

Índice general

1. Medidas para Variables Ordinales (Parte 1)	3
1.1. Principales medidas de Asociación Ordinal por pares	3
1.1.1. Medida τ_a de Kendall de Asociación Ordinal	5
1.1.2. Medida τ_b de Kendall de Asociación Ordinal	5
1.1.3. Medida τ_c de Stuart de Asociación Ordinal	6
1.1.4. Medida γ de Goodman y Kruskal	6
1.1.5. Medidas d_{yx} y d_{xy} de Somers	7
1.2. Métodos estadísticos de permutación	7
1.3. Otras medidas de Asociación Ordinal por pares	9
1.3.1. Medidas $d_{y.x}$ y $d_{x.y}$ de Kim	9
1.3.2. Medida e de Asociación Ordinal de Wilson	9
1.3.3. Medida S de Whitfield para una Variable Binaria y una Variable Ordinal	9
1.3.4. Coeficiente de correlación rango-biserial de Cureton	10
Bibliografía	13

Capítulo 1

Medidas para Variables Ordinales (Parte 1)

Las medidas de las relaciones entre dos variables de nivel ordinal suelen ser más informativas que las medidas entre simples variables de nivel nominal (categóricas), ya que las categorías disjuntas y ordenadas suelen contener más información que las categorías disjuntas y desordenadas. Las medidas de asociación para dos variables de nivel ordinal suelen ser de dos tipos: las que se basan en las diferencias entre pares, como las medidas τ_a y τ_b de Kendall y la medida γ de Goodman y Kruskal, y las que se basan en otros criterios, como la medida kappa ponderada de Cohen de concordancia entre evaluadores y el análisis rídit de Bross.

En este capítulo se incluyen las medidas de asociación ordinal τ_a y τ_b de Kendall, τ_c de Stuart, las medidas asimétricas d_{yx} y d_{xy} de Somers, las medidas $d_{y.x}$ y $d_{x.y}$ de Kim, la medida e de Wilson o el coeficiente de correlación rango-biserial de Cureton [AÑADIR LAS DEL OTRO TEMA CUANDO LOS JUNTE].

1.1. Principales medidas de Asociación Ordinal por pares

Una serie de medidas de asociación para dos variables de nivel ordinal se basan en comparaciones por pares de las diferencias entre los rangos. El estadístico de prueba S , definido por Maurice Kendall en 1938 [Kendall, 1938], desempeña un papel importante; este estadístico se expresa a menudo como $S = C - D$, donde C y D indican el número de pares concordantes y discordantes, respectivamente.

Sean dos variables ordinales clasificadas de forma cruzada en una tabla de contingencia $r \times c$, donde r y c denotan el número de filas y columnas, respectivamente. Sean $n_{i.}$, $n_{.j}$, y n_{ij} los totales de frecuencia marginal de las filas, los totales de las frecuencias marginales de las columnas y el número de individuos en la celda ij , respectivamente, para $i = 1, \dots, r$ y $j = 1, \dots, c$, y sea N el número total de individuos en la tabla de contingencia $r \times c$, es decir

$$n_{i.} = \sum_{j=1}^c n_{ij} , \quad n_{.j} = \sum_{i=1}^r n_{ij} \quad y \quad N = \sum_{i=1}^r \sum_{j=1}^c n_{ij} .$$

La siguiente tabla muestra una notación convencional para una tabla de contingencia $r \times c$ para dos variables categóricas, X_i para $i = 1, \dots, r$ e Y_j para $j = 1, \dots, c$, con el orden determinado para X por $X_1 < X_2 < \dots < X_r$ y para Y por $Y_1 < Y_2 < \dots < Y_c$:

Tabla 1.1: Clasificación cruzada de dos variables ordinales, X e Y .

$X \backslash Y$	Y_1	Y_2	\dots	Y_c	Total
X_1	n_{11}	n_{12}	\dots	n_{1c}	$n_{1.}$
X_2	n_{21}	n_{22}	\dots	n_{2c}	$n_{2.}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
X_r	n_{r1}	n_{r2}	\dots	n_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	\dots	$n_{.c}$	N

Si X e Y representan las variables de fila y columna, respectivamente, hay $N(N - 1)/2$ pares de individuos en la tabla que pueden dividirse en cinco tipos de pares mutuamente exhaustivos y exclusivos: pares concordantes, pares discordantes, pares empatados en la variable X pero no en la variable Y , pares empatados en la variable Y pero no en la variable X , y pares empatados en ambas variables.

Para una tabla de contingencia $r \times c$, los pares concordantes (pares de objetos que están clasificados en el mismo orden tanto en la variable X como en la variable Y) vienen dados por:

$$C = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{ij} \left(\sum_{k=i+1}^r \sum_{l=j+1}^c n_{kl} \right),$$

los pares discordantes (pares de objetos que se clasifican en un orden en la variable X y en el orden inverso en la variable Y) vienen dados por:

$$D = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{i,c-j+1} \left(\sum_{k=i+1}^r \sum_{l=1}^{c-j} n_{kl} \right),$$

los pares de objetos empatados en la variable X pero que difieren en la variable Y vienen dados por:

$$T_x = \sum_{i=1}^r \sum_{j=1}^{c-1} n_{ij} \left(\sum_{k=j+1}^c n_{ik} \right),$$

los pares de objetos empatados en la variable Y pero que difieren en la variable X vienen dados por:

$$T_y = \sum_{j=1}^c \sum_{i=1}^{r-1} n_{ij} \left(\sum_{k=i+1}^r n_{kj} \right),$$

y los pares de objetos empatados en la variable X y en la variable Y vienen dados por:

$$T_{xy} = \frac{1}{2} \sum_{i=1}^r \sum_{j=1}^c n_{ij}(n_{ij} - 1) .$$

Entonces,

$$C + D + T_x + T_y + T_{xy} = \frac{N(N-1)}{2} .$$

Dados C , D , T_x , T_y y N , se suelen definir seis medidas de asociación ordinal, cada una de las cuales tiene el mismo numerador, $S = C - D$, pero diferentes denominadores¹².

1.1.1. Medida τ_a de Kendall de Asociación Ordinal

La primera de estas medidas de asociación por pares propuesta fue la τ_a de Kendall [Kendall, 1948], que es una medida simétrica de asociación ordinal y se diseñó originalmente para medir la asociación entre dos conjuntos de valores de rangos no empatados, donde los dos conjuntos de puntuaciones de rango se etiquetan habitualmente como X e Y , aunque los rangos también pueden representarse en una tabla de contingencia $r \times c$ donde $n_{i.} = n_{.j} = 1$ para $i = 1, \dots, r$ y $j = 1, \dots, c$.

La τ_a de Kendall se define simplemente como la diferencia entre las proporciones de pares concordantes y discordantes, está dada por

$$\tau_a = \frac{C}{\frac{N(N-1)}{2}} - \frac{D}{\frac{N(N-1)}{2}} = \frac{C - D}{\frac{N(N-1)}{2}} = \frac{2S}{N(N-1)} .$$

La τ_a de Kendall se presenta como una alternativa al coeficiente de correlación de rango de Spearman [Kraft and van Eeden, 1968].

1.1.2. Medida τ_b de Kendall de Asociación Ordinal

Cuando existen valores empatados, la medida de asociación ordinal τ_a de Kendall no es la mejor opción, ya que ignora los dos conjuntos de valores empatados, T_x y T_y . Por esta razón, Kendall desarrolló τ_b [Kendall, 1948], una alternativa a τ_a , dada por

$$\tau_b = \frac{S}{\sqrt{(C + D + T_x)(C + D + T_y)}} .$$

La τ_b de Kendall es una medida estrictamente monótona de asociación ordinal, es decir, para cada aumento de categoría en la variable X , se espera que haya un aumento de categoría en la variable Y . Por consiguiente, τ_b sólo puede alcanzar límites de ± 1 para tablas de contingencia en las que $r = c$ y las distribuciones de frecuencias marginales de filas y columnas sean idénticas. Más concretamente, τ_b no puede alcanzar generalmente valores de ± 1 debido a la desigualdad de Cauchy:

¹El número de pares empatados en ambas variables X e Y (T_{xy}) no se utiliza en ninguna de las seis medidas.

²En realidad hay más de seis medidas de asociación ordinal basadas en comparaciones por pares; aquí sólo se tratan las seis medidas más comunes.

El cuadrado de la suma de los productos de dos conjuntos será igual o menor que el producto de las sumas al cuadrado de dos conjuntos. Formalmente, para las variables X e Y ,

$$\left(\sum_{i=1}^N x_i y_i\right)^2 \leq \sum_{i=1}^N x_i^2 \sum_{i=1}^N y_i^2 .$$

En consecuencia, el numerador de τ_b será igual o menor que el denominador, permitiendo que τ_b alcance ± 1 sólo cuando todas las observaciones se concentren en una de las dos diagonales principales de la tabla de contingencia. Si ninguna frecuencia marginal es cero, esto significa que τ_b sólo puede alcanzar ± 1 para una tabla de contingencia cuadrada con distribuciones de frecuencias marginales idénticas. Es importante señalar que, dado que las categorías están ordenadas, las distribuciones de frecuencias marginales deben ser idénticas, no simplemente equivalentes.

Por lo tanto, la τ_b de Kendall no es la medida más apropiada de asociación ordinal para la tabla de contingencia $r \times c$, con $r \neq c$.

1.1.3. Medida τ_c de Stuart de Asociación Ordinal

Alan Stuart propuso la τ_c [Stuart, 1953], que modifica la τ_b de Kendall para las tablas de contingencia en las que $r \neq c$, la medida viene dada por

$$\tau_c = \frac{2mS}{N^2(m-1)} ,$$

donde $m = \min(r, c)$.

Si N es un múltiplo de m y $r = c$ con distribuciones marginales de frecuencia idénticas, de modo que todas las observaciones caen en la diagonal de la tabla de contingencia y todas las frecuencias de las celdas son iguales, el valor máximo de la S de Kendall viene dado por

$$S_{max} = \frac{N^2(m-1)}{2m} .$$

Entonces, si $N = m$,

$$\frac{N^2(m-1)}{2m} = \frac{N^2(N-1)}{2N} = \frac{N(N-1)}{2} .$$

Sin embargo, si N no es un múltiplo de m , la expresión de S_{max} sigue siendo un límite superior que no se puede alcanzar. De ello se concluye que la τ_c de Stuart puede alcanzar a veces (y para N grande, puede alcanzar casi siempre) ± 1 .

1.1.4. Medida γ de Goodman y Kruskal

En 1954 Goodman y Kruskal desarrollaron una nueva medida de asociación simétrica para dos variables de nivel ordinal que denominaron gamma, γ [Goodman and Kruskal, 1954]. Esta medida de asociación ordinal es una medida de reducción proporcional al error que se basa únicamente en los pares no empatados, C y D , y viene dada por

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

Por tanto, de la expresión se deduce que γ es simplemente la diferencia entre las proporciones de los pares iguales y no iguales, ignorando todos los pares empatados, es decir, T_x , T_y y T_{xy} .

Según Goodman y Kruskal, existe un problema potencial con γ , y es que es inestable en varios “puntos de corte”, es decir, γ tiende a aumentar a medida que las categorías de una tabla de contingencia se colapsan porque no tiene en cuenta los pares empatados (y el número de pares empatados aumenta a medida que la tabla se colapsa). Además, γ es una medida de asociación ordinal (no estrictamente) monótona, es decir, para cada aumento (disminución) de la categoría ordenada en la variable X , la variable Y aumenta (disminuye) o permanece igual.

1.1.5. Medidas d_{yx} y d_{xy} de Somers

En 1962 el sociólogo Robert Somers [Somers, 1962] propuso dos alternativas asimétricas a la medida γ de Goodman y Kruskal, dadas por

$$d_{yx} = \frac{C - D}{C + D + T_y} = \frac{S}{C + D + T_y} ,$$

donde T_y denota el número de pares empatados en la variable Y pero no empatados en la variable X , y

$$d_{xy} = \frac{C - D}{C + D + T_x} = \frac{S}{C + D + T_x} ,$$

donde T_x denota el número de pares empatados en la variable X pero no empatados en la variable Y .

A diferencia de las cuatro medidas simétricas, τ_a , τ_b , τ_c y γ , las medidas d_{yx} y d_{xy} de Somers dependen de qué variable, Y o X , se considera que es la dependiente. Observando las ecuaciones anteriores, se ve que Somers incluyó en los denominadores de d_{yx} y d_{xy} el número de valores empatados en la variable dependiente: T_y para d_{yx} y T_x para d_{xy} . La razón para incluir los valores empatados es simplemente que cuando la variable Y es la variable dependiente (d_{yx}), si dos valores de la variable independiente, X , difieren pero los dos valores correspondientes de la variable dependiente, Y , no difieren (están empatados), hay evidencia de una falta de asociación y los empates en la variable Y (T_y) deben ser incluidos en el denominador donde actúan para disminuir el valor de d_{yx} . El mismo razonamiento es válido para la medida d_{xy} de Somers.

Por último, es evidente que la medida τ_b de Kendall de asociación ordinal es simplemente la media geométrica de las medidas d_{yx} y d_{xy} de Somers, dadas por

$$\tau_b = \sqrt{d_{yx}d_{xy}} .$$

1.2. Métodos estadísticos de permutación

Para aplicar las técnicas exactas de permutaciones en una tabla de contingencia $r \times c$, es necesario calcular la medida de asociación ordinal seleccionada para las frecuencias de celdas observadas y enumerar exhaustivamente todas las M posibles ordenaciones igualmente probables de los N individuos en las rc celdas, fijadas las distribuciones de frecuencias marginales observadas. Para cada ordenación en el conjunto de referencia de

todas las permutaciones, se calcula una medida de asociación ordinal, digamos T , y la probabilidad hipergeométrica exacta puntual bajo la hipótesis de no asociación ordinal, $p(n_{ij}|n_{i.}, n_{.j}, N)$, donde

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

Si T_o denota el valor del estadístico observado, los p-valores a una cola (P) de T_o son las sumas de los valores $p(n_{ij}|n_{i.}, n_{.j}, N)$ asociados a los valores T calculados en las posibles ordenaciones igualmente probables de las frecuencias de las celdas que son iguales o mayores que T_o cuando T_o es positivo, $Pr[T \geq T_o | \text{no asociación ordinal}]$, e iguales o menores que T_o cuando T_o es negativo, $Pr[T \leq T_o | \text{no asociación ordinal}]$. Así, la probabilidad exacta hipergeométrica de T_o cuando T es positivo viene dado por

$$P = \sum_{k=1}^M \Psi_p(T_k) p(n_{ij}|n_{i.}, n_{.j}, N) ,$$

donde

$$\Psi_p(T_k) = \begin{cases} 1 & T_k \geq T_o \\ 0 & \text{c.c.} \end{cases}$$

y la probabilidad exacta hipergeométrica de T_o cuando T es negativo viene dado por

$$P = \sum_{k=1}^M \Psi_n(T_k) p(n_{ij}|n_{i.}, n_{.j}, N) ,$$

donde

$$\Psi_n(T_k) = \begin{cases} 1 & T_k \leq T_o \\ 0 & \text{c.c.} \end{cases} .$$

Cuando el número de posibles ordenaciones de las frecuencias de las celdas es muy grande, los tests exactos de permutaciones son poco prácticos y son necesarios los métodos de Monte Carlo. Los métodos estadísticos de permutación de Monte Carlo generan una muestra aleatoria de todas las posibles ordenaciones de las frecuencias de las celdas, extraídos con reemplazo, dadas las distribuciones de frecuencias marginales observadas. Los p-valores (de cola superior e inferior) del estadístico T son simplemente las proporciones de los valores T calculados en las ordenaciones de frecuencias de celdas seleccionadas aleatoriamente que son iguales o mayores que T_o cuando T_o es positivo e iguales o menores que T_o cuando T_o es negativo. Así, el p-valor de remuestreo de Monte Carlo de T_o cuando T es positivo viene dado por

$$P(T \geq T_o | \text{No asociación ordinal}) = \frac{\text{número de veces que } T \geq T_o}{L} ,$$

donde L denota el número de ordenaciones aleatorias de los datos observados.

1.3. Otras medidas de Asociación Ordinal por pares

1.3.1. Medidas $d_{y.x}$ y $d_{x.y}$ de Kim

En 1971, Kim propuso medidas asimétricas proporcionales de reducción del error de asociación ordinal dadas por

$$d_{y.x} = \frac{C - D}{C + D + T_x} \quad y \quad d_{x.y} = \frac{C - D}{C + D + T_y}$$

[Kim, 1971]. A diferencia de las medidas d_{yx} y d_{xy} de Somers de asociación ordinal, que ajustan los empates en la variable dependiente; las medidas $d_{y.x}$ y $d_{x.y}$ de Kim ajustan los empates en la variable independiente. Es evidente que las medidas $d_{y.x}$ y $d_{x.y}$ de Kim son equivalentes a las medidas d_{xy} y d_{yx} de Somers, respectivamente.

1.3.2. Medida e de Asociación Ordinal de Wilson

En 1974, Wilson propuso otra medida de asociación ordinal que denominó e [Wilson, 1974]. Argumentando que una medida de asociación debería ajustarse para los valores empatados tanto en la variable X como en la variable Y , Wilson sugirió una medida simétrica de asociación ordinal dada por

$$e = \frac{C - D}{C + D + T_x + T_y} = \frac{S}{C + D + T_x + T_y}.$$

Como observó Wilson, e toma los valores de ± 1 si y sólo si los datos son estrictamente monótonos. Es obvio, a partir de la ecuación anterior, que la e de Wilson es equivalente a la d_{yx} de Somers cuando $T_x = 0$ y es equivalente a la d_{xy} de Somers cuando $T_y = 0$. Además, si $T_x = 0$ como $T_y = 0$, entonces $e = d_{yx} = d_{xy} = \gamma = \tau_a = \tau_b$.

1.3.3. Medida S de Whitfield para una Variable Binaria y una Variable Ordinal

En 1947 John Whitfield propuso una medida de correlación entre dos variables en la que una variable estaba compuesta por N rangos y la otra variable era dicotómica [Whitfield, 1947]. Un ejemplo servirá para ilustrar el procedimiento de Whitfield:

Considere las puntuaciones de rango que figuran en la siguiente tabla, donde las categorías de la variable dicotómica vienen dadas por “0” y “1” y las puntuaciones de rango van de 1 a 6:

Tabla 1.2: Ejemplo para la medida S de Whitfield

Casos	1	2	3	4	5	6
Binaria	0	1	0	0	0	1

Sea $n_0 = 4$ el número de casos en la categoría “0”, sea $n_1 = 2$ el número de casos en la

categoría “1” y sea $N = n_0 + n_1$.

Whitfield diseñó un procedimiento para calcular un estadístico que denominó S : teniendo en cuenta las $N = 6$ puntuaciones de la tabla, se consideran los $n_0 = 4$ casos en la categoría identificada por “0”: 1, 3, 4 y 5. Empezando por el caso 1 con la categoría “0”, no hay puntuaciones de rango con la categoría “1” a la izquierda de “0” y hay dos casos con la categoría “1” a la derecha de “0”, por lo que $0 - 2 = -2$. Para el caso 3 con la categoría “0”, hay una puntuación de rango a la izquierda de “0” con la categoría “1” (rango 2) y una puntuación de rango a la derecha de “0” con la categoría “1”; por tanto, $1 - 1 = 0$. Igualmente, para el caso 4: $1 - 1 = 0$ y para el caso 5: $1 - 1 = 0$. La suma de las diferencias entre las variables “0” y “1” es $S = -2 + 0 + 0 + 0 = -2$. De este modo, el enfoque de Whitfield se adapta a muestras con $n_0 = n_1$, así como a cualquier número de puntuaciones de rango empatadas.

Como el número de pares posibles de N enteros consecutivos viene dado por

$$\frac{N(N-1)}{2}$$

Whitfield definió y calculó la medida de asociación de rangos entre las variables “0” y “1” como

$$\tau = \frac{2S}{N(N-1)} .$$

La S de Whitfield es idéntica a la S de Kendall y está directamente relacionada con el estadístico U de suma de rangos de dos muestras de Mann y Whitney y con el estadístico W de suma de rangos de dos muestras de Wilcoxon. Las relaciones entre los estadísticos S de Whitfield y U de Mann y Whitney [Mann and Whitney, 1947] vienen dadas por

$$S = 2U - n_0n_1 \quad y \quad U = \frac{S + n_0n_1}{2}$$

y las relaciones entre la S de Whitfield y la W de Wilcoxon [Wilcoxon, 1945] vienen dadas por

$$S = n_1(N+1) - 2W \quad y \quad W = \frac{n_1(N+1) - S}{2} .$$

1.3.4. Coeficiente de correlación rango-biserial de Cureton

Sean dos variables, una representada por rangos y la otra por una dicotomía. En 1956, el psicólogo Edward Cureton propuso una nueva medida de correlación para una variable de rangos y una variable dicotómica denominada r_{rb} para la correlación rango-biserial [Cureton, 1956].

Cureton afirmó que el coeficiente de correlación debería tipificarse adecuadamente entre ± 1 y debería ser estrictamente no paramétrico, definido únicamente en términos de inversiones y concordancias entre pares de rangos, sin el uso de medias, varianzas, covarianzas o regresión. El coeficiente de correlación rango-biserial se define como:

$$r_{rb} = \frac{S}{S_{max}} ,$$

donde $S = C - D$ es el estadístico del test de Kendall [Kendall, 1938] y Whitfield [Whitfield, 1947], con C el número de pares concordantes y D el número de pares discordantes; y $S_{max} = n_0 n_1$, con n_0 el número de puntuaciones de rango en la categoría “0” y n_1 el número de puntuaciones de rango en la categoría “1”.

En 1966, Glass [Glass, 1966] dedujo una fórmula simplificada para la r_{rb} , suponiendo que no hay puntuaciones de rango empatadas, dada por

$$r_{rb} = \frac{2}{N}(\bar{y}_1 - \bar{y}_0) ,$$

donde \bar{y}_0 y \bar{y}_1 son las medias aritméticas de los valores de la variable dicotómica codificados como “0” y “1”, respectivamente.

Glass proporcionó dos fórmulas de cálculo alternativas dadas por

$$r_{rb} = \frac{2}{n_0}(\bar{y}_1 - \frac{N+1}{2}) \quad \text{ó} \quad r_{rb} = \frac{2}{n_1}(\frac{N+1}{2} - \bar{y}_0) .$$

Bibliografia

- E.E. Cureton. Rank-biserial correlation. *Psychometrika*, 21, 1956.
- G.V Glass. Note on rank-biserial correlation. *Educ. Psychol. Meas.*, 26, 1966.
- L.A. Goodman and W.H. Kruskal. Measures of association for cross classifications. *J. Am. Stat. Assoc.*, 49, 1954.
- M.G. Kendall. A new measure of rank correlation. *Biometrika*, 30, 1938.
- M.G. Kendall. *Rank Correlation Methods*. 1948.
- J.-O. Kim. Predictive measures of ordinal association. *Am. J. Soc.*, 76, 1971.
- C.A. Kraft and C. van Eeden. *A Nonparametric Introduction to Statistics*. Macmillan, 1968.
- H.B. Mann and D.R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.*, 18, 1947.
- R.H. Somers. A new asymmetric measure of association for ordinal variables. *Am. Sociol. Rev.*, 27, 1962.
- A. Stuart. The estimation and comparison of strengths of association in contingency tables. *Biometrika*, 40, 1953.
- J.W. Whitfield. Rank correlation between two variables, one of which is ranked, the other dichotomous. *Biometrika*, 34, 1947.
- F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bull.*, 1, 1945.
- T.P. Wilson. *Measures of association for bivariate ordinal hypotheses*. Blalock, H.M., 1974.