

# Índice general

<b>1. Medidas para Variables Ordinales (Parte 2)</b>	<b>3</b>
1.1. Coeficiente de correlación de rango de Spearman . . . . .	3
1.2. Medida de concordancia de la regla del pie de Spearman . . . . .	5
1.2.1. Probabilidad de la regla del pie de Spearman . . . . .	7
1.2.2. Rangos Múltiples . . . . .	7
1.3. Coeficiente de Concordancia . . . . .	8
1.3.1. Procedimientos relacionados . . . . .	9
1.4. Medida de acuerdo $u$ de Kendall . . . . .	10
1.5. Medida Kappa de Cohen . . . . .	10
1.5.1. Comparación de la ponderación lineal y cuadrática . . . . .	12
1.5.2. Kappa ponderado con múltiples jueces . . . . .	13
1.6. Análisis Ridit . . . . .	15
1.6.1. Cálculo . . . . .	16
1.6.1.1. Procedimientos de permutación exacta . . . . .	16
1.6.1.2. Procedimientos de permutación de remuestreo . . . . .	16
<b>Bibliografía</b>	<b>20</b>



# Capítulo 1

## Medidas para Variables Ordinales (Parte 2)

Este capítulo continúa la descripción de las medidas de asociación para dos variables de nivel ordinal iniciada en el capítulo 3, pero se centra las medidas que se basan en criterios distintos a las comparaciones por pares entre las puntuaciones de los rangos, aunque es inevitable cierto solapamiento.

Se incluyen métodos estadísticos de permutación exactos y de Monte Carlo para el coeficiente de correlación de Spearman, la medida de concordancia de la regla del pie de Spearman, el coeficiente de concordancia de Kendall y Babington Smith, la medida kappa ponderada de Cohen para la concordancia, y el análisis ridit de Bross.

### 1.1. Coeficiente de correlación de rango de Spearman

Se consideran dos clasificaciones de  $N$  objetos consistentes en los  $N$  primeros enteros y sean  $X_i$  e  $Y_i$  para  $i = 1, \dots, N$  la primera y la segunda clasificación, respectivamente. La correlación de rangos no está exenta de críticas, ya que la derivación de la fórmula para esta correlación implica la suposición de que las diferencias entre los rangos sucesivos son iguales.

Una medida popular de la correlación entre las dos clasificaciones es el coeficiente de correlación de orden de rango de Spearman, dado por

$$\rho = 1 - \frac{\sum_{i=1}^N d_i^2}{\frac{N(N^2-1)}{6}} = 1 - \frac{6 \sum_{i=1}^N d_i^2}{N(N^2-1)} , \quad (4.1)$$

donde  $d_i = X_i - Y_i$  para  $i = 1, \dots, N$ . Charles Spearman desarrolló  $\rho$  en el primero de los dos artículos sobre la medición de la asociación y la correlación en 1904 y 1906 que aparecieron en el *American Journal of Psychology* [Spearman, 1904] y en el *British Journal of Psychology* [Spearman, 1906], respectivamente. Con dos conjuntos de puntuaciones de rango sin valores empatados,  $X_i$  e  $Y_i$  para  $i = 1, \dots, N$ , se tiene que

$$\sum_{i=1}^N x_i = \sum_{i=1}^N y_i = \frac{N(N+1)}{2}$$

y

$$\sum_{i=1}^N x_i^2 = \sum_{i=1}^N y_i^2 = \frac{N(N+1)(2N+1)}{6} ,$$

Spearman simplemente sustituyó en la fórmula de Pearson del coeficiente de correlación producto-momento dado por

$$r_{xy} = \frac{N \sum_{i=1}^N x_i y_i - \sum_{i=1}^N x_i \sum_{i=1}^N y_i}{\sqrt{N \sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i\right)^2} \sqrt{N \sum_{i=1}^N y_i^2 - \left(\sum_{i=1}^N y_i\right)^2}}$$

y simplificó la ecuación, dando lugar a la Ecuación (4.1).

Obsérvese que el denominador del coeficiente de correlación de orden de rango de Spearman,  $N(N^2 - 1)/6$ , como se indica en la Ec. (4.1), representa la mitad del valor máximo de  $\sum_{i=1}^N d_i^2$  cuando  $X_i$  e  $Y_i$ ,  $i = 1, \dots, N$ , son puntuaciones de rango sin valores empatados y las puntuaciones de rango de  $Y_i$  son la inversa exacta de las puntuaciones de rango de  $X_i$ , es decir,  $Y_i = N - X_i + 1$  para  $i = 1, \dots, N$ . Por lo tanto, la  $\rho$  de Spearman es una medida de correlación de rango corregida al máximo y toma valores entre  $\pm 1$ , donde  $+1$  indica una asociación positiva perfecta y  $-1$  indica una asociación negativa perfecta.

Se demuestra fácilmente que el denominador de la ecuación (4.1),  $N(N^2 - 1)/6$ , es la mitad del valor máximo de  $\sum_{i=1}^N d_i^2$  cuando  $X_i$  e  $Y_i$  para  $i = 1, \dots, N$  son ambas son puntuaciones de rango no empatadas y las puntuaciones de rango  $Y_i$  son la inversa de las puntuaciones de rango  $x_i$ . Para el valor máximo de  $\sum_{i=1}^N d_i^2$ , se define:

$$\sum_{i=1}^N d_i^2 = \sum_{i=1}^N (X_i - Y_i)^2 = \sum_{i=1}^N X_i^2 + \sum_{i=1}^N Y_i^2 - 2 \sum_{i=1}^N X_i Y_i .$$

Ya que, para  $N$  valores de rango no empatados,

$$\sum_{i=1}^N X_i^2 = \sum_{i=1}^N Y_i^2 = \frac{N(N+1)(2N+1)}{6}$$

y, para  $X_i = 1, \dots, N$  e  $Y_i = N - x_i + 1$ ,  $i = 1, \dots, N$ ,

$$\sum_{i=1}^N X_i Y_i = \frac{N(N+1)(N+2)}{6} ,$$

entonces, al sustituir en la Ec. (4.1) se obtiene:

$$\sum_{i=1}^N d_i^2 = \frac{2N(N+1)(2N+1)}{6} - \frac{2N(N+1)(N+2)}{6} =$$

$$= \frac{2N(N+1)(N-1)}{6} = \frac{N(N^2-1)}{3},$$

que es el doble del valor de  $N(N^2-1)/6$ .

Kendall y Babington Smith observaron que para interpretar la significación de un valor de  $\rho$ , es necesario considerar sólo la distribución de valores obtenida a partir de las clasificaciones observadas con todas las demás permutaciones posibles de los números enteros de 1 a  $N$ , y señalaron además que en la práctica suele ser más conveniente considerar sólo la distribución de  $\sum_{i=1}^N d_i^2$  ya que  $N(N^2-1)/6$  es invariable bajo permutación [Kendall et al., 1939].

## 1.2. Medida de concordancia de la regla del pie de Spearman

Los artículos de 1904 [Spearman, 1904] y 1906 [Spearman, 1906] de Charles Spearman contenían dos nuevas medidas de correlación de rango: el conocido coeficiente de correlación de rango de Spearman,  $\rho$ , y un segundo coeficiente de correlación menos conocido que Spearman denominó “la regla del pie”. Se consideran dos clasificaciones de  $N$  objetos que consisten en los primeros  $N$  enteros y sean  $X_i$  e  $Y_i$  para  $i = 1, \dots, N$  la primera y la segunda clasificación, respectivamente. Entonces, la regla de Spearman viene dada por:

$$\mathcal{R} = 1 - \frac{\sum_{i=1}^N |X_i - Y_i|}{\frac{N^2-1}{3}} = \frac{3 \sum_{i=1}^N |X_i - Y_i|}{N^2 - 1}. \quad (4.2)$$

A diferencia del coeficiente de correlación de orden de rango de Spearman, el denominador del coeficiente de la regla del pie de Spearman,  $\sim(N^2-1)/3$ , como se da en la Ec. (4.2), no representa la mitad del valor máximo de  $\sum_{i=1}^N |X_i - Y_i|$  cuando  $X_i$  e  $Y_i$  para  $i = 1, \dots, N$  son ambas puntuaciones de rango no empatadas y las puntuaciones de rango  $Y_i$  son la inversa exacta de las puntuaciones de rango  $X_i$ , es decir,  $Y_i = N - X_i + 1$  para  $i = 1, \dots, N$ . Por lo tanto, la  $\mathcal{R}$  de Spearman no es una medida máximo-corregida de correlación de rangos y es, en cambio, una medida de concordancia corregida por el azar.

Se puede demostrar fácilmente que la  $\mathcal{R}$  de Spearman es una medida de concordancia corregida por el azar y no es, de hecho, una medida convencional de correlación, lo que explica por qué  $\mathcal{R}$  puede, en ocasiones, dar valores negativos y sólo puede alcanzar un valor de  $-1$  cuando  $N = 2$ . Para demostrar que el valor esperado de  $\sum_{i=1}^N |d_i|$  viene dado por  $(N^2-1)/3$ , se considera que:

$$\begin{aligned} \sum_{i=1}^N |d_i| &= \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N |i - j| = \frac{2}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^N (j - i) = \\ &= \frac{1}{N} \sum_{i=1}^{N-1} [N(N+1) + i^2 - i(2N+1)] = \\ &= \frac{N(N+1)}{6N} [6(N+1) + (2N-1) - 3(2N+1)] = \end{aligned}$$

$$= \frac{N^2 - 1}{3}.$$

Por lo tanto, el coeficiente de la regla del pie de Spearman, dado por

$$\mathcal{R} = a - \frac{\sum_{i=1}^N |d_i|}{\frac{N^2 - 1}{3}}$$

es una medida de concordancia corregida por el azar cuando el valor esperado de  $\sum_{i=1}^N |d_i|$  viene dado por  $(N^2 - 1)/3$ , ya que adopta la forma clásica de las medidas de concordancia corregidas por el azar dada por

$$\text{concordancia} = \frac{\text{discordancia observada}}{\text{discordancia esperada}}$$

[Krippendorff, 1970].

Existen tres limitaciones de la regla del pie de Spearman que contribuyen a su falta de uso en la investigación contemporánea [Stuart, 1977]: En primer lugar, a diferencia de otras medidas de correlación de rangos,  $\mathcal{R}$  no norma adecuadamente entre los límites de  $\pm 1$ ; en segundo lugar, al igual que la  $\rho$  de Spearman,  $\mathcal{R}$  se limita a los datos totalmente clasificados y no se adapta a las puntuaciones de rango empatadas; y en tercer lugar, debido a la suma de las diferencias absolutas entre las puntuaciones de rango, tradicionalmente ha sido algo engorroso establecer el valor de probabilidad de un valor observado de  $\mathcal{R}$ , especialmente cuando  $N$  es pequeño.

La  $\mathcal{R}$  de Spearman alcanza un valor máximo de  $+1$  cuando  $X_i$  es idéntica a  $Y_i$  para  $i = 1, \dots, N$  y no hay valores empatados. Sin embargo, si  $Y_i = N - x_i + 1$  para  $i = 1, \dots, N$ , entonces  $\mathcal{R} = -0,5$  cuando  $N$  es impar y

$$\mathcal{R} = -0.5(1 + \frac{3}{N^2 - 1})$$

cuando  $N$  es par [Kendall, 1962]. En consecuencia,  $\mathcal{R}$  no puede alcanzar un valor mínimo de  $-1$ , excepto cuando  $N = 2$ . Spearman, aparentemente sin saber que  $\mathcal{R}$  era una medida corregida por el azar y reconociendo que los valores negativos de  $\mathcal{R}$  no representaban una correlación inversa, sugirió ingenuamente que “es mejor tratar toda correlación como positiva” [Spearman, 1904]. Maurice Kendall señaló explícitamente esta aparente falta tipificación como un defecto de la regla del pie y sugirió una corrección dada por

$$\mathcal{R}' = 1 - \frac{\sum_{i=1}^N |X_i - Y_i|}{N^2},$$

que aseguraba un límite adecuado de  $+1$  cuando las dos clasificaciones estaban en completa concordancia y  $-1$  cuando las dos clasificaciones eran inversas entre sí [Kendall, 1962]. Sin embargo la corrección, aunque bien intencionada, destruyó por completo la interpretación de la regla del pie de Spearman.

### 1.2.1. Probabilidad de la regla del pie de Spearman

Cuando las variables  $X$  e  $Y$  consisten en su totalidad en puntuaciones de rango no empatadas de 1 a  $N$  y la variable  $Y$  es una permutación de las observaciones de rango en la variable  $X$ , existen métodos para determinar la probabilidad de una  $\mathcal{R}$  observada bajo la hipótesis nula de que cualquiera de los  $N!$  ordenamientos de los valores de  $X$  o  $Y$  es igualmente probable. Si

$$D = \sum_{i=1}^N |X_i - Y_i|$$

entonces, dado que  $\mathcal{R}$  es simplemente una transformación lineal de  $D$ , la probabilidad de un valor observado de  $D$  es la probabilidad de un valor observado de  $\mathcal{R}$ . Las tablas de la función de distribución acumulativa exacta de  $D$  para  $2 \leq N \leq 10$  y los valores de probabilidad aproximados basados en métodos de Monte Carlo para  $11 \leq N \leq 15$  fueron publicados por Ury y Kleinecke en 1979 [Ury and Kleinecke, 1979]. En 1988 Franklin amplió el trabajo de Ury y Kleinecke, informando de la función de distribución acumulativa exacta de  $D$  para  $11 \leq N \leq 18$ , y discutió la tasa de convergencia a una distribución normal aproximada [Franklin, 1988]. En 1990 Salama y Quade utilizaron las propiedades de la cadena de Markov para obtener la exacta función de distribución acumulativa de  $D$  para  $4 \leq N \leq 40$  y siguieron investigando aproximaciones a la distribución discreta de  $D$  [Salama and Quade, 1990]. Si la variable  $X$  o la variable  $Y$  contiene valores empatados, entonces el cálculo de un valor de probabilidad exacto es más complejo.

### 1.2.2. Rangos Múltiples

La regla del pie de Spearman, tal como se presentó originalmente en sus artículos de 1904 y 1906 en el *American Journal of Psychology* [Spearman, 1904] y el *British Journal of Psychology* [Spearman, 1906], respectivamente, se limitaba a  $N \geq 2$  puntuaciones de rango no empatadas y  $b = 2$  jueces. Sin embargo, como Berry y Mielke demostraron en 1998, la regla de Spearman puede generalizarse para incluir tanto puntuaciones de rango empatadas como no empatadas y  $b \geq 2$  conjuntos de clasificaciones [Berry and Mielke, 1998]. Sea

$$\delta = [N \binom{b}{2}]^{-1} \sum_{i=1}^N \sum_{r < s} |X_{ri} - X_{si}|$$

denota una función de distancia media basada en todas las  $\binom{b}{2}$  posibles diferencias absolutas emparejadas entre los valores de las clasificaciones por  $b$  jueces y sea

$$\mu_\delta = [N^2 \binom{b}{2}]^{-1} \sum_{i=1}^N \sum_{j=1}^N \sum_{r < s} |X_{ri} - X_{sj}|$$

denota el valor esperado de  $\delta$  donde  $b$  es el número de jueces,  $N$  es el número de objetos, y  $\sum_{r < s}$  es la suma sobre todos los  $r$  y  $s$  tales que  $1 \leq r < s \leq N$ . Entonces, la generalización de la medida de la regla del pie de Spearman viene dada por

$$\mathfrak{R} = 1 - \frac{\delta}{\mu_\delta} ,$$

donde  $\mathfrak{R}$  es una medida de concordancia corregida por el azar entre los  $b$  jueces que no se limita a las puntuaciones de rango no empatadas. Nótese que en el caso de  $b = 2$  jueces, la ecuación anterior se reduce a la regla del pie de Spearman de 1906 para  $b = 2$  jueces.

### 1.3. Coeficiente de Concordancia

Mientras que el coeficiente de correlación de rango de Spearman y las medidas  $\tau_a$  y  $\tau_b$  de Kendall expresan el grado de asociación entre dos variables medidas en puntuaciones de rango o transformadas en ellas, el coeficiente de concordancia expresa el grado de asociación entre múltiples conjuntos de puntuaciones de rango.

En 1939, Maurice Kendall y Bernard Babington Smith publicaron un artículo en *The Annals of Mathematical Statistics* sobre “The problem of  $m$  rankings” en el que desarrollaron el conocido coeficiente de concordancia [Kendall and Smith, 1939]. Sean  $N$  y  $m$  el número de puntuaciones de rango y el número de jueces, respectivamente, entonces Kendall y Babington Smith definieron el coeficiente de concordancia como

$$W = \frac{12S}{m^2(N^3 - N)} ,$$

donde  $S$  es la suma observada de los cuadrados de las desviaciones de las sumas de los rangos respecto al valor medio  $m(N + 1)/2$ .

Dado que  $m^2(N^3 - N)$  en el denominador de la ecuación anterior es invariable en todas las permutaciones de los datos observados, Kendall y Babington Smith demostraron que para comprobar si un valor observado de  $W$  es estadísticamente significativo sólo es necesario considerar la distribución de  $S$  permutando los  $N$  rangos de todas las formas posibles e igualmente posibles. Si uno de los rangos es fijo, hay  $(N!)^{m-1}$  valores posibles de  $S$ . Basándose en este procedimiento de permutación, Kendall y Babington Smith crearon cuatro tablas que proporcionan valores de probabilidad exactos para  $N = 3$  y  $m = 2, \dots, 10$ ,  $N = 4$  y  $m = 2, \dots, 6$ , y  $N = 5$  y  $m = 3$ .

$W$  también puede definirse como

$$W = \frac{12 \sum_{i=1}^N R_i^2 - 3m^2N(N + 1)}{m^2N(N^2 - 1)} ,$$

donde  $R_i$  para  $i = 1, \dots, N$  es la suma de las puntuaciones de rango para el  $i$ -ésimo de los  $N$  objetos y no hay puntuaciones de rango empatadas. También se sabe que  $W$  puede definirse como una función del valor medio de todos los coeficientes de correlación de orden de rango de Spearman por pares, dado por

$$\bar{\rho} = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^m \rho_{ij} .$$



A este respecto, Kendall y Babington Smith demostraron que  $\bar{\rho}$  es simplemente el coeficiente intraclase,  $r_I$ , para los  $m$  conjuntos de clasificaciones. Las relaciones entre  $W$  y  $\bar{\rho}$  vienen dadas por

$$\bar{\rho} = \frac{mW - 1}{m - 1} \quad y \quad W = \frac{\bar{\rho}(m - 1) + 1}{m} .$$

Si todos los ordenamientos de los conjuntos de puntuaciones de rango observadas ocurren con igual probabilidad, el valor exacto de la probabilidad del valor observado de  $W$  calculado sobre  $M$  posibles ordenamientos igualmente probables de las puntuaciones de rango observadas bajo la hipótesis nula es

$$P(W \geq W_o | H_0) = \frac{\text{número de valores } W \geq W_o}{M} ,$$

donde  $W_o$  indica el valor observado de  $W$ .

### 1.3.1. Procedimientos relacionados

Desde hace tiempo se conoce que la estructura de datos para el coeficiente de concordancia de Kendall y Babington Smith [Kendall and Smith, 1939] es la misma que la del análisis de varianza de dos vías de Friedman para rangos [Friedman, 1937] y la misma que la de la relación de correlación de Wallis para datos de puntuación de rangos [Wallis, 1939]. Mientras que la prueba de Friedman, por ejemplo, proporciona un valor de probabilidad global de las diferencias globales entre rangos, existe un procedimiento relacionado que proporciona un valor de probabilidad exacto para la suma de rangos de un solo objeto, respondiendo a la pregunta: ¿cuándo el total de un solo objeto no se debe al azar bajo la hipótesis nula de asignación aleatoria?

Supongamos que cada uno de los  $N$  jueces asigna de forma independiente  $K$  rangos distintos no empatados a  $K \geq 2$  objetos. Si  $S$  denota la suma de los  $N$  rangos para un objeto determinado bajo la hipótesis nula de que cada uno de los  $N$  jueces asigna los  $K$  rangos a los  $K$  objetos al azar, es decir, que cada objeto ocurre con probabilidad  $1/K$ , entonces la probabilidad puntual exacta de  $S$  viene dada por

$$p_S = K^{-N} C_{S-N}$$

para  $S = N, N + 1, \dots, NK$ . Sea  $m = S - N$ , entonces

$$C_m = \sum_{j=0}^v (-1)^j \binom{N}{j} \binom{m - jK + N - 1}{N - 1}$$

y  $v = \min(N, m/K)$ , es decir, el mayor número entero no negativo menor o igual que  $N$  o  $m/K$ . El valor exacto de la probabilidad unilateral de  $S$  viene dado por

$$P_1 = \sum_{j=N}^w p_j ,$$

donde  $w = \min(S, NK + N - S)$  y el valor exacto de la probabilidad bilateral de  $S$  es dado por

$$P_2 = \min(2P_1, 1) ,$$

ya que la distribución de  $S$  es simétrica respecto a  $N(K + 1)/2$  bajo la hipótesis nula.

## 1.4. Medida de acuerdo $u$ de Kendall

A veces, en lugar de pedir a un grupo de jueces que clasifiquen un conjunto de objetos, se les presenta una serie de pares de objetos y se les pide que indiquen su preferencia por uno de los dos objetos de cada par. Este procedimiento se denomina comparación por pares. Cuando se recogen datos mediante el método de las comparaciones por pares, es posible calcular el grado de concordancia entre los jueces. En 1940, Kendall y Babington Smith [Kendall and Smith, 1940] propusieron un coeficiente de concordancia para evaluar las comparaciones pareadas entre  $k$  jueces para  $N$  clasificaciones, dado por

$$u = \frac{2S}{\binom{k}{2} \binom{N}{2}} - 1$$

donde

$$S = \sum_{i=1}^N \sum_{j=1}^N \binom{a_{ij}}{2} = \sum_{i=1}^N \sum_{j=1}^N a_{ij}^2 - k \sum_{i=1}^N \sum_{j=1}^N a_{ij} + \binom{k}{2} \binom{N}{2} ,$$

$a_{ij}$  es el número de veces que un objeto asociado a la fila  $i$  de una matriz de preferencias es preferido al objeto asociado a la fila  $j$ , y  $a_{ij} \geq 0$  para  $i, j = 1, \dots, N$ . El número máximo de concordancias, que se produce cuando  $\binom{N}{2}$  celdas de la matriz de preferencias contienen  $k$  cada una, es  $\binom{N}{2} \binom{k}{2}$  y, por tanto, sólo en el caso de concordancia completa,  $u = +1$  [Kendall and Smith, 1940].

Mientras que el valor máximo de  $u$  es  $+1$  cuando hay concordancia completa entre los  $k$  jueces, el número mínimo de concordancias se produce cuando cada celda de la matriz de preferencias contiene  $k/2$  si  $k$  es par o  $(k \pm 1)/2$  si  $k$  es impar. Así, cuando  $k$  es par, el valor mínimo de  $u$  es  $-1/(k-1)$  y cuando  $k$  es impar, el valor mínimo de  $u$  es  $-1/k$ . Como el valor esperado de  $u$  es cero [44] y los valores mínimos de  $u$  son  $-1/(k-1)$  cuando  $k$  es par y  $-1/k$  cuando  $k$  es impar,  $u$  es claramente una medida de concordancia corregida por el azar, aunque aparentemente esto no fue reconocido por Kendall y Babington Smith cuando desarrollaron  $u$  en 1940.

## 1.5. Medida Kappa de Cohen

En 1960, Jacob Cohen desarrolló el estadístico kappa, una medida corregida por el azar de la concordancia entre dos jueces para un conjunto de  $c$  categorías disjuntas y no ordenadas [Cohen, 1960]. En 1968, Cohen amplió el kappa para medir la concordancia entre dos jueces para un conjunto de  $c$  categorías ordenadas y disjuntas [Cohen, 1968]. El kappa original para  $c$  categorías disjuntas y no ordenadas se conoció como kappa “no ponderado”, o  $\kappa$ , y el kappa para  $c$  categorías disjuntas y ordenadas se conoció como kappa “ponderado”, o  $\kappa_w$ . Mientras que la kappa no ponderada no distinguía entre magnitudes de

desacuerdo, la kappa ponderada incorporaba la magnitud de cada desacuerdo y otorgaba un crédito parcial a las discrepancias cuando la concordancia no era completa [Maclure and Willett, 1987]. El enfoque habitual consiste en asignar pesos a cada par de desacuerdos (pesos mayores indican un mayor desacuerdo). El kappa no ponderado para  $c$  categorías disjuntas y no ordenadas se trata en el capítulo 2. A continuación se presenta el kappa ponderado para  $c$  categorías disjuntas y ordenadas:

La medición de la concordancia es un caso especial de medición de la asociación entre dos variables de nivel ordinal. Los índices de concordancia miden el grado en que un conjunto de medidas de respuesta son idénticas a otro conjunto, es decir, concuerdan, en lugar del grado en que un conjunto de medidas de respuesta es una función lineal de otro conjunto de medidas de respuesta, es decir, correlacionan. Al igual que la regla de Spearman, la medida kappa ponderada de Cohen de la concordancia es una medida corregida por el azar, que refleja la cantidad de concordancia por encima de lo que cabría esperar por azar. Así, la kappa ponderada es igual a uno cuando se produce una concordancia perfecta, es igual a cero en caso de independencia y puede ser ligeramente negativo cuando la concordancia es inferior a la esperada por el azar [Fleiss et al., 2003].

Para simplificar, consideremos  $N \geq 2$  objetos clasificados de forma cruzada por  $b = 2$  jueces independientes en una tabla de contingencia  $c \times c$  con  $c$  categorías disjuntas y ordenadas denotadas por  $a_1, \dots, a_c$ , como en la siguiente tabla:

Tabla 1.1: Notación para una tabla de validación cruzada de  $N$  objetos por 2 jueces en  $c$  categorías disjuntas y ordenadas

Juez 1 \ Juez 2	$a_1$	$a_2$	$\dots$	$a_c$	Total
$a_1$	$n_{11}$	$n_{12}$	$\dots$	$n_{1c}$	$n_{1.}$
$a_2$	$n_{21}$	$n_{22}$	$\dots$	$n_{2c}$	$n_{2.}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$a_c$	$n_{c1}$	$n_{c2}$	$\dots$	$n_{cc}$	$n_{c.}$
<b>Total</b>	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.c}$	$N$

Donde  $n_{i.}$  es el total de la frecuencia marginal de la  $i$ -ésima fila,  $i = 1, \dots, c$ ,  $n_{.j}$  es la frecuencia marginal total de la  $j$ -ésima columna,  $j = 1, \dots, c$ ,  $n_{ij}$  son las frecuencias de las celdas y  $w_{ij}$  denotan los pesos de las celdas. Cuando las  $c$  categorías para los  $b = 2$  jueces están ordenadas de forma similar, entonces  $n_{ii}$ ,  $i = 1, \dots, c$ , y  $n_{ij}$ ,  $i \neq j$ , denotan las frecuencias de las celdas de concordancia y discordancia, respectivamente.

Aunque se han propuesto diversos esquemas de ponderación para la kappa ponderada de Cohen, el más popular es la ponderación cuadrática dada por  $w_{ij} = (i - j)^2$  para  $i, j = 1, \dots, c$ , donde las ponderaciones de discordancia de la categoría progresan geométricamente hacia fuera desde la diagonal de concordancia, es decir,  $0^2, 1^2, 2^2, 3^2$ , etc. Sin embargo, la ponderación lineal en la que  $w_{ij} = |i - j|$  para  $i, j = 1, \dots, c$ , donde las ponderaciones de la categoría de desacuerdo progresan linealmente hacia fuera desde la diagonal de concordancia, es decir,  $0, 1, 2, 3$ , y así sucesivamente.

Una fórmula sencilla para el cálculo del estadístico de la prueba kappa ponderada de Cohen con  $b = 2$  jueces viene dada por

$$\kappa_w = 1 - \frac{\frac{1}{N} \sum_{i=1}^c \sum_{j=1}^c w_{ij} n_{ij}}{\frac{1}{N^2} \sum_{i=1}^c \sum_{j=1}^c w_{ij} n_{i.} n_{.j}} .$$

Dada una tabla de concordancia  $c \times c$  con  $N$  objetos clasificados por las valoraciones de dos jueces independientes en  $c$  categorías ordenadas y disjuntas, una prueba de permutación exacta genera un conjunto de referencia de todos los  $M$  posibles ordenamientos igualmente probables de los  $N$  objetos en las  $c^2$  celdas, preservando el número total de objetos en cada categoría, es decir, las distribuciones de frecuencia marginal. Para cada ordenamiento de frecuencias de celdas con distribuciones de frecuencias marginales fijas, se calculan el estadístico kappa ponderado,  $\kappa_w$ , y el valor de probabilidad exacto,  $p(n_{ij}|n_{i.}, n_{.j}, N)$ , donde

$$p(n_{ij}|n_{i.}, n_{.j}, N) = \frac{(\prod_{i=1}^c n_{i.}!)(\prod_{j=1}^c n_{.j}!)}{N! \prod_{j=1}^c \prod_{i=1}^c n_{ij}!}$$

es el valor de probabilidad hipergeométrico convencional de una tabla de contingencia  $c \times c$ .

Sea  $\kappa_o$  el valor del estadístico kappa ponderado observado y  $M$  el número total de ordenamientos distintos de la frecuencia de las celdas de los  $N$  objetos en la tabla de clasificación  $c \times c$ , dados los totales de frecuencia marginales fijos. Entonces el valor exacto de la probabilidad de  $\kappa_o$  bajo la hipótesis nula viene dado por

$$P(\kappa_o|H_0) = \sum_{k=1}^M \Psi(\kappa_k) p(n_{ij}|n_{i.}, n_{.j}, N) ,$$

donde

$$\Psi(\kappa_k) = \begin{cases} 1 & \kappa_k \geq \kappa_o \\ 0 & c.c. \end{cases} .$$

Cuando el conjunto de referencia de todos los  $M$  ordenamientos posibles es muy grande, los análisis exactos de permutación son poco prácticos y se hacen necesarias las aproximaciones de remuestreo de Monte Carlo. Sea  $L$  una muestra aleatoria de todos los  $M$  valores posibles de  $\kappa_w$ . Entonces, bajo la hipótesis nula, el valor de la probabilidad aproximada de remuestreo para el valor observado de  $\kappa_w$ ,  $\kappa_o$ , viene dado por

$$P(\kappa_o) = \frac{1}{L} \sum_{l=1}^L \Psi_l(\kappa_w) .$$

### 1.5.1. Comparación de la ponderación lineal y cuadrática

Existe una considerable controversia sobre qué pesos deben utilizarse con el estadístico kappa ponderado de Cohen,  $\kappa_w$ . La elección de los pesos es completamente arbitraria y se puede utilizar cualquier peso de celda de desacuerdo. La ponderación lineal es quizás la

más útil, porque como las ponderaciones de las celdas de desacuerdo progresan linealmente hacia fuera desde la diagonal de concordancia, su interpretación es más fácil.

Es evidente que la ponderación lineal y la ponderación cuadrática producen los mismos resultados para las tablas de contingencia  $2 \times 2$ . También está muy claro que los valores kappa ponderados linealmente y los ponderados cuadráticamente suelen diferir muy poco para las tablas de contingencia  $3 \times 3$ . La ponderación lineal y cuadrática generalmente produce mayores diferencias con tablas de contingencia más grandes. Brenner y Kliebsch demostraron que la forma lineal del coeficiente kappa ponderado es menos sensible al número de categorías que la forma cuadrática; en consecuencia, recomendaron usar la forma lineal siempre que el número de categorías de la escala ordinal sea grande [Brenner and Kliebsch, 1996].

### 1.5.2. Kappa ponderado con múltiples jueces

Aunque la kappa ponderada de Cohen fue diseñada originalmente para  $b = 2$  jueces independientes y se limita a ellos, la kappa ponderada puede generalizarse y ampliarse para medir la concordancia entre múltiples jueces. La generalización de la kappa de Cohen a múltiples jueces ha sido controvertida durante mucho tiempo, con muchos pasos en falso y callejones sin salida en el camino. En 1988 Berry y Mielke generalizaron la medida de concordancia kappa de Cohen para dar cabida a múltiples jueces [Berry and Mielke, 1988] y en 2008 Mielke, Berry y Johnston proporcionaron un eficiente algoritmo de remuestreo de Monte Carlo para analizar los datos de concordancia con múltiples jueces [Mielke et al., 2008].

En esta sección se presenta un procedimiento algorítmico para calcular el kappa ponderado y no ponderado con múltiples calificadores. Aunque el procedimiento es apropiado para cualquier número de  $c \geq 2$  categorías disjuntas y ordenadas y  $b \geq 2$  jueces, la descripción del procedimiento se limitan a  $b = 3$  jueces independientes para simplificar la presentación, sin pérdida de generalidad.

Consideremos  $b = 3$  jueces que clasifican independientemente  $N$  objetos en  $c$  categorías ordenadas y disjuntas. La clasificación puede conceptualizarse como una tabla de contingencia  $c \times c \times c$  con  $c$  filas,  $c$  columnas y  $c$  cortes. Sean  $n_{ijk}$ ,  $R_i$ ,  $C_j$  y  $S_k$  las frecuencias de las celdas y los totales de las frecuencias marginales de las filas, columnas y cortes para  $i, j, k = 1, \dots, c$  y sea el total de frecuencias dado por

$$N = \sum_{i=1}^c \sum_{j=1}^c \sum_{k=1}^c n_{ijk} .$$

La estadística de la prueba kappa ponderada de Cohen para una tabla de contingencia de tres vías viene dada por

$$\kappa_w = \frac{N^2 \sum_{i=1}^c \sum_{j=1}^c \sum_{k=1}^c w_{ijk} n_{ijk}}{\sum_{i=1}^c \sum_{j=1}^c \sum_{k=1}^c w_{ijk} R_i C_j S_k} ,$$

donde  $w_{ijk}$  son los pesos de desacuerdo asignados a cada celda para  $i, j, k = 1, \dots, c$ . Bajo la hipótesis nula de que los jueces clasifican los  $N$  objetos de forma independiente con totales de frecuencia marginal fijos,  $E[\kappa_w] = 0$ .

Como se ha comentado anteriormente, se han propuesto diversas funciones de ponderación para el kappa ponderado para dos jueces, donde las ponderaciones de celdas arbitrarias se denotan como  $w_{ij}$  e  $i$  y  $j$  designan las  $c$  categorías para cada juez,  $i, j = 1, \dots, c$ . Normalmente, los pesos de las celdas se definen de forma que  $w_{ii} = 0$  para  $i = 1, \dots, c$  y los pesos son simétricos, es decir,  $w_{ij} = w_{ji}$  para  $i, j = 1, \dots, c$ . Algunos ejemplos de sistemas de ponderación para dos jueces son la ponderación lineal, en la que  $w_{ij} = |i - j|$ , ponderación cuadrática donde  $w_{ij} = (i - j)^2$ , y kappa no ponderado donde

$$w_{ij} = \begin{cases} 0 & i = j \\ 1 & c.c. \end{cases}.$$

Para tres jueces, las ponderaciones de desacuerdo de las celdas vienen dadas por  $w_{ijk}$ , donde  $i, j$ , y  $k$  designan las  $c$  categorías de cada juez. De forma análoga a  $w_{ij}$ ,  $w_{ijk}$  puede ser definida de forma que  $w_{iii} = 0$  para  $i = 1, \dots, c$  y las ponderaciones son simétricas, es decir,  $w_{ijk} = w_{ikj} = w_{jik} = w_{jki} = w_{kji} = w_{kij}$  para  $i, j, k = 1, \dots, c$ . Ejemplos de sistemas de ponderación para tres jueces incluyen la ponderación lineal donde

$$w_{ijk} = |i - j| + |i - k| + |j - k|$$

y la ponderación cuadrática donde

$$w_{ijk} = (i - j)^2 + (i - k)^2 + (j - k)^2$$

para  $i, j, k = 1, \dots, c$ .

El kappa ponderado para tres jueces se reduce al kappa no ponderado cuando

$$w_{ijk} = \begin{cases} 0 & i = j = k \\ 1 & c.c. \end{cases}.$$

Dada una tabla de contingencia  $c \times c \times c$  con  $N$  objetos clasificados de forma cruzada por tres jueces independientes, una prueba de permutación exacta implica generar todos los ordenamientos posibles, igualmente probables, de los  $N$  objetos en las  $c^3$  celdas, preservando las distribuciones de frecuencia marginal observadas. Para cada ordenamiento en el conjunto de referencia de todas las permutaciones de las frecuencias de las celdas, se calcula el estadístico kappa ponderado,  $\kappa_w$ , y el valor de la probabilidad puntual hipergeométrica exacta bajo la hipótesis nula,  $p(n_{ijk}|R_i, C_j, S_k, N)$ , donde

$$p(n_{ij}|n_{i.}, n_{.j}, N) = \frac{(\prod_{i=1}^c R_i!)(\prod_{j=1}^c C_j!)(\prod_{k=1}^c S_k!)}{(N!)^2 \prod_{j=1}^c \prod_{i=1}^c \prod_{k=1}^c n_{ijk}!}$$

[Mielke and Berry, 1988].

Si  $\kappa_o$  denota el valor del estadístico kappa ponderado observado, el valor de probabilidad exacto de  $\kappa_o$  bajo la hipótesis nula viene dado por

$$P(\kappa_o|H_0) = \sum_{l=1}^M \Psi_l(n_{ijk}|R_i, C_j, S_k, N),$$

donde

$$\Psi_l(n_{ijk}|R_i, C_j, S_k, N) = \begin{cases} p(n_{ijk}|R_i, C_j, S_k, N) & \kappa_w \geq \kappa_o \\ 0 & c.c. \end{cases}$$

y  $M$  denota el número total de posibles ordenamientos de frecuencia de celdas igualmente probables dados los totales de frecuencia marginal observados. Cuando el conjunto de referencia de  $M$  ordenamientos posibles es muy grande, como es típico en las tablas de contingencia multidireccionales, las pruebas exactas son poco prácticas y se hacen necesarios los procedimientos de remuestreo de Monte Carlo. En el remuestreo, una muestra aleatoria de tamaño  $L$  extraída de los  $M$  posibles ordenamientos de frecuencias de celdas permite comparar los estadísticos de prueba  $\kappa_w$  calculados en las  $L$  tablas aleatorias con el estadístico de prueba  $\kappa_w$  calculado en la tabla observada.

## 1.6. Análisis Ridit

En 1958, I.D.J. Bross introdujo la puntuación ridit para el análisis de datos categóricos ordenados, donde “ridit” es un acrónimo de *Relative to an Identified Distribution* (relativo a una distribución identificada) y la “it” representa un tipo de transformación similar a logit y probit [Bross, 1958]. Son comunes dos aplicaciones del análisis ridit:

- La primera compara los grupos de tratamiento y de control, donde el grupo de control observado sirve como grupo de referencia y los ridits se calculan para las  $c$  categorías disjuntas y ordenadas del grupo de control y se aplican a las  $c$  categorías disjuntas y ordenadas del grupo de tratamiento.

En la primera aplicación, el grupo de control y los ridits correspondientes se tratan como una población infinita y parámetros de población, respectivamente.

- La segunda aplicación compara dos grupos de tratamiento independientes en los que ninguno de los grupos de tratamiento se considera un grupo de referencia y se calculan los ridits para las  $c$  frecuencias de categorías ordenadas y disjuntas de cada grupo de tratamiento; y se aplican a las  $c$  categorías ordenadas y disjuntas del otro grupo de tratamiento. En esta aplicación, los  $k = 2$  grupos de tratamiento se consideran muestras finitas independientes, sin que ninguno se identifique como grupo de referencia. El supuesto de la segunda aplicación de que los grupos son finitos es, en realidad, más realista. En 2009 Mielke, Long, Berry y Johnston generalizaron el análisis ridit para  $k \geq 2$  grupos de tratamiento independientes [Mielke et al., 2009].

Se considera una tabla de contingencia de clasificación cruzada  $c \times k$  con  $c$  categorías de respuesta disjuntas y ordenadas y  $k$  grupos de tratamiento desordenados. Siguiendo la notación de Bross,  $m_{ij}$  denota la frecuencia de celda observada de la  $i$ -ésima fila y  $j$ -ésima columna para  $i = 1, \dots, c$  y  $j = 1, \dots, k$ , sean

$$M_j = \sum_{i=1}^c m_{ij}$$

los totales de frecuencia de tratamiento no ordenados para  $j = 1, \dots, k$ , y sea

$$N = \sum_{i=1}^c \sum_{j=1}^k m_{ij}$$

el total de la frecuencia de la tabla para todas las  $ck$  celdas. Las puntuaciones ridit para el  $j$ -ésimo tratamiento observado,  $j = 1, \dots, k$ , vienen dadas por

$$R_{1j} = \frac{m_{1j}}{2M_j}, \quad R_{2j} = \frac{m_{1j} + \frac{m_{2j}}{2}}{M_j}, \quad \dots, \quad R_{cj} = \frac{m_{1j} + \dots + m_{c-1,j} + \frac{m_{cj}}{2}}{M_j}.$$

Así, la puntuación ridit  $R_{ij}$  para la  $i$ -ésima de las  $c$  categorías en el  $j$ -ésimo de los  $k$  tratamientos es la proporción de observaciones en las categorías inferiores a la  $i$ -ésima categoría en el  $j$ -ésimo tratamiento mas la mitad de la proporción de observaciones en la  $i$ -ésima categoría del  $j$ -ésimo tratamiento.

### 1.6.1. Cálculo

Se defina el estadístico de prueba  $T$  como

$$T = \sum_{i=1}^{k-1} \sum_{j=i+1}^k |x_{ij} - x_{ji}|,$$

donde

$$x_{ij} = \sum_{k=1}^c \frac{R_{ki} m_{kj}}{M_j}$$

para  $i, j = 1, \dots, k$ .

En el contexto de un análisis ridit de  $k$  tratamientos, los procedimientos de permutación exacta examinan todas las posibles asignaciones igualmente probables de los  $N$  sujetos a las  $c$  categorías ordenadas y disjuntas. Alternativamente, los procedimientos de permutación de remuestreo de Monte Carlo examinan un subconjunto aleatorio seleccionado de entre todas las posibles asignaciones de los  $N$  sujetos a las  $c$  categorías ordenadas y separadas. La hipótesis nula de una prueba de permutación específica que todos los resultados posibles del análisis ridit son igualmente probables.

#### 1.6.1.1. Procedimientos de permutación exacta

Los  $M_j$  sujetos del  $j$ -ésimo grupo de tratamiento,  $j = 1, \dots, k$ , se clasifican en  $c$  categorías disjuntas y ordenadas. Entre las  $c^N$  configuraciones de asignación igualmente probables bajo la hipótesis nula, hay

$$W = \prod_{j=1}^k \binom{M_j + c - 1}{c - 1}$$

particiones distintas de las  $c^N$  configuraciones de asignación de los  $k$  grupos de tratamiento. En una aplicación típica,  $W$  y  $c^N$  suelen ser muy grandes.

Por lo tanto, un análisis de permutación exacto no suele ser práctico para los análisis ridit con  $k > 2$  tratamientos y se recomiendan los procedimientos de permutación de Monte Carlo.

#### 1.6.1.2. Procedimientos de permutación de remuestreo

Un procedimiento de permutación de remuestreo de Monte Carlo genera  $L$  conjuntos de  $N$  asignaciones aleatorias seleccionadas con reemplazo de las  $c^N$  configuraciones de asignación igualmente probables de los  $k$  grupos de tratamiento. En general,  $L = 1.000.000$



es suficiente para garantizar una precisión de tres decimales [Johnston et al., 2007]. Para cada uno de los  $L$  conjuntos, los contadores de las  $c$  categorías disjuntas y ordenadas indexadas por  $i = 1, \dots, c$  se ponen a cero y se generan  $j$  variables aleatorias uniformes independientes,  $U_j$ , sobre  $[0, 1)$ , para  $j = 1, \dots, N$ . Si  $U_j$  pertenece a  $[\frac{i-1}{c}, \frac{i}{c})$ , el  $i$ -ésimo de los contadores  $c$  se incrementa en 1. A continuación se calcula el estadístico  $T$  de la prueba ridit para cada uno de los  $L$  conjuntos de  $N$  asignaciones aleatorias de las frecuencias de las categorías ordenadas. Sea  $T_o$  el valor observado de  $T$ . Entonces, dados los estadísticos ridit del remuestreo  $T_1, \dots, T_L$ , el valor de la probabilidad de la cola superior del remuestreo de  $T_o$  bajo la hipótesis nula viene dado por

$$P = \frac{1}{L} \sum_{i=1}^L \Psi(T_i) ,$$

donde

$$\Psi(T_i) = \begin{cases} 1 & T_i \geq T_o \\ 0 & c.c. \end{cases} .$$



# Bibliografia

- K.J. Berry and P.W. Mielke. A generalization of cohen's kappa agreement measure to interval measurement and multiple raters. *Educ. Psychol. Meas.*, 48, 1988.
- K.J. Berry and P.W. Mielke. Extension of spearman's footrule to multiple rankings. *Psychol. Rep.*, 82, 1998.
- H. Brenner and U. Kliebsch. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, 7, 1996.
- I.D.J. Bross. How to use riddit analysis. 14, 1958.
- J. Cohen. *A coefficient of agreement for nominal scales*, volume 20. 1960.
- J. Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.*, 70, 1968.
- J.L. Fleiss, B. Levin, and M.C. Paik. *Statistical Methods for Rates and Proportions*. 5 edition, 2003.
- L.A. Franklin. Exact tables of spearman's footrule for  $n = 11(1)18$  with estimate of convergence and errors for the normal approximation. *Stat. Probab. Lett.*, 6, 1988.
- M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.*, 32, 1937.
- J.E. Johnston, K.J. Berry, and P.W. Mielke. Permutation tests: Precision in estimating probability values. *Percept. Motor Skill*, 105, 2007.
- M.G. Kendall. *Rank Correlation Methods*, 3rd edn. 1962.
- M.G. Kendall and B. Babington Smith. The problem of  $m$  rankings. *Ann.Math. Stat.*, 10, 1939.
- M.G. Kendall and B. Babington Smith. On the method of paired comparisons. *Biometrika*, 31, 1940.
- M.G. Kendall, S.F.H. Kendall, and B. Babington Smith. The distribution of spearman's coefficient of rank correlation in a universe in which all rankings occur an equal number of times. *Biometrika*, 30, 1939.
- K. Krippendorff. *Bivariate agreement coefficients for reliability of data*. Borgatta, E.F. (ed.) Sociological Methodology, 1970.

- M. Maclure and W.C. Willett. Misinterpretation and misuse of the kappa statistic. *Am. J. Epidemiol.*, 126, 1987.
- P.W. Mielke and K.J. Berry. Cumulant methods for analyzing independence of r-way contingency tables and goodness-of-fit frequency data. *Biometrika*, 75, 1988.
- P.W. Mielke, K.J. Berry, and J.E. Johnston. Resampling probability values for weighted kappa with multiple raters. *Psychol. Rep.*, 102, 2008.
- P.W. Mielke, M.A. Long, K.J. Berry, and J.E. Johnston. g-treatment rdit analysis: Resampling permutation methods. *Stat. Methodol.*, 6, 2009.
- I.A. Salama and D. Quade. A note on spearman's footrule. *Commun. Stat. Simul. C*, 19, 1990.
- C.E. Spearman. The proof and measurement of association between two things. *Am. J. Psychol.*, 15, 1904.
- C.E. Spearman. 'footrule' for measuring correlation. *Brit. J. Psychol.*, 2, 1906.
- A. Stuart. Spearman-like computation of kendall's tau. *Brit. J. Math. Stat. Psy.*, 30, 1977.
- H.K. Ury and D.C. Kleinecke. Tables of the distribution of spearman's footrule. *J. R. Stat. Soc. C Appl.*, 28, 1979.
- W.A. Wallis. The correlation ratio for ranked data. *J. Am. Stat. Assoc.*, 34, 1939.