada uno de los sujetos".

Referencias bibliográficas

Las referencias básicas sobre K, el estadístico kappa, y otros índices de acuerdo para datos en escalas nominales, se encuentran en Scott (1955), Cohen (1960) y Fleiss (1971). Cohen (1968) generalizó su índice a situaciones en las cuales las categorías están ponderadas por alguna función objetiva o subjetiva. Otras generalizaciones se hallan en Fleiss (1971), quien incluyó un índice de acuerdo con un criterio (como T_c) y en el trabajo de Light (1971). En Bishop y colaboradores (1975) el lector podrá consultar otras explicaciones útiles.

VARIABLES ORDENADAS Y EL ESTADÍSTICO GAMMA *G*

Función

Hemos examinado con cierto detenimiento medidas útiles para evaluar la relación entre dos variables ordenadas. Estas medidas incluyeron la correlación r_s de Spearman de rangos ordenados y la correlación T de Kendall de rangos ordenados. Aunque tales estadísticos son apropiados para su uso con variables que están en rangos, son menos útiles y apropiados cuando existen muchos empates o en cualquier situación en la que sea significativo colocar los datos en forma de una tabla de contingencia. Se han propuesto numerosas medidas de asociación para variables ordenadas en tablas de contingencia. El índice que se presentará aquí es muy útil, relativamente fácil de calcular y está relacionado a otras medidas que hemos estudiado (en particular, la T de Kendall). El *estadístico gamma G* es apropiado para medir la relación entre dos variables en escalas ordinales. El estadístico gamma fue examinado primero ampliamente por Goodman y Kruskal.

Racionalización

La racionalización del estadístico gamma es muy similar a la de la T de Kendall. Supongamos que tenemos dos variables, A y B, que son ambas variables ordenadas. Asumiremos que la variable A puede tomar los valores $A_1, A_2, \ldots, A_k$. Más aún, asumiremos que las variables están ordenadas en magnitud por sus subíndices esto es, $A_1 < A_2 < \ldots < A_k$. De manera similar, suponemos que la variable B está ordenada de un modo semejante, $B_1 < B_2 < \ldots < B_r$. En la población de la cual las variables A y B se derivan, definimos el parámetro de la población como una función del acuerdo en el ordenamiento de *pares de observación* seleccionados aleatoriamente. El lector notará que una observación consta de dos datos: una observación de la variable A y una observación de la variable B. El parámetro es entonces la diferencia en la probabilidad de que dentro de un par de observaciones A y B no concuerden en su ordenamiento, dando como resultado que no existen empates en los datos. Esto es,

$$\gamma = \frac{P[A \text{ y } B \text{ concuerdan en el orden}] - P[A \text{ y } B \text{ no concuerdan en el orden}]}{1 - P[A \text{ y } B \text{ están ligados}]}$$

$$= \frac{P[A \text{ y } B \text{ concuerdan en el orden}] - P[A \text{ y } B \text{ no concuerdan en el orden}]}{P[A \text{ y } B \text{ concuerdan en el orden}] + P[A \text{ y } B \text{ no concuerdan en el orden}]}$$

Ya que raramente conocemos las probabilidades en la población, debemos estimarlas a partir de los datos; así, debemos usar el estadístico G para estimar γ.

Método

Para calcular el estadístico gamma G de dos conjuntos de variables ordinales, digamos $A_1, A_2, \ldots, A_k$ y $B_1, B_2, \ldots, B_r$, arreglamos las frecuencias en una tabla de contingencias:

	A_1	A_2	$\cdots$	A_k	*Total*
B_1	n_{11}	n_{12}	$\cdots$	n_{1k}	R_1
B_2	n_{21}	n_{22}	$\cdots$	n_{2k}	R_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
B_r	n_{r1}	n_{r2}	$\cdots$	n_{rk}	R_r
Total	C_1	C_2	$\cdots$	C_k	N

Los datos pueden consistir en cualquier número de categorías. Esto es, se puede calcular el estadístico gamma para datos de una tabla de 2×2, una de 2×5, una de 4×4, una de 3×7 o cualquier tabla $r \times k$.

El estadístico gamma G se define como sigue:

$$G = \frac{\#\text{ de acuerdos} - \#\text{ de desacuerdos}}{\#\text{ de acuerdos} + \#\text{ de desacuerdos}}$$

$$= \frac{\#(+) - \#(-)}{\#(+) + \#(-)} \tag{8.32}$$

donde $\#(+)$ y $\#(-)$ denotan el número de acuerdos y el número de desacuerdos, respectivamente, en los rangos. El lector debe notar las similitudes entre G y T examinadas anteriormente en este capítulo. (Si no existen observaciones empatadas, esto es, si todas las frecuencias en la tabla de contingencia son iguales a uno o cero, entonces $G = T$.) El lector interesado debe revisar la sección correspondiente al coeficiente T de Kendall para detalles sobre el cálculo del número de acuerdos y de desacuerdos de los datos "crudos". La expresión proporcionada en dicha sección es una fórmula computacional perfectamente buena; sin embargo, una aproximación alternativa puede simplificar mucho el cálculo de G, en especial si los datos se han colocado en una tabla de contingencia. Debemos primero proporcionar una aproximación "formal" para el cálculo; ésta será seguida por una aproximación heurística que es extremadamente simple.

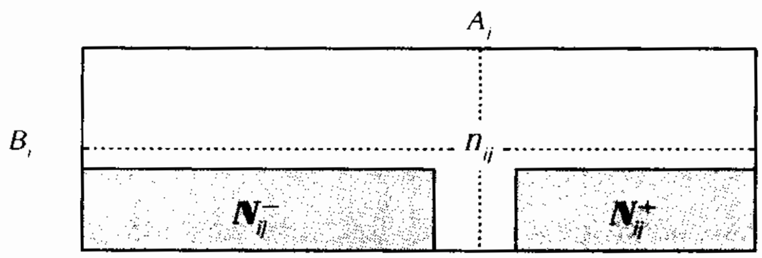
Necesitamos primero un modo simple de calcular el número de acuerdos y el número de desacuerdos para ordenar cada observación cuando los datos han sido agregados en una tabla de contingencia. Podemos hacerlo como sigue:

$$\begin{aligned}\#(+) &= \#\text{ de acuerdos}\\ &= \sum_{i=1}^{r-1}\sum_{j=1}^{k-1} n_{ij} \sum_{p=i+1}^{r}\sum_{q=j+1}^{k} n_{pq} \qquad (8.33a)\\ &= \sum_{i,j} n_{ij}\,\mathbf{N}^{+}_{ij} \qquad i = 1, 2, \ldots, r-1\\ &\qquad\qquad\qquad\quad j = 1, 2, \ldots, k-1 \qquad (8.33b)\end{aligned}$$

donde $\mathbf{N}^{+}_{ij}$ es la *suma* de todas las frecuencias *abajo y a la derecha* de la ij-ésima celda.

$$\begin{aligned}\#(-) &= \#\text{ de desacuerdos}\\ &= \sum_{i=1}^{r-1}\sum_{j=2}^{k} n_{ij} \sum_{p=i+1}^{r}\sum_{q=1}^{j-1} n_{pq} \qquad (8.34a)\\ &= \sum_{i,j} n_{ij}\,\mathbf{N}^{-}_{ij} \qquad i = 1, 2, \ldots, r-1 \qquad (8.34b)\\ &\qquad\qquad\qquad\quad j = 2, \ldots, k\end{aligned}$$

donde $\mathbf{N}^{-}_{ij}$ es la *suma* de todas las frecuencias *abajo y a la izquierda* de la ij-ésima celda en la tabla de contingencia. Gráficamente, podemos representar la expresión como sigue:



En esta tabla, $\mathbf{N}^{+}_{ij}$ y $\mathbf{N}^{-}_{ij}$ son las sumas de las frecuencias en las porciones correspondientes de la tabla. Con estas sumas, y ponderándolas por la frecuencia en la ij-ésima celda, contamos los acuerdos y desacuerdos para cada par de datos en la tabla entera. (Hemos contado acuerdos y desacuerdos considerando cada par sólo una vez.)

Como una ilustración del cálculo del estadístico gamma, considérense los datos de la tabla 8.16. La variable A puede tomar valores sobre $k = 4$ y la variable B puede tomar valores sobre $r = 3$. Se tomaron un total de $N = 70$ observaciones y se colocaron los datos en una tabla de contingencia. Para calcular el número de acuerdos, $\#(+)$ y el número de desacuerdos, $\#(-)$, debemos encontrar diferentes valores de $\mathbf{N}_{ij}^{+}$ y $\mathbf{N}_{ij}^{-}$:

$$\begin{aligned}
\mathbf{N}_{11}^{+} &= 9 + 7 + 1 + 6 + 8 + 9 \\
&= 40 \\
\mathbf{N}_{12}^{+} &= 7 + 1 + 8 + 9 \\
&= 25 \\
\mathbf{N}_{12}^{-} &= 8 + 2 \\
&= 10 \\
\mathbf{N}_{14}^{-} &= 8 + 9 + 7 + 2 + 6 + 8 \\
&= 40
\end{aligned}$$

Tabla 8.16. Datos ficticios para el cálculo del estadístico gamma G.

	Variable A				
Variable B	A_1	A_2	A_3	A_4	Total
B_1	10	5	2	3	20
B_2	8	9	7	1	25
B_3	2	6	8	9	25
Total	20	20	17	13	70

Con estos valores (y los otros valores de $\mathbf{N}_{ij}^{+}$ requeridos), calculamos

$$\begin{aligned}
\#(+) &= \sum_{i,j} n_{ij}\,\mathbf{N}_{ij}^{+} \qquad i = 1, 2 \\
&\qquad\qquad\qquad\quad j = 1, 2, 3 \\
&= (10)(40) + (5)(25) + (2)(10) \\
&\quad + (8)(23) + (9)(17) + (7)(9) \\
&= 945
\end{aligned} \tag{8.33b}$$

y $$\#(-) = \sum_{i,j} n_{ij}\, \mathbf{N}_{ij}^{-} \qquad i = 1, 2 \qquad j = 2, 3, 4 \tag{8.34b}$$

$$= (5)(10) + (2)(25) + (3)(40) + (9)(2) + (7)(8) + (1)(16)$$

$$= 310$$

Con estos valores encontramos

$$G = \frac{\#(+) - \#(-)}{\#(+) + \#(-)} \tag{8.32}$$

$$= \frac{945 - 310}{945 + 310}$$

$$= 0.51$$

Así, podemos concluir que existe un moderado acuerdo (o correlación) entre las dos variables.

El estadístico gamma G es igual a uno si las frecuencias en la tabla de contingencia están concentradas en la diagonal desde la parte superior izquierda hasta la parte inferior derecha de la tabla de contingencia. (Recuérdese que las variables A y B están ordenadas por la magnitud de sus subíndices.) $G = -1$ si todas las frecuencias descansan en la diagonal desde la esquina superior derecha hasta la esquina inferior izquierda de la tabla de contingencia. Existen otros casos para los que $G = 1$. Con tal que no existan desacuerdos en el ordenamiento de las variables, $G = 1$, esto es, si $\#(-) = 0$. De manera similar, si no existen acuerdos en el ordenamiento $[\#(-) = 0]$, $G = -1$. Por ejemplo, en cada una de las siguientes tablas $G = 1$:

	A_1	A_2	A_3
B_1	X	X	
B_2		X	X
B_3			X

	A_1	A_2	A_3
B_1	X		
B_2	X		
B_3	X	X	X

donde X denota cualquier entrada diferente de cero. Si las variables A y B son independientes, entonces $\gamma = 0$. Sin embargo, excepto cuando la tabla de contingencia es 2×2, $\gamma = 0$ no implica independencia.

Ejemplo. Ha habido numerosos estudios en años recientes relacionados con la conducta de fumar y a la habilidad de los individuos que desean dejar de fumar. Un factor que afecta muchos estudios es la variedad de características de la muestra estudiada. En un estudio reciente, un investigador examinó la relación entre la habilidad de dejar de fumar (habilidad de cesación) y el número de años que esa persona había estado fumando.[23] Todos los sujetos fueron enfermeras quienes estaban conscientes de los beneficios que trae consigo dejar de fumar. Más aún, debido a que los sujetos compartían la misma ocupación, deberían ser similares el estrés del trabajo para continuar fumando, así como también las presiones de salud.

Las enfermeras en el estudio eran todas personas que habían dejado de fumar o habían tratado de dejar el hábito. Así, se asignó a cada una de ellas a tres categorías: cesación exitosa, cesación en proceso y cesación no exitosa. Además, los sujetos fueron categorizados por el número de años que habían estado fumando: desde uno a más de 25 años. Los años de fumar se combinaron en siete categorías. Una cuestión importante es si el éxito que se tiene en cesación está relacionado con el número de años que se tienen de fumar.

Estos datos se resumen en la tabla 8.17 para la muestra de $N = 240$ enfermeras. Ya que ambas variables están ordenadas, el estadístico gamma G es una medida apropiada de asociación.

Tabla 8.17. Habilidad de eliminación por cantidad de tiempo fumado.

	Años de fumar							
	1	*2-4*	*5-9*	*10-14*	*15-19*	*20-25*	*>25*	*Total*
Éxito en la eliminación	13	29	26	22	9	8	8	115
Eliminación en proceso	5	2	6	2	1	3	0	19
Eliminación no exitosa	1	9	16	14	21	16	29	106
Total	19	40	48	38	31	27	37	240

Para calcular G, necesitamos calcular el número de acuerdos y desacuerdos en las ordenaciones de las variables en la tabla. Nótese que existen $r = 3$ filas en la tabla correspondiente al estatus actual de fumar y existen $k = 7$ columnas correspondientes al número de años que el sujeto había estado fumando.

$$\#(+) = \sum_{i,j} n_{ij} \mathbf{N}_{ij}^{+} \qquad i = 1, 2 \qquad j = 1, 2, \ldots, 6 \tag{8.33b}$$

$$= (13)(119) + (29)(108) + (26)(86) + \ldots + (1)(45) + (3)(29)$$

$$= 10\,580$$

[23] Wagner, T. J., "Smoking behavior of nurses in western New York", en *Nursing Research*, núm. 34 1985, págs. 58-60.

y $$\#(-) = \sum_{i,j} n_{ij}\ \mathbf{N}_{ij}^{-} \qquad i = 1, 2 \qquad j = 2, 3, \ldots, 7 \tag{8.34b}$$

$$= (29)(6) + (26)(17) + (22)(39) + \ldots + (3)(61) + (0)(77)$$

$$= 3\,690$$

En seguida, calculamos el valor de G:

$$G = \frac{\#(+) - \#(-)}{\#(+) + \#(-)} \tag{8.32}$$

$$= \frac{10\,580 - 3\,690}{10\,580 + 3\,690}$$

$$= 0.483$$

Así, para los datos de la cesación de fumar, existe una asociación positiva entre la inhabilidad de dejar de fumar y el número de años que una persona ha estado fumando, esto es, mientras más tiempo ha estado fumando una persona, menos probable será su éxito en dejar el hábito de fumar.

Prueba de significación de *G*

Para probar la significación de G, debemos recurrir a una aproximación que requiere grandes muestras. Si N es relativamente grande, la distribución de G es aproximadamente normal con media γ. Aunque la expresión para la varianza es complicada, se puede escribir un límite superior para la varianza regularmente sencillo:

$$\operatorname{var}(G) \leq \frac{N(1 - G^2)}{\#(+) + \#(-)} \tag{8.35}$$

Por tanto, la cantidad

$$z = (G - \gamma)\ \sqrt{\frac{\#(+) + \#(-)}{N(1 - G^2)}} \tag{8.36}$$

se distribuye de manera aproximadamente normal con media cero y desviación estándar uno. Ya que la varianza de G proporcionada por la ecuación (8.35) es un límite superior, la prueba de significación al usar la ecuación (8.36) es conservadora; esto es, podemos inferir que el nivel de significación "real" es *al menos* el obtenido mediante la ecuación (8.36) usando una tabla de la distribución normal (por ejemplo, la tabla A del Apéndice I).

Ejemplo. En el estudio de cesación de fumar, encontramos que $G = 0.483$. Aunque esta asociación parece ser grande, quisiéramos probar la hipótesis H_0: $\gamma = 0$ contra la hipótesis H_1: $\gamma \neq 0$. Se requiere una prueba bidireccional debido a que el investigador no tiene una hipótesis *a priori* acerca de la dirección de la asociación. Eligiremos $\alpha = 0.01$ como el nivel de significación. Primero calculamos z:

$$z = (G - \gamma) \sqrt{\frac{\#(+) + \#(-)}{N(1 - G^2)}} \tag{8.36}$$

$$= (0.483 - 0) \sqrt{\frac{10\,580 + 3\,690}{(240)(1 - 0.483^2)}}$$

$$= (0.483)(8.81)$$

$$= 4.24$$

Ya que el valor de z excede el valor crítico para $\alpha = 0.01$ ($z = 2.58$, para un contraste bidireccional), podemos rechazar la hipótesis de que $\gamma = 0$ y concluir que las variables no son independientes en la población.

Resumen del procedimiento

Éstos son los pasos en el cálculo del estadístico gamma G:

1. Coloque las N frecuencias observadas en una tabla de contingencia $r \times k$, donde r es el número de categorías en las que una variable es medida y k es el número de categorías en las que la otra variable es medida. Ya que las variables están ordenadas, la variable de la columna debe ser arreglada con el propósito de incrementar la magnitud a través de las columnas; de manera similar, la variable de la fila debe arreglarse con el propósito de incrementar la magnitud a lo largo de las filas.
2. Use las ecuaciones (8.33*b*) y (8.34*b*) para calcular el número de acuerdos en el ordenamiento, $\#(+)$, y el número de desacuerdos en el ordenamiento, $\#(-)$. Sustituya estos valores en la ecuación (8.32) para calcular G.
3. Si N es moderada o grande, pruebe la hipótesis H_0: $\gamma = 0$ (o la hipótesis H_0: $\gamma = \gamma_0$, si es apropiado), usando la ecuación (8.36) para calcular una desviación normal. Determine la significación de la probabilidad mediante la tabla A del Apéndice I. La probabilidad obtenida es un estimador conservador de la probabilidad de significación "real".

Referencias bibliográficas

Las referencias anteriores en este capítulo son relevantes también para esta sección. Un análisis cabal del estadístico gamma se puede consultar en una serie de escritos de Goodman y Kruskal (1954, 1959, 1963, 1972). Goodman y Kruskal (1963) proporcionan un estimador de la varianza de G más preciso, pero mucho más complejo de calcular. También de interés es el trabajo de Somers (1980), quien proporciona expresiones alternativas para la varianza muestral de G. El escrito de Goodman y Kruskal (1954) proporciona una racionalización para un "gamma parcial", que es similar al coeficiente de correlación parcial $T_{xy.z}$ de Kendall de rangos ordenados examinado anteriormente en este capítulo.

ASOCIACIÓN ASIMÉTRICA Y EL ESTADÍSTICO LAMBDA L_B

Función y racionalización

En la primera sección de este capítulo estudiamos el coeficiente C de Cramér como un índice de asociación para una tabla $r \times k$. Aunque ese índice es muy útil, tiene algunas limitaciones que ya se señalaron. Una de esas limitaciones es que C no mide la asociación que puede existir diferencialmente entre las variables de fila y de columna; en lugar de ello, es un índice del grado de dependencia (o no independencia) entre las dos variables. El coeficiente descrito en esta sección puede usarse cuando queremos medir la asociación entre dos variables. Un ejemplo sería cuando hemos observado una secuencia de conductas y hemos codificado algunas que son antecedentes a una conducta particular y algunas que son consecuentes. Así, los datos consisten en pares antecedente-consecuente. De particular interés para el investigador podría ser la relación entre los antecedentes y los consecuentes (o el grado en el que los consecuentes están relacionados con los antecedentes). En tales situaciones, el coeficiente de Cramér no es sensible a las diferencias en la dependencia que el investigador desea evaluar.

El *estadístico lambda* L_B desarrollado por Goodman y Kruskal es un índice adecuado de asociación cuando se desea evaluar la relación entre una variable y otra. El estadístico lambda hace pocas suposiciones acerca de las categorías que definen las variables orginales. Supone que los datos son sólo categóricos o nominales, esto es, que las variables no están ordenadas. Ya que el estadístico lambda es una medida de la relación asimétrica entre las variables, existen dos índices diferentes, uno basado en filas y otro basado en columnas. En el ejemplo de la conducta secuencial descrito anteriormente, el investigador puede estar interesado en qué tanto la variable A (un antecedente) "predice" la variable B (un consecuente). Sin embargo, la relación inversa entre las dos variables puede ser de menor (o de ningún) interés. El estadístico está diseñado para evaluar el relativo decremento en la impredictibilidad de una variable (por ejemplo, un consecuente), cuando la otra variable (por ejemplo, un antecedente) es conocida; esto es, es una medida de la reducción relativa del error al predecir una variable cuando otra se conoce.

La racionalización del estadístico lambda es relativamente directa. Supongamos que en la población, P[error] es la probabilidad de un error al predecir B y P[error | A] es la probabilidad condicional de un error al predecir B cuando se conoce la variable A; la forma general del índice puede escribirse como

$$\lambda_B = \frac{P[\text{error}] - P[\text{error} \mid A]}{P[\text{error}]}$$

Para calcular λ_B, necesitamos encontrar las dos probabilidades P[error] y P[error | A]. Intuitivamente, la mejor suposición de B cuando el antecedente es desconocido, es elegir esa B_i con la probabilidad de ocurrencia más grande. De manera similar, si se conoce el antecedente A_j, se podría elegir ese consecuente que tuviera la probabilidad de ocurrencia más grande dado A_j. Sin embargo, raramente se conocen estas probabilidades. Por tanto, deben estimarse y, así, estimamos λ_B usando el estadístico L_B.

Método

Para calcular el estadístico lambda L_B de dos conjuntos de variable categóricas, por decir $A_1, A_2, \ldots, A_k$ y $B_1, B_2, \ldots, B_r$, arreglamos las frecuencias en una tabla de contingencia:

	A_1	A_2	$\cdots$	A_k	*Total*
B_1	n_{11}	n_{12}	$\cdots$	n_{1k}	R_1
B_2	n_{21}	n_{22}	$\cdots$	n_{2k}	R_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
B_r	n_{r1}	n_{r2}	$\cdots$	n_{rk}	R_r
Total	C_1	C_2	$\cdots$	C_k	N

Los datos pueden consistir en cualquier número de categorías. Esto es, se puede calcular el estadístico lambda para datos de una tabla de 2×2, una de 2×5, una de 4×4, una de 3×7 o cualquier tabla de $r \times k$.

El estadístico lambda L_B se calcula de una tabla de contingencia como sigue:

$$L_B = \frac{\sum_{j=1}^{k} n_{Mj} - \max(R_i)}{N - \max(R_i)} \tag{8.37}$$

donde n_{Mj} es la frecuencia *más grande* en la j-ésima columna y $\max(R_i)$ es la fila total *más grande*.

Para ilustrar el cálculo de L_B, en la tabla 8.18 se resumirá un conjunto de datos artificiales. Los datos consisten en 60 pares antecedentes-consecuentes. Para estos datos, el total de fila la más grande es 17, tal que $\max(R_i) = 17$. En seguida, necesitamos sumar las frecuencias más grandes en cada columna:

$$\sum_{j=1}^{k} n_{Mj} = 10 + 12 + 8 = 30$$

Entonces, el valor del estadístico lambda es

$$L_B = \frac{\sum_{j=1}^{k} n_{Mj} - \max(R_i)}{N - \max(R_i)} \tag{8.37}$$

$$= \frac{30 - 17}{60 - 17}$$

$$= 0.30$$

Tabla 8.18. Datos ficticios para el cálculo del L_B.

	Antecedente			
Consecuente	A_1	A_2	A_3	*Total*
B_1	10	1	4	15
B_2	5	3	6	14
B_3	3	12	2	17
B_4	3	3	8	14
Total	21	19	20	60

Este valor puede interpretarse de la siguiente manera: cuando conocemos el antecedente (variable A), existe un 30 % de reducción en el error al predecir el valor de la variable B.

Prueba de significación de L_B

Es posible probar hipótesis concernientes a λ_B. Sin embargo, la distribución muestral es relativamente complicada y no es posible probar la hipótesis de que $\lambda_B = 0$ o $\lambda_B = 1$. Podemos probar la hipótesis de que la reducción en el error es un valor particular, esto es, podemos probar la hipótesis H_0: $\lambda_B = \lambda_{B0}$.

Cuando N es relativamente grande, L_B se distribuye de manera aproximadamente normal con media λ_{B0} y varianza

$$\operatorname{var}(L_B) = \frac{\left(N - \sum_{j=1}^{k} n_{Mj}\right)\left(\sum_{j=1}^{k} n_{Mj} + \max(R_i) - 2\Sigma' n_{Mj}\right)}{[N - \max(R_i)]^3} \tag{8.38}$$

donde $\Sigma' n_{Mj}$ es la suma de todas las frecuencias *máximas* que ocurren en la fila asociada con $\max(R_i)$. Si existe sólo un máximo en esa fila, entonces $\Sigma' n_{Mj} = n_{Mj}$. Como una ilustración en el ejemplo proporcionado anteriormente

$$\operatorname{var}(L_B) = \frac{(60 - 30)[30 + 17 - (2)(12)]}{(60 - 17)^3}$$

$$= 0.00868$$

Supóngase que tenemos una hipótesis nula H_0: $\lambda_{B0} = 0.10$, con un nivel de significación de $\alpha = 0.05$ y los datos de la tabla 8.18. Entonces

$$z = \frac{0.30 - 0.10}{\sqrt{0.00868}}$$

$$= 2.15$$

Así, podemos rechazar la hipótesis H_0 de que el valor de λ_B es 0.10; esto es, podemos concluir que el decremento en el error en la predictibilidad de B cuando se conoce A, excede el 10 %.

Propiedades de L_B

Aunque λ_B comparte algunas propiedades con el coeficiente de Cramér, tiene distintas ventajas debido a sus propiedades asimétricas. Algunas de las propiedades de λ_B son las siguientes:

1. Puede variar desde cero hasta uno. Un valor de cero significa que la variable A no tiene valor para predecir la variable B, mientras que un valor de 1 implica una predictibilidad perfecta de la variable B a partir de la variable A.
2. Es igual a cero si y sólo si la variable A no es de ayuda para predecir la variable B.
3. Es igual a uno sólo si existe una completa predictibilidad desde la variable A hasta la variable B. Esto es, si $\lambda_B = 1$, entonces el conocimiento de la variable A nos permitirá predecir la variable B *perfectamente*. Si $\lambda_B = 1$, entonces, para cada valor de la variable A existe sólo *un* posible valor para la variable B. Así, si $L_B = 1$, entonces existe sólo una entrada diferente de cero en cada columna de la tabla de contingencia.
4. Si las variables A y B son independientes, entonces $\lambda_B = 0$. Sin embargo, $\lambda_B = 0$ no implica que las variables A y B sean independientes.
5. El valor de λ_B no está afectado por las permutaciones de filas (o columnas) en la tabla de contingencia. Esto refleja el hecho de que el estadístico no supone ningún ordenamiento de los valores de cualesquiera de las variables.

Debe notarse que aunque existen muchas ventajas para las medidas de asociación asimétricas, un defecto es que las medidas confunden frecuentemente al investigador principiante. Muchos de nosotros estamos tan acostumbrados a pensar sobre las medidas usuales de asociación (simétricas), que resulta difícil interpretar un índice asimétrico.

Predicción de las columnas a partir de las filas: L_A

En nuestro análisis del estadístico lambda, nos hemos centrado en L_B, que se usa para medir la reducción en el error de predicción de la variable B cuando se conoce la variable A. Existe un índice correspondiente para evaluar la reducción en el error de predicción de la variable A cuando se conoce la variable B. Aunque podemos intercambiar las filas y las columnas y calcular el estadístico lambda mediante la ecuación (8.37), por lo general es más conveniente usar una ecuación que no requiera rearreglo de las entradas en la tabla de frecuencia:

$$L_A = \frac{\sum_{j=1}^{r} n_{iM} - \max(C_j)}{N - \max(C_j)} \tag{8.39}$$

donde n_{iM} es la frecuencia *más grande* en la i-ésima fila, y $\max(C_j)$ es la columna total *más grande*. Naturalmente, debe reescribirse la expresión de la varianza de L_A de un modo similar:

$$\operatorname{var}(L_A) = \frac{\left(N - \sum_{i=1}^{r} n_{iM}\right)\left(\sum_{i=1}^{r} n_{iM} + \max(C_j) - 2\Sigma' n_{iM}\right)}{[N - \max(C_j)]^3} \tag{8.40}$$

donde $\Sigma' n_{iM}$ es la suma de todas las frecuencias *máximas* en la columna asociada con $\max(C_j)$. Si existe sólo un máximo en esa columna, entonces $\Sigma' n_{iM} = n_{iM}$.

En general, $L_A \neq L_B$. El lector puede verificar, como un ejercicio, que para los datos de la tabla 8.18, $L_A = 0.38$. De hecho, es posible que L_A (o L_B) sea igual a uno (predictibilidad perfecta), mientras que L_B (o L_A) pueda ser muy pequeño.

En la primera sección de este capítulo se notó que, si el coeficiente de Cramér era igual a uno y la tabla de contingencia *no* era cuadrada, entonces había asociación "perfecta" en sólo una dirección. El estadístico lambda L_B (o L_A) será igual a uno cuando $C = 1$. Si la tabla es cuadrada, entonces si un índice es igual a uno, el otro será igual a uno también.

Resumen del procedimiento

Éstos son los pasos en el cálculo del estadístico lambda L_B:

1. Coloque las N frecuencias observadas en una tabla de contingencia $r \times k$ como la tabla 8.18, donde r es el número de categorías en las que se mide una variable, y k es el número de categorías en las que se mide la otra variable. Calcule los totales marginales de las filas y columnas.

2. Determine la frecuencia máxima en cada columna de la tabla de contingencia (denotada n_{Mj}) y el máximo total de la fila [denotado $\max(R_i)$]. Use estos valores para calcular el valor de L_B usando la ecuación (8.37).
3. Para probar la significación de L_B, use la ecuación (8.38) a fin de calcular la varianza y emplee este valor para calcular una puntuación z. Cuando N es grande, la significación de z (y de aquí la de L_B) puede determinarse mediante la tabla A del Apéndice I. Si el valor observado de z excede el valor crítico, podemos rechazar H_0: $\lambda_B = \lambda_{B0}$.
4. Para calcular L_A y probar hipótesis acerca de λ_A, siga los pasos 1 a 3 usando las ecuaciones (8.39) y (8.40).

Referencias bibliográficas

Se pueden encontrar análisis del estadístico lambda en la serie de escritos de Goodman y Kruskal (1954, 1959, 1963, 1972). Una explicación general de la aplicación de L_B y L_A con énfasis en el análisis de datos secuenciales se halla en Castellan (1979). En este último también se examinan los intervalos de confianza y las pruebas para comparar dos o más lambdas. Todas las referencias anteriores analizan, así mismo, un índice L_{AB}, que es una medida de la reducción en el error de predecir a partir ya sea de la variable A o de la variable B.

ASOCIACIÓN ASIMÉTRICA PARA VARIABLES ORDENADAS: d_{BA} DE SOMERS

Función y racionalización

El estadístico gamma que estudiamos en la sección anterior es un índice apropiado para medir la asociación entre variables ordenadas. Como con el coeficiente de Cramér, que mide la asociación entre dos variables que son categóricas, el estadístico gamma no es sensible a la relación *diferencial* entre dos variables. Cuando las variables son categóricas en escalas nominales, el estadístico lambda es un índice adecuado de asociación asimétrica entre una variable y otra. Cuando las variables están ordenadas, existe algunas veces una necesidad de medir el grado de asociación entre una variable particular y otra. Un ejemplo sería cuando una de las variables está diseñada como una variable independiente y la otra como una variable dependiente. Otro caso sería cuando estamos estudiando secuencias de conductas: ¿están las conductas antecedentes relacionadas con las conductas consecuentes? La Δ de Somers es un índice asimétrico apropiado de relación entre dos variables ordenadas. Siguiendo el rotulamiento de la sección previa, supongamos que la variable A es una variable en una escala ordinal para la cual $A_1 < A_2 < \ldots < A_k$, y que puede considerarse como una variable independiente. Más aún, supongamos que la variable B es una variable en una escala ordinal para la cual $B_1 < B_2 < \ldots < B_r$, y que puede considerarse como una variable dependiente. Esto es, suponemos que A y B están ordenadas en magnitud por sus subíndices. Entonces, Δ_{BA} es un índice asimétrico de asociación entre las variables. Si los papeles de las dos variables se invierten, entonces el índice se denota Δ_{AB}. En una muestra, los estadísticos correspondientes serían d_{BA} y d_{AB} respectivamente.

El parámetro Δ_{BA} es la diferencia entre la probabilidad de que dentro de un par de observaciones, A y B estén en el mismo orden y la probabilidad de que dentro de un par de observaciones A y B no concuerden en su orden, condicionado a *no* empates en la variable A. Una expresión para este parámetro es

$$\Delta_{BA} = \frac{P[A \text{ y } B \text{ concuerdan en el orden}] - P[A \text{ y } B \text{ no concuerdan en el orden}]}{P[\text{un par de observaciones no estén ligadas en } A]}$$

En forma similar

$$\Delta_{AB} = \frac{P[A \text{ y } B \text{ concuerdan en el orden}] - P[A \text{ y } B \text{ no concuerdan en el orden}]}{P[\text{un par de observaciones no ligadas en } B]}$$

Ya que raramente conocemos las probabilidades en la población, debemos estimarlas a partir de los datos; así, debemos usar el estadístico d_{BA} y d_{AB} para estimar Δ_{BA} y Δ_{AB}, respectivamente.

Método

Para calcular la d de Sommers de los dos conjuntos de varianza ordinales, digamos $A_1, A_2, \ldots, A_k$ y $B_1, B_2, \ldots, B_r$, arreglamos las frecuencias en una tabla de contingencia:

	A_1	A_2	$\cdots$	A_k	*Total*
B_1	n_{11}	n_{12}	$\cdots$	n_{1k}	R_1
B_2	n_{21}	n_{22}	$\cdots$	n_{2k}	R_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
B_r	n_{r1}	n_{r2}	$\cdots$	n_{rk}	R_r
Total	C_1	C_2	$\cdots$	C_k	N

Los datos pueden consistir en cualquier número de categorías. Esto es, se puede calcular el estadístico d de Somers para datos de una tabla de 2×2, de una de 2×5 o cualquier tabla de $r \times k$.

Como con el estadístico gamma, empezamos calculando el número de acuerdos y desacuerdos entre los pares de variables; la diferencia entre d y G está en el denominador, ya que debemos omitir las ligas en la variable A. Para calcular d_{BA}, la ecuación es

$$d_{BA} = \frac{\#\text{ acuerdos } - \#\text{ desacuerdos}}{\#\text{ de pares no ligados en la variable } A}$$
$$= \frac{2[\#(+) - \#(-)]}{N^2 - \sum_{j=1}^{k} C_j^2} \tag{8.41}$$

donde # (+) y # (−) con el número de acuerdos y desacuerdos en los ordenamientos, respectivamente, como son definidos en las ecuaciones (8.33) y (8.34). Los procedimientos para calcular estas cantidades de una tabla de contingencia se muestran en la sección correspondiente al coeficiente de correlación r_s de Spearman, en este capítulo. N es el número de observaciones y C_j es la frecuencia marginal del j-ésimo valor de la variable A. Aunque puede no parecer que el denominador cuente los pares y omita los empates en la variable A, si lo hace. Si contamos cada uno de los posibles apareamientos de observaciones, habría $\frac{1}{2} N^2$ apareamientos. (Incluimos aquí la posibilidad de aparear una observación consigo misma, pero dividida por dos debido a que deseamos contar solo apareamientos únicos.) Entonces, existen $\frac{1}{2} C_1^2$ pares para el primer valor de la variable A, esto es, A_1, $\frac{1}{2} C_2^2$ es el número de empates en A_2, etc. Sustraemos estos empates del número total de pares.

Si deseamos calcular el índice asimétrico d_{AB}, la fórmula es

$$d_{AB} = \frac{\#\text{ acuerdos } - \#\text{ desacuerdos}}{\#\text{ pares no ligados en la variable } B}$$
$$= \frac{2[\#(+) - \#(-)]}{N^2 - \sum_{i=1}^{r} R_i^2} \tag{8.42}$$

donde R_i es la frecuencia marginal para el valor B_i.

Para ilustrar el cálculo de d_{BA}, calcularemos el estadístico para los datos de la tabla 8.16. El uso del d_{BA} de Somers sería apropiado si suponemos que la variable A es una variable independiente y que la variable B es una variable dependiente, y que deseamos evaluar la asociación de A a B. En la sección correspondiente al estadístico gamma encontramos $\#(+) = 945$ y $\#(-) = 310$. Usando estos valores y los totales marginales de la columna de la tabla, encontramos

$$d_{BA} = \frac{2[\#(+) - \#(-)]}{N^2 - \sum_{j=1}^{k} C_j^2} \tag{8.41}$$

$$= \frac{2(945 - 310)}{70^2 - (20^2 + 20^2 + 17^2 + 13^2)} = \frac{2(635)}{3\,642} = 0.35$$

Este valor de d_{BA} indica que existe una moderada relación o asociación asimétrica desde la variable A hasta la variable B. (Nótese que no hemos encontrado d_{AB}. Esto se deja como un ejercicio para el lector.)

Ejemplo. Con el desarrollo de un código de barras exploratorio para usar en supermercados y muchas otras tiendas, ha habido una tendencia hacia la omisión de marcar los precios en los reactivos individuales. Los comerciantes están bastante entusiasmados e interesados en no marcar los precios individuales. Dos de las más importantes razones son las siguientes: 1. el ahorro de trabajo resultante de no tener que marcar cada artículo, y 2. la habilidad de reimprimir los artículos rápidamente en respuesta a los cambios en el costo, ventas especiales, etc. Por otra parte, los compradores se han vuelto a acostumbrar a tener los precios marcados sobre los artículos individuales. Las ventajas de los precios unitarios que los compradores citan incluyen la habilidad de 1. comparar fácilmente los precios en diferentes ramas de un producto particular, 2. revisar el costo total de artículos en una canasta marcada y 3. asegurar los cargos correctos al pagar. Si los comerciantes quieren cambiar hacia la omisión de marcar los precios, los especialistas de mercado argumentan que deben mantenerse las campañas de relaciones públicas para educar al público acerca de las ventajas de tales omisiones. Para tener una campaña efectiva, es importante conocer las actitudes actuales y qué tipo de compradores tienen las mayores resistencias a la omisión de precios. En un estudio de compradores en una gran ciudad del Oeste Medio en Estados Unidos[24] se obtuvieron las actitudes hacia la omisión de precios individuales y se relacionaron con un número de variables demográficas tales como edad, sueldo, educación, etcétera.

En una investigación, las variables demográficas pueden considerarse variables independientes y la respuesta a una pregunta de actitud es la variable dependiente. Una de las variables demográficas fue la educación, y los investigadores querían determinar cómo ésta afectaba la actitud. Ya que las variables educación y actitud son ambas variables ordinales y debido a que estamos interesados principalmente en el efecto de la educación sobre la actitud, el d_{BA} de Somers es una medida apropiada. En la tabla 8.19 se resumen las respuestas de $N = 165$ mujeres compradoras. Para determinar la asociación, se calculará el d_{BA} de Somers.

Tabla 8.19. Actitud hacia la omisión de precios en los artículos para diferentes niveles educativos.

	Educación				
Actitud	*Menor de secundaria*	*Secundaria*	*Comercio*	*Bachillerato*	*Total*
Muy mala a mala	22	39	19	8	88
Indiferente	6	8	6	14	34
Buena a muy buena	5	16	12	10	43
Total	33	63	37	32	165

[24] Langrehr, F. W. y Langrehr, V. B., "Consumer acceptance of item price removal: A survey study of Milwaukee shoppers", en *Journal of Consumer Affairs*, núm. 17, 1983, págs. 149-171.

Primero necesitamos determinar el número de acuerdos y desacuerdos en el ordenamiento de las dos variables:

$$\#(+) = (22)(66) + (39)(42) + (19)(24) + \ldots + (8)(22) + (6)(10)$$
$$= 4\,010$$

y
$$\#(-) = (39)(11) + (19)(35) + (8)(53) + \ldots + (6)(21) + (14)(33)$$
$$= 2\,146$$

En seguida, calculamos el d_{BA} de Somers:

$$d_{BA} = \frac{2[\#(+) - \#(-)]}{N^2 - \sum_{j=1}^{k} C_j^2} \tag{8.41}$$

$$= \frac{2[4\,010 - 2\,146]}{165^2 - (33^2 + 63^2 + 37^2 + 32^2)}$$

$$= \frac{2(1\,864)}{19\,774}$$

$$= 0.189$$

Con base en este análisis concluimos que la educación tiene una pequeña relación con la omisión de los precios en los artículos. La tabla muestra una tendencia de que las mujeres con mayor educación tienen actitudes más positivas hacia la omisión de los precios en los artículos y que las mujeres con menor educación tienen actitudes más negativas. Posteriormente se examinará si esta tendencia es significativa.

Interpretación del d_{BA} de Somers

Ya que Δ_{BA} "ignora" los empates entre las variables de la columna, es un índice de la asociación entre dos pares de observaciones que están en dos diferentes columnas (esto es, no existen empates en la variable A). Considérense dos observaciones seleccionadas aleatoriamente $(A - B)$ y $(A' - B')$, en las que A y A' son diferentes. Δ_{BA} de Somers es la diferencia en la probabilidad de que A y A' estén en el mismo orden en que están B y B' (siendo $B = B'$ considerado como un acuerdo en el orden), menos la probabilidad de que A y A' estén en un orden diferente que B y B', todo ello condicionado a que $A \neq A'$.

El índice $d_{BA} = 1$ si y sólo si $\#(-) = 0$ (no existen desacuerdos en el orden) y cada fila tiene al menos una celda diferente de cero. La apariencia de tal tabla de contingencia tendría las celdas diferentes de cero descendiendo desde el extremo superior izquierdo hasta el extremo inferior derecho, como una escalera. De modo similar, $d_{BA} = -1$ si las celdas diferentes de cero ascienden desde el extremo inferior izquierdo hasta el extremo superior derecho.

El índice $d_{BA} = 0$ si las variables (en la muestra) son independientes; sin embargo, $d_{BA} = 0$ no implica independencia, a menos que la tabla de contingencia sea de 2×2. El lector notará que en la población, si las variables A y B son independientes, $\Delta_{BA} = 0$ mientras que $\Delta_{BA} = 0$ no implica independencia.

Si el investigador se centra en d_{BA}, entonces se pueden hacer los argumentos correspondientes; sin embargo, el papel de hacer los argumentos correspondientes; sin embargo, el papel de las filas y de las columnas debe intercambiarse.

Prueba de significación de d_{BA}

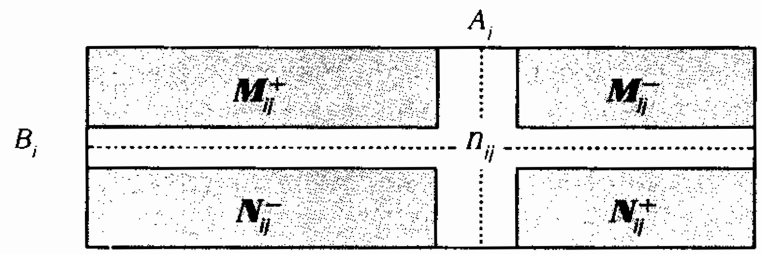
Como muchas de las medidas de asociación proporcionadas en este capítulo, la distribución muestral de d_{BA} es relativamente complicada. Sin embargo, existen algunas simplificaciones posibles que pueden hacer más fácil la prueba de significación.

Recuérdese que cuando calculamos el número de acuerdos $\#(+)$ y el número de desacuerdos $\#(-)$, incluimos sólo apareamientos únicos de los datos en los recuentos. Cuando se calcula la varianza de d_{BA}, necestamos contar todos los acuerdos que ocurren con cada dato. Para hacer esto requerimos alguna notación adicional. Cuando describimos el cálculo de $\#(+)$ y $\#(-0)$, usamos los símbolos $\mathbf{N}^+_{ij}$ y $\mathbf{N}^-_{ij}$ para denotar la *suma* de las frecuencias abajo y a la derecha y la *suma* de las frecuencias abajo y a la izquierda de la ij-ésima celda, respectivamente. Para calcular la varianza de d_{BA} necesitaremos las frecuencias arriba y a la izquierda y arriba y a la derecha de la ij-ésima celda. Denotaremos esto como $\mathbf{M}^+_{ij}$ y $\mathbf{M}^-_{ij}$, respectivamente. Estas dos variables pueden ser definidas usando la siguiente notación:

$$\mathbf{M}^+_{ij} = \sum_{p=1}^{i-1} \sum_{q=1}^{j-1} n_{pq} \tag{8.43}$$

$$\mathbf{M}^-_{ij} = \sum_{p=1}^{i-1} \sum_{q=j+1}^{k} n_{pq} \tag{8.44}$$

Gráficamente, podemos describir la expresión como sigue:



Con el uso de estas sumas, junto con $\mathbf{N}^+_{ij}$ y $\mathbf{N}^-_{ij}$, y ponderándolas por la frecuencia en la ij-ésima celda, podemos contar los acuerdos y desacuerdos para cada par de datos en la tabla entera. (Hemos contado acuerdos y desacuerdos considerando cada objeto como cualquier otro objeto: cada par ha sido contado dos veces.)

Todos estos términos se usan para calcular la varianza de d_{BA} según la hipótesis H_0: $\Delta_{BA} = 0$:

$$\mathrm{var}(d_{BA}) = \frac{4\sum_{i=1}^{r}\sum_{j=1}^{k} n_{ij}\,(\mathbf{N}_{ij}^{+} + \mathbf{M}_{ij}^{+} - \mathbf{N}_{ij}^{-} - \mathbf{M}_{ij}^{-})^2}{\left[N^2 - \sum_{j=1}^{k} C_j^2\right]^2} \tag{8.45}$$

Si suponemos que la muestra ha sido extraída de una población con una distribución uniforme sobre todas las celdas en la tabla, la ecuación (8.45) se simplifica a

$$\mathrm{var}(d_{BA}) = \frac{4(r^2 - 1)(k + 1)}{9Nr^2\,(k - 1)} \tag{8.46a}$$

La ecuación (8.46*a*) también parece ser un estimador razonable de var (d_{BA}), aun cuando la muestra no sea multinomial. Debido a su facilidad de cálculo, puede usarse la ecuación (8.46*a*) cuando el investigador puede suponer un muestreo multinomial uniforme. En muchos casos, el investigador no tiene control sobre las probabilidades de la muestra en, al menos, las columnas y podría ejercer algún control adicional eligiendo cuidadosamente las categorías de B; por tanto, puede ser razonable para estas situaciones la suposición de muestreo multinomial uniforme.

Para probar la hipótesis H_0: $\Delta_{AB} = 0$ contra la alterna uni o bidireccional se usa el siguiente estadístico:

$$z = \frac{d_{BA}}{\sqrt{\mathrm{var}(d_{BA})}} \tag{8.47}$$

Este valor se distribuye de manera aproximadamente normal con media cero y desviación estándar uno. Esta significación de z y, de aquí, la de d_{BA}, puede determinarse consultando la tabla A del Apéndice I.

Si el investigador quiere probar la hipótesis acerca de Δ_{AB}, entonces la varianza var(d_{AB}) podría calcularse mediante la ecuación (8.45), excepto que el denominador fuera reemplazado por

$$\left(N^2 - \sum_{i=1}^{r} R_i^2\right)^2$$

Si la varianza fuera a estimarse mediante la ecuación (8.46*a*), las variables r y k serían intercambiadas:

$$\mathrm{var}(d_{AB}) = \frac{4(k^2 - 1)(r + 1)}{9Nk^2\,(r - 1)} \tag{8.46b}$$

Debe notarse que la varianza proporcionada por la ecuación (8.45) no puede usarse para determinar intervalos de confianza o para probar otras hipótesis diferentes de H_0: $\Delta_{BA} = 0$. En las referencias al final de esta sección se proporcionan las varianzas para otras situaciones.

Ejemplo. En el estudio de actitud del ejemplo previo, encontramos que $d_{BA} = 0.189$. No podemos hablar acerca de la magnitud de d_{BA} sola si el valor observado es significativamente diferente de 0. Probaremos la hipótesis H_0: $\Delta_{BA} = 0$ contra la hipótesis H_1: $\Delta_{BA} \neq 0$. Se usará una prueba bidireccional debido a que los autores no tienen nociones *a priori* acerca de la relación entre la educación y la actitud. Empezamos calculando var(d_{BA}):

$$\mathrm{var}(d_{BA}) = \frac{4\sum_{i=1}^{r}\sum_{j=1}^{k} n_{ij}\,(\mathbf{N}_{ij}^{+} + \mathbf{M}_{ij}^{+} - \mathbf{N}_{ij}^{-} - \mathbf{M}_{ij}^{-})^2}{\left(N^2 - \sum_{j=1}^{k} C_j^2\right)^2} \tag{8.45}$$

$$= \frac{4[(22)(66 - 0) + (39)(42 - 11) + \ldots + (12)(75 - 22) + (10)(100 - 0)]}{[165^2 - (33^2 + 63^2 + 37^2 + 32^2)]^2}$$

$$= \frac{4(389\,112)}{19\,774^2}$$

$$= 0.00398$$

Usando este valor para la varianza, podemos calcular

$$z = \frac{d_{BA}}{\sqrt{\mathrm{var}(d_{BA})}} \tag{8.47}$$

$$= \frac{0.189}{\sqrt{0.00398}}$$

$$= 3.00$$

Ya que este valor excede el valor crítico (bidireccional) de z para $\alpha = 0.05$, podemos rechazar la hipótesis de que la educación no tiene relación con la actitud. Nótese, sin embargo, que *no* hemos probado si existe una asociación entre la educación y la actitud. Hemos considerado sólo la relación asimétrica de la relación *con* la actitud.

Finalmente, como una verificación sobre la aproximación de var(d_{BA}) para un muestreo multinomial uniforme, calcularemos ese estimado:

$$\mathrm{var}(d_{BA}) = \frac{4(r^2 - 1)(k + 1)}{9Nr^2\,(k - 1)} \tag{8.46a}$$

$$= \frac{4(3^3 - 1)(4 + 1)}{9(165)(3^2)(4 - 1)}$$

$$= 0.00399$$

Este valor es muy cercano al obtenido mediante la ecuación (8.45). Aunque los valores fueron extremadamente cercanos en este ejemplo, no existe seguridad de que siempre será así. No obstante, los estudios de Monte Carlo realizados por Somers han comprobado que la diferencia es relativamente pequeña en muchos casos.

Resumen del procedimiento

Éstos son los pasos en el cálculo del d_{BA} de Somers:

1. Coloque las N frecuencias observadas en una tabla de contingencia de $r \times k$, donde r es el número de categorías en las que una variable es evaluada y k es el número de categorías en las que la otra variable es evaluada. Para las variables de la fila, los valores deben tabularse en orden de magnitud creciente a través de las columnas. De manera similar, las variables de la columna deben ser ordenadas en magnitud creciente abajo de las filas. Denote las variables de la columna como A y las variables de la fila como B.
2. Use las ecuaciones (8.33b) y (8.34b) para calcular el número de acuerdos en el ordenamiento, $\#(+)$, y el número de desacuerdos en el ordenamiento, $\#(-)$. Sustituya esos valores en la ecuación (8.41) [o (8.42)] para determinar d_{BA} (o d_{AB}).
3. Si N es moderada o grande, pruebe la hipótesis H_0: $\Delta_{BA} = 0$ (o, si es apropiado, H_0: $\Delta_{AB} = 0$), usando la ecuación (8.47) para calcular una desviación z normal. Use la tabla A del Apéndice I para determinar la significación de z.

Referencias bibliográficas

El índice asimétrico de asociación Δ_{BA} fue propuesto por Somers (1962), quien también ha considerado formas alternativas para su distribución muestral (1980). Son también relevantes las referencias de las dos secciones previas, particularmente las de Goodman y Kruskal (1963, 1972).

ANÁLISIS

En este capítulo hemos presentado numerosas técnicas no paramétricas para medir el grado de asociación de variables en una muestra. Para cada una de éstas se presentaron pruebas de significación de la asociación observada.

Asociación para variables en escalas nominales

Cuatro de estas técnicas, el coeficiente C de Cramér, el coeficiente phi r_ϕ, el estadístico kappa K y el coeficiente lambda L_B, pueden aplicarse cuando los datos son categóricos y descansan en una escala nominal. Esto es, si la medida es tal que las clasificaciones implicadas no están relacionadas dentro de cualquier conjunto y, por tanto, no pueden ordenarse significativamente, entonces estos coeficientes proporcionan medidas útiles del grado de asociación en los datos.

El coeficiente C de Cramér es una de las medidas más simples de asociación para variables categóricas. Aunque proporciona información mínima acerca de la asociación entre las variables, puede ser una opción poco práctica. El coeficiente phi r_ϕ es un índice de asociación apropiado cuando existen dos niveles de cada variable y la información está resumida en una tabla de 2×2.

El estadístico kappa K es un índice útil cuando varios evaluadores han categorizado cada uno de un grupo de objetos o sujetos dentro de categorías nominales. K es un índice de acuerdo entre los evaluadores.

El coeficiente lambda L_B es un índice de asociación asimétrico, que es una medida de la predictibilidad de una de las variables categóricas cuando se conoce el valor de la otra variable. Existen dos medidas: L_B, donde se mide la predictibilidad de la variable B a partir de la variable A; y L_A, donde se mide la predictibilidad de la variable A a partir de la variable B. En general, $L_B \neq L_A$. Como resultado, se debe tener especial cautela al interpretar el estadístico.

Asociación para variables en escalas ordinales

Si las variables en estudio han sido medidas en al menos una escala ordinal, se puede aún usar una de las cuatro medidas categóricas de asociación; sin embargo, una de las diferentes medidas de correlación *por rangos* utilizará la información ordenada en los datos y es, por tanto, preferible.

Si los datos son al menos ordinales, los dos coeficientes de correlación de rangos ordenados la r_s de Spearman y la T de Kendall, son apropiadas. La r_s de Spearman es un tanto más fácil de calcular. La T de Kendall tiene la ventaja adicional de ser generalizable a un coeficiente de correlación parcial $T_{xy.z}$.

El coeficiente de correlación parcial $T_{xy.z}$ de Kendall de rangos ordenados mide el grado de relación entre dos variables X y Y, mientras que una tercera variable Z se mantiene constante (de la cual la asociación entre X y Y pudiera depender lógicamente). $T_{xy.z}$ es el equivalente no paramétrico del coeficiente de correlación parcial producto-momento. De acuerdo con suposiciones razonables, pueden probarse hipótesis acerca del correspondiente parámetro de la población.

Si existen varios conjuntos de rangos u ordenamientos para ser analizados, existen dos medidas de concordancia o acuerdo que pueden utilizarse entre los diferentes conjuntos de rangos. El coeficiente de concordancia W de Kendall y el coeficiente de acuerdo u de Kendall miden la extensión de asociación entre varios (k) conjuntos de rangos de N entidades. Cada uno de ellos es útil para determinar el acuerdo entre varios jueces o la asociación entre tres o más variables. El coeficiente de concordancia W de Kendall está relacionado linealmente con la r_s de Spearman. El otro índice, el coeficiente de acuerdo u de Kendall, está linealmente relacionado con la T de Kendall.

El coeficiente de acuerdo de Kendall también puede ser generalizado a una medida de la concordancia entre varios jueces y un rango criterio T_c. También puede usarse el coeficiente de acuerdo para proporcionar un método estándar para ordenar entidades de acuerdo con el consenso cuando no están disponibles o no se conocen *a priori* ordenamientos objetivos de los objetos.

El coeficiente de acuerdo *u* de Kendall también tiene la ventaja de ser un índice de asociación apropiado cuando los datos se han recabado por el método de comparaciones apareadas, más que por el de asignarles rangos. Para ciertos diseños experimentales, las comparaciones apareadas pueden ser datos más apropiados que los rangos. El índice aún puede usarse si las comparaciones no son consistentes o transitivas.

El estadístico gamma G de Goodman y Kruskal y el d_{BA} de Somers son medidas apropiadas de asociación cuando las observaciones de dos variables ordenadas están reunidas en una tabla de contingencia o cuando las variables son rangos para los que existen muchos empates. El d_{BA} de Somers proporciona una medida de asociación cuando una de las dos variables es de particular importancia o existe una distinción especial entre las variables, por ejemplo, cuando una es una variable dependiente y la otra es una variable independiente. Como el estadístico lambda, el d_{BA} de Somers es asimétrico y se debe tener cuidado en su interpretación.

Existen muchas medidas de asociación que se han desarrollado para usarse con datos categóricos y ordinales. En este capítulo no ha sido posible presentarlas todas. Nuestras elecciones fueron motivadas por un deseo de proporcionar una cobertura de aquellas técnicas que creemos que son las más útiles a los investigadores. Algunas de éstas, tales como el coeficiente de Cramér, la r_s de Spearman y la T de Kendall son familiares a muchos investigadores. Otras, como el estadístico kappa y el coeficiente de acuerdo de Kendall, son menos comunes. Todas ellas resultan útiles si se aplican de manera adecuada.

Apéndice I. Tablas

A. Probabilidades asociadas con el lado superior de la distribución normal.
A_{II} Valores críticos de z para las $\#c$ comparaciones múltiples.
A_{III} Valores críticos de $q(\alpha, \#c)$ para las $\#c$ comparaciones múltiples dependientes.
B. Valores críticos de la distribución t de Student.
C. Valores críticos de la distribución *ji* cuadrada.
D. Tabla de las probabilidades asociadas con valores tan pequeños (o más pequeños) que los valores observados de k en la prueba binomial.
E. La distribución binomial.
F. Valores críticos de D en la prueba de Kolmogorov-Smirnov para una muestra.
G. Valores críticos de r en la prueba de series aleatorias.
H. Valores críticos de T^+ para la prueba de los rangos asignados de Wilcoxon.
I. Probabilidades para tablas de cuatro entradas, prueba exacta de Fisher, $N \leq 15$.
J. Probabilidades del lado inferior y superior para W_x, el estadístico de la suma de rangos de Wilcoxon-Mann-Whitney.
K. Valores críticos de $\acute{U}$ para la prueba poderosa de rangos ordenados.
L_I Prueba de Kolmogorov-Smirnov para dos muestras (contrastes unidireccionales).
L_{II} Prueba de Kolmogorov-Smirnov para dos muestras (contrastes bidireccionales).
L_{III} Valores críticos de $D_{m,n}$ para la prueba de Kolmogorov-Smirnov para dos muestras (muestras grandes, contrastes bidireccionales).
M. Valores críticos para la prueba estadística del análisis de varianza bifactorial por rangos de Friedman, F_r.
N. Valores críticos de estadístico L de la prueba de Page.
O. Valores críticos para el análisis de varianza unifactorial por rangos de Kruskal-Wallis.
P. Valores críticos del estadístico J, de la prueba de Jonckheere.
Q. Valores críticos del coeficiente de correlación R_s de Spearman de rangos ordenados.
R_I Probabilidades del lado superior para T, del coeficiente de correlación de Kendall de rangos ordenados ($N \leq 10$).
R_{II} Valores críticos de T, del coeficiente de correlación de Kendall de rangos ordenados.
S. Valores críticos de $T_{xy \cdot z}$, del coeficiente de correlación parcial de Kendall de rangos ordenados.
T. Valores críticos de W, el coeficiente de acuerdos de Kendall.
U. Probabilidades del lado superior de u, de acuerdos al coeficiente de Kendall cuando los datos corresponden a comparaciones apareadas.
V. Probabilidades del lado superior de T_c, la correlación de k rangos con un criterio de ordenamiento por rangos.
W. Factoriales.
X. Coeficientes binomiales.

Tabla A. Probabilidades asociadas con el lado superior de la distribución normal. El cuerpo de la tabla proporciona las probabilidades unidireccionales según H_0 de z. Las columnas marginales de la izquierda proporcionan diferentes valores de z hasta para un decimal. La fila superior proporciona diferentes valores hasta el segundo lugar decimal. Así, por ejemplo, la probabilidad unidireccional para $z \geq 0.11$ o $z \leq -0.11$ es $p = 0.4562$.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	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									
3.3	0.0005									
3.4	0.0003									
3.5	0.00023									
3.6	0.00016									
3.7	0.00011									
3.8	0.00007									
3.9	0.00005									
4.0	0.00003									

Tabla A. (*Continuación*)

Niveles de significación seleccionados para la distribución normal									
Bidireccional α	0.20	0.10	0.05	0.02	0.01	0.002	0.001	0.0001	0.00001
Unidireccional α	0.10	0.05	0.025	0.01	0.005	0.001	0.0005	0.00005	0.000005
z	1.282	1.645	1.960	2.326	2.576	3.090	3.291	3.891	4.417

Tabla A_{II}. Valores críticos de z para # c* comparaciones múltiples.
Las entradas en la tabla para un # c dado y un nivel de significación α es el punto de la distribución normal estándar tal que la probabilidad del lado superior sea igual a $\frac{1}{2}\alpha / \# c$. Para valores de # c fuera de los rangos incluidos en la tabla, se puede encontrar z usando la tabla A del Apéndice.

	α						
	Bidireccional	0.30	0.25	0.20	0.15	0.10	0.05
# c	Unidireccional	0.15	0.125	0.10	0.075	0.05	0.025
1		1.036	1.150	1.282	1.440	1.645	1.960
2		1.440	1.534	1.645	1.780	1.960	2.241
3		1.645	1.732	1.834	1.960	2.128	2.394
4		1.780	1.863	1.960	2.080	2.241	2.498
5		1.881	1.960	2.054	2.170	2.326	2.576
6		1.960	2.037	2.128	2.241	2.394	2.638
7		2.026	2.100	2.189	2.300	2.450	2.690
8		2.080	2.154	2.241	2.350	2.498	2.734
9		2.128	2.200	2.287	2.394	2.539	2.773
10		2.170	2.241	2.326	2.432	2.576	2.807
11		2.208	2.278	2.362	2.467	2.608	2.838
12		2.241	2.301	2.394	2.498	2.638	2.866
15		2.326	2.394	2.475	2.576	2.713	2.935
21		2.450	2.515	2.593	2.690	2.823	3.038
28		2.552	2.615	2.690	2.785	2.913	3.125

* # c es el número de comparaciones.

Tabla A_{III}. Valores críticos de $q(\alpha, \# c)$ para las $\# c$ comparaciones múltiples dependientes.*†‡

Las entradas en la tabla para un $\# c$ dado y un nivel de significación α son los valores críticos para los máximos valores absolutos de $\# c$ variables aleatorias normales estándar con correlación 0.5 para la prueba bidireccional y los valores críticos para el lado superior de $\# c$ variables aleatorias normales estándar con correlación común 0.5 para una prueba unidireccional.

		Bidireccional		*Unidireccional*	
$\# c$	α:	0.05	0.01	0.05	0.01
1		1.96	2.58	1.65	2.33
2		2.21	2.79	1.92	2.56
3		2.35	2.92	2.06	2.69
4		2.44	3.00	2.16	2.77
5		2.51	3.06	2.24	2.84
6		2.57	3.11	2.29	2.89
7		2.61	3.15	2.34	2.94
8		2.65	3.19	2.38	2.97
9		2.69	3.22	2.42	3.00
10		2.72	3.25	2.45	3.03
11		2.74	3.27	2.48	3.06
12		2.77	3.29	2.50	3.08
15		2.83	3.35	2.57	3.14
20		2.91	3.42	2.64	3.21

* $\# c$ es el número de comparaciones.

† Las entradas bidireccionales se adaptaron de Dunnett, C. W., "New tables for multiple comparisons with a control, *Biometrics*", núm. 20, 1964, págs. 482-491. (Con autorización del autor y el editor de *Biometrics*.)

‡ Las entradas unidireccionales se adaptaron de Gupta, S. S., "Probability integrals of multivariate normal and multivariate t", en *Annals of Mathematical Statistics*, núm. 34, 1963, págs. 792-828. (Con autorización del autor y de los editores de *Annals of Mathematical Statistics*.)

Tabla B. Valores críticos de la distribución *t* de Student.*

gl	*Nivel de significación para pruebas unidireccionales*					
	0.10	0.05	0.025	0.01	0.005	0.0005
	Nivel de significación para pruebas bidireccionales					
	0.20	0.10	0.05	0.02	0.01	0.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

* La tabla B es una condensación de tabla III de Fisher y Yates, *Statiscal tables for biological, agricultural, and medical research*, Longman Group UK Ltd., Londres (previamente publicada por Oliver y Boyd Ltd., Edimburgo) y con autorización de los autores y los editores.

Tabla C. Valores críticos de la distribución ji cuadrada.*

df	Probabilidad según H_0 de que $\chi^2 \geq X^2$													
	0.99	0.98	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	0.00016	0.00063	0.0039	0.016	0.064	0.15	0.46	1.07	1.64	2.71	3.84	5.41	6.64	10.83
2	0.02	0.04	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	7.82	9.21	13.82
3	0.12	0.18	0.35	0.58	1.00	1.42	2.37	3.66	4.64	6.25	7.82	9.84	11.34	16.27
4	0.30	0.43	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	4.49	11.67	13.28	18.46
5	0.55	0.75	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	13.39	15.09	20.52
6	0.87	1.13	1.64	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	15.03	16.81	22.46
7	1.24	1.56	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	16.62	18.48	24.32
8	1.65	2.03	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	18.17	20.09	26.12
9	2.09	2.53	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	19.68	21.67	27.88
10	2.56	3.06	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	21.16	23.21	29.59
11	3.05	3.61	4.58	5.58	6.99	8.15	10.34	12.90	14.63	17.28	19.68	22.62	24.72	31.26
12	3.57	4.18	5.23	6.30	7.81	9.03	11.34	14.01	15.81	18.55	21.03	24.05	26.22	32.91
13	4.11	4.76	5.89	7.04	8.63	9.93	12.34	15.12	16.98	19.81	22.36	25.47	27.69	34.53
14	4.66	5.37	6.57	7.79	9.47	10.82	13.34	16.22	18.15	21.06	23.68	26.87	29.14	36.12
15	5.23	5.98	7.26	8.55	10.31	11.72	14.34	17.32	19.31	22.31	25.00	28.26	30.58	37.70
16	5.81	6.61	7.96	9.31	11.15	12.62	15.34	18.42	20.46	23.54	26.30	29.63	32.00	39.29
17	6.41	7.26	8.67	10.08	12.00	13.53	16.34	19.51	21.62	24.77	27.59	31.00	33.41	40.75
18	7.02	7.91	9.39	10.86	12.86	14.44	17.34	20.60	22.76	25.99	28.87	32.35	34.80	42.31
19	7.63	8.57	10.12	11.65	13.72	15.35	18.34	21.69	23.90	27.20	30.14	33.69	36.19	43.82
20	8.26	9.24	10.85	12.44	14.58	16.27	19.34	22.78	25.04	28.41	31.41	35.02	37.57	45.32
21	8.90	9.92	11.59	13.24	15.44	17.18	20.34	23.86	26.17	29.62	32.67	36.34	38.93	46.80
22	9.54	10.60	12.34	14.04	16.31	18.10	21.24	24.94	27.30	30.81	33.92	37.66	40.29	48.27
23	10.20	11.29	13.09	14.85	17.19	19.02	22.34	26.02	28.43	32.01	35.17	38.97	41.64	49.73
24	10.86	11.99	13.85	15.66	18.06	19.94	23.34	27.10	29.55	33.20	36.42	40.27	42.98	51.18
25	11.52	12.70	14.61	16.47	18.94	20.87	24.34	28.17	30.68	34.38	37.65	41.57	44.31	52.62
26	12.20	13.41	15.38	17.29	19.82	21.79	25.34	29.25	31.80	35.56	38.88	42.86	45.64	54.05
27	12.88	14.12	16.15	18.11	20.70	22.72	26.34	30.32	32.91	36.74	40.11	44.14	46.96	55.48
28	13.56	14.85	16.93	18.94	21.59	23.65	27.34	31.39	34.03	37.92	41.34	45.42	48.28	56.89
29	14.26	15.57	17.71	19.77	22.48	24.58	28.34	32.46	35.14	39.09	42.56	46.69	49.59	58.80
30	14.95	16.31	18.49	20.60	23.36	25.51	29.34	33.53	36.25	40.26	43.77	47.96	50.89	59.70

* La tabla C es una condensación de la tabla IV de Fischer y Yates, *Statistical tables for biological, agricultural, and medical research*, Longman Group UK Ltd., Londres (previamente publicada por Oliver y Boyd Ltd., Edimburgo) y con autorización de los autores y editores.

Tabla D. Tabla de probabilidades asociadas con valores tan pequeños (o más pequeños) que los valores observados de k en la prueba binomial.

En el cuerpo de la tabla se proporcionan las probabilidades unidireccionales H_0 para la prueba binomial cuando $p = q = \frac{1}{2}$.

Las entradas son $P[Y \leq k]$. Nótese que las entradas también pueden leerse como $P[Y \geq N - k]$.

	k																	
N	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
4	062	312	688	938	1.0													
5	031	188	500	812	969	1.0												
6	016	109	344	656	891	984	1.0											
7	008	062	227	500	773	938	992	1.0										
8	004	035	145	363	637	855	965	996	1.0									
9	002	020	090	254	500	746	910	980	998	1.0								
10	001	011	055	172	377	623	828	945	989	999	1.0							
11		006	033	113	274	500	726	887	967	994	999+	1.0						
12		003	019	073	194	387	613	806	927	981	997	999+	1.0					
13		002	011	046	133	291	500	709	867	954	989	998	999+	1.0				
14		001	006	029	090	212	395	605	788	910	971	994	999	999+	1.0			
15			004	018	059	151	304	500	696	849	941	982	996	999+	999+	1.0		
16			002	011	038	105	227	402	598	773	895	962	989	998	999+	999+	1.0	
17			001	006	025	072	166	315	500	685	834	928	975	994	999	999+	999+	1.0
18			001	004	015	048	119	240	407	593	760	881	952	985	996	999	999+	999+
19				002	010	032	084	180	324	500	676	820	916	968	990	998	999+	999+
20				001	006	021	058	132	252	412	588	748	868	942	979	994	999	999+
21				001	004	013	039	095	192	332	500	668	808	905	961	987	996	999
22					002	008	026	067	143	262	416	584	738	857	933	974	992	998
23					001	005	017	047	105	202	339	500	661	798	895	953	983	995
24					001	003	011	032	076	154	271	419	581	729	846	924	968	989
25						002	007	022	054	115	212	345	500	655	788	885	946	978
26						001	005	014	038	084	163	279	423	577	721	837	916	962
27						001	003	010	026	061	124	221	351	500	649	779	876	939
28							002	006	018	044	092	172	286	425	575	714	828	908
29							001	004	012	031	068	132	229	356	500	644	771	868
30							001	003	008	021	049	100	181	292	428	572	708	819
31								002	005	015	035	075	141	237	360	500	640	763
32								001	004	010	025	055	108	189	298	430	570	702
33								001	002	007	018	040	081	148	243	364	500	636
34									001	005	012	029	061	115	196	304	432	568
35									001	003	008	020	045	088	155	250	368	500

Nota: Los puntos decimales y los valores menores que 0.0005 se omiten.

Tabla E. La distribución binomial.*

$$P[Y = k] = \binom{N}{k} p^k (1 - p)^{N-k}$$

Los puntos decimales se han omitido. Todas las entradas deben leerse como 0.*nnnn*. Para valores de $p \leq 0.5$, úsese la fila superior para p y la columna izquierda para k. Para valores de $p > 0.5$, úsese la fila inferior para p y la columna derecha para k.

		p												
N	k	0.01	0.05	0.10	0.15	0.20	0.25	0.30	1/3	0.40	0.45	0.50		
2	0	9801	9025	8100	7225	6400	5625	4900	4444	3600	3025	2500	2	2
	1	198	950	1800	2550	3200	3750	4200	4444	4800	4950	5000	1	
	2	1	25	100	225	400	625	900	1111	1600	2025	2500	0	
3	0	9703	8574	7290	6141	5120	4219	3430	2963	2160	1664	1250	3	3
	1	294	1354	2430	3251	3840	4219	4410	4444	4320	4084	3750	2	
	2	3	71	270	574	960	1406	1890	2222	2880	3341	3750	1	
	3	0	1	10	34	80	156	270	370	640	911	1250	0	
4	0	9606	8145	6561	5220	4096	3164	2401	1975	1296	915	625	4	4
	1	388	1715	2916	3685	4096	4219	4116	3951	3456	2995	2500	3	
	2	6	135	486	975	1536	2109	2646	2963	3456	3675	3750	2	
	3	0	5	36	115	256	469	756	988	1536	2005	2500	1	
	4	0	0	1	5	16	39	81	123	256	410	625	0	
5	0	9510	7738	5905	4437	3277	2373	1681	1317	778	503	312	5	5
	1	480	2036	3280	3915	4096	3955	3602	3292	2592	2059	1562	4	
	2	10	214	729	1382	2048	2637	3087	3292	3456	3369	3125	3	
	3	0	11	81	244	512	879	1323	1646	2304	2757	3125	2	
	4	0	0	4	22	64	146	283	412	768	1125	1562	1	
	5	0	0	0	1	3	10	24	41	102	185	312	0	
6	0	9415	7351	5314	3771	2621	1780	1176	878	467	277	156	6	6
	1	571	2321	3543	3993	3932	3560	3025	2634	1866	1359	938	5	
	2	14	305	984	1762	2458	2966	3241	3292	3110	2780	2344	4	
	3	0	21	146	415	819	1318	1852	2195	2765	3032	3125	3	
	4	0	1	12	55	154	330	595	823	1382	1861	2344	2	
	5	0	0	1	4	15	44	102	165	369	609	938	1	
	6	0	0	0	0	1	2	7	14	41	83	156	0	
7	0	9321	6983	4783	3206	2097	1335	824	585	280	152	78	7	7
	1	659	2573	3720	3960	3670	3115	2471	2048	1306	872	547	6	
	2	20	406	1240	2097	2753	3115	3177	3073	2613	2140	1641	5	
	3	0	36	230	617	1147	1730	2269	2561	2903	2918	2734	4	
	4	0	2	26	109	287	577	972	1280	1935	2388	2734	3	
	5	0	0	2	12	43	115	250	384	774	1172	1641	2	
	6	0	0	0	1	4	13	36	64	172	320	547	1	
	7	0	0	0	0	0	1	2	5	16	37	78	0	
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	2/3	0.60	0.55	0.50	k	N
		p												

Tabla E. (*Continuación*)

N	k	p = 0.01	0.05	0.10	0.15	0.20	0.25	0.30	1/3	0.40	0.45	0.50		
8	0	9227	6634	4305	2725	1678	1001	576	390	168	84	39	8	8
	1	746	2793	3826	3847	3355	2670	1977	1561	896	548	312	7	
	2	26	515	1488	2376	2936	3115	2965	2731	2090	1569	1094	6	
	3	1	54	331	839	1468	2076	2541	2731	2787	2568	2188	5	
	4	0	4	46	185	459	865	1361	1707	2322	2627	2734	4	
	5	0	0	4	26	92	231	467	683	1239	1719	2188	3	
	6	0	0	0	2	11	38	100	171	413	703	1094	2	
	7	0	0	0	0	1	4	12	24	79	164	312	1	
	8	0	0	0	0	0	0	1	2	7	17	39	0	
9	0	9135	6302	3874	2316	1342	751	404	260	101	46	20	9	9
	1	830	2985	3874	3679	3020	2253	1556	1171	605	339	176	8	
	2	34	629	1722	2597	3020	3003	2668	2341	1612	1110	703	7	
	3	1	77	446	1069	1762	2336	2668	2731	2508	2119	1641	6	
	4	0	6	74	283	661	1168	1715	2048	2508	2600	2461	5	
	5	0	0	8	50	165	389	735	1024	1672	2128	2461	4	
	6	0	0	1	6	28	87	210	341	743	1160	1641	3	
	7	0	0	0	0	3	12	39	73	212	407	703	2	
	8	0	0	0	0	0	1	4	9	35	83	176	1	
	9	0	0	0	0	0	0	0	1	3	8	20	0	
10	0	9044	5987	3487	1969	1074	563	282	173	60	25	10	10	10
	1	914	3151	3874	3474	2684	1877	1211	867	403	207	98	9	
	2	42	746	1937	2759	3020	2816	2335	1951	1209	763	439	8	
	3	1	105	574	1298	2013	2503	2668	2601	2150	1665	1172	7	
	4	0	10	112	401	881	1460	2001	2276	2508	2384	2051	6	
	5	0	1	15	85	264	584	1029	1366	2007	2340	2461	5	
	6	0	0	1	12	55	162	368	569	1115	1596	2051	4	
	7	0	0	0	1	8	31	90	163	425	746	1172	3	
	8	0	0	0	0	1	4	14	30	106	229	439	2	
	9	0	0	0	0	0	0	1	3	16	42	98	1	
	10	0	0	0	0	0	0	0	0	1	3	10	0	
15	0	8601	4633	2059	874	352	134	47	23	5	1	0	15	15
	1	1303	3658	3432	2312	1319	668	305	171	47	16	5	14	
	2	92	1348	2669	2856	2309	1559	916	599	219	90	32	13	
	3	4	307	1285	2184	2501	2252	1700	1299	634	318	139	12	
	4	0	49	428	1156	1876	2252	2186	1948	1268	780	417	11	
	5	0	6	105	449	1032	1651	2061	2143	1859	1404	916	10	
	6	0	0	19	132	430	917	1472	1786	2066	1914	1527	9	
	7	0	0	3	30	138	393	811	1148	1771	2013	1964	8	
	8	0	0	0	5	35	131	348	574	1181	1647	1964	7	
	9	0	0	0	1	7	34	116	223	612	1048	1527	6	
	10	0	0	0	0	1	7	30	67	245	515	916	5	
	11	0	0	0	0	0	1	6	15	74	191	417	4	
	12	0	0	0	0	0	0	1	3	16	52	139	3	
	13	0	0	0	0	0	0	0	0	3	10	32	2	
	14	0	0	0	0	0	0	0	0	0	1	5	1	
	15	0	0	0	0	0	0	0	0	0	0	0	0	
		p = 0.99	0.95	0.90	0.85	0.80	0.75	0.70	2/3	0.60	0.55	0.50	k	N

Tabla E. (*Continuación*)

		p												
N	*k*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	1/3	0.40	0.45	0.50		
20	0	8179	3585	1216	388	115	32	8	3	0	0	0	20	20
	1	1652	3474	2702	1368	576	211	68	30	5	1	0	19	
	2	159	1887	2852	2293	1369	669	278	143	31	8	2	18	
	3	0	596	1901	2428	2054	1339	716	429	123	40	11	17	
	4	0	133	898	1821	2182	1897	1304	911	350	139	46	16	
	5	0	22	319	1028	1746	2023	1789	1457	746	365	148	15	
	6	0	3	89	454	1091	1686	1916	1821	1244	746	370	14	
	7	0	0	20	160	545	1124	1643	1821	1659	1221	739	13	
	8	0	0	4	46	222	609	1144	1480	1797	1623	1201	12	
	9	0	0	1	11	74	271	654	987	1597	1771	1602	11	
	10	0	0	0	2	20	99	308	543	1171	1593	1762	10	
	11	0	0	0	0	5	30	120	247	710	1185	1602	9	
	12	0	0	0	0	1	8	39	92	355	727	1201	8	
	13	0	0	0	0	0	2	10	28	146	366	739	7	
	14	0	0	0	0	0	0	2	7	49	150	370	6	
	15	0	0	0	0	0	0	0	1	13	49	148	5	
	16	0	0	0	0	0	0	0	0	3	13	46	4	
	17	0	0	0	0	0	0	0	0	0	2	11	3	
	18	0	0	0	0	0	0	0	0	0	0	2	2	
	19	0	0	0	0	0	0	0	0	0	0	0	1	
	20	0	0	0	0	0	0	0	0	0	0	0	0	
25	0	7778	2774	718	172	38	8	1	0	0	0	0	25	25
	1	1964	3650	1994	759	236	63	14	5	0	0	0	24	
	2	238	2305	2659	1607	708	251	74	30	4	1	0	23	
	3	18	930	2265	2174	1358	641	243	114	19	4	1	22	
	4	1	269	1384	2110	1867	1175	572	313	71	18	4	21	
	5	0	60	646	1564	1960	1645	1030	658	199	63	16	20	
	6	0	10	239	920	1633	1828	1472	1096	442	172	53	19	
	7	0	1	72	441	1108	1654	1712	1487	800	381	143	18	
	8	0	0	18	175	623	1241	1651	1673	1200	701	322	17	
	9	0	0	4	58	294	781	1336	1580	1511	1084	609	16	
	10	0	0	1	16	118	417	916	1264	1612	1419	974	15	
	11	0	0	0	4	40	189	536	862	1465	1583	1328	14	
	12	0	0	0	1	12	74	268	503	1140	1511	1550	13	
	13	0	0	0	0	3	25	115	251	760	1236	1550	12	
	14	0	0	0	0	1	7	42	108	434	867	1328	11	
	15	0	0	0	0	0	2	13	40	212	520	974	10	
	16	0	0	0	0	0	0	4	12	88	266	609	9	
	17	0	0	0	0	0	0	1	3	31	115	322	8	
	18	0	0	0	0	0	0	0	1	9	42	143	7	
	19	0	0	0	0	0	0	0	0	2	13	53	6	
	20	0	0	0	0	0	0	0	0	0	3	16	5	
	21	0	0	0	0	0	0	0	0	0	1	4	4	
	22	0	0	0	0	0	0	0	0	0	0	1	3	
	23	0	0	0	0	0	0	0	0	0	0	0	2	
	24	0	0	0	0	0	0	0	0	0	0	0	1	
	25	0	0	0	0	0	0	0	0	0	0	0	0	
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	2/3	0.60	0.55	0.50	*k*	*N*
		p												

Tabla E. (*Continuación*)

		p												
N	*k*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	1/3	0.40	0.45	0.50		
30	0	7397	2146	424	76	12	2	0	0	0	0	0	30	30
	1	2242	3389	1413	404	93	18	3	1	0	0	0	29	
	2	328	2586	2277	1034	337	86	18	6	0	0	0	28	
	3	31	1270	2361	1703	785	269	72	26	3	0	0	27	
	4	2	451	1771	2028	1325	604	208	89	12	2	0	26	
	5	0	124	1023	1861	1723	1047	464	232	41	8	1	25	
	6	7	27	474	1368	1795	1455	829	484	115	29	6	24	
	7	0	5	180	828	1538	1662	1219	829	263	81	19	23	
	8	0	1	58	420	1106	1593	1501	1192	505	191	55	22	
	9	0	0	16	181	676	1298	1573	1457	823	382	133	21	
	10	0	0	4	67	355	909	1416	1530	1152	656	280	20	
	11	0	0	1	22	161	551	1103	1391	1396	976	509	19	
	12	0	0	0	6	64	291	749	1101	1474	1265	805	18	
	13	0	0	0	1	22	134	444	762	1360	1433	1115	17	
	14	0	0	0	0	7	54	231	463	1101	1424	1354	16	
	15	0	0	0	0	2	19	106	247	783	1242	1445	15	
	16	0	0	0	0	0	6	42	116	489	953	1354	14	
	17	0	0	0	0	0	2	15	48	269	642	1115	13	
	18	0	0	0	0	0	0	5	17	129	379	805	12	
	19	0	0	0	0	0	0	1	5	54	196	509	11	
	20	0	0	0	0	0	0	0	1	20	88	280	10	
	21	0	0	0	0	0	0	0	0	6	34	133	9	
	22	0	0	0	0	0	0	0	0	1	12	55	8	
	23	0	0	0	0	0	0	0	0	0	3	19	7	
	24	0	0	0	0	0	0	0	0	0	1	6	6	
	25	0	0	0	0	0	0	0	0	0	0	1	5	
	26	0	0	0	0	0	0	0	0	0	0	0	4	
	27	0	0	0	0	0	0	0	0	0	0	0	3	
	28	0	0	0	0	0	0	0	0	0	0	0	2	
	29	0	0	0	0	0	0	0	0	0	0	0	1	
	30	0	0	0	0	0	0	0	0	0	0	0	0	
		0.99	0.95	0.90	0.85	0.80	0.75	0.70	2/3	0.60	0.55	0.50	*k*	*N*
		p												

* Reproducida de Hammond, K. R., Householder, J. E. y Castellan, N. J. Jr., *Introduction to the statistical method*, 2a. ed., A. A. Knopf, Nueva York, 1970, con autorización de los autores y los editores.

Tabla F. Valores críticos de D en la prueba de Kolmogorov-Smirnov para una muestra.*

Tamaño de la muestra (N)	Nivel de significación para D = máximo $\mid F_o(X) - S_N(X) \mid$				
	0.20	0.15	0.10	0.05	0.01
1	0.900	0.925	0.950	0.975	0.995
2	0.684	0.726	0.776	0.842	0.929
3	0.565	0.597	0.642	0.708	0.828
4	0.494	0.525	0.564	0.624	0.733
5	0.446	0.474	0.510	0.565	0.669
6	0.410	0.436	0.470	0.521	0.618
7	0.381	0.405	0.438	0.486	0.577
8	0.358	0.381	0.411	0.457	0.543
9	0.339	0.360	0.388	0.432	0.514
10	0.322	0.342	0.368	0.410	0.490
11	0.307	0.326	0.352	0.391	0.468
12	0.295	0.313	0.338	0.375	0.450
13	0.284	0.302	0.325	0.361	0.433
14	0.274	0.292	0.314	0.349	0.418
15	0.266	0.283	0.304	0.338	0.404
16	0.258	0.274	0.295	0.328	0.392
17	0.250	0.266	0.286	0.318	0.381
18	0.244	0.259	0.278	0.309	0.371
19	0.237	0.252	0.272	0.301	0.363
20	0.231	0.246	0.264	0.294	0.356
25	0.21	0.22	0.24	0.27	0.32
30	0.19	0.20	0.22	0.24	0.29
35	0.18	0.19	0.21	0.23	0.27
Más de 35	$\frac{1.07}{\sqrt{N}}$	$\frac{1.14}{\sqrt{N}}$	$\frac{1.22}{\sqrt{N}}$	$\frac{1.36}{\sqrt{N}}$	$\frac{1.63}{\sqrt{N}}$

* Adaptada de Massey, F. J. Jr., "The Kolmogorov-Smirnov test for goodness of fit", en *Journal of the American Statistical Association*, núm. 46, 1951, pág. 70, por cortesía del autor y los editores.

Tabla G. Valores críticos de r en la prueba de series.*
Los diferentes valores críticos de r están proporcionados en las tablas para valores de m y n menores o iguales a 20. Para la prueba de series de una muestra, cualquier valor observado de r que sea menor o igual al valor más pequeño, o que sea mayor o igual al valor más grande en un par, es significativo en el nivel $\alpha = 0.05$.

m \ n	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2											2	2	2	2	2	2	2	2	2
											-	-	-	-	-	-	-	-	-
3					2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
					-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4				2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4
				9	9	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5			2	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5
			9	10	10	11	11	-	-	-	-	-	-	-	-	-	-	-	-
6		2	2	3	3	3	3	4	4	4	4	5	5	5	5	5	5	6	6
		-	9	10	11	12	12	13	13	13	13	-	-	-	-	-	-	-	-
7		2	2	3	3	3	4	4	5	5	5	5	5	6	6	6	6	6	6
		-	-	11	12	13	13	14	14	14	14	15	15	15	-	-	-	-	-
8		2	3	3	3	4	4	5	5	5	6	6	6	6	6	7	7	7	7
		-	-	11	12	13	14	14	15	15	16	16	16	16	17	17	17	17	17
9		2	3	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8
		-	-	-	13	14	14	15	16	16	16	17	17	18	18	18	18	18	18
10		2	3	3	4	5	5	5	6	6	7	7	7	7	8	8	8	8	9
		-	-	-	13	14	15	16	16	17	17	18	18	18	19	19	19	20	20
11		2	3	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9
		-	-	-	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
	-	-	-	-	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
	-	-	-	-	-	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
	-	-	-	-	-	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
	-	-	-	-	-	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
	-	-	-	-	-	-	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
	-	-	-	-	-	-	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
	-	-	-	-	-	-	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
	-	-	-	-	-	-	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
	-	-	-	-	-	-	17	18	20	21	22	23	24	25	25	26	27	27	28

* Adaptado de Swed y Eisenhart, C., "Tables for testing randomness of grouping in a sequence of alternatives", en *Annals of Mathematical Statistics*, núm. 14, 1943, págs. 83-86, por cortesía de los autores y los editores.

Tabla H. Valores críticos de T^+ para la prueba de rangos asignados de Wilcoxon. Las entradas de la tabla para una N dada es $P[T^+ \geq c]$, la probabilidad de que T^+ sea mayor o igual a la suma de c.

	N						
c	3	4	5	6	7	8	9
3	0.6250						
4	0.3750						
5	0.2500	0.5625					
6	0.1250	0.4375					
7		0.3125					
8		0.1875	0.5000				
9		0.1250	0.4063				
10		0.0625	0.3125				
11			0.2188	0.5000			
12			0.1563	0.4219			
13			0.0938	0.3438			
14			0.0625	0.2813	0.5313		
15			0.0313	0.2188	0.4688		
16				0.1563	0.4063		
17				0.1094	0.3438		
18				0.0781	0.2891	0.5273	
19				0.0469	0.2344	0.4727	
20				0.0313	0.1875	0.4219	
21				0.0156	0.1484	0.3711	
22					0.1094	0.3203	
23					0.0781	0.2734	0.5000
24					0.0547	0.2305	0.4551
25					0.0391	0.1914	0.4102
26					0.0234	0.1563	0.3672
27					0.0156	0.1250	0.3262
28					0.0078	0.0977	0.2852
29						0.0742	0.2480
30						0.0547	0.2129
31						0.0391	0.1797
32						0.0273	0.1504
33						0.0195	0.1250
34						0.0117	0.1016
35						0.0078	0.0820
36						0.0039	0.0645
37							0.0488
38							0.0371
39							0.0273
40							0.0195
41							0.0137
42							0.0098
43							0.0059
44							0.0039
45							0.0020

	N					
c	10	11	12	13	14	15
28	0.5000					
29	0.4609					
30	0.4229					
31	0.3848					
32	0.3477					
33	0.3125	0.5171				
34	0.2783	0.4829				
35	0.2461	0.4492				
36	0.2158	0.4155				
37	0.1875	0.3823				
38	0.1611	0.3501				
39	0.1377	0.3188	0.5151			
40	0.1162	0.2886	0.4849			
41	0.0967	0.2598	0.4548			
42	0.0801	0.2324	0.4250			
43	0.0654	0.2065	0.3955			
44	0.0527	0.1826	0.3667			
45	0.0420	0.1602	0.3386			
46	0.0322	0.1392	0.3110	0.5000		
47	0.0244	0.1201	0.2847	0.4730		
48	0.0186	0.1030	0.2593	0.4463		
49	0.0137	0.0874	0.2349	0.4197		
50	0.0098	0.0737	0.2119	0.3934		
51	0.0068	0.0615	0.1902	0.3677		
52	0.0049	0.0508	0.1697	0.3424		
53	0.0029	0.0415	0.1506	0.3177	0.5000	
54	0.0020	0.0337	0.1331	0.2939	0.4758	
55	0.0010	0.0269	0.1167	0.2709	0.4516	
56		0.0210	0.1018	0.2487	0.4276	
57		0.0161	0.0881	0.2274	0.4039	
58		0.0122	0.0757	0.2072	0.3804	
59		0.0093	0.0647	0.1879	0.3574	
60		0.0068	0.0549	0.1698	0.3349	0.5110
61		0.0049	0.0461	0.1527	0.3129	0.4890
62		0.0034	0.0386	0.1367	0.2915	0.4670
63		0.0024	0.0320	0.1219	0.2708	0.4452
64		0.0015	0.0261	0.1082	0.2508	0.4235
65		0.0010	0.0212	0.0955	0.2316	0.4020
66		0.0005	0.0171	0.0839	0.2131	0.3808
67			0.0134	0.0732	0.1955	0.3599
68			0.0105	0.0636	0.1788	0.3394
69			0.0081	0.0549	0.1629	0.3193
70			0.0061	0.0471	0.1479	0.2997
71			0.0046	0.0402	0.1338	0.2807
72			0.0034	0.0341	0.1206	0.2622
73			0.0024	0.0287	0.1083	0.2444
74			0.0017	0.0239	0.0969	0.2271
75			0.0012	0.0199	0.0863	0.2106
76			0.0007	0.0164	0.0765	0.1947
77			0.0005	0.0133	0.0676	0.1796
78			0.0002	0.0107	0.0594	0.1651

Tabla H. (*Continuación*)

	N		
c	13	14	15
79	0.0085	0.0520	0.1514
80	0.0067	0.0453	0.1384
81	0.0052	0.0392	0.1262
82	0.0040	0.0338	0.1147
83	0.0031	0.0290	0.1039
84	0.0023	0.0247	0.0938
85	0.0017	0.0209	0.0844
86	0.0012	0.0176	0.0757
87	0.0009	0.0148	0.0677
88	0.0006	0.0123	0.0603
89	0.0004	0.0101	0.0535
90	0.0002	0.0083	0.0473
91	0.0001	0.0067	0.0416
92		0.0054	0.0365
93		0.0043	0.0319
94		0.0034	0.0277
95		0.0026	0.0240
96		0.0020	0.0206
97		0.0015	0.0177
98		0.0012	0.0151
99		0.0009	0.0128
100		0.0006	0.0108
101		0.0004	0.0090
102		0.0003	0.0075
103		0.0002	0.0062
104		0.0001	0.0051
105		0.0001	0.0042
106			0.0034
107			0.0027
108			0.0021
109			0.0017
110			0.0013
111			0.0010
112			0.0008
113			0.0006
114			0.0004
115			0.0003
116			0.0002
117			0.0002
118			0.0001
119			0.0001
120			0.0000+

Tabla I. Probabilidades para tablas de cuatro entradas, prueba de Fisher, $N \leq 15$.*

X	-	S_1
-	-	-
S_2	-	N

N es el tamaño total de la muestra, S_1 es el total marginal más pequeño, S_2 es el siguiente más pequeño y X es la frecuencia en la celda correspondiente a los dos totales más pequeños. Para un conjunto dado de N, los valores posibles S_1 y S_2 de X son 0, 1, 2, . . . , S_1. Bajo la línea de cada conjunto está un valor de X tal que $X/S_1 \leq (S_2 - X)/(N - S_1)$, mientras que para valores mayores $X/S_1 > (S_2 - X)/(N - S_1)$. Estos puntos de corte definen direcciones *iguales* y *opuestas* de igualdad de las proporciones en las dos muestras. La probabilidad acumulativa de una desviación tan grande o más grande en la *misma* dirección de la igualdad de las proporciones, está en la columna rotulada "Obs.", mientras que la probabilidad de una desviación tan grande o más grande en la dirección opuesta de la igualdad de las proporciones está en la columna rotulada "Otra". El tamaño de la desviación aquí se mide por el valor absoluto de $X_1/S_1 - (S_2 - X)/(N - S_1)$.

Estas tablas se tomaron de tablas más extensas preparadas por Donald Goyette y M. Ray Mickey, *Health Science Computing Facility*, UCLA.

				Probabilidad							Probabilidad							Probabilidad		
N	S_1	S_2	X	Obs.	Otra	Totales	N	S_1	S_2	X	Obs.	Otra	Totales	N	S_1	S_2	X	Obs.	Otra	Totales
2	1	1	0	0.500	0.500	1.000	7	2	2	0	0.476	0.048	0.524	9	1	1	0	0.889	0.111	1.000
			1	0.500	0.500	1.000				1	0.524	0.476	1.000				1	0.111	0.000	0.111
3	1	1	0	0.667	0.333	1.000				2	0.048	0.000	0.048	9	1	2	0	0.778	0.222	1.000
			1	0.333	0.000	0.333	7	2	3	0	0.286	0.143	0.429				1	0.222	0.000	0.222
4	1	1	0	0.750	0.250	1.000				1	0.714	0.286	1.000	9	1	3	0	0.667	0.333	1.000
			1	0.250	0.000	0.250				2	0.143	0.000	0.143				1	0.333	0.000	0.333
4	1	2	0	0.500	0.500	1.000	7	3	3	0	0.114	0.029	0.143	9	1	4	0	0.556	0.444	1.000
			1	0.500	0.500	1.000				1	0.629	0.371	1.000				1	0.444	0.000	0.444
4	2	2	0	0.167	0.167	0.333				2	0.371	0.114	0.486	9	2	2	0	0.583	0.417	1.000
			1	0.833	0.833	1.000				3	0.029	0.000	0.029				1	0.417	0.000	0.417
			2	0.167	0.167	0.333	8	1	1	0	0.875	0.125	1.000				2	0.028	0.000	0.028
5	1	1	0	0.800	0.200	1.000				1	0.125	0.000	0.125	9	2	3	0	0.417	0.083	0.500
			1	0.200	0.000	0.200	8	1	2	0	0.750	0.250	1.000				1	0.583	0.417	1.000
5	1	2	0	0.600	0.400	1.000				1	0.250	0.000	0.250				2	0.083	0.000	0.083
			1	0.400	0.000	0.400	8	1	3	0	0.625	0.375	1.000	9	2	4	0	0.278	0.167	0.444
5	2	2	0	0.300	0.100	0.400				1	0.375	0.000	0.375				1	0.722	0.278	1.000
			1	0.700	0.300	1.000	8	1	4	0	0.500	0.500	1.000				2	0.167	0.000	0.167
			2	0.100	0.000	0.100				1	0.500	0.500	1.000	9	3	3	0	0.238	0.226	0.464
6	1	1	0	0.833	0.167	1.000	8	2	2	0	0.536	0.464	1.000				1	0.774	0.774	1.000
			1	0.167	0.000	0.167				1	0.464	0.536	1.000				2	0.226	0.238	0.464
6	1	2	0	0.667	0.333	1.000				2	0.036	0.000	0.036				3	0.012	0.000	0.012
			1	0.333	0.000	0.333	8	2	3	0	0.357	0.107	0.464	9	3	4	0	0.119	0.048	0.167
6	1	3	0	0.500	0.500	1.000				1	0.643	0.357	1.000				1	0.595	0.450	1.000
			1	0.500	0.500	1.000				2	0.107	0.000	0.107				2	0.405	0.119	0.524
6	2	2	0	0.400	0.067	0.467	8	2	4	0	0.214	0.214	0.429				3	0.048	0.000	0.048
			1	0.600	0.400	1.000				1	0.786	0.786	1.000	9	4	4	0	0.040	0.008	0.048
			2	0.067	0.000	0.067				2	0.214	0.214	0.429				1	0.357	0.167	0.524
6	2	3	0	0.200	0.200	0.400	8	3	3	0	0.179	0.018	0.196				2	0.643	0.357	1.000
			1	0.800	0.800	1.000				1	0.714	0.286	1.000				3	0.167	0.040	0.206
			2	0.200	0.200	0.400				2	0.286	0.179	0.464				4	0.008	0.000	0.008
6	3	3	0	0.050	0.050	0.100				3	0.018	0.000	0.018	10	1	1	0	0.900	0.100	1.000
			1	0.500	0.500	1.000	8	3	4	0	0.071	0.071	0.143				1	0.100	0.000	0.100
			2	0.500	0.500	1.000				1	0.500	0.500	1.000	10	1	2	0	0.800	0.200	1.000
			3	0.050	0.050	0.100				2	0.500	0.500	1.000				1	0.200	0.000	0.200
7	1	1	0	0.857	0.143	1.000				3	0.071	0.071	0.143	10	1	3	0	0.700	0.300	1.000
			1	0.143	0.000	0.143	8	4	4	0	0.014	0.014	0.029				1	0.300	0.000	0.300
7	1	2	0	0.714	0.286	1.000				1	0.243	0.243	0.486	10	1	4	0	0.600	0.400	1.000
			1	0.286	0.000	0.286				2	0.757	0.757	1.000				1	0.400	0.000	0.400
7	1	3	0	0.571	0.429	1.000				3	0.243	0.243	0.486	10	1	5	0	0.500	0.500	1.000
			1	0.429	0.000	0.429				4	0.014	0.014	0.029				1	0.500	0.500	1.000

* Reproducida de la tabla A-9e en Dixon, W. J. y Massey, F. J. Jr., *Introduction to a statistical analysis*, 4a. ed., McGraw-Hill, Nueva York, 1983. Con autorización de los editores. Estamos también agradecidos con el doctor M. R. Mickey y con la UCLA por la autorización para reproducir estas tablas.

Tabla I. (*Continuación*)

N	S1	S2	X	Probabilidad Obs.	Otra	Totales	N	S1	S2	X	Probabilidad Obs.	Otra	Totales	N	S1	S2	X	Probabilidad Obs.	Otra	Totales
10	2	2	0	0.622	0.378	1.000	11	3	4	0	0.212	0.024	0.236				3	0.091	0.091	0.182
			1	0.378	0.000	0.378				1	0.721	0.279	1.000	12	4	4	0	0.141	0.067	0.208
			2	0.022	0.000	0.022				2	0.279	0.212	0.491				1	0.594	0.406	1.000
10	2	3	0	0.467	0.067	0.533				3	0.024	0.000	0.024				2	0.406	0.141	0.547
			1	0.533	0.467	1.000	11	3	5	0	0.121	0.061	0.182				3	0.067	0.000	0.067
			2	0.067	0.000	0.067				1	0.576	0.424	1.000				4	0.002	0.000	0.002
10	2	4	0	0.333	0.133	0.467				2	0.424	0.121	0.545	12	4	5	0	0.071	0.010	0.081
			1	0.667	0.333	1.000				3	0.061	0.000	0.061				1	0.424	0.152	0.576
			2	0.133	0.000	0.133	11	4	4	0	0.106	0.088	0.194				2	0.576	0.424	1.000
10	2	5	0	0.222	0.222	0.444				1	0.530	0.470	1.000				3	0.152	0.171	0.222
			1	0.778	0.778	1.000				2	0.470	0.106	0.576				4	0.010	0.000	0.010
			2	0.222	0.222	0.444				3	0.088	0.000	0.088	12	4	6	0	0.030	0.030	0.061
10	3	3	0	0.292	0.183	0.475				4	0.003	0.000	0.003				1	0.273	0.273	0.545
			1	0.708	0.292	1.000	11	4	5	0	0.045	0.015	0.061				2	0.727	0.727	1.000
			2	0.183	0.000	0.183				1	0.348	0.197	0.545				3	0.273	0.273	0.545
			3	0.008	0.000	0.008				2	0.652	0.348	1.000				4	0.030	0.030	0.061
10	3	4	0	0.167	0.033	0.200				3	0.197	0.045	0.242	12	5	5	0	0.027	0.001	0.028
			1	0.667	0.333	1.000				4	0.015	0.000	0.015				12	0.247	0.045	0.293
			2	0.333	0.167	0.500	11	5	5	0	0.013	0.002	0.015				2	0.689	0.311	1.000
			3	0.033	0.000	0.033				1	0.175	0.067	0.242				3	0.311	0.247	0.558
10	3	5	0	0.083	0.083	0.167				2	0.608	0.392	1.000				4	0.045	0.027	0.072
			1	0.500	0.500	1.000				3	0.392	0.175	0.567				5	0.001	0.000	0.001
			2	0.500	0.500	1.000				4	0.067	0.013	0.080	12	5	6	0	0.008	0.008	0.015
			3	0.083	0.083	0.167				5	0.002	0.000	0.002				1	0.121	0.121	0.242
10	4	4	0	0.071	0.005	0.076	12	1	1	0	0.917	0.083	1.000				2	0.500	0.500	1.000
			1	0.452	0.119	0.571				1	0.083	0.000	0.083				3	0.500	0.500	1.000
			2	0.548	0.452	1.000	12	1	2	0	0.833	0.167	1.000				4	0.121	0.121	0.242
			3	0.119	0.071	0.190				1	0.167	0.000	0.167				5	0.008	0.008	0.015
			4	0.005	0.000	0.005	12	1	3	0	0.750	0.250	1.000	12	6	6	0	0.001	0.001	0.002
10	4	5	0	0.024	0.024	0.048				1	0.250	0.000	0.250				1	0.040	0.040	0.080
			1	0.262	0.262	0.524	12	1	4	0	0.667	0.333	1.000				2	0.284	0.284	0.567
			2	0.738	0.738	1.000				1	0.333	0.000	0.333				3	0.716	0.716	1.000
			3	0.262	0.262	0.524	12	1	5	0	0.583	0.417	1.000				4	0.284	0.284	0.567
			4	0.024	0.024	0.048				1	0.417	0.000	0.471				5	0.040	0.040	0.080
10	5	5	0	0.004	0.004	0.008	12	1	6	0	0.500	0.500	1.000				6	0.001	0.001	0.002
			1	0.103	0.103	0.206				1	0.500	0.500	1.000	13	1	1	0	0.923	0.077	1.000
			2	0.500	0.500	1.000	12	2	2	0	0.682	0.318	1.000				1	0.077	0.000	0.077
			3	0.500	0.500	1.000				1	0.318	0.000	0.318	13	1	2	0	0.846	0.154	1.000
			4	0.103	0.103	0.206				2	0.015	0.000	0.015				1	0.154	0.000	0.154
			5	0.004	0.004	0.008	12	2	3	0	0.545	0.455	1.000	13	1	3	0	0.769	0.231	1.000
11	1	1	0	0.909	0.091	1.000				1	0.455	0.545	1.000				1	0.231	0.000	0.231
			1	0.091	0.000	0.091				2	0.045	0.000	0.045	13	1	4	0	0.692	0.308	1.000
11	1	2	0	0.818	0.182	1.000	12	2	4	0	0.424	0.091	0.515				1	0.308	0.000	0.308
			1	0.182	0.000	0.182				1	0.576	0.424	1.000	13	1	5	0	0.615	0.385	1.000
11	1	3	0	0.727	0.273	1.000				2	0.091	0.000	0.091				1	0.385	0.000	0.385
			1	0.273	0.000	0.273	12	2	5	0	0.318	0.152	0.470	13	1	6	0	0.538	0.462	1.000
11	1	4	0	0.636	0.364	1.000				1	0.682	0.318	1.000				1	0.462	0.000	0.462
			1	0.364	0.000	0.364				2	0.152	0.000	0.152	13	2	2	0	0.705	0.295	1.000
11	1	5	0	0.545	0.455	1.000	12	2	6	0	0.227	0.227	0.455				1	0.295	0.000	0.295
			1	0.455	0.000	0.455				1	0.773	0.773	1.000				2	0.013	0.000	0.013
11	2	2	0	0.655	0.345	1.000				2	0.227	0.227	0.455	13	2	3	0	0.577	0.423	1.000
			1	0.345	0.000	0.345	12	3	3	0	0.382	0.127	0.509				1	0.423	0.000	0.423
			2	0.018	0.000	0.018				1	0.618	0.382	1.000				2	0.038	0.000	0.038
11	2	3	0	0.509	0.055	0.564				2	0.127	0.000	0.127	13	2	4	0	0.462	0.077	0.538
			1	0.491	0.509	1.000				3	0.005	0.000	0.005				1	0.538	0.462	1.000
			2	0.055	0.000	0.055	12	3	4	0	0.255	0.236	0.491				2	0.077	0.000	0.077
11	2	4	0	0.382	0.109	0.491				1	0.764	0.764	1.000	13	2	5	0	0.359	0.128	0.487
			1	0.618	0.382	1.000				2	0.236	0.255	0.491				1	0.641	0.359	1.000
			2	0.109	0.000	0.109				3	0.018	0.000	0.018				2	0.128	0.000	0.128
11	2	5	0	0.273	0.182	0.455	12	3	5	0	0.159	0.045	0.205	13	2	6	0	0.269	0.192	0.462
			1	0.727	0.273	1.000				1	0.636	0.364	1.000				1	0.731	0.269	1.000
			2	0.182	0.000	0.182				2	0.364	0.159	0.523				2	0.192	0.000	0.192
11	3	3	0	0.339	0.152	0.491				3	0.045	0.000	0.045	13	3	3	0	0.420	0.108	0.528
			1	0.661	0.339	1.000	12	3	6	0	0.091	0.091	0.182				1	0.580	0.420	0.100
			2	0.152	0.000	0.152				1	0.500	0.500	1.000				2	0.108	0.000	0.108
			3	0.006	0.000	0.006				2	0.500	0.500	1.000				3	0.003	0.000	0.003

N	S_1	S_2	X	Probabilidad Obs.	Probabilidad Otra	Probabilidad Totales
13	3	4	0	0.294	0.203	0.497
			1	0.706	0.294	1.000
			2	0.203	0.000	0.203
			3	0.014	0.000	0.014
13	3	5	0	0.196	0.035	0.231
			1	0.685	0.315	1.000
			2	0.315	0.196	0.510
			3	0.035	0.000	0.035
13	3	6	0	0.122	0.070	0.192
			1	0.563	0.437	1.000
			2	0.437	0.122	0.559
			3	0.070	0.000	0.070
13	4	4	0	0.176	0.052	0.228
			1	0.646	0.354	1.000
			2	0.354	0.176	0.530
			3	0.052	0.000	0.052
			4	0.001	0.000	0.001
13	4	5	0	0.098	0.007	0.105
			1	0.490	0.119	0.608
			2	0.510	0.490	1.000
			3	0.119	0.098	0.217
			4	0.007	0.000	0.007
13	4	6	0	0.049	0.021	0.070
			1	0.343	0.217	0.559
			2	0.657	0.343	1.000
			3	0.217	0.049	0.266
			4	0.021	0.000	0.021
13	5	5	0	0.044	0.032	0.075
			1	0.315	0.249	0.565
			2	0.685	0.315	1.000
			3	0.249	0.044	0.293
			4	0.032	0.000	0.032
			5	0.001	0.000	0.001
13	5	6	0	0.016	0.005	0.021
			1	0.179	0.086	0.266
			2	0.587	0.413	1.000
			3	0.413	0.179	0.592
			4	0.086	0.016	0.103
			5	0.005	0.000	0.005
13	6	6	0	0.004	0.001	0.005
			1	0.078	0.025	0.103
			2	0.383	0.209	0.592
			3	0.617	0.383	1.000
			4	0.209	0.078	0.286
			5	0.025	0.004	0.029
			6	0.001	0.000	0.001
14	1	1	0	0.929	0.071	1.000
			1	0.071	0.000	0.071
14	1	2	0	0.857	0.143	1.000
			1	0.143	0.000	0.143
14	1	3	0	0.786	0.214	1.000
			1	0.214	0.000	0.214
14	1	4	0	0.714	0.286	1.000
			1	0.286	0.000	0.286
14	1	5	0	0.643	0.357	1.000
			1	0.357	0.000	0.357
14	1	6	0	0.571	0.429	1.000
			1	0.429	0.000	0.429
14	1	7	0	0.500	0.500	1.000
			1	0.500	0.500	1.000
14	2	2	0	0.725	0.275	1.000
			1	0.275	0.000	0.275
			2	0.011	0.000	0.011
14	2	3	0	0.604	0.396	1.000
			1	0.396	0.000	0.396
			2	0.033	0.000	0.033
14	2	4	0	0.495	0.066	0.560
			1	0.505	0.495	1.000
			2	0.066	0.000	0.066
14	2	5	0	0.396	0.110	0.505
			1	0.604	0.396	1.000
			2	0.110	0.000	0.110
14	2	6	0	0.308	0.165	0.473
			1	0.692	0.308	1.000
			2	0.165	0.000	0.165
14	2	7	0	0.231	0.231	0.462
			1	0.769	0.769	1.000
			1	0.231	0.231	0.462
14	3	3	0	0.453	0.093	0.547
			1	0.547	0.453	1.000
			2	0.093	0.000	0.093
			3	0.003	0.000	0.003
14	3	4	0	0.330	0.176	0.505
			1	0.670	0.330	1.000
			2	0.176	0.000	0.176
			3	0.011	0.000	0.011
14	3	5	0	0.231	0.027	0.258
			1	0.725	0.275	1.000
			2	0.275	0.231	0.505
			3	0.027	0.000	0.027
14	3	6	0	0.154	0.055	0.209
			1	0.615	0.385	1.000
			2	0.385	0.154	0.538
			3	0.055	0.000	0.055
14	3	7	0	0.096	0.096	0.192
			1	0.500	0.500	1.000
			2	0.500	0.500	1.000
			3	0.096	0.096	0.192
14	4	4	0	0.210	0.041	0.251
			1	0.689	0.311	1.000
			2	0.311	0.210	0.520
			3	0.041	0.000	0.041
			4	0.001	0.000	0.001
14	4	5	0	0.126	0.095	0.221
			1	0.545	0.455	1.000
			2	0.455	0.126	0.580
			3	0.095	0.000	0.095
			4	0.005	0.000	0.005
14	4	6	0	0.070	0.015	0.085
			1	0.406	0.175	0.580
			2	0.594	0.406	1.000
			3	0.175	0.070	0.245
			4	0.015	0.000	0.015
14	4	7	0	0.035	0.035	0.070
			1	0.280	0.280	0.559
			2	0.720	0.720	1.000
			3	0.280	0.280	0.559
			4	0.035	0.035	0.070
14	5	5	0	0.063	0.023	0.086
			1	0.378	0.203	0.580
			2	0.622	0.378	1.000
			3	0.203	0.063	0.266
			4	0.023	0.000	0.023
			5	0.000	0.000	0.000
14	5	6	0	0.028	0.003	0.031
			1	0.238	0.063	0.301
			2	0.657	0.343	1.000
			3	0.343	0.238	0.580
			4	0.063	0.028	0.091
			5	0.003	0.000	0.003
14	5	7	0	0.010	0.010	0.021
			1	0.133	0.133	0.266
			2	0.500	0.500	1.000
			3	0.500	0.500	1.000
			4	0.133	0.133	0.266
			5	0.010	0.010	0.021
14	6	6	0	0.009	0.000	0.010
			1	0.121	0.016	0.138
			2	0.471	0.156	0.627
			3	0.529	0.471	1.000
			4	0.156	0.121	0.277
			5	0.016	0.009	0.026
			6	0.000	0.000	0.000
14	6	7	0	0.002	0.002	0.005
			1	0.051	0.051	0.103
			2	0.296	0.296	0.592
			3	0.704	0.704	1.000
			4	0.296	0.296	0.592
			5	0.051	0.051	0.103
			6	0.002	0.002	0.005
14	7	7	0	0.000	0.000	0.001
			1	0.015	0.015	0.029
			2	0.143	0.143	0.286
			3	0.500	0.500	1.000
			4	0.500	0.500	0.100
			5	0.143	0.143	0.286
			6	0.015	0.015	0.029
			7	0.000	0.000	0.001
15	1	1	0	0.933	0.067	1.000
			1	0.067	0.000	0.067
15	1	2	0	0.867	0.133	1.000
			1	0.133	0.000	0.133
15	1	3	0	0.800	0.200	1.000
			1	0.200	0.000	0.200
15	1	4	0	0.733	0.267	1.000
			1	0.267	0.000	0.267
15	1	5	0	0.667	0.333	1.000
			1	0.333	0.000	0.333
15	1	6	0	0.600	0.400	1.000
			1	0.400	0.000	0.400
15	1	7	0	0.533	0.467	1.000
			1	0.467	0.000	0.467
15	2	2	0	0.743	0.257	1.000
			1	0.257	0.000	0.257
			2	0.010	0.000	0.010
15	2	3	0	0.629	0.371	1.000
			1	0.371	0.000	0.371
			2	0.029	0.000	0.029
15	2	4	0	0.524	0.057	0.581
			1	0.476	0.524	1.000
			2	0.057	0.000	0.057
15	2	5	0	0.429	0.095	0.524
			1	0.571	0.429	1.000
			2	0.095	0.000	0.095
15	2	6	0	0.343	0.143	0.486
			1	0.657	0.343	1.000
			2	0.143	0.000	0.143
15	2	7	0	0.267	0.200	0.467
			1	0.733	0.267	1.000
			2	0.200	0.000	0.200
15	3	3	0	0.484	0.081	0.565
			1	0.516	0.484	1.000
			2	0.081	0.000	0.081
			3	0.002	0.000	0.002
15	3	4	0	0.363	0.154	0.516
			1	0.637	0.363	1.000
			2	0.154	0.000	0.154
			3	0.009	0.000	0.009

Tabla I. (*Continuación*)

				Probabilidad							Probabilidad							Probabilidad		
N	S_1	S_2	X	*Obs.*	*Otra*	*Totales*	N	S_1	S_2	X	*Obs.*	*Otra*	*Totales*	N	S_1	S_2	X	*Obs.*	*Otra*	*Totales*
15	3	5	0	0.264	0.242	0.505				2	0.538	0.462	1.000				4	0.100	0.019	0.119
			1	0.758	0.758	1.000				3	0.143	0.092	0.235				5	0.007	0.000	0.007
			2	0.242	0.264	0.505				4	0.011	0.000	0.011	15	6	6	0	0.017	0.011	0.028
			3	0.022	0.000	0.022	15	4	7	0	0.051	0.026	0.077				1	0.168	0.119	0.287
15	3	6	0	0.185	0.044	0.229				1	0.338	0.231	0.569				2	0.545	0.455	1.000
			1	0.659	0.341	1.000				2	0.662	0.338	1.000				3	0.455	0.168	0.622
			2	0.341	0.185	0.525				3	0.231	0.051	0.282				4	0.119	0.017	0.136
			3	0.044	0.000	0.044				4	0.026	0.000	0.026				5	0.011	0.000	0.011
15	3	7	0	0.123	0.077	0.200	15	5	5	0	0.084	0.017	0.101				6	0.000	0.000	0.000
			1	0.554	0.446	1.000				1	0.434	0.167	0.600	15	6	7	0	0.006	0.001	0.007
			2	0.446	0.123	0.569				2	0.566	0.434	1.000				1	0.084	0.035	0.119
			3	0.077	0.000	0.077				3	0.167	0.084	0.251				2	0.378	0.231	0.608
15	4	4	0	0.242	0.033	0.275				4	0.017	0.000	0.017				3	0.622	0.378	1.000
			1	0.725	0.275	1.000				5	0.000	0.000	0.000				4	0.231	0.084	0.315
			2	0.275	0.242	0.516	15	5	6	0	0.042	0.047	0.089				5	0.035	0.006	0.041
			3	0.033	0.000	0.033				1	0.294	0.287	0.580				6	0.001	0.000	0.001
			4	0.001	0.000	0.001				2	0.713	0.713	1.000	15	7	7	0	0.001	0.000	0.001
15	4	5	0	0.154	0.077	0.231				3	0.287	0.294	0.580				1	0.032	0.009	0.041
			1	0.593	0.407	1.000				4	0.047	0.042	0.089				2	0.214	0.100	0.315
			2	0.407	0.154	0.560				5	0.002	0.000	0.002				3	0.595	0.405	1.000
			3	0.077	0.000	0.077	15	5	7	0	0.019	0.007	0.026				4	0.405	0.214	0.619
			4	0.004	0.000	0.004				1	0.182	0.100	0.282				5	0.100	0.032	0.132
15	4	6	0	0.092	0.011	0.103				2	0.573	0.427	1.000				6	0.009	0.001	0.010
			1	0.462	0.143	0.604				3	0.427	0.182	0.608				7	0.000	0.000	0.000

Tabla J. Probabilidades de los lados inferior y superior para W_x, el estadístico de la suma de rangos de Wilcoxon-Mann-Whitney.

Las entradas son $P\,[W_x \leq c_L]$ y $P[W_x \leq c_U]$. W_x es la suma de rangos para el grupo más pequeño.

$m = 3$																				
c_L	$n = 3$	c_U	$n = 4$	c_U	$n = 5$	c_U	$n = 6$	c_U	$n = 7$	c_U	$n = 8$	c_U	$n = 9$	c_U	$n = 10$	c_U	$n = 11$	c_U	$n = 12$	c_U
6	0.0500	15	0.0286	18	0.0179	21	0.0119	24	0.0083	27	0.0061	30	0.0045	33	0.0035	36	0.0027	39	0.0022	42
7	0.1000	14	0.0571	17	0.0357	20	0.0238	23	0.0167	26	0.0121	29	0.0091	32	0.0070	35	0.0055	38	0.0044	41
8	0.2000	13	0.1143	16	0.0714	19	0.0476	22	0.0333	25	0.0242	28	0.0182	31	0.0140	34	0.0110	37	0.0088	40
9	0.3500	12	0.2000	15	0.1250	18	0.0833	21	0.0583	24	0.0424	27	0.0318	30	0.0245	33	0.0192	36	0.0154	39
10	0.5000	11	0.3143	14	0.1964	17	0.1310	20	0.0917	23	0.0667	26	0.0500	29	0.0385	32	0.0302	35	0.0242	38
11	0.6500	10	0.4286	13	0.2857	16	0.1905	19	0.1333	22	0.0970	25	0.0727	28	0.0559	31	0.0440	34	0.0352	37
12	0.8000	9	0.5714	12	0.3929	15	0.2738	18	0.1917	21	0.1394	24	0.1045	27	0.0804	30	0.0632	33	0.0505	36
13	0.9000	8	0.6857	11	0.5000	14	0.3571	17	0.2583	20	0.1879	23	0.1409	26	0.1084	29	0.0852	32	0.0681	35
14	0.9500	7	0.8000	10	0.6071	13	0.4524	16	0.3333	19	0.2485	22	0.1864	25	0.1434	28	0.1126	31	0.0901	34
15	1.0000	6	0.8857	9	0.7143	12	0.5476	15	0.4167	18	0.3152	21	0.2409	24	0.1853	27	0.1456	30	0.1165	33
16			0.9429	8	0.8036	11	0.6429	14	0.5000	17	0.3879	20	0.3000	23	0.2343	26	0.1841	29	0.1473	32
17			0.9714	7	0.8750	10	0.7262	13	0.5833	16	0.4606	19	0.3636	22	0.2867	25	0.2280	28	0.1824	31
18			1.0000	6	0.9286	9	0.8095	12	0.6667	15	0.5394	18	0.4318	21	0.3462	24	0.2775	27	0.2242	30
19					0.9643	8	0.8690	11	0.7417	14	0.6121	17	0.5000	20	0.4056	23	0.3297	26	0.2681	29
20					0.9821	7	0.9167	10	0.8083	13	0.6848	16	0.5682	19	0.4685	22	0.3846	25	0.3165	28
21					1.0000	6	0.9524	9	0.8667	12	0.7515	15	0.6364	18	0.5315	21	0.4423	24	0.3670	27
22							0.9762	8	0.9083	11	0.8121	14	0.7000	17	0.5944	20	0.5000	23	0.4198	26
23							0.9881	7	0.9417	10	0.8606	13	0.7591	16	0.6538	19	0.5577	22	0.4725	25
24							1.0000	6	0.9667	9	0.9030	12	0.8136	15	0.7133	18	0.6154	21	0.5275	24

Tabla J. *(Continuación)*

								$m = 4$									
c_L	$n = 4$	c_U	$n = 5$	c_U	$n = 6$	c_U	$n = 7$	c_U	$n = 8$	c_U	$n = 9$	c_U	$n = 10$	c_U	$n = 11$	c_U	$n = 12$	c_U
10	0.0143	26	0.0079	30	0.0048	34	0.0030	38	0.0020	42	0.0014	46	0.0010	50	0.0007	54	0.0005	58
11	0.0286	25	0.0159	29	0.0095	33	0.0061	37	0.0040	41	0.0028	45	0.0020	49	0.0015	53	0.0011	57
12	0.0571	24	0.0317	28	0.0190	32	0.0121	36	0.0081	40	0.0056	44	0.0040	48	0.0029	52	0.0022	56
13	0.1000	23	0.0556	27	0.0333	31	0.0212	35	0.0141	39	0.0098	43	0.0070	47	0.0051	51	0.0038	55
14	0.1714	22	0.0952	26	0.0571	30	0.0364	34	0.0242	38	0.0168	42	0.0120	46	0.0088	50	0.0066	54
15	0.2429	21	0.1429	25	0.0857	29	0.0545	33	0.0364	37	0.0252	41	0.0180	45	0.0132	49	0.0099	53
16	0.3429	20	0.2063	24	0.1286	28	0.0818	32	0.0545	36	0.0378	40	0.0270	44	0.0198	48	0.0148	52
17	0.4429	19	0.2778	23	0.1762	27	0.1152	31	0.0768	35	0.0531	39	0.0380	43	0.0278	47	0.0209	51
18	0.5571	18	0.3651	22	0.2381	26	0.1576	30	0.1071	34	0.0741	38	0.0529	42	0.0388	46	0.0291	50
19	0.6571	17	0.4524	21	0.3048	25	0.2061	29	0.1414	33	0.0993	37	0.0709	41	0.0520	45	0.0390	49
20	0.7571	16	0.5476	20	0.3810	24	0.2636	28	0.1838	32	0.1301	36	0.0939	40	0.0689	44	0.0516	48
21	0.8286	15	0.6349	19	0.4571	23	0.3242	27	0.2303	31	0.1650	35	0.1199	39	0.0886	43	0.0665	47
22	0.9000	14	0.7222	18	0.5429	22	0.3939	26	0.2848	30	0.2070	34	0.1518	38	0.1128	42	0.0852	46
23	0.9429	13	0.7937	17	0.6190	21	0.4636	25	0.3414	29	0.2517	33	0.1868	37	0.1399	41	0.1060	45
24	0.9714	12	0.8571	16	0.6952	20	0.5364	24	0.4040	28	0.3021	32	0.2268	36	0.1714	40	0.1308	44
25	0.9857	11	0.9048	15	0.7619	19	0.6061	23	0.4667	27	0.3552	31	0.2697	35	0.2059	39	0.1582	43
26	1.0000	10	0.9444	14	0.8238	18	0.6758	22	0.5333	26	0.4126	30	0.3177	34	0.2447	38	0.1896	42
27			0.9683	13	0.8714	17	0.7364	21	0.5960	25	0.4699	29	0.3666	33	0.2857	37	0.2231	41
28			0.9841	12	0.9143	16	0.7939	20	0.6586	24	0.5301	28	0.4196	32	0.3304	36	0.2604	40
29			0.9921	11	0.9429	15	0.8424	19	0.7152	23	0.5874	27	0.4725	31	0.3766	35	0.2995	39
30			1.0000	10	0.9667	14	0.8848	18	0.7697	22	0.6448	26	0.5275	30	0.4256	34	0.3418	38
31					0.9810	13	0.9182	17	0.8162	21	0.6979	25	0.5804	29	0.4747	33	0.3852	37
32					0.9905	12	0.9455	16	0.8586	20	0.7483	24	0.6334	28	0.5253	32	0.4308	36
33					0.9952	11	0.9636	15	0.8929	19	0.7930	23	0.6823	27	0.5744	31	0.4764	35
34					1.0000	10	0.9788	14	0.9232	18	0.8350	22	0.7303	26	0.6234	30	0.5236	34

						m = 5						
c_L	$n = 5$	c_U	$n = 6$	c_U	$n = 7$	c_U	$n = 8$	c_U	$n = 9$	c_U	$n = 10$	c_U
15	0.0040	40	0.0022	45	0.0013	50	0.0008	55	0.0005	60	0.0003	65
16	0.0079	39	0.0043	44	0.0025	49	0.0016	54	0.0010	59	0.0007	64
17	0.0159	38	0.0087	43	0.0051	48	0.0031	53	0.0020	58	0.0013	63
18	0.0278	37	0.0152	42	0.0088	47	0.0054	52	0.0035	57	0.0023	62
19	0.0476	36	0.0260	41	0.0152	46	0.0093	51	0.0060	56	0.0040	61
20	0.0754	35	0.0411	40	0.0240	45	0.0148	50	0.0095	55	0.0063	60
21	0.1111	34	0.0628	39	0.0366	44	0.0225	49	0.0145	54	0.0097	59
22	0.1548	33	0.0887	38	0.0530	43	0.0326	48	0.0210	53	0.0140	58
23	0.2103	32	0.1234	37	0.0745	42	0.0466	47	0.0300	52	0.0200	57
24	0.2738	31	0.1645	36	0.1010	41	0.0637	46	0.0415	51	0.0276	56
25	0.3452	30	0.2143	35	0.1338	40	0.0855	45	0.0559	50	0.0376	55
26	0.4206	29	0.2684	34	0.1717	39	0.1111	44	0.0734	49	0.0496	54
27	0.5000	28	0.3312	33	0.2159	38	0.1422	43	0.0949	48	0.0646	53
28	0.5794	27	0.3961	32	0.2652	37	0.1772	42	0.1199	47	0.0823	52
29	0.6548	26	0.4654	31	0.3194	36	0.2176	41	0.1489	46	0.1032	51
30	0.7262	25	0.5346	30	0.3775	35	0.2618	40	0.1818	45	0.1272	50
31	0.7897	24	0.6039	29	0.4381	34	0.3108	39	0.2188	44	0.1548	49
32	0.8452	23	0.6688	28	0.5000	33	0.3621	38	0.2592	43	0.1855	48
33	0.8889	22	0.7316	27	0.5619	32	0.4165	37	0.3032	42	0.2198	47
34	0.9246	21	0.7857	26	0.6225	31	0.4716	36	0.3497	41	0.2567	46
35	0.9524	20	0.8355	25	0.6806	30	0.5284	35	0.3986	40	0.2970	45
36	0.9722	19	0.8766	24	0.7348	29	0.5835	34	0.4491	39	0.3393	44
37	0.9841	18	0.9113	23	0.7841	28	0.6379	33	0.5000	38	0.3839	43
38	0.9921	17	0.9372	22	0.8283	27	0.6892	32	0.5509	37	0.4296	42
39	0.9960	16	0.9589	21	0.8662	26	0.7382	31	0.6014	36	0.4765	41
40	1.0000	15	0.9740	20	0.8990	25	0.7824	30	0.6503	35	0.5235	40

Tabla J. (*Continuación*)

$m = 6$										
c_L	$n = 6$	c_U	$n = 7$	c_U	$n = 8$	c_U	$n = 9$	c_U	$n = 10$	c_U
21	0.0011	57	0.0006	63	0.0003	69	0.0002	75	0.0001	81
22	0.0022	56	0.0012	62	0.0007	68	0.0004	74	0.0002	80
23	0.0043	55	0.0023	61	0.0013	67	0.0008	73	0.0005	79
24	0.0076	54	0.0041	60	0.0023	66	0.0014	72	0.0009	78
25	0.0130	53	0.0070	59	0.0040	65	0.0024	71	0.0015	77
26	0.0206	52	0.0111	58	0.0063	64	0.0038	70	0.0024	76
27	0.0325	51	0.0175	57	0.0100	63	0.0060	69	0.0037	75
28	0.0465	50	0.0256	56	0.0147	62	0.0088	68	0.0055	74
29	0.0660	49	0.0367	55	0.0213	61	0.0128	67	0.0080	73
30	0.0898	48	0.0507	54	0.0296	60	0.0180	66	0.0112	72
31	0.1201	47	0.0688	53	0.0406	59	0.0248	65	0.0156	71
32	0.1548	46	0.0903	52	0.0539	58	0.0332	64	0.0210	70
33	0.1970	45	0.1171	51	0.0709	57	0.0440	63	0.0280	69
34	0.2424	44	0.1474	50	0.0906	56	0.0567	62	0.0363	68
35	0.2944	43	0.1830	49	0.1142	55	0.0723	61	0.0467	67
36	0.3496	42	0.2226	48	0.1412	54	0.0905	60	0.0589	66
37	0.4091	41	0.2669	47	0.1725	53	0.1119	59	0.0736	65
38	0.4686	40	0.3141	46	0.2068	52	0.1361	58	0.0903	64
39	0.5314	39	0.3654	45	0.2454	51	0.1638	57	0.1099	63
40	0.5909	38	0.4178	44	0.2864	50	0.1942	56	0.1317	62
41	0.6504	37	0.4726	43	0.3310	49	0.2280	55	0.1566	61
42	0.7056	36	0.5274	42	0.3773	48	0.2643	54	0.1838	60
43	0.7576	35	0.5822	41	0.4259	47	0.3035	53	0.2139	59
44	0.8030	34	0.6346	40	0.4749	46	0.3445	52	0.2461	58
45	0.8452	33	0.6859	39	0.5251	45	0.3878	51	0.2811	57
46	0.8799	32	0.7331	38	0.5741	44	0.4320	50	0.3177	56
47	0.9102	31	0.7774	37	0.6227	43	0.4773	49	0.3564	55
48	0.9340	30	0.8170	36	0.6690	42	0.5227	48	0.3962	54
49	0.9535	29	0.8526	35	0.7136	41	0.5680	47	0.4374	53
50	0.9675	28	0.8829	34	0.7546	40	0.6122	46	0.4789	52
51	0.9794	27	0.9097	33	0.7932	39	0.6555	45	0.5211	51

				$m = 7$				
c_L	$n = 7$	c_U	$n = 8$	c_U	$n = 9$	c_U	$n = 10$	c_U
28	0.0003	77	0.0002	84	0.0001	91	0.0001	98
29	0.0006	76	0.0003	83	0.0002	90	0.0001	97
30	0.0012	75	0.0006	82	0.0003	89	0.0002	96
31	0.0020	74	0.0011	81	0.0006	88	0.0004	95
32	0.0035	73	0.0019	80	0.0010	87	0.0006	94
33	0.0055	72	0.0030	79	0.0017	86	0.0010	93
34	0.0087	71	0.0047	78	0.0026	85	0.0015	92
35	0.0131	70	0.0070	77	0.0039	84	0.0023	91
36	0.0189	69	0.0103	76	0.0058	83	0.0034	90
37	0.0265	68	0.0145	75	0.0082	82	0.0048	89
38	0.0364	67	0.0200	74	0.0115	81	0.0068	88
39	0.0487	66	0.0270	73	0.0156	80	0.0093	87
40	0.0641	65	0.0361	72	0.0209	79	0.0125	86
41	0.0825	64	0.0469	71	0.0274	78	0.0165	85
42	0.1043	63	0.0603	70	0.0356	77	0.0215	84
43	0.1297	62	0.0760	69	0.0454	76	0.0277	83
44	0.1588	61	0.0946	68	0.0571	75	0.0351	82
45	0.1914	60	0.1159	67	0.0708	74	0.0439	81
46	0.2279	59	0.1405	66	0.0869	73	0.0544	80
47	0.2675	58	0.1678	65	0.1052	72	0.0665	79
48	0.3100	57	0.1984	64	0.1261	71	0.0806	78
49	0.3552	56	0.2317	63	0.1496	70	0.0966	77
50	0.4024	55	0.2679	62	0.1755	67	0.1148	76
51	0.4508	54	0.3063	61	0.2039	68	0.1349	75
52	0.5000	53	0.3472	60	0.2349	67	0.1574	74
53	0.5492	52	0.3894	59	0.2680	66	0.1819	73
54	0.5976	51	0.4333	58	0.3032	65	0.2087	72
55	0.6448	50	0.4775	57	0.3403	64	0.2374	71
56	0.6900	49	0.5225	56	0.3788	63	0.2681	70
57	0.7325	48	0.5667	55	0.4185	62	0.3004	69
58	0.7721	47	0.6106	54	0.4591	61	0.3345	68
59	0.8086	46	0.6528	53	0.5000	60	0.3698	67
60	0.8412	45	0.6937	52	0.5409	59	0.4063	66
61	0.8703	44	0.7321	51	0.5815	58	0.4434	65
62	0.8957	43	0.7683	50	0.6212	57	0.4811	64
63	0.9175	42	0.8016	49	0.6597	56	0.5189	63

Tabla J. (*Continuación*)

			$m = 8$			
c_L	$n = 8$	c_U	$n = 9$	c_U	$n = 10$	c_U
36	0.0001	100	0.0000	108	0.0000	116
37	0.0002	99	0.0001	107	0.0000	115
38	0.0003	98	0.0002	106	0.0001	114
39	0.0005	97	0.0003	105	0.0002	113
40	0.0009	96	0.0005	104	0.0003	112
41	0.0015	95	0.0008	103	0.0004	111
42	0.0023	94	0.0012	102	0.0007	110
43	0.0035	93	0.0019	101	0.0010	109
44	0.0052	92	0.0028	100	0.0015	108
45	0.0074	91	0.0039	99	0.0022	107
46	0.0103	90	0.0056	98	0.0031	106
47	0.0141	89	0.0076	97	0.0043	105
48	0.0190	88	0.0103	96	0.0058	104
49	0.0249	87	0.0137	95	0.0078	103
50	0.0325	86	0.0180	94	0.0103	102
51	0.0415	85	0.0232	93	0.0133	101
52	0.0524	84	0.0296	92	0.0171	100
53	0.0652	83	0.0372	91	0.0217	99
54	0.0803	82	0.0464	90	0.0273	98
55	0.0974	81	0.0570	89	0.0338	97
56	0.1172	80	0.0694	88	0.0416	96
57	0.1393	79	0.0836	87	0.0506	95
58	0.1641	78	0.0998	86	0.0610	94
59	0.1911	77	0.1179	85	0.0729	93
60	0.2209	76	0.1383	84	0.0864	92
61	0.2527	75	0.1606	83	0.1015	91
62	0.2869	74	0.1852	82	0.1185	90
63	0.3227	73	0.2117	81	0.1371	89
64	0.3605	72	0.2404	80	0.1577	88
65	0.3992	71	0.2707	79	0.1800	87
66	0.4392	70	0.3029	78	0.2041	86
67	0.4796	69	0.3365	77	0.2299	85
68	0.5204	68	0.3715	76	0.2574	84
69	0.5608	67	0.4074	75	0.2863	83
70	0.6008	66	0.4442	74	0.3167	82
71	0.6395	65	0.4813	73	0.3482	81
72	0.6773	64	0.5187	72	0.3809	80
73	0.7131	63	0.5558	71	0.4143	79
74	0.7473	62	0.5926	70	0.4484	78
75	0.7791	61	0.6285	69	0.4827	77
76	0.8089	60	0.6635	68	0.5173	76

$m = 9$									
c_L	$n = 9$	c_U	$n = 10$	c_U	c_L	$n = 9$	c_U	$n = 10$	c_U
45	0.0000	126	0.0000	135	68	0.0680	103	0.0394	112
46	0.0000	125	0.0000	134	69	0.0807	102	0.0474	111
47	0.0001	124	0.0000	133	70	0.0951	101	0.0564	110
48	0.0001	123	0.0001	132	71	0.1112	100	0.0667	109
49	0.0002	122	0.0001	131	72	0.1290	99	0.0782	108
50	0.0004	121	0.0002	130	73	0.1487	98	0.0912	107
51	0.0006	120	0.0003	129	74	0.1701	97	0.1055	106
52	0.0009	119	0.0005	128	75	0.1933	96	0.1214	105
53	0.0014	118	0.0007	127	76	0.2181	95	0.1388	104
54	0.0020	117	0.0011	126	77	0.2447	94	0.1577	103
55	0.0028	116	0.0015	125	78	0.2729	93	0.1781	102
56	0.0039	115	0.0021	124	79	0.3024	92	0.2001	101
57	0.0053	114	0.0028	123	80	0.3332	91	0.2235	100
58	0.0071	113	0.0038	122	81	0.3652	90	0.2483	99
59	0.0094	112	0.0051	121	82	0.3981	89	0.2745	98
60	0.0122	111	0.0066	120	83	0.4317	88	0.3019	97
61	0.0157	110	0.0086	119	84	0.4657	87	0.3304	96
62	0.0200	109	0.0110	118	85	0.5000	86	0.3598	95
63	0.0252	108	0.0140	117	86	0.5343	85	0.3901	94
64	0.0313	107	0.0175	116	87	0.5683	84	0.4211	93
65	0.0385	106	0.0217	115	88	0.6019	83	0.4524	92
66	0.0470	105	0.0267	114	89	0.6348	82	0.4841	91
67	0.0567	104	0.0326	113	90	0.6668	81	0.5159	90

Tabla J. (*Continuación*)

$m = 10$					
c_L	$n = 10$	c_U	c_L	$n = 10$	c_U
55	0.0000	155	81	0.0376	129
56	0.0000	154	82	0.0446	128
57	0.0000	153	83	0.0526	127
58	0.0000	152	84	0.0615	126
59	0.0001	151	85	0.0716	125
60	0.0001	150	86	0.0827	124
61	0.0002	149	87	0.0952	123
62	0.0002	148	88	0.1088	122
63	0.0004	147	89	0.1237	121
64	0.0005	146	90	0.1399	120
65	0.0008	145	91	0.1575	119
66	0.0010	144	92	0.1763	118
67	0.0014	143	93	0.1965	117
68	0.0019	142	94	0.2179	116
69	0.0026	141	95	0.2406	115
70	0.0034	140	96	0.2644	114
71	0.0045	139	97	0.2894	113
72	0.0057	138	98	0.3153	112
73	0.0073	137	99	0.3421	111
74	0.0093	136	100	0.3697	110
75	0.0116	135	101	0.3980	109
76	0.0144	134	102	0.4267	108
77	0.0177	133	103	0.4559	107
78	0.0216	132	104	0.4853	106
79	0.0262	131	105	0.5147	105
80	0.0315	130			

Tabla K. Valores críticos de U para la prueba poderosa de rangos ordenados.†

					n						
α	3	4	5	6	7	8	9	10	11	12	m
.10	**2.347**	**1.732**	**1.632**	**1.897**	**1.644**	**1.500**	**1.575**	**1.611**	**1.638**	**1.616**	
.05	∞*	3.273	2.324	2.912	2.605	2.777	2.353	2.553	2.369	2.449	
.025		∞*	**4.195**	**5.116**	**6.037**	**4.082**	**3.566**	**3.651**	**3.503**	**3.406**	3
.01			∞*	∞*	∞*	6.957	7.876	8.795	5.831	5.000	
		1.586	**1.500**	**1.434**	**1.428**	**1.371**	**1.434**	**1.466**	**1.448**	**1.455**	
		2.502	2.160	2.247	2.104	2.162	2.057	2.000	2.067	2.096	
		4.483	**3.265**	**3.021**	**3.295**	**2.868**	**2.683**	**2.951**	**2.776**	**2.847**	4
		∞*	∞*	6.899	4.786	4.252	4.423	4.276	4.017	3.904	
			1.447	**1.362**	**1.308**	**1.378**	**1.361**	**1.361**	**1.340**	**1.369**	
			2.063	1.936	1.954	1.919	1.893	1.900	1.891	1.923	
			2.859	**2.622**	**2.465**	**2.556**	**2.536**	**2.496**	**2.497**	**2.479**	5
			7.187	3.913	4.246	3.730	3.388	3.443	3.435	3.444	
				1.335	**1.326**	**1.327**	**1.338**	**1.339**	**1.320**	**1.330**	
				1.860	1.816	1.796	1.845	1.829	1.833	1.835	
				2.502	**2.500**	**2.443**	**2.349**	**2.339**	**2.337**	**2.349**	6
				3.712	3.519	3.230	3.224	3.164	3.161	3.151	
					1.333	**1.310**	**1.320**	**1.313**	**1.302**	**1.318**	
					1.804	1.807	1.790	1.776	1.796	1.787	
					2.331	**2.263**	**2.287**	**2.248**	**2.240**	**2.239**	7
					3.195	3.088	2.967	3.002	2.979	2.929	
						1.295	**1.283**	**1.284**	**1.290**	**1.293**	
						1.766	1.765	1.756	1.746	1.759	
						2.251	**2.236**	**2.209**	**2.205**	**2.198**	8
						2.954	2.925	2.880	2.856	2.845	
							1.294	**1.304**	**1.288**	**1.299**	
							1.744	1.742	1.744	1.737	
							2.206	**2.181**	**2.172**	**2.172**	9
							2.857	2.802	2.798	2.770	
								1.295	**1.284**	**1.284**	
								1.723	1.726	1.720	
								2.161	**2.152**	**2.144**	10
								2.770	2.733	2.718	
									1.289	**1.290**	
									1.716	1.708	
									2.138	**2.127**	11
									2.705	2.683	
										1.283	
										1.708	
										2.117	12
										2.661	

Los valores tabulados se encuentran en renglones sucesivos para $\alpha = 0.10, 0.05, 0.025, 0.01$ para distintos valores de m y n.
Nota: para esta prueba, m es la muestra de menor tamaño y n es la de mayor tamaño.
* El mayor tamaño de $\hat{U}$ se utiliza cuando V_x o V_y es igual a cero, o cuando $\hat{U}$ no está definida.

† Adaptada de Fligner, M. A. y Policello, G. E. II, "Robust tank procedures for the Behrens-Fisher problem", en *Journal of the American Statistical Association*, 76, 1981, págs. 162-168. Con autorización de los autores y del editor.

Tabla L_I. Prueba de Kolmogorov-Smirnov para dos muestras.*
Valores críticos para la región de rechazo unidireccional de $mnD_{m,n} \geq c$. Los valores superior, medio e inferior indican $c_{0.10}$, $c_{0.05}$ y $c_{0.01}$ para cada entrada (m, n).

	m																						
n	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
3	**9**	**10**	**11**	**15**	**15**	**16**	**21**	**19**	**22**	**24**	**25**	**26**	**30**	**30**	**32**	**36**	**36**	**37**	**42**	**40**	**43**	**45**	**46**
	9	10	13	15	16	19	21	22	25	27	28	31	33	34	35	39	40	41	45	46	47	51	52
	**	**	**	**	*19*	*22*	*27*	*28*	*31*	*33*	*34*	*37*	*42*	*43*	*43*	*48*	*49*	*52*	*54*	*55*	*58*	*63*	*64*
4	**10**	**16**	**13**	**16**	**18**	**24**	**21**	**24**	**26**	**32**	**29**	**32**	**34**	**40**	**37**	**40**	**41**	**48**	**45**	**48**	**49**	**56**	**53**
	10	16	16	18	21	24	25	28	29	36	33	38	38	44	44	46	49	52	52	56	57	60	61
	**	**	*17*	*22*	*25*	*32*	*29*	*34*	*37*	*40*	*41*	*46*	*46*	*52*	*53*	*56*	*57*	*64*	*64*	*66*	*69*	*76*	*73*
5	**11**	**13**	**20**	**19**	**21**	**23**	**26**	**30**	**30**	**32**	**35**	**37**	**45**	**41**	**44**	**46**	**47**	**55**	**51**	**54**	**56**	**58**	**65**
	13	16	20	21	24	26	28	35	35	36	40	42	50	46	49	51	56	60	60	62	65	67	75
	**	*17*	*25*	*26*	*29*	*33*	*36*	*40*	*41*	*46*	*48*	*51*	*60*	*56*	*61*	*63*	*67*	*75*	*75*	*76*	*81*	*82*	*90*
6	**15**	**16**	**19**	**24**	**24**	**26**	**30**	**32**	**33**	**42**	**37**	**42**	**45**	**48**	**49**	**54**	**54**	**56**	**60**	**62**	**63**	**72**	**67**
	15	18	21	30	25	30	33	36	38	48	43	48	51	54	56	66	61	66	69	70	73	78	78
	**	*22*	*26*	*36*	*31*	*38*	*42*	*44*	*49*	*54*	*54*	*60*	*63*	*66*	*68*	*78*	*77*	*80*	*84*	*88*	*91*	*96*	*96*
7	**15**	**18**	**21**	**24**	**35**	**28**	**32**	**34**	**38**	**40**	**44**	**49**	**48**	**51**	**54**	**56**	**59**	**61**	**70**	**68**	**70**	**72**	**74**
	16	21	24	25	35	34	36	40	43	45	50	56	56	58	61	64	68	72	77	77	79	83	85
	19	*25*	*29*	*31*	*42*	*42*	*46*	*50*	*53*	*57*	*59*	*70*	*70*	*71*	*75*	*81*	*85*	*87*	*98*	*97*	*99*	*103*	*106*
8	**16**	**24**	**23**	**26**	**28**	**40**	**33**	**40**	**41**	**48**	**47**	**50**	**52**	**64**	**57**	**62**	**64**	**72**	**71**	**74**	**76**	**88**	**81**
	19	24	26	30	34	40	40	44	48	52	53	58	60	72	65	72	73	80	81	84	89	96	95
	22	*32*	*33*	*38*	*42*	*48*	*49*	*56*	*59*	*64*	*66*	*72*	*75*	*88*	*81*	*88*	*91*	*100*	*100*	*106*	*107*	*120*	*118*

9	**21**	**21**	**26**	**30**	**32**	**33**	**45**	**43**	**45**	**51**	**51**	**54**	**60**	**61**	**65**	**72**	**70**	**73**	**78**	**79**	**82**	**87**	**88**
	21	25	28	33	36	40	54	46	51	57	57	63	69	68	74	81	80	83	90	91	94	99	101
	27	*29*	*36*	*42*	*46*	*49*	*63*	*61*	*62*	*69*	*73*	*77*	*84*	*86*	*92*	*99*	*99*	*103*	*111*	*111*	*117*	*123*	*124*
10	**19**	**24**	**30**	**32**	**34**	**40**	**43**	**50**	**48**	**52**	**55**	**60**	**65**	**66**	**69**	**72**	**74**	**90**	**80**	**86**	**88**	**92**	**100**
	22	28	35	36	40	44	46	60	57	60	62	68	75	76	77	82	85	100	91	98	101	106	110
	28	*34*	*40*	*44*	*50*	*56*	*61*	*70*	*69*	*74*	*78*	*84*	*90*	*94*	*97*	*104*	*104*	*120*	*118*	*120*	*125*	*130*	*140*
11	**22**	**26**	**30**	**33**	**38**	**41**	**45**	**48**	**66**	**54**	**59**	**63**	**66**	**69**	**72**	**76**	**79**	**84**	**85**	**99**	**95**	**98**	**100**
	25	29	35	38	43	48	51	57	66	64	67	72	76	80	83	87	92	95	101	110	108	111	116
	31	*37*	*41*	*49*	*53*	*59*	*62*	*69*	*88*	*77*	*85*	*89*	*95*	*100*	*104*	*108*	*114*	*117*	*124*	*143*	*132*	*138*	*143*
12	**24**	**32**	**32**	**42**	**40**	**48**	**51**	**52**	**54**	**72**	**61**	**68**	**72**	**76**	**77**	**84**	**85**	**92**	**93**	**98**	**100**	**108**	**106**
	27	36	36	48	45	52	57	60	64	72	71	78	84	88	89	96	98	104	108	110	113	132	120
	33	*40*	*46*	*54*	*57*	*64*	*69*	*74*	*77*	*96*	*92*	*94*	*102*	*108*	*111*	*120*	*121*	*128*	*132*	*138*	*138*	*156*	*153*
13	**25**	**29**	**35**	**37**	**44**	**47**	**51**	**55**	**59**	**61**	**78**	**72**	**75**	**79**	**81**	**87**	**89**	**95**	**97**	**100**	**105**	**109**	**111**
	28	33	40	43	50	53	57	62	67	71	91	78	86	90	94	98	102	108	112	117	120	124	131
	34	*41*	*48*	*54*	*59*	*66*	*73*	*78*	*85*	*92*	*104*	*102*	*106*	*112*	*118*	*121*	*127*	*135*	*138*	*143*	*150*	*154*	*160*
14	**26**	**32**	**37**	**42**	**49**	**50**	**54**	**60**	**63**	**68**	**72**	**84**	**80**	**84**	**87**	**92**	**94**	**100**	**112**	**108**	**110**	**116**	**119**
	31	38	42	48	56	58	63	68	72	78	78	98	92	96	99	104	108	114	126	124	127	132	136
	37	*46*	*51*	*60*	*70*	*72*	*77*	*84*	*89*	*94*	*102*	*112*	*111*	*120*	*124*	*130*	*135*	*142*	*154*	*152*	*157*	*164*	*169*
15	**30**	**34**	**45**	**45**	**48**	**52**	**60**	**65**	**66**	**72**	**75**	**80**	**90**	**87**	**91**	**99**	**100**	**110**	**111**	**111**	**117**	**123**	**130**
	33	38	50	51	56	60	69	75	76	84	86	92	105	101	105	111	113	125	126	130	134	141	145
	42	*46*	*60*	*63*	*70*	*75*	*84*	*90*	*95*	*102*	*106*	*111*	*135*	*120*	*130*	*138*	*142*	*150*	*156*	*160*	*165*	*174*	*180*
16	**30**	**40**	**41**	**48**	**51**	**64**	**61**	**66**	**69**	**76**	**79**	**84**	**87**	**112**	**94**	**100**	**104**	**112**	**114**	**118**	**122**	**136**	**130**
	34	44	46	54	58	72	68	76	80	88	90	96	101	112	109	116	120	128	130	136	140	152	148
	43	*52*	*56*	*66*	*71*	*88*	*86*	*94*	*100*	*108*	*112*	*120*	*120*	*144*	*139*	*142*	*149*	*156*	*162*	*168*	*174*	*184*	*185*

Tabla L$_1$. (*Continuación*)
Valores críticos para la región de rechazo unidireccional de $mnD_{m,n} \geq c$. Los valores superior, medio e inferior indican $c_{0.10}$, $c_{0.05}$ y $c_{0.01}$ para cada entrada (m, n).

n	m: 3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
17	**32**	**37**	**44**	**49**	**54**	**57**	**65**	**69**	**72**	**77**	**81**	**87**	**91**	**94**	**119**	**102**	**108**	**113**	**118**	**122**	**128**	**132**	**137**
	35	44	49	56	61	65	74	77	83	89	94	99	105	109	136	118	125	130	135	141	146	150	156
	43	*53*	*61*	*68*	*75*	*81*	*92*	*97*	*104*	*111*	*118*	*124*	*130*	*139*	*153*	*150*	*157*	*162*	*168*	*175*	*181*	*187*	*192*
18	**36**	**40**	**46**	**54**	**56**	**62**	**72**	**72**	**76**	**84**	**87**	**92**	**99**	**100**	**102**	**126**	**116**	**120**	**126**	**128**	**133**	**144**	**142**
	39	46	51	66	64	72	81	82	87	96	98	104	111	116	118	144	127	136	144	148	151	162	161
	48	*56*	*63*	*78*	*81*	*88*	*99*	*104*	*108*	*120*	*121*	*130*	*138*	*142*	*150*	*180*	*160*	*170*	*177*	*184*	*189*	*198*	*201*
19	**36**	**41**	**47**	**54**	**59**	**64**	**70**	**74**	**79**	**85**	**89**	**94**	**100**	**104**	**108**	**116**	**133**	**125**	**128**	**132**	**137**	**142**	**148**
	40	49	56	61	68	73	80	85	92	98	102	108	113	120	125	127	152	144	147	151	159	162	168
	49	*57*	*67*	*77*	*85*	*91*	*99*	*104*	*114*	*121*	*127*	*135*	*142*	*149*	*157*	*160*	*190*	*171*	*183*	*189*	*197*	*204*	*211*
20	**37**	**48**	**55**	**56**	**61**	**72**	**73**	**90**	**84**	**92**	**95**	**100**	**110**	**112**	**113**	**120**	**125**	**140**	**134**	**138**	**143**	**152**	**155**
	41	52	60	66	72	80	83	100	95	104	108	114	125	128	130	136	144	160	154	160	163	172	180
	52	*64*	*75*	*80*	*87*	*100*	*103*	*120*	*117*	*128*	*135*	*142*	*150*	*156*	*162*	*170*	*171*	*200*	*193*	*196*	*203*	*212*	*220*
21	**42**	**45**	**51**	**60**	**70**	**71**	**78**	**80**	**85**	**93**	**97**	**112**	**111**	**114**	**118**	**126**	**128**	**134**	**147**	**142**	**147**	**156**	**158**
	45	52	60	69	77	81	90	91	101	108	112	126	126	130	135	144	147	154	168	163	170	177	182
	54	*64*	*75*	*84*	*98*	*100*	*111*	*118*	*124*	*132*	*138*	*154*	*156*	*162*	*168*	*177*	*183*	*193*	*210*	*205*	*212*	*222*	*225*
22	**40**	**48**	**54**	**62**	**68**	**74**	**79**	**86**	**99**	**98**	**100**	**108**	**111**	**118**	**122**	**128**	**132**	**138**	**142**	**176**	**151**	**158**	**163**
	46	56	62	70	77	84	91	98	110	110	117	124	130	136	141	148	151	160	163	198	173	182	188
	55	*66*	*76*	*88*	*97*	*106*	*111*	*120*	*143*	*138*	*143*	*152*	*160*	*168*	*175*	*184*	*189*	*196*	*205*	*242*	*217*	*228*	*234*

23	**43**	**49**	**56**	**63**	**70**	**76**	**82**	**88**	**95**	**100**	**105**	**110**	**117**	**122**	**128**	**133**	**137**	**143**	**147**	**151**	**184**	**160**	**169**
	47	57	65	73	79	89	94	101	108	113	120	127	134	140	146	151	159	163	170	173	207	183	194
	58	*69*	*81*	*91*	*99*	*107*	*117*	*125*	*132*	*138*	*150*	*157*	*165*	*174*	*181*	*189*	*197*	*203*	*212*	*217*	*253*	*228*	*242*
24	**45**	**56**	**58**	**72**	**72**	**88**	**87**	**92**	**98**	**108**	**109**	**116**	**123**	**136**	**132**	**144**	**142**	**152**	**156**	**158**	**160**	**192**	**178**
	51	60	67	78	83	96	99	106	111	132	124	132	141	152	150	162	162	172	177	182	183	216	204
	63	*76*	*82*	*96*	*103*	*120*	*123*	*130*	*138*	*156*	*154*	*164*	*174*	*184*	*187*	*198*	*204*	*212*	*222*	*228*	*228*	*264*	*254*
25	**46**	**53**	**65**	**67**	**74**	**81**	**88**	**100**	**100**	**106**	**111**	**119**	**130**	**130**	**137**	**142**	**148**	**155**	**158**	**163**	**169**	**178**	**200**
	52	61	75	78	85	95	101	110	116	120	131	136	145	148	156	161	168	180	182	188	194	204	225
	64	*73*	*90*	*96*	*106*	*118*	*124*	*140*	*143*	*153*	*160*	*169*	*180*	*185*	*192*	*201*	*211*	*220*	*225*	*234*	*242*	*254*	*275*

* Adaptada de Gail, M. H. y Green, S. B., "Critical values for the one-sided two-sample Kolmogorov-Smirnov statistic", en *Journal of the American Statistical Association*, núm. 71, 1976, págs. 757-760. Con autorización de los autores y del editor.

** Estadísticos que no alcanzan este nivel de significación.

Tabla L$_{II}$. Prueba de Kolmogorov-Smirnov para dos muestras.*
Valores críticos para la región de rechazo bidireccional de $mnD_{m,n} \geq c$. Los valores superior, medio e inferior indican $c_{0.10}$, $c_{0.05}$ y $c_{0.01}$ para cada entrada (m, n).

	m																								
n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1																			**19**	**20**	**21**	**22**	**23**	**24**	**25**
2					**10**	**12**	**14**	**16**	**18**	**18**	**20**	**22**	**24**	**24**	**26**	**28**	**30**	**32**	**32**	**34**	**36**	**38**	**38**	**40**	**42**
								16	18	20	22	24	26	26	28	30	32	34	36	38	38	40	42	44	46
																			38	*40*	*42*	*44*	*46*	*48*	*50*
3			**9**	**12**	**15**	**15**	**18**	**21**	**21**	**24**	**27**	**27**	**30**	**33**	**33**	**36**	**36**	**39**	**42**	**42**	**45**	**48**	**48**	**51**	**54**
					15	18	21	21	24	27	30	30	33	36	36	39	42	45	45	48	51	51	54	57	60
									27	*30*	*33*	*36*	*39*	*42*	*42*	*45*	*48*	*51*	*54*	*57*	*57*	*60*	*63*	*66*	*69*
4			**12**	**16**	**16**	**18**	**21**	**24**	**27**	**28**	**29**	**36**	**35**	**38**	**40**	**44**	**44**	**46**	**49**	**52**	**52**	**56**	**57**	**60**	**63**
				16	20	20	24	28	28	30	33	36	39	42	44	48	48	50	53	60	59	62	64	68	68
						24	*28*	*32*	*36*	*36*	*40*	*44*	*48*	*48*	*52*	*56*	*60*	*60*	*64*	*68*	*72*	*72*	*76*	*80*	*84*
5		**10**	**15**	**16**	**20**	**24**	**25**	**27**	**30**	**35**	**35**	**36**	**40**	**42**	**50**	**48**	**50**	**52**	**56**	**60**	**60**	**63**	**65**	**67**	**75**
			15	20	25	24	28	30	35	40	39	43	45	46	55	54	55	60	61	65	69	70	72	76	80
					25	*30*	*35*	*35*	*40*	*45*	*45*	*50*	*52*	*56*	*60*	*64*	*68*	*70*	*71*	*80*	*80*	*83*	*87*	*90*	*95*
6		**12**	**15**	**18**	**24**	**30**	**28**	**30**	**33**	**36**	**38**	**48**	**46**	**48**	**51**	**54**	**56**	**66**	**64**	**66**	**69**	**70**	**73**	**78**	**78**
			18	20	24	30	30	34	39	40	43	48	52	54	57	60	62	72	70	72	75	78	80	90	88
				24	*30*	*36*	*36*	*40*	*45*	*48*	*54*	*60*	*60*	*64*	*69*	*72*	*73*	*84*	*83*	*88*	*90*	*92*	*97*	*102*	*107*
7		**14**	**18**	**21**	**25**	**28**	**35**	**34**	**36**	**40**	**44**	**46**	**50**	**56**	**56**	**59**	**61**	**65**	**69**	**72**	**77**	**77**	**80**	**84**	**86**
			21	24	28	30	42	40	42	46	48	53	56	63	62	64	68	72	76	79	91	84	89	92	97
				28	*35*	*36*	*42*	*48*	*49*	*53*	*59*	*60*	*65*	*77*	*75*	*77*	*84*	*87*	*91*	*93*	*105*	*103*	*108*	*112*	*115*

8	**16**	**21**	**24**	**27**	**30**	**34**	**40**	**40**	**44**	**48**	**52**	**54**	**58**	**60**	**72**	**68**	**72**	**74**	**80**	**81**	**84**	**89**	**96**	**95**
	16	21	28	30	34	40	48	46	48	53	60	62	64	67	80	77	80	82	88	89	94	98	104	104
			32	*35*	*40*	*48*	*56*	*55*	*60*	*64*	*68*	*72*	*76*	*81*	*88*	*88*	*94*	*98*	*104*	*107*	*112*	*115*	*128*	*125*
9	**18**	**21**	**27**	**30**	**33**	**36**	**40**	**54**	**50**	**52**	**57**	**59**	**63**	**69**	**69**	**74**	**81**	**80**	**84**	**90**	**91**	**94**	**99**	**101**
	18	24	28	35	39	42	46	54	53	59	63	65	70	75	78	82	90	89	93	99	101	106	111	114
		27	*36*	*40*	*45*	*49*	*55*	*63*	*63*	*70*	*75*	*78*	*84*	*90*	*94*	*99*	*108*	*107*	*111*	*117*	*122*	*126*	*132*	*135*
10	**18**	**24**	**28**	**35**	**36**	**40**	**44**	**50**	**60**	**57**	**60**	**64**	**68**	**75**	**76**	**79**	**82**	**85**	**100**	**95**	**98**	**101**	**106**	**110**
	20	27	30	40	40	46	48	53	70	60	66	70	74	80	84	89	92	94	110	105	108	114	118	125
		30	*36*	*45*	*48*	*53*	*60*	*63*	*80*	*77*	*80*	*84*	*90*	*100*	*100*	*106*	*108*	*113*	*130*	*126*	*130*	*137*	*140*	*150*
11	**20**	**27**	**29**	**35**	**38**	**44**	**48**	**52**	**57**	**66**	**64**	**67**	**73**	**76**	**80**	**85**	**88**	**92**	**96**	**101**	**110**	**108**	**111**	**117**
	22	30	33	39	43	48	53	59	60	77	72	75	82	84	89	93	97	102	107	112	121	119	124	129
		33	*40*	*45*	*54*	*59*	*64*	*70*	*77*	*88*	*86*	*91*	*96*	*102*	*106*	*110*	*118*	*122*	*127*	*134*	*143*	*142*	*150*	*154*
12	**22**	**27**	**36**	**36**	**48**	**46**	**52**	**57**	**60**	**64**	**72**	**71**	**78**	**84**	**88**	**90**	**96**	**99**	**104**	**108**	**110**	**113**	**132**	**120**
	24	30	36	43	48	53	60	63	66	72	84	81	86	93	96	100	108	108	116	120	124	125	144	138
		36	*44*	*50*	*60*	*60*	*68*	*75*	*80*	*86*	*96*	*95*	*104*	*108*	*116*	*119*	*126*	*130*	*140*	*141*	*148*	*149*	*168*	*165*
13	**24**	**30**	**35**	**40**	**46**	**50**	**54**	**59**	**64**	**67**	**71**	**91**	**78**	**87**	**91**	**96**	**99**	**104**	**108**	**113**	**117**	**120**	**125**	**131**
	26	33	39	45	52	56	62	65	70	75	81	91	89	96	101	105	110	114	120	126	130	135	140	145
		39	*48*	*52*	*60*	*65*	*72*	*78*	*84*	*91*	*95*	*117*	*104*	*115*	*121*	*127*	*131*	*138*	*143*	*150*	*156*	*161*	*166*	*172*
14	**24**	**33**	**38**	**42**	**48**	**56**	**58**	**63**	**68**	**73**	**78**	**78**	**98**	**92**	**96**	**100**	**104**	**110**	**114**	**126**	**124**	**127**	**132**	**136**
	26	36	42	46	54	63	64	70	74	82	86	89	112	98	106	111	116	121	126	140	138	142	146	150
		42	*48*	*56*	*64*	*77*	*76*	*84*	*90*	*96*	*104*	*104*	*126*	*123*	*126*	*134*	*140*	*148*	*152*	*161*	*164*	*170*	*176*	*182*
15	**26**	**33**	**40**	**50**	**51**	**56**	**60**	**69**	**75**	**76**	**84**	**87**	**92**	**105**	**101**	**105**	**111**	**114**	**125**	**126**	**130**	**134**	**141**	**145**
	28	36	44	55	57	62	67	75	80	84	93	96	98	120	114	116	123	127	135	138	144	149	156	160

Tabla L_{II}. (*Continuación*)
Valores críticos para la región de rechazo bidireccional de $mnD_{m,n} \geq c$. Los valores superior, medio e inferior indican $c_{0.10}$, $c_{0.05}$ y $c_{0.01}$ para cada entrada (m, n).

	m																								
n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
16		**28**	**36**	**44**	**48**	**54**	**59**	**72**	**69**	**76**	**80**	**88**	**91**	**96**	**101**	**112**	**109**	**116**	**120**	**128**	**130**	**136**	**141**	**152**	**149**
		30	39	48	54	60	64	80	78	84	89	96	101	106	114	128	124	128	133	140	145	150	157	168	167
			45	*56*	*64*	*72*	*77*	*88*	*94*	*100*	*106*	*116*	*121*	*126*	*133*	*160*	*143*	*154*	*160*	*168*	*173*	*180*	*187*	*200*	*199*
17		**30**	**36**	**44**	**50**	**56**	**61**	**68**	**74**	**79**	**85**	**90**	**96**	**100**	**105**	**109**	**136**	**118**	**126**	**132**	**136**	**142**	**146**	**151**	**156**
		32	42	48	55	62	68	77	82	89	93	100	105	111	116	124	136	133	141	146	151	157	163	168	173
			48	*60*	*68*	*73*	*84*	*88*	*99*	*106*	*110*	*119*	*127*	*134*	*142*	*143*	*170*	*164*	*166*	*175*	*180*	*187*	*196*	*203*	*207*
18		**32**	**39**	**46**	**52**	**66**	**65**	**72**	**81**	**82**	**88**	**96**	**99**	**104**	**111**	**116**	**118**	**144**	**133**	**136**	**144**	**148**	**152**	**162**	**162**
		34	45	50	60	72	72	80	90	92	97	108	110	116	123	128	133	162	142	152	159	164	170	180	180
			51	*60*	*70*	*84*	*87*	*94*	*108*	*108*	*118*	*126*	*131*	*140*	*147*	*154*	*164*	*180*	*176*	*182*	*189*	*196*	*204*	*216*	*216*
19	**19**	**32**	**42**	**49**	**56**	**64**	**69**	**74**	**80**	**85**	**92**	**99**	**104**	**110**	**114**	**120**	**126**	**133**	**152**	**144**	**147**	**152**	**159**	**164**	**168**
		36	45	53	61	70	76	82	89	94	102	108	114	121	127	133	141	142	171	160	163	169	177	183	187
		38	*54*	*64*	*71*	*83*	*91*	*98*	*107*	*113*	*122*	*130*	*138*	*148*	*152*	*160*	*166*	*176*	*190*	*187*	*199*	*204*	*209*	*218*	*224*
20	**20**	**34**	**42**	**52**	**60**	**66**	**72**	**80**	**84**	**100**	**96**	**104**	**108**	**114**	**125**	**128**	**132**	**136**	**144**	**160**	**154**	**160**	**164**	**172**	**180**
		38	48	60	65	72	79	88	93	110	107	116	120	126	135	140	146	152	160	180	173	176	184	192	200
		40	*57*	*68*	*80*	*88*	*93*	*104*	*111*	*130*	*127*	*140*	*143*	*152*	*160*	*168*	*175*	*182*	*187*	*220*	*199*	*212*	*219*	*228*	*235*
21	**21**	**36**	**45**	**52**	**60**	**69**	**77**	**81**	**90**	**95**	**101**	**108**	**113**	**126**	**126**	**130**	**136**	**144**	**147**	**154**	**168**	**163**	**171**	**177**	**182**
		38	51	59	69	75	91	89	99	105	112	120	126	140	138	145	151	159	163	173	189	183	189	198	202
		42	*57*	*72*	*80*	*90*	*105*	*107*	*117*	*126*	*134*	*141*	*150*	*161*	*168*	*173*	*180*	*189*	*199*	*199*	*231*	*223*	*227*	*237*	*244*

22	**22**	**38**	**48**	**56**	**63**	**70**	**77**	**84**	**91**	**98**	**110**	**110**	**117**	**124**	**130**	**136**	**142**	**148**	**152**	**160**	**163**	**198**	**173**	**182**	**189**
		40	51	62	70	78	84	94	101	108	121	124	130	138	144	150	157	164	169	176	183	198	194	204	209
		44	*60*	*72*	*83*	*92*	*103*	*112*	*122*	*130*	*143*	*148*	*156*	*164*	*173*	*180*	*187*	*196*	*204*	*212*	*223*	*242*	*237*	*242*	*250*
23	**23**	**38**	**48**	**57**	**65**	**73**	**80**	**89**	**94**	**101**	**108**	**113**	**120**	**127**	**134**	**141**	**146**	**152**	**159**	**164**	**171**	**173**	**207**	**183**	**195**
		42	54	64	72	80	89	98	106	114	119	125	135	142	149	157	163	170	177	184	189	194	230	205	216
		46	*63*	*76*	*87*	*97*	*108*	*115*	*126*	*137*	*142*	*149*	*161*	*170*	*179*	*187*	*196*	*204*	*209*	*219*	*227*	*237*	*253*	*249*	*262*
24	**24**	**40**	**51**	**60**	**67**	**78**	**84**	**96**	**99**	**106**	**111**	**132**	**125**	**132**	**141**	**152**	**151**	**162**	**164**	**172**	**177**	**182**	**183**	**216**	**204**
		44	57	68	76	90	92	104	111	118	124	144	140	146	156	168	168	180	183	192	198	204	205	240	225
		48	*66*	*80*	*90*	*102*	*112*	*128*	*132*	*140*	*150*	*168*	*166*	*176*	*186*	*200*	*203*	*216*	*218*	*228*	*237*	*242*	*249*	*288*	*262*
25	**25**	**42**	**54**	**63**	**75**	**78**	**86**	**95**	**101**	**110**	**117**	**120**	**131**	**136**	**145**	**149**	**156**	**162**	**168**	**180**	**182**	**189**	**195**	**204**	**225**
		46	60	68	80	88	97	104	114	125	129	138	145	150	160	167	173	180	187	200	202	209	216	225	250
		50	*69*	*84*	*95*	*107*	*115*	*125*	*135*	*150*	*154*	*165*	*172*	*182*	*195*	*199*	*207*	*216*	*224*	*235*	*244*	*250*	*262*	*262*	*300*

* Adaptada de la tabla 55 en Pearson, E. S. y Hartley, H. O., *Biometrika tables for statisticians*, vol. 2, Cambridge University Press, Cambridge, 1972. Por cortesía del consejo de administración de *Biometrika*.

Tabla L_{III}. Valores críticos de $D_{m,n}$ para la prueba de Kolmogorov-Smirnov para dos muestras (muestras grandes, bidireccional).*

Nivel de significación	Valor de $D_{m,n}$ que indica el rechazo de H_0 en el nivel de significación indicado, donde $D_{m,n}$ = máximo $\lvert S_m(X) - S_n(X) \rvert$
0.10	$1.22 \sqrt{\frac{m+n}{mn}}$
0.05	$1.36 \sqrt{\frac{m+n}{mn}}$
0.025	$1.48 \sqrt{\frac{m+n}{mn}}$
0.01	$1.63 \sqrt{\frac{m+n}{mn}}$
0.005	$1.73 \sqrt{\frac{m+n}{mn}}$
0.001	$1.95 \sqrt{\frac{m+n}{mn}}$

* Adaptada de Smirnov, N., "Tables for estimating the goodness of fit of empirical distributions", en *Annals of Mathematical Statistics*, núm. 19, 1948, págs. 280-281, por cortesía del editor.

Tabla M. Valores críticos para la prueba estadística de análisis de varianza bifactorial por rangos de Friedman, F_r.*

k	N	$\alpha \leq .10$	$\alpha \leq .05$	$\alpha \leq .01$
3	3	6.00	6.00	–
	4	6.00	6.50	8.00
	5	5.20	6.40	8.40
	6	5.33	7.00	9.00
	7	5.43	7.14	8.86
	8	5.25	6.25	9.00
	9	5.56	6.22	8.67
	10	5.00	6.20	9.60
	11	4.91	6.54	8.91
	12	5.17	6.17	8.67
	13	4.77	6.00	9.39
	∞	4.61	5.99	9.21
4	2	6.00	6.00	–
	3	6.60	7.40	8.60
	4	6.30	7.80	9.60
	5	6.36	7.80	9.96
	6	6.40	7.60	10.00
	7	6.26	7.80	10.37
	8	6.30	7.50	10.35
	∞	6.25	7.82	11.34
5	3	7.47	8.53	10.13
	4	7.60	8.80	11.00
	5	7.68	8.96	11.52
	∞	7.78	9.49	13.28

* Algunas entradas fueron adaptadas y reproducidas con autorización de los editores Charles Griffin & Co. Ltd., 16 Pembridge Road, Londres W11 3HL, de la tabla del Apéndice 5 de Kendall, M. G., *Rank correlation methods*, 4a. ed., 1970. Otras entradas se adaptaron de la tabla A.15 de Hollander, M. y Wolfe, D. A., *Nonparametric statistics*, 1973, J. Wiley, Nueva York, 1973. Reproducida con autorización de los autores y el editor.

Tabla N. Valores críticos del estadístico L de la prueba de Page.*
*Los valores tabulados son L_a, $P\,[L \geq L_a] = \alpha$.

	$k = 3$			$k = 4$			$k = 5$			$k = 6$		
	α			α			α			α		
N	0.05	0.01	0.001	0.05	0.01	0.001	0.05	0.01	0.001	0.05	0.01	0.001
2	28			58	60		103	106	109	166	173	178
3	41	42		84	87	89	150	155	160	244	252	260
4	54	55	56	111	114	117	197	204	210	321	331	341
5	66	68	70	137	141	145	244	251	259	397	409	420
6	79	81	83	163	167	172	291	299	307	474	486	499
7	91	93	96	189	193	198	338	346	355	550	563	577
8	104	106	109	214	220	225	384	393	403	625	640	655
9	116	119	121	240	246	252	431	441	451	701	717	733
10	128	131	134	266	272	278	477	487	499	777	793	811
11	141	144	147	292	298	305	523	534	546	852	869	888
12	153	156	160	317	324	331	570	581	593	928	949	965
13	165	169	172									
14	178	181	185									
15	190	194	197									
16	202	206	210									
17	215	218	223									
18	227	231	235									
19	239	243	248									
20	251	256	260									

	$k = 7$			$k = 8$			$k = 9$			$k = 10$		
	α			α			α			α		
N	0.05	0.01	0.001	0.05	0.01	0.001	0.05	0.01	0.001	0.05	0.01	0.001
2	252	261	269	362	376	388	500	520	544	670	696	726
3	370	382	394	532	549	567	736	761	790	987	1019	1056
4	487	501	516	701	722	743	971	999	1032	1301	1339	1382
5	603	620	637	869	893	917	1204	1236	1273	1614	1656	1704
6	719	737	757	1037	1063	1090	1436	1472	1512	1927	1972	2025
7	835	855	876	1204	1232	1262	1668	1706	1750	2238	2288	2344
8	950	972	994	1371	1401	1433	1900	1940	1987	2549	2602	2662
9	1065	1088	1113	1537	1569	1603	2131	2174	2223	2859	2915	2980
10	1180	1205	1230	1703	1736	1773	2361	2407	2459	3169	3228	3296
11	1295	1321	1348	1868	1905	1943	2592	2639	2694	3478	3541	3612
12	1410	1437	1465	2035	2072	2112	2822	2872	2929	3788	3852	3927

* Adaptada de Page E. B., "Ordered hypotheses for multiple treatments: a significance test for linear ranks", en *Journal of the American Statistical Association*, núm. 58, 1963, págs. 216-230. Con autorización del autor y el editor.

Tabla O. Valores críticos para el análisis de varianza unifactorial por rangos de Kruskal-Wallis, *KW*.

Tamaños de las muestras			α				
n_1	n_2	n_3	0.10	0.05	0.01	0.005	0.001
2	2	2	4.25				
3	2	1	4.29				
3	2	2	4.71	4.71			
3	3	1	4.57	5.14			
3	3	2	4.56	5.36			
3	3	3	4.62	5.60	7.20	7.20	
4	2	1	4.50				
4	2	2	4.46	5.33			
4	3	1	4.06	5.21			
4	3	2	4.51	5.44	6.44	7.00	
4	3	3	4.71	5.73	6.75	7.32	8.02
4	4	1	4.17	4.97	6.67		
4	4	2	4.55	5.45	7.04	7.28	
4	4	3	4.55	5.60	7.14	7.59	8.32
4	4	4	4.65	5.69	7.66	8.00	8.65
5	2	1	4.20	5.00			
5	2	2	4.36	5.16	6.53		
5	3	1	4.02	4.96			
5	3	2	4.65	5.25	6.82	7.18	
5	3	3	4.53	5.65	7.08	7.51	8.24
5	4	1	3.99	4.99	6.95	7.36	
5	4	2	4.54	5.27	7.12	7.57	8.11
5	4	3	4.55	5.63	7.44	7.91	8.50
5	4	4	4.62	5.62	7.76	8.14	9.00
5	5	1	4.11	5.13	7.31	7.75	
5	5	2	4.62	5.34	7.27	8.13	8.68
5	5	3	4.54	5.71	7.54	8.24	9.06
5	5	4	4.53	5.64	7.77	8.37	9.32
5	5	5	4.56	5.78	7.98	8.72	9.68
Muestras grandes			4.61	5.99	9.21	10.60	13.82

Nota: La ausencia de una entrada en los extremos indica que la distribució·· puede no tomar los valores extremos necesarios.

Adaptada de la tabla F en Kraft, C. H. y van Eeden, C., *A nonparametric introduction to statistics*, Macmillan, Nueva York, 1968, con autorización del editor.

Tabla P. Valores críticos del estadístico J, de la prueba de Jonckheere
Las entradas son $P(J \geq$ valor de tablas) para $k = 3$ y $r_i \leq 8$ e iguales n $(2 \leq n \leq 6)$ para $k = 4, 5, 6, 7, 8$.

Tamaños de las muestras			α 0.10	0.05	0.01	0.005
2	2	2	10	11	12	--
2	2	3	13	14	15	16
2	2	4	16	17	19	20
2	2	5	18	20	22	23
2	2	6	21	23	25	27
2	2	7	24	26	29	30
2	2	8	27	29	32	33
2	3	3	16	18	19	20
2	3	4	20	21	23	25
2	3	5	23	25	27	29
2	3	6	26	28	31	33
2	3	7	30	32	35	37
2	3	8	33	35	39	41
2	4	4	24	25	28	29
2	4	5	27	29	33	34
2	4	6	31	34	37	39
2	4	7	35	38	42	44
2	4	8	39	42	46	49
2	5	5	32	34	38	40
2	5	6	36	39	43	45
2	5	7	41	44	48	51
2	5	8	45	48	53	56
2	6	6	42	44	49	51
2	6	7	47	50	55	57
2	6	8	52	55	61	64
2	7	7	52	56	61	64
2	7	8	58	62	68	71
2	8	8	64	68	75	78
3	3	3	20	22	24	25
3	3	4	24	26	29	30

Nota: Los valores críticos de la tabla se han escogido de tal forma que coincidan con niveles de significación redondeados; por ejemplo, un valor de J con una probabilidad ≤ 0.0149 es la entrada tabulada para el nivel de significación $\alpha = 0.01$.

Tabla P. (*Continuación*)

Tamaños de las muestras			α			
			0.10	0.05	0.01	0.005
3	3	5	28	30	33	35
3	3	6	32	34	38	40
3	3	7	36	38	42	44
3	3	8	40	42	47	49
3	4	4	29	31	34	36
3	4	5	33	35	39	41
3	4	6	38	40	44	46
3	4	7	42	45	49	52
3	4	8	47	50	55	57
3	5	5	38	41	45	47
3	5	6	43	46	51	53
3	5	7	48	51	57	59
3	5	8	53	57	63	65
3	6	6	49	52	57	60
3	6	7	54	58	64	67
3	6	8	60	64	70	73
3	7	7	61	64	71	74
3	7	8	67	71	78	81
3	8	8	74	78	86	89
4	4	4	34	36	40	42
4	4	5	39	41	45	48
4	4	6	44	47	51	54
4	4	7	49	52	57	60
4	4	8	54	57	63	66
4	5	5	44	47	52	55
4	5	6	50	53	58	61
4	5	7	56	59	65	68
4	5	8	61	65	71	75
4	6	6	56	60	66	69
4	6	7	62	66	73	76
4	6	8	68	73	80	83
4	7	7	69	73	81	84
4	7	8	76	80	88	92
4	8	8	83	88	97	100

Tabla P. (*Continuación*)

Tamaños de las muestras	α			
	0.10	0.05	0.01	0.005
5 5 5	50	54	59	62
5 5 6	57	60	66	69
5 5 7	63	67	73	76
5 5 8	69	73	80	84
5 6 6	63	67	74	77
5 6 7	70	74	82	85
5 6 8	77	81	89	93
5 7 7	77	82	90	94
5 7 8	85	89	98	102
5 8 8	92	98	107	111
6 6 6	71	75	82	86
6 6 7	78	82	91	94
6 6 8	85	90	99	103
6 7 7	86	91	100	103
6 7 8	94	99	109	113
6 8 8	102	108	118	122
7 7 7	94	99	109	113
7 7 8	102	108	119	123
7 8 8	111	117	129	133
8 8 8	121	127	139	144
2 2 2 2	18	19	21	22
2 2 2 2 2	28	30	33	34
2 2 2 2 2 2	40	43	46	49
3 3 3 3	37	39	43	45
3 3 3 3 3	58	62	68	70
3 3 3 3 3 3	85	89	97	101
4 4 4 4	63	66	72	76
4 4 4 4 4	100	105	115	119
4 4 4 4 4 4	146	153	166	171
5 5 5 5	95	100	109	113
5 5 5 5 5	152	159	173	178
5 5 5 5 5 5	223	233	251	258
6 6 6 6	134	140	153	158
6 6 6 6 6	215	225	243	250
6 6 6 6 6 6	316	329	353	362

Adaptada de Odeh, R. E., "On Jonckheere's *k*-sample test against ordered alternatives", en *Technometrics*, núm. 13, 1971, págs. 912-918, con autorización del autor y del editor; y de Jonckheere, A. R., "A distribution-free *k*-sample test against ordered alternatives", en *Biometrika*, núm. 41, 1951, págs. 133-145, con autorización del consejo de administración de *Biometrika*.

Tabla Q. Valores críticos del coeficiente de correlación de Spearman r_s, de rangos ordenados.

N	α 0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005	*unidireccional*
	α 0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001	*bidireccional*
4	0.600	1.000	1.000							
5	0.500	0.800	0.900	1.000	1.000					
6	0.371	0.657	0.829	0.886	0.943	1.000	1.000			
7	0.321	0.571	0.714	0.786	0.893	0.929	0.964	1.000	1.000	
8	0.310	0.524	0.643	0.738	0.833	0.881	0.905	0.952	0.976	
9	0.267	0.483	0.600	0.700	0.783	0.833	0.867	0.917	0.933	
10	0.248	0.455	0.564	0.648	0.745	0.794	0.830	0.879	0.903	
11	0.236	0.427	0.536	0.618	0.709	0.755	0.800	0.845	0.873	
12	0.224	0.406	0.503	0.587	0.671	0.727	0.776	0.825	0.860	
13	0.209	0.385	0.484	0.560	0.648	0.703	0.747	0.802	0.835	
14	0.200	0.367	0.464	0.538	0.622	0.675	0.723	0.776	0.811	
15	0.189	0.354	0.443	0.521	0.604	0.654	0.700	0.754	0.786	
16	0.182	0.341	0.429	0.503	0.582	0.635	0.679	0.732	0.765	
17	0.176	0.328	0.414	0.485	0.566	0.615	0.662	0.713	0.748	
18	0.170	0.317	0.401	0.472	0.550	0.600	0.643	0.695	0.728	
19	0.165	0.309	0.391	0.460	0.535	0.584	0.628	0.677	0.712	
20	0.161	0.299	0.380	0.447	0.520	0.570	0.612	0.662	0.696	
21	0.156	0.292	0.370	0.435	0.508	0.556	0.599	0.648	0.681	
22	0.152	0.284	0.361	0.425	0.496	0.544	0.586	0.634	0.667	
23	0.148	0.278	0.353	0.415	0.486	0.532	0.573	0.622	0.654	
24	0.144	0.271	0.344	0.406	0.476	0.521	0.562	0.610	0.642	
25	0.142	0.265	0.337	0.398	0.466	0.511	0.551	0.598	0.630	
26	0.138	0.259	0.331	0.390	0.457	0.501	0.541	0.587	0.619	
27	0.136	0.255	0.324	0.382	0.448	0.491	0.531	0.577	0.608	
28	0.133	0.250	0.317	0.375	0.440	0.483	0.522	0.567	0.598	
29	0.130	0.245	0.312	0.368	0.433	0.475	0.513	0.558	0.589	
30	0.128	0.240	0.306	0.362	0.425	0.467	0.504	0.549	0.580	
31	0.126	0.236	0.301	0.356	0.418	0.459	0.496	0.541	0.571	
32	0.124	0.232	0.296	0.350	0.412	0.452	0.489	0.533	0.563	
33	0.121	0.229	0.291	0.345	0.405	0.446	0.482	0.525	0.554	
34	0.120	0.225	0.287	0.340	0.399	0.439	0.475	0.517	0.547	
35	0.118	0.222	0.283	0.335	0.394	0.433	0.468	0.510	0.539	
36	0.116	0.219	0.279	0.330	0.388	0.427	0.462	0.504	0.533	
37	0.114	0.216	0.275	0.325	0.383	0.421	0.456	0.497	0.526	
38	0.113	0.212	0.271	0.321	0.378	0.415	0.450	0.491	0.519	
39	0.111	0.210	0.267	0.317	0.373	0.410	0.444	0.485	0.513	
40	0.110	0.207	0.264	0.313	0.368	0.405	0.439	0.479	0.507	
41	0.108	0.204	0.261	0.309	0.364	0.400	0.433	0.473	0.501	
42	0.107	0.202	0.257	0.305	0.359	0.395	0.428	0.468	0.495	
43	0.105	0.199	0.254	0.301	0.355	0.391	0.423	0.463	0.490	
44	0.104	0.197	0.251	0.298	0.351	0.386	0.419	0.458	0.484	
45	0.103	0.194	0.248	0.294	0.347	0.382	0.414	0.453	0.479	
46	0.102	0.192	0.246	0.291	0.343	0.378	0.410	0.448	0.474	
47	0.101	0.190	0.243	0.288	0.340	0.374	0.405	0.443	0.469	
48	0.100	0.188	0.240	0.285	0.336	0.370	0.401	0.439	0.465	
49	0.098	0.186	0.238	0.282	0.333	0.366	0.397	0.434	0.460	
50	0.097	0.184	0.235	0.279	0.329	0.363	0.393	0.430	0.456	

Fuente: Zar, J. H., "Significance testing of the Spearman rank correlation coefficient", en *Journal of the American Statistical Association*, núm. 67, 1972, págs. 578-580. Adaptada con autorización del autor y del editor.

Tabla $\mathbf{R_1}$. Probabilidades del lado superior para T, del coeficiente de correlación de Kendall de rangos ordenados ($N \leq 10$).*
Las entradas son $p = P[T \geq \text{valor de tabla}]$.

N	T	P	N	T	P	N	T	P	N	T	P
4	0.000	0.625	7	0.048	0.500	9	0.000	0.540	10	0.022	0.500
	0.333	0.375		0.143	0.386		0.056	0.460		0.067	0.431
	0.667	0.167		0.238	0.281		0.111	0.381		0.111	0.364
	1.000	0.042		0.333	0.191		0.167	0.306		0.156	0.300
				0.429	0.119		0.222	0.238		0.200	0.242
5	0.000	0.592		0.524	0.068		0.278	0.179		0.244	0.190
	0.200	0.408		0.619	0.035		0.333	0.130		0.289	0.146
	0.400	0.242		0.714	0.015		0.389	0.090		0.333	0.108
	0.600	0.117		0.810	0.005		0.444	0.060		0.378	0.078
	0.800	0.042		0.905	0.001		0.500	0.038		0.422	0.054
	1.000	0.008		1.000	0.000		0.556	0.022		0.467	0.036
							0.611	0.012		0.511	0.023
6	0.067	0.500	8	0.000	0.548		0.667	0.006		0.556	0.014
	0.200	0.360		0.071	0.452		0.722	0.003		0.600	0.008
	0.333	0.235		0.143	0.360		0.778	0.001		0.644	0.005
	0.467	0.136		0.214	0.274		0.833	0.000		0.689	0.002
	0.600	0.068		0.286	0.199		0.889	0.000		0.733	0.001
	0.733	0.028		0.357	0.138		0.944	0.000		0.778	0.000
	0.867	0.008		0.429	0.089		1.000	0.000		0.822	0.000
	1.001	0.001		0.500	0.054					0.867	0.000
				0.571	0.031					0.911	0.000
				0.643	0.016					0.956	0.000
				0.714	0.007					1.000	0.000
				0.786	0.003						
				0.857	0.001						
				0.929	0.000						
				1.000	0.000						

* Adaptada y reproducida con autorización de los editores Charles Griffin & Co. Ltd., 16 Pembridge Road, Londres W11 3HL, de la tabla 5 del Apéndice de Kendall, M. G., *Rank correlation methods*, 4a. ed., 1970.

Tabla R_{II}. Valores críticos para T, el coeficiente de correlación de Kendall* de rangos ordenados.
Las entradas son valores de T tales que $P[T \geq \text{valor de tabla}] \leq \alpha$.

N	α unidireccional: 0.100	0.050	0.025	0.010	0.005
	α bidireccional: 0.200	0.100	0.050	0.020	0.010
11	0.345	0.418	0.491	0.564	0.600
12	0.303	0.394	0.455	0.545	0.576
13	0.308	0.359	0.436	0.513	0.564
14	0.275	0.363	0.407	0.473	0.516
15	0.276	0.333	0.390	0.467	0.505
16	0.250	0.317	0.383	0.433	0.483
17	0.250	0.309	0.368	0.426	0.471
18	0.242	0.294	0.346	0.412	0.451
19	0.228	0.287	0.333	0.392	0.439
20	0.221	0.274	0.326	0.379	0.421
21	0.210	0.267	0.314	0.371	0.410
22	0.195	0.253	0.295	0.344	0.378
23	0.202	0.257	0.296	0.352	0.391
24	0.196	0.246	0.290	0.341	0.377
25	0.193	0.240	0.287	0.333	0.367
26	0.188	0.237	0.280	0.329	0.360
27	0.179	0.231	0.271	0.322	0.356
28	0.180	0.228	0.265	0.312	0.344
29	0.172	0.222	0.261	0.310	0.340
30	0.172	0.218	0.255	0.301	0.333

* Adaptada y reproducida con autorización de los editores Charles Griffin & Co. Ltd., 16 Pembridge Road, Londres W11 3HL, de la tabla 5 del Apéndice de Kendall, M. G., *Rank correlation methods*, 4a. ed., 1970.

Tabla S. Valores críticos para el coeficiente de correlación parcial $T_{xy.z}$ de Kendall de rangos ordenados.*

	α							
N	0.25	0.20	0.10	0.05	0.025	0.01	0.005	0.001
3	0.500	1.000						
4	0.447	0.500	0.707	0.707	1.000			
5	0.333	0.408	0.534	0.667	0.802	0.816	1.000	
6	0.277	0.327	0.472	0.600	0.667	0.764	0.866	1.000
7	0.233	0.282	0.421	0.527	0.617	0.712	0.761	0.901
8	0.206	0.254	0.382	0.484	0.565	0.648	0.713	0.807
9	0.187	0.230	0.347	0.443	0.515	0.602	0.660	0.757
10	0.170	0.215	0.325	0.413	0.480	0.562	0.614	0.718
11	0.162	0.202	0.305	0.387	0.453	0.530	0.581	0.677
12	0.153	0.190	0.288	0.465	0.430	0.505	0.548	0.643
13	0.145	0.180	0.273	0.347	0.410	0.481	0.527	0.616
14	0.137	0.172	0.260	0.331	0.391	0.458	0.503	0.590
15	0.133	0.166	0.251	0.319	0.377	0.442	0.485	0.570
16	0.125	0.157	0.240	0.305	0.361	0.423	0.466	0.549
17	0.121	0.151	0.231	0.294	0.348	0.410	0.450	0.532
18	0.117	0.147	0.222	0.284	0.336	0.395	0.434	0.514
19	0.114	0.141	0.215	0.275	0.326	0.382	0.421	0.498
20	0.111	0.139	0.210	0.268	0.318	0.374	0.412	0.488
25	0.098	0.122	0.185	0.236	0.279	0.329	0.363	0.430
30	0.088	0.110	0.167	0.213	0.253	0.298	0.329	0.390
35	0.081	0.101	0.153	0.196	0.232	0.274	0.303	0.361
40	0.075	0.094	0.142	0.182	0.216	0.255	0.282	0.335
45	0.071	0.088	0.133	0.171	0.203	0.240	0.265	0.316
50	0.067	0.083	0.126	0.161	0.192	0.225	0.250	0.298
60	0.060	0.075	0.114	0.147	0.174	0.206	0.227	0.270
70	0.056	0.070	0.106	0.135	0.160	0.190	0.210	0.251
80	0.052	0.065	0.098	0.126	0.150	0.178	0.197	0.235
90	0.049	0.061	0.092	0.119	0.141	0.167	0.185	0.221

* Adaptada de Maghsoodloo, S., "Estimates of the quantiles of Kendall's partial rank correlation coefficient", en *Journal of Statistical Computing and Simulation*, núm. 4, 1975, págs. 155-164; y Maghsoodloo, S. y Pallos, L. L., "Asymptotic behavior of Kendall's partial rank correlation coefficient and additional quantile estimates", en *Journal of Statistical Computing and Simulation*, núm. 13, 1981, págs. 41-48, por cortesía del autor y el editor.

Tabla T. Valores críticos del coeficiente de acuerdos W de Kendall.*

$N = 3$		
K	α 0.05	0.01
8	0.376	0.522
9	0.333	0.469
10	0.300	0.425
12	0.250	0.359
14	0.214	0.311
15	0.200	0.291
16	0.187	0.274
18	0.166	0.245
20	0.150	0.221

	$N = 4$		$N = 5$		$N = 6$		$N = 7$	
K	α 0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
3	–	–	0.716	0.840	0.660	0.780	0.624	0.737
4	0.619	0.768	0.552	0.683	0.512	0.629	0.484	0.592
5	0.501	0.644	0.449	0.571	0.417	0.524	0.395	0.491
6	0.421	0.553	0.378	0.489	0.351	0.448	0.333	0.419
8	0.318	0.429	0.287	0.379	0.267	0.347	0.253	0.324
10	0.256	0.351	0.231	0.309	0.215	0.282	0.204	0.263
15	0.171	0.240	0.155	0.211	0.145	0.193	0.137	0.179
20	0.129	0.182	0.117	0.160	0.109	0.146	0.103	0.136

Nota: Para $N = 3$ y $k < 8$ no existe valor de W en el lado superior que tenga una probabilidad de ocurrencia menor que 0.05.

* Adaptada y reproducida con autorización de los editores Charles Griffin & Co. Ltd., 16 Pembridge Road, Londres W11 3HL, de la tabla del Apéndice de Kendall, M. G., *Rank correlation methods*, 4a. ed., 1970.

Tabla U. Probabilidades del lado superior del coeficiente de acuerdos *u* de Kendall cuando los datos corresponden a comparaciones apareadas.

$K = 3$			
N	S	u	p
2	1	−0.333	1.0000
	3	1.000	0.2500
3	5	0.111	0.5781
	7	0.556	0.1563
	9	1.000	0.0156
4	10	0.111	0.4661
	12	0.333	0.1694
	14	0.556	0.0376
	16	0.778	0.0046
	18	1.000	0.0002
5	16	0.067	0.4744
	18	0.200	0.2241
	20	0.333	0.0781
	22	0.467	0.0197
	24	0.600	0.0035
	26	0.733	0.0004
6	23	0.022	0.5387
	25	0.111	0.3135
	27	0.200	0.1484
	29	0.289	0.0566
	31	0.378	0.0173
	33	0.467	0.0042
	35	0.556	0.0008
	37	0.644	0.0001
7	33	0.048	0.4334
	35	0.111	0.2564
	37	0.175	0.1299
	39	0.238	0.0561
	41	0.302	0.0206
	43	0.365	0.0064
	45	0.429	0.0017
	47	0.492	0.0004
	49	0.556	0.0001

$K = 3$			
N	S	u	p
8	42	0.000	0.5721
	44	0.048	0.4003
	46	0.095	0.2499
	48	0.143	0.1385
	50	0.190	0.0679
	52	0.238	0.0294
	54	0.286	0.0112
	56	0.333	0.0038
	58	0.381	0.0011
	60	0.429	0.0003
	62	0.476	0.0001

$K = 4$			
N	S	u	p
2	2	−0.333	1.0000
	3	0.000	0.6250
	6	1.000	0.1250
3	9	0.000	0.4551
	10	0.111	0.3301
	11	0.222	0.2773
	12	0.333	0.1367
	14	0.556	0.0430
	15	0.667	0.0254
	18	1.000	0.0020
4	18	0.000	0.5242
	19	0.056	0.4097
	20	0.111	0.2779
	21	0.167	0.1853
	22	0.222	0.1372
	23	0.278	0.0877
	24	0.333	0.0438
	25	0.389	0.0271
	26	0.444	0.0188
	27	0.500	0.0079
	28	0.556	0.0030
	29	0.611	0.0025
	30	0.667	0.0011
	32	0.778	0.0002
	33	0.833	0.0001

Tabla U. (*Continuación*)

		$K = 4$	
N	S	u	p
5	30	0.000	0.5137
	31	0.033	0.4126
	32	0.067	0.3266
	33	0.100	0.2491
	34	0.133	0.1795
	35	0.167	0.1271
	36	0.200	0.0903
	37	0.233	0.0604
	38	0.267	0.0376
	39	0.300	0.0242
	40	0.333	0.0156
	41	0.367	0.0088
	42	0.400	0.0048
	43	0.433	0.0030
	44	0.467	0.0017
	45	0.500	0.0007
	46	0.533	0.0004
	47	0.567	0.0002
	48	0.600	0.0001
6	45	0.000	0.5134
	46	0.022	0.4310
	47	0.044	0.3532
	48	0.067	0.2837
	49	0.089	0.2231
	50	0.111	0.1708
	51	0.133	0.1277
	52	0.156	0.0939
	53	0.178	0.0676
	54	0.200	0.0472
	55	0.222	0.0324
	56	0.244	0.0219
	57	0.267	0.0145
	58	0.289	0.0092
	59	0.311	0.0058
	60	0.333	0.0037
	61	0.356	0.0022
	62	0.378	0.0013
	63	0.400	0.0008
	64	0.422	0.0004
	65	0.444	0.0002
	66	0.467	0.0001
	67	0.489	0.0001

		$K = 4$	
N	S	u	p
7	63	0.000	0.5111
	64	0.016	0.4413
	65	0.032	0.3746
	66	0.048	0.3124
	67	0.063	0.2562
	68	0.079	0.2066
	69	0.095	0.1637
	70	0.111	0.1275
	71	0.127	0.0977
	72	0.143	0.0736
	73	0.159	0.0545
	74	0.175	0.0397
	75	0.190	0.0285
	76	0.206	0.0201
	77	0.222	0.0139
	78	0.238	0.0095
	79	0.254	0.0064
	80	0.270	0.0042
	81	0.286	0.0028
	82	0.302	0.0018
	83	0.317	0.0011
	84	0.333	0.0007
	85	0.349	0.0004
	86	0.365	0.0003
	87	0.381	0.0002
	88	0.397	0.0001
	89	0.413	0.0001
8	84	0.000	0.5098
	85	0.012	0.4490
	86	0.024	0.3903
	87	0.036	0.3348
	88	0.048	0.2833
	89	0.060	0.2366
	90	0.071	0.1949
	91	0.083	0.1585
	92	0.095	0.1271
	93	0.107	0.1006
	94	0.119	0.0786
	95	0.131	0.0606
	96	0.143	0.0461

Tabla U. (*Continuación*)

		$K = 4$	
N	S	u	p
8	97	0.155	0.1346
	98	0.167	0.0257
	99	0.179	0.0188
	100	0.190	0.0136
	101	0.202	0.0097
	102	0.214	0.0068
	103	0.226	0.0048
	104	0.238	0.0033
	105	0.250	0.0022
	106	0.262	0.0015
	107	0.274	0.0010
	108	0.286	0.0007
	109	0.298	0.0004
	110	0.310	0.0003
	111	0.321	0.0002
	112	0.333	0.0001
	113	0.345	0.0001

		$K = 5$	
N	S	u	p
2	4	−0.200	1.0000
	6	0.200	0.3750
	10	1.000	0.0620
3	16	0.067	0.3896
	18	0.200	0.2065
	20	0.333	0.1028
	22	0.467	0.0295
	24	0.600	0.0112
	26	0.733	0.0039
	30	1.000	0.0002
4	30	0.000	0.5381
	32	0.067	0.3533
	34	0.133	0.2080
	36	0.200	0.1074
	38	0.267	0.0528
	40	0.333	0.0238
	42	0.400	0.0093
	44	0.467	0.0039
	46	0.533	0.0012
	48	0.600	0.0004
	50	0.667	0.0001

		$K = 5$	
N	S	u	p
5	52	0.040	0.3838
	54	0.080	0.2544
	56	0.120	0.1579
	58	0.160	0.0918
	60	0.200	0.0500
	62	0.240	0.0257
	64	0.280	0.0124
	66	0.320	0.0057
	68	0.360	0.0025
	70	0.400	0.0010
	72	0.440	0.0004
	74	0.480	0.0001
6	76	0.013	0.4663
	78	0.040	0.3453
	80	0.067	0.2428
	82	0.093	0.1623
	84	0.120	0.1034
	86	0.147	0.0628
	88	0.173	0.0364
	90	0.200	0.0202
	92	0.227	0.0108
	94	0.253	0.0055
	96	0.280	0.0027
	98	0.307	0.0013
	100	0.333	0.0006
	102	0.360	0.0002
	104	0.387	0.0001
7	106	0.010	0.4718
	108	0.029	0.3674
	110	0.048	0.2750
	112	0.067	0.1980
	114	0.086	0.1372
	116	0.105	0.0916
	118	0.124	0.0589
	120	0.143	0.0366
	122	0.162	0.0220
	124	0.181	0.0128
	126	0.200	0.0072
	128	0.219	0.0039
	130	0.238	0.0021

Tabla U. (*Continuación*)

	$K = 5$		
N	S	u	p
7	132	0.257	0.0011
	134	0.276	0.0005
	136	0.295	0.0003
	138	0.314	0.0001
	140	0.333	0.0001
8	140	0.000	0.5233
	142	0.014	0.4291
	144	0.029	0.3411
	146	0.043	0.2629
	148	0.057	0.1965
	150	0.071	0.1425
	152	0.086	0.1003
	154	0.100	0.0686
	156	0.114	0.0456
	158	0.129	0.0294
	160	0.143	0.0185
	162	0.157	0.0113
	164	0.171	0.0068
	166	0.186	0.0039
	168	0.200	0.0022
	170	0.214	0.0012
	172	0.229	0.0007
	174	0.243	0.0003
	176	0.257	0.0002
	178	0.271	0.0001

	$k = 6$		
N	S	u	p
2	6	−0.200	1.0000
	7	−0.067	0.6875
	10	0.333	0.2188
	15	1.000	0.0312
3	23	0.022	0.4682
	24	0.067	0.3034
	26	0.156	0.1798
	27	0.200	0.1469
	28	0.244	0.0883
	29	0.289	0.0608
	30	0.333	0.0402

	$K = 6$		
N	S	u	p
3	31	0.378	0.0336
	32	0.422	0.0226
	35	0.556	0.0062
	36	0.600	0.0029
	37	0.644	0.0020
	40	0.778	0.0006
4	45	0.000	0.4656
	46	0.022	0.4094
	47	0.044	0.3374
	48	0.067	0.2569
	49	0.089	0.2086
	50	0.111	0.1746
	51	0.133	0.1332
	52	0.156	0.0970
	53	0.178	0.0725
	54	0.200	0.0566
	55	0.222	0.0433
	56	0.244	0.0289
	57	0.267	0.0198
	58	0.289	0.0160
	59	0.311	0.0114
	60	0.333	0.0072
	61	0.356	0.0049
	62	0.378	0.0034
	63	0.400	0.0025
	64	0.422	0.0016
	65	0.444	0.0008
	66	0.467	0.0007
	67	0.489	0.0005
	68	0.511	0.0003
	69	0.533	0.0002
	70	0.556	0.0001
	71	0.578	0.0001
5	75	0.000	0.4841
	76	0.013	0.4258
	77	0.027	0.3665
	78	0.040	0.3085
	79	0.053	0.2600
	80	0.067	0.2190
	81	0.080	0.1800

Tabla U. (*Continuación*)

		$K = 6$				$K = 6$	
N	S	u	p	N	S	u	p
5	82	0.093	0.1452	6	128	0.138	0.0352
	83	0.107	0.1173		129	0.147	0.0280
	84	0.120	0.0949		130	0.156	0.0221
	85	0.133	0.0753		131	0.164	0.0173
	86	0.147	0.0583		132	0.173	0.0135
	87	0.160	0.0452		133	0.182	0.0105
	88	0.173	0.0355		134	0.191	0.0081
	89	0.187	0.0272		135	0.200	0.0062
	90	0.200	0.0202		136	0.209	0.0047
	91	0.213	0.0151		137	0.218	0.0036
	92	0.227	0.0115		138	0.227	0.0027
	93	0.240	0.0085		139	0.236	0.0020
	94	0.253	0.0062		140	0.244	0.0015
	95	0.267	0.0044		141	0.253	0.0011
	96	0.280	0.0033		142	0.262	0.0008
	97	0.293	0.0024		143	0.271	0.0006
	98	0.307	0.0017		144	0.280	0.0004
	99	0.320	0.0011		145	0.289	0.0003
	100	0.333	0.0008		146	0.298	0.0002
	101	0.347	0.0006		147	0.307	0.0002
	102	0.360	0.0004		148	0.316	0.0001
	103	0.373	0.0003		149	0.324	0.0001
	104	0.387	0.0002		150	0.333	0.0001
	105	0.400	0.0001				
	106	0.413	0.0001	7	158	0.003	0.4694
	107	0.427	0.0001		159	0.010	0.4258
					160	0.016	0.3838
6	113	0.004	0.4640		161	0.022	0.3436
	114	0.013	0.4126		162	0.029	0.3057
	115	0.022	0.3637		163	0.035	0.2703
	116	0.031	0.3186		164	0.041	0.2375
	117	0.040	0.2768		165	0.048	0.2074
	118	0.049	0.2380		166	0.054	0.1800
	119	0.058	0.2030		167	0.060	0.1553
	120	0.067	0.1723		168	0.067	0.1332
	121	0.076	0.1451		169	0.073	0.1136
	122	0.084	0.1209		170	0.079	0.0963
	123	0.093	0.1000		171	0.086	0.0812
	124	0.102	0.0824		172	0.092	0.0680
	125	0.111	0.0674		173	0.098	0.0567
	126	0.120	0.0546		174	0.105	0.0470
	127	0.129	0.0439		175	0.111	0.0388

Tabla U. (*Continuación*)

		$K = 6$				$K = 6$	
N	S	u	p	N	S	u	p
7	176	0.117	0.0318	8	226	0.076	0.0742
	177	0.124	0.0260		227	0.081	0.0633
	178	0.130	0.0211		228	0.086	0.0538
	179	0.137	0.0170		229	0.090	0.0455
	180	0.143	0.0137		230	0.095	0.0383
	181	0.149	0.0110		231	0.100	0.0321
	182	0.156	0.0087		232	0.105	0.0268
	183	0.162	0.0069		233	0.110	0.0223
	184	0.168	0.0054		234	0.114	0.0185
	185	0.175	0.0043		235	0.119	0.0152
	186	0.181	0.0033		236	0.124	0.0125
	187	0.187	0.0026		237	0.129	0.0102
	188	0.194	0.0020		238	0.133	0.0083
	189	0.200	0.0015		239	0.138	0.0068
	190	0.206	0.0012		240	0.143	0.0055
	191	0.213	0.0009		241	0.148	0.0044
	192	0.219	0.0007		242	0.152	0.0035
	193	0.225	0.0005		243	0.157	0.0028
	194	0.232	0.0004		244	0.162	0.0022
	195	0.238	0.0003		245	0.167	0.0018
	196	0.244	0.0002		246	0.171	0.0014
	197	0.251	0.0002		247	0.176	0.0011
	198	0.257	0.0001		248	0.181	0.0009
	199	0.263	0.0001		249	0.186	0.0007
	200	0.270	0.0001		250	0.190	0.0005
					251	0.195	0.0004
8	210	0.000	0.4930		252	0.200	0.0003
	211	0.005	0.4545		253	0.205	0.0002
	212	0.010	0.4169		254	0.210	0.0002
	213	0.014	0.3805		255	0.214	0.0001
	214	0.019	0.3455		256	0.219	0.0001
	215	0.024	0.3122		257	0.224	0.0001
	216	0.029	0.2807		258	0.229	0.0001
	217	0.033	0.2511				
	218	0.038	0.2235				
	219	0.043	0.1980				
	220	0.048	0.1745				
	221	0.052	0.1531				
	222	0.057	0.1337				
	223	0.062	0.1162				
	224	0.067	0.1005				
	225	0.071	0.0866				

Nota: Los valores presentados corresponden a las probabilidades $c \geq 0.0001$ (redondeadas). Así, las probabilidades del lado superior para valores grandes de u tienen probabilidades < 0.00005.

Tabla V. Probabilidades del lado superior de T_c, la correlación de k rangos con un criterio de ordenamiento por rangos.*

$k = 2$			$k = 3$		
N	T_c	p	N	T_c	p
2	0.000	0.750	2	0.333	0.500
	1.000	0.250		1.000	0.125
3	0.000	0.639	3	0.111	0.500
	0.333	0.361		0.333	0.278
	0.667	0.139		0.556	0.116
	1.000	0.028		0.778	0.033
				1.000	0.005
4	0.000	0.592	4	0.000	0.576
	0.167	0.408		0.111	0.424
	0.333	0.241		0.222	0.282
	0.500	0.118		0.333	0.167
	0.667	0.045		0.444	0.086
	0.833	0.012		0.556	0.038
	1.000	0.002		0.667	0.014
				0.778	0.004
				0.889	0.001
				1.000	0.000+
5	0.000	0.567	5	0.000	0.556
	0.100	0.433		0.067	0.445
	0.200	0.306		0.113	0.338
	0.300	0.198		0.200	0.243
	0.400	0.116		0.267	0.164
	0.500	0.060		0.333	0.103
	0.600	0.027		0.400	0.060
	0.700	0.010		0.467	0.032
	0.800	0.003		0.533	0.016
	0.900	0.001		0.600	0.007
	1.000	0.000+		0.667	0.003
				0.733	0.001
				0.800	0.000+

* Adaptada de Stilson, D. W. y Campbell, V. N., "A note on calculating tau and average tau and on the sampling distribution of average tau with a criterion ranking", en *Journal of the American Statistical Association*, núm. 57, 1962, págs. 567-571. Con autorización del autor y del editor.

Tabla W. Factoriales.

N	$N!$
0	1
1	1
2	2
3	6
4	24
5	120
6	720
7	5040
8	40320
9	362880
10	3628800
11	39916800
12	479001600
13	6227020800
14	87178291200
15	1307674368000
16	20922789888000
17	355687428096000
18	6402373705728000
19	121645100408832000
20	2432902008176640000

Tabla X. Coeficientes binomiales.

N	$\binom{N}{0}$	$\binom{N}{1}$	$\binom{N}{2}$	$\binom{N}{3}$	$\binom{N}{4}$	$\binom{N}{5}$	$\binom{N}{6}$	$\binom{N}{7}$	$\binom{N}{8}$	$\binom{N}{9}$	$\binom{N}{10}$
0	1										
1	1	1									
2	1	2	1								
3	1	3	3	1							
4	1	4	6	4	1						
5	1	5	10	10	5	1					
6	1	6	15	20	15	6	1				
7	1	7	21	35	35	21	7	1			
8	1	8	28	56	70	56	28	8	1		
9	1	9	36	84	126	126	84	36	9	1	
10	1	10	45	120	210	252	210	120	45	10	1
11	1	11	55	165	330	462	462	330	165	55	11
12	1	12	66	220	495	792	924	792	495	220	66
13	1	13	78	286	715	1287	1716	1716	1287	715	286
14	1	14	91	364	1001	2002	3003	3432	3003	2002	1001
15	1	15	105	455	1365	3003	5005	6435	6435	5005	3003
16	1	16	120	560	1820	4368	8008	11440	12870	11440	8008
17	1	17	136	680	2380	6188	12376	19448	24310	24310	19448
18	1	18	153	816	3060	8568	18564	31824	43758	48620	43758
19	1	19	171	969	3876	11628	27132	50388	75582	92378	92378
20	1	20	190	1140	4845	15504	38760	77520	125970	167960	184756

Apéndice II. Programas

1. Caso de una muestra: prueba para la simetría.
2. Una muestra, dos medidas: prueba de las permutaciones para pares replicados.
3. Dos muestras independientes: prueba de las permutaciones para dos muestras independientes.
4. k muestras independientes: prueba ji cuadrada para tablas de contingencia $r \times k$ con partición.
5. k muestras independientes: prueba de Jonckheere para niveles ordenados de la variable.

Programa 1

Caso de una muestra: prueba para la simetría.

```
100 REM
110 REM Randles, R. H., Fligner, M. A., Policello, G. E., and Wolfe, D. A.,
120 REM An Asymptotically Distribution-Free Test for Symmetry Versus Asymmetry,
130 REM Journal of the American Statistical Association, 1980, 75, 168-172.
140 REM Coded by N. J. Castellan, Jr., Copyright 1982.
150 PRINT"An Asymptotically Distribution-Free Test for Symmetry Versus Asymmetry"
160 INPUT "How many observations (N)"; N
170 DIM X(N),T1(N),T2(N,N)
180 REM -- read data
190 PRINT "Enter the data values one at a time."
200 FOR I=1 TO N : INPUT X(I) : NEXT I
210 REM -- begin computations
220 REM -- this program assumes that all variables and arrays
230 REM -- are initialized to zero's
240 FOR I=1 TO N-2
250      FOR J=I+1 TO N-1
260           FOR K=J+1 TO N
270                AVE = (X(I)+X(J)+X(K))/3
280                MIN=X(I) : MED=X(J) : MAX=X(K)
290                IF MIN>MED THEN SWAP MIN,MED
300                IF MED>MAX THEN SWAP MED,MAX
310                IF MIN>MED THEN SWAP MIN,MED
320                IF AVE>MED THEN RL=1 ELSE IF AVE<MED THEN RL=-1 ELSE RL=0
330                T = T+RL
340                T1(I)=T1(I)+RL
350                T1(J)=T1(J)+RL
360                T1(K)=T1(K)+RL
370                T2(I,J)=T2(I,J)+RL
380                T2(I,K)=T2(I,K)+RL
390                T2(J,K)=T2(J,K)+RL
400           NEXT K
410      NEXT J
420 NEXT I
430 B1=T1(N)^2 : B2=0
440 FOR I=1 TO N-1
450      B1=B1+T1(I)^2
460      FOR J=I+1 TO N
470           B2=B2+T2(I,J)^2
480      NEXT J
490 NEXT I
500 PRINT "N =";N;", Sum B2(i) =";B1;", Sum B2(ij) =";B2
510 VAR = B1*(N-3)*(N-4)/((N-1)*(N-2)) + B2*(N-3)/(N-4) + N*(N-1)*(N-2)/6
520 VAR = VAR - (1 - (N-3)*(N-4)*(N-5)/(N*(N-1)*(N-2))*T^2
530 PRINT "T =";T;", Var(T) =";VAR;", z = T/sqr(var(T)) =";T/SQR(VAR)
540 END
```

Programa 2

Una muestra, dos medidas: prueba de las permutaciones para pares replicados.

```
' THE PERMUTATION TEST FOR PAIRED REPLICATES.
' Coded in QuickBASIC.  Copyright 1987 N. John Castellan, Jr.
' Algorithm will work if number of pairs of data (N) < 15
' Note: This limit is not checked by the program.
' For larger sample sizes remove the DEFINT statement.
' (Removing DEFINT allows larger samples sizes at the expense of
' increased execution time.)
' This version of the program has not been optimized to minimize
' the number of iterations.  (This was done to make program more readable.)
' Integrated package version is optimized and handles large N.
DEFINT I,N,W,U ' Remove this statement if N>14.
UPPERTAIL=0 : NPERM=0: CRIT=0
INPUT "What is the sample size";N
DIM D(N,2),INDEX(N)
PRINT "Input the data, pair by pair (two entries separated by a comma):"
' The following data are from example in Section 5.4
DATA 82,63, 69,42, 73,74, 43,37, 58,51, 56,43, 76,80, 85,82
FOR I=1 TO N
    ' After debugging, insert a ' before the following READ statement and
    ' delete the ' from the next line so data may be entered from keyboard.
    READ D1,D2:D(I,1)=D1-D2:D(I,2)=D2-D1
    ' INPUT D1,D2 : D(I,1)=D1-D2 : D(I,2)=D2-D1
    CRIT=CRIT+D(I,1)
    INDEX(I)=1
NEXT I
LOOP1:
      SUM=0
      FOR I=1 TO N
          SUM=SUM+D(I,INDEX(I))
      NEXT I
      NPERM=NPERM+1
      IF SUM>=CRIT THEN UPPERTAIL=UPPERTAIL+1
      I=N
      WHILE I>0
            IF INDEX(I)=1 THEN INDEX(I)=2 : GOTO LOOP1
            INDEX(I)=1 ' Reset index
            I=I-1
      WEND
' Calculations done, print summary
PRINT "PERMUTATION TEST FOR PAIRED REPLICATES"
PRINT USING "Observed sum of differences = #####.##";CRIT
PRINT USING "Number of sums >= observed sum: ##### out of ##### sums.";UPPERTAIL
,NPERM
PRINT USING "Upper Tail Probability = #.####";UPPERTAIL/NPERM
END
```

Salida de muestra para el programa 2

Los datos siguientes son del ejemplo de la página 123.

```
What is the sample size? 8
Input the data, pair by pair (two entries separated by a comma):
? 82.63
? 69.42
? 73.74
? 43.37
? 58.51
? 56.43
? 76.80
? 85.82

PERMUTATION TEST FOR PAIRED REPLICATES

Observed sum of differences =    70.00

Number of sums >= observed sum:     6 out of   256 sums.

Upper Tail Probability = 0.0234


Normal termination. Press any key.
```

Programa 3

Dos muestras independientes: prueba de las permutaciones para dos muestras independientes.

```
' PERMUTATION TEST FOR TWO INDEPENDENT SAMPLES.
' Coded in QuickBASIC.  Copyright 1987 N. J. Castellan, Jr.
' Algorithm will work if number of permutations is less than 32768
' Limits:  M + N < 18
'          or (M=7, N<12), (M=6, N<14), (M=5, N<18), (M=4, N<28), (M=3, N<57)
' Note: These limits are not checked by program.
' Program will run significantly faster if the smaller group is entered first.
' This version of the program has not been optimized to minimize
' the number of iterations.  (This was done to make program more readable.)
' Integrated package version is optimized and handles large M and N.
DEFINT I,M,N,U
INPUT "What are the samples sizes";M,N
MN=M+N
DIM X(MN), INDEX(MN)
PRINT "Input the data for Group 1. (One datum at a time.)"
FOR I=1 TO M : INPUT X(I) : NEXT I
PRINT "Input the data for Group 2. (One datum at a time.)"
FOR I=1 TO N : INPUT X(M+I) : NEXT I
' Get totals and set indexes
FOR I=1 TO MN
     SUM = SUM + X(I)
     IF I=M THEN CRIT1=SUM
     INDEX(I)=I
NEXT I
NPERM=1 : UPPERTAIL=1 ' NPERM = # permutations, UPPERTAIL = # in uppertail
LOOP1:
        I=M
        LOOP2:
                IF INDEX(I)=MN THEN I=I-1 : IF I=0 GOTO WRAPUP ELSE GOTO LOOP2
                INDEX(I)=INDEX(I)+1
                IF I<M THEN ' get next element of permutation
                        I=I+1
                        INDEX(I)=INDEX(I-1)
                        GOTO LOOP2
                    ELSE ' Evaluate current sum
                        NPERM=NPERM+1
                        SUM1=0
                        FOR I=1 TO M
                             SUM1=SUM1+X(INDEX(I))
                        NEXT I
                        IF SUM1 >= CRIT1 THEN UPPERTAIL=UPPERTAIL+1
                        GOTO LOOP1
                END IF
WRAPUP: ' Computation is done, print results
PRINT "                 PERMUTATION TEST"
PRINT "       Group:      1         2"
PRINT USING "Sample Size:  ###          ###";M,N
PRINT USING "Sample Sums: #####.##  #####.##";CRIT1,SUM-CRIT1
PRINT USING "Number of Sums >= Group 1 Sum: ##### out of ##### sums.";UPPERTAIL,
NPERM
PRINT USING "Upper Tail Probability = #.####";UPPERTAIL/NPERM
END
```

Salida de muestra para el programa 3

Los datos siguientes son de la página 182.

```
What are the samples sizes? 5,4
Input the data for Group 1. (One datum at a time.)
? 22
? 19
? 16
? 29
? 24
Input the data for Group 2. (One datum at a time.)
? 11
? 12
? 20
? 0
              PERMUTATION TEST
      Group:        1         2
Sample Size:        5         4
Sample Sums:    110.00     43.00
Number of Sums >= Group 1 Sum:     3 out of   126 sums.
Upper Tail Probability = 0.0238



Normal termination.  Press any key.
```

Programa 4

k muestras independientes: prueba ji cuadrada para tablas de contingencia $r \times k$ con partición.

```
100 REM
110 REM Coded by N. J. Castellan, Jr., Copyright 1984, 1985.
120 PRINT "Routine to calculate chi-square and partitioned chi-square"
130 PRINT "for general r by k contingency table."
140 PRINT : PRINT"    You must enter the size of the contingency table,"
150 PRINT "    followed by the cell frequencies." : PRINT
160 INPUT "How many rows";R
170 INPUT "How many columns";K
180 DIM X(R,K), ROW(R), COL(K), E(R,K)
190 REM Read Data -- READ X(I,J) at line 240 may be changed to INPUT X(I,J)
200 REM                 Remove the REM at the beginning of line 250
210 PRINT : PRINT"Now enter the data, cell by cell."
220 FOR I=1 TO R
230     FOR J=1 TO K
240     READ X(I,J)
250     REM PRINT "Enter the data for cell";I;",";J; : INPUT X(I,J)
260     NEXT J
270 NEXT I
280 REM Calculate marginal frequencies
290 FOR I=1 TO R
300     FOR J=1 TO K
310         ROW(I)=ROW(I)+X(I,J)
320         COL(J)=COL(J)+X(I,J)
330         N=N+X(I,J)
340     NEXT J
350 NEXT I
360 REM Find expected values and calculate chi-square (X2)
370 FOR I=1 TO R
380     FOR J=1 TO K
390         E(I,J)=ROW(I)*COL(J)/N
400         X2=X2+(X(I,J)^2)/E(I,J)
410     NEXT J
420 NEXT I
430 X2=X2-N
440 PRINT : PRINT "Chi-square =";X2;" with ";(R-1)*(K-1);" degrees of freedom."
450 REM Begin partitioning procedure
460 PRINT : PRINT"Partition  cell(i,j)   Chi-Square"
470 FOR J=2 TO K
480     UR=X(1,J) : UL=0 : LL=0 : LR=0
490     FOR JJ=1 TO J-1 : UL=UL+X(1,JJ) : NEXT JJ
500     SR=0 : SC=SC+COL(J-1)
510     FOR I=2 TO R
520         UL=UL+LL
530         UR=UR+LR
540         LL=0 : FOR JJ=1 TO J-1 : LL=LL+X(I,JJ) : NEXT JJ
550         LR=X(I,J)
560         SR=SR+ROW(I-1)
570         XT=N*(COL(J)*(ROW(I)*UL - SR*LL) - SC*(ROW(I)*UR - LR*SR))^2
580         XT=XT/(COL(J)*ROW(I)*SC*(SC+COL(J))*SR*(SR+ROW(I)))
590         T=(R-1)*(J-2)+I-1
600         PRINT USING "  ###        ##:##      ###.###";T,I,J,XT
610     NEXT I
620 NEXT J
630 STOP : '-------------------------------------------------------------
640 REM Data from Sample Problem
650 REM k = 4 Groups, r = 3 levels or rows
660 DATA 13,8,10,3, 20,23,27,18, 11,12,12,21
670 END
```

Salida de muestra para el programa 4

Los siguientes datos son de los ejemplos de las páginas 225 y 229.

```
Routine to calculate chi-square and partitioned chi-square
for general r by k contingency table.

     You must enter the size of the contingency table,
     followed by the cell frequencies.

How many rows?  3
How many columns?  4


Chi-square = 12.778 with 6 degrees of freedom.

Partition  cell(i,j)  Chi-Square
    1         2: 2        1.620
    2         3: 2        0.085
    3         2: 3        0.415
    4         3: 3        0.055
    5         2: 4        1.840
    6         3: 4        8.762
```

Programa 5

k muestras independientes: prueba de Jonckheere para niveles ordenados de la variable.

```
100 REM
110 REM Coded by N. J. Castellan, Jr., Copyright 1982.
120 PRINT "Routine to Calculate the Jonckheere Test for Ordered Alternatives"
130 INPUT "How many groups";K
140 DIM N(K), U(K,K)
150 REM Read Group sizes and Calculate terms for Mean and Variance
160 REM -- This program assumes that all variables and arrays
170 REM -- are initialized to zero's
180 FOR I = 1 TO K
190     PRINT "How many observations in group ";I; : INPUT N(I)
200     N1 = N1 + N(I)
210     N2 = N2 + N(I)^2
220     N3 = N3 + N(I)^3
230 NEXT I
240 REM Read Data -- READ X(IJ) may be changed to INPUT X(IJ)
250 DIM X(N1)
260 FOR I = 1 TO K
270     FOR J = 1 TO N(I)
280           IJ=IJ+1
290           READ X(IJ)
300     NEXT J
310 NEXT I
320 REM Calculate Mean and Variance
330 MEAN = (N1^2 - N2)/4
340 VARIANCE = ((N1^2)*(2*N1 + 3) - (2*N3 + 3*N2))/72
350 REM Calculate Mann-Whitney U-counts
360 ILOW-0 : IHIGH=0
370 FOR I = 1 TO K-1
380     ILOW = IHIGH + 1
390     IHIGH = ILOW + N(I) - 1
400     FOR IX = ILOW TO IHIGH
410         JHIGH = IHIGH
420         FOR J = I+1 TO K
430             JLOW = JHIGH + 1
440             JHIGH = JLOW + N(J) - 1
450             FOR JX = JLOW TO JHIGH
460                 IF X(IX) < X(JX) THEN U(I,J) = U(I,J) + 1
470                 IF X(IX) = X(JX) THEN U(I,J) = U(I,J) + .5
480             NEXT JX
490         NEXT J
500     NEXT IX
510 NEXT I
520 PRINT "Group          Group        U(i,j)"
530 FOR I=1 TO K-1
540     FOR J=I+1 TO K
550         PRINT I,J,U(I,J)
560         JS = JS + U(I,J)
570     NEXT J
580 NEXT I
590 PRINT : PRINT "Jonckheere Statistic: J = ";JS
600 PRINT : PRINT "Mean = "; MEAN; ", Variance = "; VARIANCE
610 PRINT "Standard Normal Approximation: J* = "; (JS - MEAN)/SQR(VARIANCE)
620 STOP : '--------------------------------------------------------------
630 REM Data from Sample Problem
640 REM k = 4 Groups, n(1)=12, n(2)=9, n(3)=8, n(4)=6
650 DATA  8.82, 11.27, 15.78, 17.39, 24.99, 39.05, 47.54, 48.85
660 DATA 71.66, 72.77, 90.38, 103.13
670 DATA 13.53, 28.42, 48.11, 48.64, 51.40, 59.91, 67.98, 79.13, 103.05
680 DATA 19.23, 67.83, 73.68, 75.22, 77.71, 83.67, 86.83, 93.25
690 DATA 73.51, 85.25, 85.82, 88.88, 90.33, 118.11
700 END
```

Salida de muestra para el programa 5

Los siguientes datos son del ejemplo de la página 253.

```
How many observations in group 1 ? 12
How many observations in group 2 ? 9
How many observations in group 3 ? 8
How many observations in group 4 ? 6

Group         Group         U(i,j)
 1             2              66
 1             3              73
 1             4              62
 2             3              52
 2             4              48
 3             4              36

Jonckheere Statistic: J =  337

Mean =  255 . Variance =  1140
Standard Normal Approximation: J* =  3.31715
```

Apéndice III. Pruebas estadísticas no paramétricas

		Caso de dos muestras		Caso de k muestras		
Nivel de medición	*Caso de una muestra (cap. 3)*	*Muestras relacionadas o apareadas (cap. 4)*	*Muestras independientes (cap. 5)*	*Muestras relacionadas (cap. 6)*	*Muestras independientes (cap. 7)*	*Medidas de asociación (cap. 8)*
Nominal o categórica	Prueba binomial Prueba ji cuadrada de la bondad de ajuste	Prueba del cambio de McNemar	Prueba exacta de Fisher para tablas de 2×2 Prueba ji cuadrada para tablas de $r \times 2$	Prueba Q de Cochran	Prueba ji cuadrada para tablas $r \times k$	Coeficiente C de Cramér Coeficiente phi r_ϕ Coeficiente kappa de acuerdo, K Estadístico lambda de asociación asimétrica, L_B
Ordinal u ordenada	Prueba de Kolmogorov-Smirnov de una muestra, $D_{m,n}$ Prueba de series aleatorias de una muestra	Prueba de los signos Prueba de rangos asignados de Wilcoxon, T^+	Prueba de la mediana Prueba de Wilcoxon-Mann-Whitney, W_x Prueba poderosa de rangos ordenados, $\grave{U}$	Análisis de varianza bifactorial por rangos de Friedman, F_r Prueba de Page para alternativas ordenadas, L	Extensión de la prueba de la mediana Análisis de varianza de una forma por rangos de Kruskal-Wallis, KW	Coeficiente de correlación r_s, de Spearman de rangos ordenados Coeficiente de correlación T, de Kendall de

	Prueba del momento del cambio		Prueba de Kolmogorov-Smirnov para dos muestras, $D_{m,n}$ Prueba de Siegel-Tukey para diferencias en la escala		Prueba de Jonckheere para alternativas ordenadas, J	rangos ordenados Coeficiente de correlación parcial $T_{xy \cdot z}$, de Kendall de rangos ordenados Coeficiente de concordancia W de Kendall Coeficiente de acuerdo u de Kendall Correlación entre k jueces y un criterio T_C
Intervalo		Prueba de las permutaciones para pares replicados	Prueba de las permutaciones para dos muestras independientes Prueba para rangos ligados de Moses para diferencias en la escala			Estadístico gamma G Índice de Somers d_{BA} de asociación asimétrica

NOTA: En cada columna se enumeran, acumulativamente hacia abajo, las pruebas aplicables para el nivel de medición dado. Por ejemplo, en el caso de k muestras relacionadas, cuando las variables están ordenadas, tanto el análisis de varianza bifactorial de Friedman como la prueba Q de Cochran son aplicables. Sin embargo, véase el texto para una explicación acerca de lo apropiado de una prueba particular para un determinado tipo de datos.

Bibliografía

Bailey, D. E., *Probability and statistics: models for research*, J. Wiley, Nueva York, 1971.

Bishop, Y. M. M., S. E., Feinberg, y P. W., Holland, *Discrete multivariate analysis: theory and practice*, MIT Press, Cambridge, MA, 1975.

Bradley, J. V., *Distribution-free statistical tests*, Prentice-Hall, Englewood Cliffs, N. J., 1968.

Castellan, N. J., Jr., "On the partitioning of contingency tables", en *Psychological Bulletin*, núm. **64**, 1965, págs. 330-338.

———, "The analysis of behavior sequences", en R. B. Cairns (ed.), *The analysis of social interactions: methods, issues, and illustrations*, L. Erlbaum, Hillsdale, N. J., 1979, págs. 81-116.

Chacko, V. J., "Testing homogeneity against ordered alternatives", en *Annals of Mathematical Statistics*, núm. **34**, 1963, págs. 945-956.

Cochran, W. G., "The comparison of percentages in matched samples", en *Biometrika*, núm. **37**, 1950, págs. 256-266.

———, "The χ^2 test of goodness of fit", en *Annals of Mathematical Statistics, núm.* **23**, 1952, págs. 315-345.

———, "Some methods for strengthening the common χ^2 tests", en *Biometrics*, núm. **10**, 1954, págs. 417-451.

Cohen, J., "A coefficient of agreement for nominal scales", en *Educational and Psychological Measurement*, núm. **20**, 1960, págs. 37-46.

———, "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit", en *Psychological Bulletin*, núm. **70**, 1968, págs. 213-220.

Davidson, D., P. Suppes, y S. Siegel, *Decision making: an experimental approach*, Stanford University Press, Stanford, CA, 1957.

Delucchi, K. L., "The use and misuse of chi-square: Lewis and Burke revisited", en *Psychological Bulletin*, núm. **94**, 1983, págs. 166-176.

Dixon, W. J. y F. J., Massey, *Introduction to statistical analysis*, McGraw-Hill, Nueva York, 1983.

Edwards, A. L., *Statistical methods*, 2a. ed., Holt, Rinehart and Winston, Nueva York, 1967.

Ehrenberg, A. S. C., "On sampling from a population of rankers", en *Biometrika*, núm. **39**, 1952, págs. 82-87.

Everitt, B. S., *The analysis of contingency tables*, Chapman and Hall, Londres, 1977.

Feigin, P. D. y A., Cohen, "On a model for concordance between judges", en *Journal Royal Statistical Society* (Series B), núm. **40**, 1978, págs. 203-221.

Fisher, R. A., *Statistical methods for research workers*, 14a. ed., Hafner, Nueva York, 1973.

Fleiss, J. L., "Measuring nominal scale agreement among many raters", en *Psychological Bulletin*, núm. **76**, 1971, págs. 378-382.

Fligner, M. A. y G. E., III, Policello, "Robust rank procedures for the Behrens-Fisher problem", en *Journal of the American Statistical Association*, núm. **76**, 1981, págs. 162-168.

Fraser, C. O., "Measurement in psychology", en *British Journal of Psychology*, núm. **71**, 1980, págs. 23-24.

Friedman, M., "The use of ranks to avoid the assumption of normality implicit in the analysis of variance", en *Journal of the American Statistical Association*, núm. **32**, 1937, págs. 675-701.

——, "A comparison of alternative tests of significance for the problem of *m* rankings", en *Annals of Mathematical Statistics*, núm. **11**, 1940, págs. 86-92.

Gibbons, J. D., *Nonparametric methods for quantitative analysis*, Holt, Rinehart and Winston, Nueva York, 1976.

——, *Nonparametric statistical inference*, 2a. ed., Marcel Dekker, Nueva York, 1985.

Goodman, L. A., Kolmogorov-Smirnov tests for psychological research", en *Psychological Bulletin*, núm. **51**, 1954, págs. 160-168.

—— y W. H., Kruskal, "Measures of association for cross classifications", en *Journal of the American Statistical Association*, núm. **49**, 1954, págs. 732-764.

—— y ——, "Measures of association for cross classifications. II: Further discussion and references", en *Journal of the American Statistical Association*, núm. **54**, 1959, págs. 123-163.

—— y ——, "Measures of association for cross classifications. III: Approximate sampling theory", en *Journal of the American Statistical Association*, núm. **58**, 1963, págs. 310-364.

—— y ——, "Measures of association for cross classifications. IV: Simplification of asymptotic variances", en *Journal of the American Statistical Association*, núm. **67**, 1972, págs. 415-421.

Haberman, S. J., "The analysis of residuals in cross-classified tables", en *Biometrics*, núm. **29**, 1973, págs. 205-220.

Hammond, K. R., J. E. Householder, y N. J., Jr., Castellan, *Introduction to the statistical method*, 2a. ed. A. A. Knopf, Nueva York, 1970.

Hays, W. L., "A note on average tau as a measure of concordance", en *Journal of the American Statistical Association*, núm. **55**, 1960, págs. 331-341.

——, *Statistics*, 3a. ed. Holt, Rinehart and Winston, Nueva York, 1981.

Hettmansperger, T. P., *Statistical inference based on ranks*, J. Wiley, Nueva York, 1984.

Hollander, M., "Asymptotic efficiency of two nonparametric competitors of Wilcoxon's two sample test", en *Journal of the American Statistical Association*, núm. **62**, 1967, págs. 939-949.

—— y D. A., Wolfe, *Nonparametric statistical methods*, J. Wiley, Nueva York, 1973.

Johnson, N. S., "Nonnull properties of Kendall's partial rank correlation coefficient", en *Biometrika*, núm. **66**, 1979, págs. 333-338.

Jonckheere, A. R., "A distribution-free *k*-sample test against ordered alternatives", en *Biometrika*, núm. **41**, 1954, págs. 133-145.

Kendall, M. G., *Rank correlation methods*, 4a. ed., Londres, 1970.

Kolmogorov, A., "Confidence limits for an unknown distribution function", en *Annals of Mathematical Statistics*, núm. **12**, 1941, págs. 461-463.

Kruskal, W. H., "A nonparametric test for the several sample problem", en *Annals of Mathematical Statistics*, núm. **23**, 1952, págs. 525-540.

—— y W. A., Wallis, "Use of ranks in one-criterion variance analysis", en *Journal of the American Statistical Association*, núm. **47**, 1952, págs. 583-621.

Lehmann, E. L., *Nonparametrics: statistical methods based on ranks*, Holden-Day, San Francisco, 1975.

Lewis, D. y C. J., Burke, "The use and misuse of the chi-square test", en *Psychological Bulletin*, núm. **46**, 1949, págs. 433-489.
Lienert, G. A. y P., Netter, "Nonparametric analysis of treatment-response tables by bipredictive configural frequency analysis", en *Methods of Information in Medicine*; núm. **26**, 1987, págs. 89-92.
Light, R. J., "Measures of response agreement for qualitative data: some generalizations and alternatives", en *Psychological Bulletin*, núm. **76**, 1971, págs. 365-377.
Mack, G. A. y D. A., Wolfe, "*k*-sample rank tests for umbrella alternatives", en *Journal of the American Statistical Association*, núm. **76**, 1981, págs. 175-181.
Maghsoodloo, S., "Estimates of the quantiles of Kendall's partial rank correlation coefficient", en *Journal of Statistical Computing and Simulation*, núm. **4**, 1975, págs. 155-164.
——— y L. L., Pallos, "Asymptotic behavior of Kendall's partial rank correlation coefficient and additional quantile estimates", en *Journal of Statistical Computing and Simulation*, núm. **13**, 1981, págs. 41-48.
Mann, H. B. y D. R., Whitney, "On a test of whether one of two random variables is stochastically larger than the other", en *Annals of Mathematical Statistics*, núm. **18**, 1947, págs. 50-60.
Marascuilo, L. A. y M., McSweeney, *Nonparametric and distribution-free methods for the social sciences*, Brooks/Cole, Monterey, CA, 1977.
McNemar, Q., *Psychological statistics*, 4a. ed., J. Wiley, Nueva York, 1969.
Mood, A. M., *Introduction to the theory of statistics*, McGraw-Hill, Nueva York, 1950.
Moran, P. A. P., "Partial and multiple rank correlation", en *Biometrika*, núm. **38**, 1951, págs. 26-32.
Moses, L. E., "Non-parametric statistics for psychological research", en *Psychological Bulletin*, núm. **49**, 1952, págs. 122-143.
———, "Rank tests of dispersion", en *Annals of Mathematical Statistics*, núm. **34**, 1963, págs. 973-983.
Mosteller, F., "A *k*-sample slippage test for an extreme population", en *Annals of Mathematical Statistics*, núm. **19**, 1948, págs. 58-65.
——— y J. W., Tukey, "Significance levels for a *k*-sample slippage test", en *Annals of Mathematical Statistics*, núm. **21**,1950, págs. 120-123.
Page, E. B., "Ordered hypotheses for multiple treatments: A significance test for linear ranks", en *Journal of the American Statistical Association*, núm. **58**, 1963, págs. 216-230.
Page, E. S., "A test for a change in a parameter ocurring at an unknown point", en *Biometrika*, núm. **42**, 1955, págs. 523-527.
Patil, K. D., "Cochran's Q test: Exact distribution", en *Journal of the American Statistical Association*, núm. **70**, 1975, págs. 186-189.
Pettitt, A. N., "A non-parametric approach to the change-point problem", en *Applied Statistics*, núm. **28**, págs. 126-135.
Pitman, E. J. G., "Significance tests which may be applied to samples from any populations", Suplemento de *Journal of the Royal Statistical Society*, núm. **4**, 1937*a*, págs. 119-130.
———, "Significance tests which may be applied to samples from any populations. II. The correlation coefficient test", Suplemento de *Journal of the Royal Statistical Society*, núm. **4**, 1937*b*, págs. 225-232.
———, "Significance tests which may be applied to samples from any populations. III. The analysis of variance test", en *Biometrika*, núm. **29**, 1937*c*, págs. 322-335.
Potter, R. W. y G. W., Strum, "The power of Jonckheere's test", en *Journal of the American Statistical Association*, núm. **35**, 1981, págs. 249-250.
Puri, M. L., "Some distribution-free *k*-sample rank tests for homogeneity against ordered alternatives", *Communications Pure Applied Mathematics*, núm. **18**, 1965, págs. 51-63.
Randles, R. H., M. A., Fligner, G. E., III, Policello, y D. A., Wolfe, "An asymptotically distribution-free test for symmetry versus asymmetry", en *Journal of the American Statistical Association, núm.* **75**, 1980, págs. 168-172.

—— y D. A., Wolfe, *Introduction to the theory of nonparametric statistics*, J. Wiley, Nueva York, 1979.

Scheffé, H. V., "Statistical inference in the non-parametric case", en *Annals of Mathematical Statistics*, núm. **14**, 1943, págs. 305-332.

Schorak, G. R., "Testing and estimating ratios of scale parameters", en *Journal of the American Statistical Association*, núm. **64**, 1969, págs. 999-1 013.

Scott, W. A., "Reliability of content analysis: The case of nominal scale coding", en *Public Opinion Quarterly* núm. **19**, 1955, págs. 321-325.

Shaffer, J. P., "Defining and testing hypotheses in multidimensional contingency tables", en *Psychological Bulletin*, núm. **79**, 1973, págs. 127-141.

Siegel, S. y J. W., Tukey, "A nonparametric sum of ranks procedure for relative spread in unpaired samples", en *Journal of the American Statistical Association*, núm. **55**, 1960, págs. 429-445 (Corrección en *Journal of the American Statistical Association*, núm. **56**, 1961, pág. 1005).

Smirnov, N. V., "Table for estimating the goodness of fit of empirical distributions", en *Annals of Mathematical Statistics*, núm. **19**, 1948, págs. 279-281.

Somers, R. H., "A new asymmetric measure of association for ordinal variables", en *American Sociological Review*, núm. **27**, 1962, págs. 799-811.

——, "Simple approximations to null sampling variances: Goodman and Kruskal's gamma, Kendall's tau, and Somers's d_{yx}", en *Sociological Methods and Research*, núm. **9**, 1980, págs. 115-126.

Stilson, D. W. y V. N., Campbell, "A note on calculating tau and average tau and on the sampling distribution of average tau with a criterion ranking", en *Journal of the American Statistical Association*, núm. **57**, 1962, págs. 567-571.

Swed, F. S. y C., Eisenhart, "Tables for testing randomness of grouping in a sequence of alternatives", en *Annals of Mathematical Statistics*, núm. **14**, 1943, págs. 83-86.

Terpstra, T. J. "The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking", en *Indagationes Mathematicae*, núm. **14**, 1952, págs. 327-333.

Townsend, J. T. y F. G., Ashby, "Measurement scales and statistics: The misconception misconceived", en *Psychological Bulletin*, núm. **96**, 1984, págs. 394-401.

Whitney, D. R., "A comparison of the power of non-parametric tests and tests based on the normal distribution under non-normal alternatives", Tesis doctoral inédita, Ohio State University.

Wilcoxon, F., "Individual comparisons by ranking methods", en *Biometrics*, núm. **1**, 1945, págs. 80-83.

——, "Probability tables for individual comparisons by ranking methods", en *Biometrics*, núm. **3**, 1947, págs. 119-122.

——, *Some rapid approximate statistical procedures*, American Cyanamid, Stamford, CT, 1949.

Yates, F., "Contingency tables involving small numbers and the χ^2 test", en *Journal of the Royal Statistical Society Supplement*, núm. **1**, 1934, págs. 217-235.

Índice analítico

La publicación de esta obra la realizó
Editorial Trillas, S. A. de C. V.

División Administrativa, Av. Río Churubusco 385,
Col. Pedro María Anaya, C. P. 03340, México, D. F.
Tel. 6884233, FAX 6041364

División Comercial, Calz. de la Viga 1132, C. P. 09439
México, D. F. Tel. 6330995, FAX 6330870

Se terminó de imprimir
el 4 de mayo 1998 (TRSS),
en los talleres de Litográfica Ingramex, S. A. de C. V.
Se encuadernó en Ediciones Pegaso, S. A. de C. V.
BM2 80