Medidas de asociación para variables ordinales

Las variables ordinales son aquellas cuyas categorías se presentan jerarquizadas, ya que reconocen una disposición inalterable (por ejemplo: la variable "edad" considerada en un nivel de medición ordinal podría presentar las siguientes categorías: niño, joven, adulto, anciano. Decimos que la disposición de las categorías es inalterable ya que no se puede ser adulto antes de ser joven, anciano antes que niño, etc.). Dada esta característica es posible observar la existencia —o no— de asociación en el comportamiento de dos variables de este tipo, esto es, verificar si ante la concentración de casos (frecuencia) en un extremo de una variable ordinal *A*, dicha concentración se verifica en un mismo extremo de una variable ordinal *B* (asociación directa o positiva, con valor cercano a 1) o en el extremo opuesto (asociación inversa o negativa, con valor cercano a –1). Si la variación de una no afecta la variación de la otra, hay independencia estadísticas entre ambas variables (valor cercano a 0).

PARES DE VALORES

Para realizar este tipo de observaciones es necesario registrar como se comportan los valores de cada variable, considerados de a pares (dentro de cada variable). Para ello debemos saber el total de pares de valores que es posible encontrar (*T*), sin repeticiones, siendo N el total de casos. Se calculan de acuerdo a la siguiente ecuación:

$$T = \frac{N\,(N-1)}{2}$$

Si N = 4, *T* = 6. Los pares serían AB, AC, AD, BC, BD, CD, es decir, pares en los cuales difiere al menos uno de sus elementos.

Realizando esta operación para cada variable, la ordenación de pares entre ambas variables tiene cinco posibilidades, que observaremos sobre un ejemplo, a fin de hacerlo más comprensible.

Ingresos	Nivel educativo			
	Alto	Medio	Bajo	Total
Alto	45 ^{s}	23	12 ^{d}	80
Medio	37	11	18	66
Bajo	22	19	34	75
Total	104	53	64	221

a) *Pares semejantes o concordantes (N_s)*: son pares que se encuentran ordenados en sentido idéntico en cada variable (celda marcada con *s*).

b) *Pares desemenjantes o discordantes (N_d)*: son pares que se encuentran ordenados en sentido opuesto en cada variable (celda marcada con *d*)

c) *Pares empatados solo en la variable independiente*: se trata de pares que ocupan la misma posición, es decir, el mismo valor ordinal; pero solo en la variable independiente, no en la dependiente. En este caso, "Nivel educativo". Se lo representa como T_x.

d) *Pares empatados solo en la variable dependiente*: igual caso que el anterior, pero en la variable dependiente. En este caso, "Ingresos". Se lo representa como T_y.

e) *Pares empatados en ambas variables*: situación combinada de los casos c) y d). Se lo representa como T_{xy}.

La diagonal positiva (la que contiene las celdas "alto–alto" y "bajo–bajo") va del extremo superior izquierdo al extremo inferior derecho. La diagonal negativa (la que contiene las celdas "alto–bajo" y "bajo–alto") va del extremo inferior izquierdo al extremo superior derecho. La celda *s* es el final de la diagonal positiva, y *d*, el final de la diagonal negativa.

Hagamos el cálculo de *T*.

$$T = \frac{221\ (221 - 1)}{2}$$

$T = 24.310$

Ahora calculemos los distintos tipos de pares:

a) Pares semejante o concordantes (N_s): surgen de la sumatoria del producto de las celdas extremas, partiendo de *s*, y la sumatoria de las frecuencias de las celdas internas. Vamos a observarlo gráficamente.

45		
	11	18
	19	34

45 x (11 + 18 + 19 + 34) = 3.690

	23	
		18
		34

23 x (18 + 34) = 1.196

37		
	19	34

37 x (19 + 34) = 1.961

	11	
		34

11 x 34 = 374

N_s = 3690 + 1196 + 1961 + 374 = 7221

b) Pares desemejantes o discordantes (N_d): es igual procedimiento al anterior, pero partiendo de la celda *d*.

		12
37	11	
22	19	

12 x (37 + 11 + 22 + 19) = 1.068

	23	
37		
22		

23 x (37 + 22) = 1.357

		18
22	19	

18 x (22 + 19) = 738

	11	
22		

11 x 22 = 242

N_d = 1068 + 1357 + 738 + 242 = 3405

c) Cálculo de pares empatados sólo en la variable independiente (T_x):

45		
37		
22		

45 x (37 + 22) = 2.655

37		
22		

37 x 22 = 814

	23	
	11	
	19	

23 x (11 + 19) = 690

	11	
	19	

11 x 19 = 209

		12
		18
		34

		18
		34

12 x (18 + 34) = 624

18 x 34 = 612

T_x = 2655 + 814 + 690 + 209 + 624 + 612 = 5604

d) Cálculo de pares empatados sólo en la variable independiente (T_y):

45	23	12

37	11	18

45 x (23 + 12) = 1.575

37 x (11 + 18) = 1.073

	23	12

	11	18

23 x 12 = 276

11 x 18 = 198

22	19	34

	19	34

22 x (19 + 34) = 1.166

19 x 34 = 646

$T_y = 1575 + 1073 + 276 + 198 + 1166 + 646 = 4934$

e) Cálculo de pares empatados en ambas variables (T_{xy}): se calcula mediante una ecuación, en la que f es la frecuencia de cada celda.

$$T_{xy} = \sum \frac{f(f-1)}{2}$$

45 x (45 – 1) / 2 = 990	37 x (37 – 1) / 2 = 666	22 x (22 – 1) / 2 = 231
23 x (23 – 1) / 2 = 253	11 x (11 – 10) / 2 = 55	19 x (19 – 1) / 2 = 171
12 x (12 – 1) / 2 = 66	18 x (18 – 1) / 2 = 153	34 x (34 – 1) / 2 = 561

$T_{xy} = 990 + 666 + 231 + 253 + 55 + 171 + 66 + 153 + 561 = 2255$

Estamos ya en condiciones de aplicar coeficientes para determinar el grado de asociación entre variables ordinales. En todos los coeficientes veremos que el numerador es $N_s - N_d$, la diferencia entre los pares semejantes y los discordantes. Esta diferencia nos indicará el sentido de la asociación: positivo si $N_s > N_d$, negativo si $N_s < N_d$.

Coeficiente *tau–a* de Kendall

Es el más sencillo de los coeficientes, y se define como la razón entre la diferencia de pares concordantes y discordantes con la totalidad de pares posibles (T). La ecuación es la siguiente:

$$t_a = \frac{N_s - N_d}{T}$$

En el ejemplo que hemos tomado:

$$t_a = \frac{7221 - 3405}{24310} = 0{,}157$$

El tau–a varía entre 1 y –1. El signo indica el sentido de la asociación, y el 0 la independencia estadística. No tiene limitaciones de cantidad de celdas (es decir, de cantidad de categorías de cada variable), y es una medida simétrica, esto es, que no importa cuál sea la variable dependiente y cuál la independiente, ya que en ambos sentidos tiene los mismos resultados.

Una limitación importante es que en ciertos casos no alcanza el 1 (o –1) aún existiendo asociación perfecta, ya que para ello requiere que todas las celdas centrales tengan frecuencia 0.

Coeficiente *gamma* de Goodman y Kruskal

Para paliar este problema es necesario eliminar los casos centrales no solo del numerador, sino también del denominador. Esto es lo que corrige el coeficiente gamma (γ) de Goodman y Kruskal. Es también una medida simétrica, como el tau–a, pero, a diferencia de ésta, siempre puede alcanzar los extremos (1 y –1). La fórmula es la siguiente:

$$\gamma = \frac{N_s - N_d}{N_s + N_d}$$

En el ejemplo que presentamos, el gamma se calcula de la siguiente manera:

$$\gamma = \frac{7221 - 3405}{7221 + 3405} = 0{,}359$$

Se observan dos cuestiones. En primer lugar, que es más sensible que el tau–a para una misma situación. En segundo término, que si tenemos dos variables dicotómicas, es idéntico al *Q* de Yule. De allí que algunos autores lo consideren como una versión general del coeficiente *Q* de Yule, pero para tablas de más de cuatro celdas.

Coeficiente *rho* de Spearman

Este es uno de los coeficientes más utilizados para medir asociación entre variables ordinales. Tiene una lógica diferente a los que hemos visto. Esta medida está indicada para cuando se ordenan atributos (de una variable dependiente) en dos categorías de una variable independiente (que puede ser nominal). Su cálculo requiere de un ordenamiento por alguna medida resumen, como por ejemplo la media. Veamos su imp0lementación con un ejemplo.

Se le pregunta a estudiantes y docentes que puntúen de 0 a 10 a una grupo de asignaturas, siendo 0 la mayor dificultad para el proceso enseñanza–aprendizaje, y 10, la menor dificultad en dicho proceso. Obtenidas las respuestas, se calculan las medias para cada una de las asignaturas, según hayan respondido estudiantes o docentes, y se construye el siguiente cuadro:

Asignatura	Docentes		Estudiantes		d	d^2
	Media	Orden	Media	Orden		
Matemática	1,34	1	2,33	1	0	0
Historia	6,58	5	8,44	7	–2	4
Geografía	5,43	3	6,70	5	–2	4
Castellano	7,10	6	7,94	6	0	0
Inglés	7,11	7	4,58	4	3	9
Física	6,24	4	3,25	2	2	4
Química	5,12	2	4,11	3	–1	1
				Σ	0	22

Lo que hicimos fue calcular las medias de las respuestas para cada asignatura y ordenarlas en función de dichas medias. Así, para los docentes, la mayor dificultad está en Matemática, con 1,34 de media, seguido de Química con una media de 5,12; etc. Para los estudiantes también el primer lugar en el grado de dificultad está en la Matemática (media de 2,33), pero en orden de dificultades la sigue Física (3,25 de media), etc.

Para calcular el γ lo que hacemos es ver la variación en el orden obtenido, observamos esa diferencia en los lugares que han asignado los dos grupos sociales de nuestro interés (estudiantes y docentes) a cada una de las asignaturas. Así vemos que en algunas asignaturas no hay diferencias (Matemática y Castellano), mientras que en otras sí existen diferencias. Calculamos las diferencias (columna *d*) y luego elevamos esa diferencia al cuadrado (columna d^2). Esto se realiza porque, como sabemos (por propiedades de las medias) la sumatoria de las diferencias es siempre igual a cero.

El γ se obtiene de la siguiente manera:

$$\gamma = 1 - \frac{6 \,\Sigma\, d^2}{n\,(n^2 - 1)}$$

En el ejemplo tomado tenemos que el primer miembro es $6\,\Sigma\, d^2 = 6 \times 22 = 132$; el segundo miembro es $n\,(n^2 - 1) = 7\,(49 - 1) = 336$. La razón entre ellos (132 / 336) es = 0,39. Finalmente, el γ = 1 – 0,39 = 0,61.

En este caso encontramos una asociación de 0,61 (entre media y alta) en la dificultad para el proceso enseñanza–aprendizaje tanto para docentes como para estudiantes. Si opinaran exactamente lo mismo el γ = 1. Si opinaran exactamente lo contrario, hubiésemos obtenido un γ = –1.