Capítulo 8

MEDIDAS DE ASOCIACION PARA VARIABLES NOMINALES Y ORDINALES

Son muy variadas las medidas de asociación de que puede disponer un sociólogo interesado en el estudio de relaciones bivariables. En el capítulo anterior tuvimos ocasión de estudiar algunas de ellas basadas en el valor de delta, o diferencia entre la frecuencia observada y la frecuencia esperada. Pero algunos de los coeficientes estudiados en dicho capítulo no son de interés para el investigador social, ya que no están «normatizados» y, por lo tanto, no está recomendada su utilización comparativa entre diferentes tablas, y menos aún la interpretación del carácter de la asociación. En el presente capítulo vamos a estudiar las medidas de asociación basadas en el criterio de «reducción proporcional del error», por ser las más utilizadas por los sociólogos, y ello para las relaciones entre variables medidas a nivel nominal y a nivel ordinal. En el próximo capítulo continuaremos con el estudio de las medidas basadas en el mismo criterio de reducción del error, pero para el caso de variables de intervalo, con lo que abordaremos uno de los temas centrales de la estadística, el estudio de la regresión simple.

Dado el carácter introductorio del presente libro, no vamos a estudiar las medidas de asociación apropiadas para situaciones especiales, porque esperamos que, con el bagaje de técnicas estadísticas que se presentan aquí, el estudiante de sociología puede pasar a realizar por sí mismo una investigación empírica sólida. Por ello remitimos al lector interesado en medidas de asociación especiales a otros libros, tales como el de Freeman (1971), y algunos otros trabajos que se citan en la bibliografía, para que pueda estudiar y conocer las mismas.

8.1. Medidas de asociación basadas en el criterio de «reducción proporcional del error» (RPE)

Un simple repaso al estudio de las diferentes medidas de asociación disponibles para el estudio de datos pone rápidamente de manifiesto la dificultad de encontrar un principio lógico consistente que sea

capaz de suministrar una visión integral de la asociación a todos los niveles de medición. Como señalan Leik y Gove (1971, pág. 279), al avanzar los datos del nivel nominal al ordinal y de éste al de intervalo, las medidas de asociación debieran simplemente incorporar las propiedades matemáticas que se van acumulando al tipo de expresión o fórmula utilizado en los niveles más bajos. Si esto se cumpliera, las medidas de asociación entre variables ordinales serían las mismas que para medidas nominales, pero utilizando datos ordenados jerárquicamente. E, igualmente, las medidas de asociación para variables de intervalo serían las mismas que las empleadas con medidas ordinales, pero utilizando el grado de distancia.

Sin embargo, ésta no es la situación actual con los procedimientos de que dispone el investigador que desea analizar unos datos sociológicos determinados. Se han intentado diversos procedimientos para establecer un principio lógico básico que dé coherencia a los diferentes tipos de medidas de asociación, pero todos presentan algún tipo de limitación. Con todo, es preciso recurrir a algún tipo de lógica para ordenar la presentación de las diferentes medidas de asociación, ya que, de lo contrario, se corre el peligro de que el estudiante de estadística y de sociología se desoriente ante la diversidad existente de índices.

Desde luego, ese principio lógico ordenador no se puede encontrar en los coeficientes basados en chi-cuadrado, porque, como ya señaló Blalock hace algún tiempo (1960, pág. 230), «todas las medidas basadas en chi-cuadrado son de naturaleza un tanto arbitraria, y su interpretación deja mucho que desear». En efecto, ya vimos en el capítulo anterior que el propio coeficiente de chi-cuadrado está relacionado con el tamaño de la muestra y con los grados de libertad, lo que dificulta su comparación para tablas de tamaño diferente.

Mejores perspectivas presenta el principio de «reducción proporcional del error» (*RPE*), sugerido por Costner (1965), inspirándose en los trabajos de Goodman y Kruskal (1954) y Guttman (1941), y desarrollado por Kim (1971). Las medidas de tipo *RPE* consisten en simples cocientes o *ratios* de la cantidad de error cometido al predecir la variable dependiente en dos situaciones: primeramente, la predicción se realiza cuando no se conoce más que la distribución de la propia variable dependiente y, en segundo lugar, la predicción se realiza cuando se dispone del conocimiento adicional de una variable independiente y de la forma en que la variable dependiente se distribuye dentro de las categorías de dicha variable independiente. Lo que realizan las medidas tipo *RPE* es simplemente formular la proporción en que se puede reducir el error cometido en la primera de las situaciones descritas, al utilizar la información que suministra la segunda de las situaciones. Es decir:

$$RPE = \frac{\text{Reducción del error con más información}}{\text{Cantidad original de error}}$$

Más recientemente, Leik y Gove (1971, págs. 279-301) han propuesto un principio lógico diferente, ya que, según estos autores, al asumir las medidas tipo *RPE* se introduce todavía una cierta diversidad en la forma en que se especifican las reglas de predicción. El nuevo principio lógico se basa en la predicción de pares de valores, en lugar de la predicción de valores únicos. Pero, dado que este nuevo modelo no se ha impuesto universalmente, vamos a basar nuestra presentación de las medidas de asociación en el criterio *RPE*, por su mayor implantación en el trabajo de análisis que realizan en la actualidad los sociólogos.

Por otro lado, y tal como señalan acertadamente Loether y McTavish (1974, pág. 212), el problema de la predicción es común a todas las ciencias; de ahí que parece adecuado basar una medida de asociación en la idea de realizar predicciones precisas de los valores de alguna variable dependiente. Así, si nuestro conocimiento teórico y empírico previo nos indica que las personas más religiosas tienden a votar con mayor frecuencia que las no religiosas a partidos políticos de derecha, lo que estamos diciendo realmente es que el conocimiento de las diferencias de puntuación en el nivel de religiosidad nos va a permitir realizar predicciones más precisas sobre el tipo de partido que se va a votar. Si fuera posible eliminar todos los errores de predicción del partido por el que se va a votar, al basar nuestras predicciones en el nivel de religiosidad, en tal caso existiría una asociación perfecta entre ambas variables. Si, por otro lado, y tal como ocurre en la realidad, la asociación entre ambas variables no es perfecta, aunque sí bastante alta, la medida de asociación que se obtenga expresará la proporción de los errores predictivos originales que se pueden evitar, gracias al conocimiento adicional del nivel de religiosidad.

Según sea el nivel de medición de las variables cuya asociación tratamos de conocer, así será el tipo de valor que se trata de predecir. Cuando disponemos de variables nominales, lo que interesa habitualmente predecir es la categoría o puntuación exacta de la variable dependiente, siendo suficiente a menudo predecir el valor modal o típico de la variable dependiente. Si el análisis de asociación se basa en variables ordinales, lo más probable es que pretendamos predecir el orden del rango de pares de valores en la variable dependiente, aunque también se puede tratar de predecir la mediana u otro percentil. Por último, cuando las variables vienen dadas al nivel de intervalo, el interés se dirigirá a predecir el valor de la media aritmética de la variable dependiente.

Tal como se ha dicho anteriormente, la predicción de la variable se realiza en dos situaciones o siguiendo dos reglas. La predicción I se realiza bajo la regla de la mínima suposición, es decir, cuando no se conoce más que la distribución de la variable dependiente, y la segunda predicción II se realiza bajo condiciones más favorables, al conocerse la distribución de las categorías de la variable independiente y de la distribución en cada una de ellas de las correspondientes categorías de la

variable dependiente. Pues bien, las medidas de asociación que vamos a estudiar a continuación consisten simplemente en un contraste entre los errores cometidos al realizar la primera de las predicciones y los errores cometidos al utilizar la segunda predicción al predecir la moda (variables nominales), el orden del rango (variables ordinales) o la media (variables de intervalo) buscadas. Para cada caso, el contraste se forma como sigue:

$$\text{Medida asociación } RPE = \frac{\text{Errores cometidos con predicción I} - \text{Errores cometidos con predicción II}}{\text{Errores cometidos con predicción I}} \qquad [8.1]$$

8.2. Medidas de asociación para variables nominales

En el capítulo anterior hemos estudiado los coeficientes basados en chi-cuadrado y el coeficiente Q de Yule, que pueden utilizarse para calcular medidas de asociación entre variables nominales, aunque de hecho no se suelen utilizar por los problemas de normatización que presentan, estando más aconsejado el empleo del coeficiente chi-cuadrado en la estadística inferencial para contrastar hipótesis. Por ello, vamos ahora a presentar otros dos coeficientes que, al estar basados en el criterio de la reducción proporcional del error de la moda, se encuentran normatizados y resulta más significativa la interpretación de los resultados obtenidos mediante su empleo en el análisis de datos sociológicos.

8.2.1. *El coeficiente Lambda*

El coeficiente Lambda, λ_{xy}, llamado también «coeficiente de predictibilidad de Guttman», es una medida asimétrica de asociación especialmente creada para analizar distribuciones bivariables en las que ambas variables son del tipo nominal. Además, se trata de una medida que ilustra perfectamente la lógica subyacente a las medidas *RPE*.

La fórmula para Lambda se puede expresar, en términos de la reducción proporcional en el error cometido al predecir la moda, de la siguiente manera:

$$\lambda_{yx} = \frac{(N - M_y) - (N - \Sigma m_y)}{N - M_y} = \frac{\Sigma m_y - M_y}{N - M_y} \qquad [8.2]$$

en donde el primer término del numerador expresa el número de errores que se cometen mediante la predicción I y el segundo término es el número de errores que se cometen al utilizar la predicción II. Por lo que se refiere al contenido de cada término, N es el tamaño total de la muestra; M_y es la frecuencia modal global de la variable dependiente Y, y Σm_y es la suma de las frecuencias modales de la variable dependiente Y *dentro* de cada categoría, por separado, de la variable independiente X. Al simplificar la expresión original, la fórmula de Lambda queda tal como aparece en la segunda parte de [8.2], que se puede leer del siguiente modo: el numerador es el número de no-errores cometidos bajo la predicción II (Σm_y) menos el número de errores cometidos bajo la predicción I (M_y), siendo el denominador el número de errores cometidos bajo la predicción I.

Como ya se ha observado, el símbolo que se utiliza para representar el coeficiente Lambda es la correspondiente letra griega minúscula, acompañada de dos subfijos, x e y, que representan, respectivamente, la variable independiente, x, y la variable dependiente, y. El subfijo que ocupa el primer lugar representa la variable dependiente, y el que ocupa el segundo lugar la variable independiente, es decir, λ_{yx}.

Antes de pasar a discutir más propiedades del coeficiente Lambda nos detendremos en el estudio de un ejemplo práctico, con el fin de fijar los conceptos hasta ahora introducidos.

Supongamos que estamos estudiando la situación matrimonial de los cabezas de familia españoles y que hemos obtenido, a partir de una muestra representativa de la población, los datos que se presentan en la tabla 8.1. Nuestro interés concreto consiste en realizar predicciones sobre la situación matrimonial de las personas que son cabezas de familia. A partir de la información que se contiene en dicha tabla, nos va a resultar más fácil predecir, por ejemplo, qué cabezas de familia están casados. Así, si conocemos que el valor modal de la variable situación matrimonial es «casado», entonces el valor que más racionalmente se

TABLA 8.1

Distribución de frecuencias absolutas de la situación matrimonial de una muestra de cabezas de familia, según el tipo de familia *

Situación matrimonial del cabeza de familia	*Tipo de familia*				*Total*
	El cabeza de familia es varón		*El cabeza de familia es mujer*		
	Hay niños menores de 15 años	*No hay niños menores de 15 años*	*Hay niños menores de 15 años*	*No hay niños menores de 15 años*	
Casado	*6.444*	*4.804*	78	50	11.376
Separado	20	126	250	106	502
Divorciado	19	237	*284*	276	816
Viudo	47	300	236	1.614	2.197
Total	6.530	5.467	848	2.046	14.891

* Datos ficticios.

puede predecir en relación a un cabeza de familia es que se encuentre casado, ya que si elegimos dicha categoría acertaremos con mayor frecuencia que si hubiéramos elegido el resto de las categorías.

Esto es, si hubiéramos supuesto, antes de visitar a cada uno de los 14.891 cabezas de familia entrevistados, que cada uno de ellos estaba casado, habríamos acertado 11.376 veces y nos hubiéramos equivocado en 3.515 ocasiones (14.891−11.376=3.515). Esta última cantidad representaría el número total de errores de predicción que se cometerían si predijéramos simplemente la moda global de la situación matrimonial de la variable cabeza de familia.

Sin embargo, si en lugar de predecir la moda global, predijéramos la moda para cada una de las categorías consideradas de la variable «tipo de familia», que en este caso la consideramos como una varible independiente en relación a la variable dependiente «situación matrimonial», se produciría una reducción en el error de predicción de la moda. Veamos en cuánto se puede reducir dicho error. Es decir, vamos a ver cuántas veces acertaríamos y cuántas veces nos equivocaríamos al predecir la situación matrimonial del cabeza de familia si, al ir a visitar a cada entrevistado, conociéramos previamente el número de cabezas de familia que son varones o mujeres y que tienen o no viviendo en el hogar niños menores de quince años.

Si supiéramos que el cabeza de familia es un varón y que tiene hijos menores de quince años, al predecir su situación matrimonial como que se encuentra casado, acertaríamos 6.444 veces, en 6.530 visitas, y nos equivocaríamos en 86 ocasiones (6.530−6.444=86). Si supiéramos que el cabeza de familia no tiene viviendo en su casa niños menores de quince años, y predijéramos que está casado, acertaríamos 4.804 veces, de 5.467, y nos equivocaríamos en 663 ocasiones (5.467−4.804=663). El saber que el cabeza de familia es una mujer y que tiene en la casa viviendo niños de quince años nos conduciría a predecir con más facilidad que su situación matrimonial es la de estar divorciada, ya que ésta es la categoría modal para ese tipo de familia. Acertaríamos en 284 ocasiones de 848. Finalmente, si supiéramos que el cabeza de familia es igualmente una mujer, pero que no tiene viviendo con ella a niños de quince años, la mejor predicción sería para la categoría «viuda», acertando en 1.614 ocasiones de 2.046.

Ahora ya podemos calcular cuánto hemos mejorado nuestra capacidad predictiva al añadir la anterior información. El cálculo lo realizaremos sumando las predicciones correctas realizadas dentro de cada categoría de la variable independiente (predicción tipo II) y contrastando dicho resultado con la frecuencia global de la categoría modal de la variable dependiente.

Tenemos que la suma de las categorías modales dentro de cada categoría de la variable dependiente $\Sigma m_y = 6.444+4.804+284+1.614=13.146$ predicciones correctas, lo que representa 13.146−11.376=1.770 errores

menos que se han cometido que si hubiéramos calculado la moda global de la situación matrimonial. Esto significa una reducción del 33,3 por 100 en los errores realizados al predecir la situación matrimonial de los cabezas de familia. Este valor es precisamente Lambda, que se obtiene simplemente sustituyendo los errores totales y las reducciones parciales de error en la fórmula [8.2]:

$$\lambda_{yx}=\frac{m_y-M_y}{N-M_x}=\frac{13.146-11.376}{14.891-11.376}=\frac{1.170}{3.515}=0{,}333$$

El numerador expresa, pues, la reducción de error conseguido al mejorar la información que suministra la variable independiente, y el denominador expresa el error cometido al disponer del mínimo de información que suministra el solo conocimiento de la variable dependiente. El resultado del cociente es 0,333 o, en términos porcentuales, el 33,3 por 100, y expresa, como se ha dicho antes, la reducción proporcional de error lograda.

El coeficiente λ_{yx} varía en magnitud desde el valor 0,0 al valor +1,0, y ello con independencia del tamaño de la tabla y de la muestra. A partir del supuesto de que existe, globalmente, un cierto recorrido de las puntuaciones en la variable dependiente, se define una asociación perfecta como una condición en la que todos los casos en cada categoría de la variable independiente se concentran en una única categoría (la categoría modal) de la variable dependiente. En tal caso, el valor de Lambda es la unidad. Por el contrario, el valor de Lambda es cero cuando se realiza la misma predicción modal dentro de todas las categorías de la variable independiente que la que realizaríamos si se predijera la moda global. Esto es, en tal caso la información adicional suministrada por la variable independiente no añade ningún valor predictivo adicional a la predicción de la moda de la variable dependiente. En la tabla 8.2 se contiene un ejemplo para el que Lambda es cero.

En efecto, se observa que las modas se concentran en todos los casos en la misma categoría de la variable dependiente, programa «cine», para cada una de las categorías de la variable independiente o grupos de edad. Obsérvese, sin embargo, que el hecho de que Lambda sea cero no significa en absoluto que no exista ningún tipo de asociación entre las dos variables. De hecho, si nos fijamos en las distribuciones porcentuales que se contienen en la tabla 8.2, se observa un cierto grado de asociación entre el tipo de programa preferido y la edad, al comparar las diferencias entre los porcentajes de las columnas. Esto viene a ilustrar la necesidad de seleccionar medidas que sean sensibles a los rasgos deseados de los datos. Así, mientras desde el punto de vista de la predicción de la moda el valor de la medida de la asociación es cero, desde el punto de vista de la diferencia de los porcentajes de la columna la asociación sí existe y, por tanto, es diferente de cero.

TABLA 8.2

Distribución porcentual del tipo de programa de televisión preferido según la edad

Tipo de programa	*Edad (años)*				
	15-20	*21-25*	*26-30*	*Más de 30*	*Total*
● Noticias	8	12	20	25	18
● Musicales	20	24	12	10	15
● Divulgación	10	15	20	18	16
● Cine	62	49	48	47	51
Total	100	100	100	100	100
$\lambda_{yx}=$	0,00				

* Datos ficticios.

Estas consideraciones ponen de manifiesto una limitación del coeficiente Lambda, y es que, aunque ofrece una medida bastante sensible de la fuerza de la asociación, no ofrece información sobre la naturaleza de la asociación. Si el investigador desea analizar la naturaleza de la asociación, lo mejor será analizar las diferencias porcentuales entre las columnas, tal como se ha hecho en el capítulo anterior.

Ya hemos dicho que Lambda es una medida asimétrica. Por ello, antes de proceder a su cálculo se hace preciso definir previamente qué variable es la independiente y cuál es la dependiente. Si en lugar de haber utilizado el tipo de familia como predictor de la situación matrimonial hubiéramos estado interesados en el valor predictivo de la situación matrimonial de cara al tipo de familia, los papeles de ambas variables se intercambiarían y se obtendría un valor distinto de Lambda y unas conclusiones diferentes. Fijándonos de nuevo en los datos que se contienen en la tabla 8.1, al tratar de predecir las modas en la situación matrimonial tanto globalmente como dentro de cada categoría de la variable tipo de familia se obtienen los siguientes resultados:

$$M_x=6.530$$

m_x
6.444
250
284
1.614
$\Sigma m_x=8.592$

$$\lambda_{xy}=\frac{8.592-6.530}{14.891-6.530}=\frac{2.062}{8.361}=0,246$$

La situación matrimonial permite una reducción proporcional de error del 24,6 por 100 al predecir el tipo de familia, porcentaje que es menor que en el caso contrario. Al utilizar el coeficiente Lambda se puede conocer, pues, la variable que permite una reducción mayor del error cometido al predecir las modas de una variable dependiente determinada. Nótese también que cuanto más precisa sea la medición de la variable independiente o predictora, mejor será la predicción. Así, si se quiere predecir una variable dependiente que consta de cuatro categorías mediante una variable predictora que sólo tiene tres categorías, en realidad sólo se podrán predecir tres modas diferentes, y no cuatro. De ahí que los investigadores prefieran habitualmente, y en general, conservar el mayor número de categorías en el análisis estadístico, ya que de esta forma el análisis ofrece mayores posibilidades de cara a la reducción del error con un número grande que con un número pequeño de categorías.

8.2.2. *El coeficiente Tau-y de Goodman y Kruskal*

Se trata de otra medida de la asociación para variables nominales, pero que se basa en una regla de predicción diferente de la utilizada por el coeficiente Lambda. Al igual que Lambda, el coeficiente Tau-y de Goodman y Kruskal es una medida asimétrica que varía entre el valor 0,0, para la ausencia de reducción en el error, y el valor 1,0, que representa una reducción perfecta del error. El coeficiente Tau-y ha sido ideado para tratar el problema de la *predicción de la distribución* de la variable dependiente *Y*. En esto difiere del coeficiente Lambda, que está indicado para predecir un valor óptimo de la variable dependiente, la moda.

Para el caso del coeficiente Tau-y, la predicción tipo I, o suposición con el mínimo de información, consiste en la asignación aleatoria de casos a las categorías de la variable dependiente, de tal manera que la distribución marginal de los casos no cambie. Volviendo a la tabla 8.1, podemos comprobar que esto significa que asignaríamos aleatoriamente 11.376 casos de la categoría de «casado», 502 a la categoría de «separado», 816 a la de «divorciado» y 2.197 a la de «viudo». Esta asignación de los 14.891 casos implicaría, naturalmente, algún tipo de error, y la cantidad esperada de error por dicha asignación aleatoria puede calcularse para cada categoría de la variable dependiente y, a continuación, sumarse para dar lugar al error esperado bajo la predicción tipo I. Utilizando los propios datos de la tabla 8.1, el procedimiento a seguir sería el siguiente:

En esta tabla, 11.376 casos se encuentran en la categoría de «casado», de un total de 14.891 unidades, dejando la diferencia, 3.515 casos, fuera de la categoría «casado». Cabe esperar que la proporción 3.515/14.891 de los 11.376 casos de la categoría «casado», se clasifiquen de

forma incorrecta si se asignaran aleatoriamente 11.376 casos a dicha categoría del total de casos. La idea que subyace a este razonamiento es como sigue. Se supone que se clasificarán de forma incorrecta por puro azar una cierta proporción de casos, y que esta proporción, para cualquier categoría, es simplemente la proporción de casos que no pertenecen a dicha categoría en relación a los casos que sí pertenecen a ella, basado en la distribución marginal de la variable dependiente. De este modo, si todos los casos se encontraran en una categoría, no se produciría error alguno al predecir sólo dicha categoría. Pero, en tanto que los casos se distribuyen en más de una categoría, existe alguna probabilidad de que la asignación al azar será correcta, y también otra probabilidad de que se cometan errores. Volviendo a los datos del ejemplo, todo ello significa que el número de errores esperados asciende a:

$$\frac{3.515}{14.891}(11.376)=2.684{,}7 \text{ errores esperados}$$

A este número se le añaden los errores esperados que resultan al asignar al azar los casos al resto de las categorías, errores que se calculan de idéntico modo; esto es:

Error esperado en una categoría, con asignación aleatoria = Proporción que no se encuentra en la categoría dada × La frecuencia de dicha categoría

Simbólicamente, se puede expresar la suma de los errores esperados para todas las categorías de la variable dependiente del siguiente modo:

$$E_1=\sum_{i=1}^{k}\left[\frac{N-f_i}{N}(f_i)\right]$$

en donde f_i es la frecuencia de la categoría i de la variable dependiente, y K es el número de categorías de la misma variable.

Siguiendo esta notación, los errores que se cometerían al predecir la situación matrimonial a partir de los datos de la tabla 8.1, se calculan de la siguiente forma:

$$\frac{14.891-11.376}{14.891}(11.376)=2.684{,}7$$

$$\frac{14.891-502}{14.891}(502)=\ 485{,}0$$

$$\frac{14.891-816}{14.891}(816)=\ 771{,}3$$

$$\frac{14.891-2.197}{14.891}(2.197)=1.867{,}5$$

$$E_1=5.807{,}5$$

Para realizar ahora la predicción tipo II de la distribución exacta de la variable dependiente se hace uso de la información que suministra la distribución de la variable dependiente dentro de las categorías de la variable independiente. Los procedimientos de cálculo son idénticos a los anteriores; sólo que ahora se realizan para cada una de las columnas correspondientes a las categorías de la variable independiente, esto es, el anterior sumatorio hay que realizarlo para cada categoría y sumar, a continuación, los resultados globales. Simbólicamente, la expresión del error esperado al realizar la predicción tipo II se escribe así:

$$E_2=\sum_{i=1}^{c}\sum^{k}\left|\frac{N_i-n_i}{N_i}(n_i)\right|$$

en donde n_i es la frecuencia de la celdilla en la categoría i de la variable dependiente, dentro de cada una de las c categorías de la variable independiente, y N_i es el total parcial de casos en cada una de las categorías de la variable independiente. Obtenidas las sumas para cada categoría, se suman todas ellas entre sí para obtener E_2. Con los datos de la tabla 8.1, el cálculo de E_2 sería como sigue:

- Error esperado para la categoría cabeza de familia varón con niños menores de 15 años ... 170,39
- Error esperado para la categoría cabeza de familia varón sin niños menores de 15 años ... 1.224,96
- Error esperado para la categoría cabeza de familia mujer con niños menores de 15 años ... 606,31
- Error esperado para la categoría cabeza de familia mujer sin niños menores de 15 años ... 728,74

$E_2=$ 2.730,40

Conocidos E_1 y E_2, el coeficiente Tau-y de Goodman y Kruskal se calcula a partir de la siguiente fórmula:

$$\text{Tau-y}=\frac{E_1-E_2}{E_1} \qquad [8.3]$$

Aplicando los valores obtenidos anteriormente para E_1 y E_2 en [8.3], se obtiene:

$$\text{Tau-y}=\frac{E_1-E_2}{E_1}=\frac{5.807{,}5-2.730{,}4}{5.807{,}5}=0{,}53$$

Así, pues, el coeficiente Tau-y obtenido nos indica que se han reducido en un 53 por 100 los errores cometidos al predecir la colocación de los casos en las categorías de la variable dependiente, mediante la información que aporta la distribución de los casos en la variable independiente. Naturalmente, si en lugar de haber considerado como inde-

pendiente la variable «tipo de familia» hubiéramos estado interesados en la predicción de esta variable a partir de la distribución de la variable «situación matrimonial», se hubiera obtenido un valor de Tau-y diferente, ya que, tal como se ha apuntado anteriormente, se trata de una medida asimétrica.

8.3. Medidas de asociación para variables ordinales

La predicción de valores en las variables ordinales es diferente del tipo de predicción que hemos estudiado anteriormente para el caso de las variables nominales. Como sabemos, una variable se llama ordinal cuando se puede ordenar a lo largo de ella una serie de casos u objetos, de tal manera que podamos saber cuál es el primero, cuál es el segundo, etc., pero sin poder atribuirles auténticos números, ya que no se conoce la distancia que hay entre dos casos u objetos. Como señalan Loether y McTavish (*op. cit.*, pág. 221), dado que el interés con las variables ordinales se centra en la ordenación de los valores, resulta útil considerar *pares de observaciones*, ya que hay que disponer al menos de dos valores o puntuaciones para poder «ordenar».

Si de lo que se trata es de obtener una medida de asociación para dos variables ordinales, el interés se centrará en la ordenación de pares de casos u objetos entre las variables, ya que lo que se pretende saber es si el conocimiento de la ordenación de los casos en una variable resulta útil para la predicción de la ordenación de los casos en otra variable. Si tal conocimiento no es de ninguna utilidad para predecir la ordenación de los casos en la segunda variable, entonces la medida de asociación ordinal debería ser igual a cero, mientras que si resulta de alguna utilidad diremos que sí existe asociación entre ambas variables, teniendo que distinguir en tal caso entre la «asociación positiva» y la «asociación negativa». Diremos que existe asociación positiva cuando el tipo de ordenación de los casos en la primera variable permite predecir en alguna medida la misma ordenación de los casos en la segunda variable. La asociación resulta de carácter negativo cuando la ordenación de los casos en la primera variable ayuda a predecir un ordenamiento inverso de los casos en la segunda variable. Así, por ejemplo, si un individuo *A* tiene un nivel de educación mayor que el individuo *B* se podría predecir que el nivel de ingresos de ambos guardarán el mismo orden, ya que existe una asociación positiva entre las variables nivel de educación y nivel de ingresos. Por el contrario, se puede predecir que los niveles de anomia de ambos individuos guardan una ordenación inversa a la de sus respectivos grados de interés por la política, porque sabemos que las variables nivel de anomia e interés por la política se encuentran negativamente relacionadas.

Antes de pasar a estudiar las medidas de asociación que más se suelen utilizar en el análisis sociológico, nos detendremos unos momentos en la exposición de algunas precisiones terminológicas.

8.3.1. *Tipos y cálculo de pares*

Recordemos en primer lugar que el número total de los pares de casos posibles, sin repetición, que se pueden formar a partir de N casos viene dado por:

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

Así, si disponemos de 10 casos, es decir, que $N=10$, se pueden formar 45 pares de casos que difieran en uno, al menos, de sus elementos. Además, si los T pares diferentes se miden en dos variables ordinales, existen cinco posibles formas de ordenación en ambas variables: a) *Pares semejantes o concordantes* (N_s); son pares que se encuentran distribuidos con idéntico orden en ambas variables. b) *Pares desemejantes o discordantes* (N_d); son pares que se encuentran ordenados en orden opuesto. c) *Pares empatados * sólo en la variable independiente* (X), pero no empatados en la variable dependiente (Y); se representan mediante el símbolo T_x. d) *Pares empatados sólo en la variable dependiente* (Y), pero no empatados en la variable independiente (X); se representan mediante el símbolo T_y. e) *Pares empatados en ambas variables;* se representan mediante el símbolo T_{xy}.

Estos cinco tipos de pares representan todas las posibilidades de formación de pares a partir de N casos, y su suma es igual, por tanto, a T, que es el número total de pares que difieren en uno, al menos, de sus elementos. Veamos ahora la forma de calcular estos pares, a partir de una tabla que recoja la tabulación cruzada de dos variables ordinales.

Supongamos que en un estudio sobre estratificación social, realizado en base a los resultados obtenidos en una encuesta realizada con una muestra de jóvenes, se encontraron los siguientes datos que relacionan el nivel de educación alcanzado por los jóvenes con el nivel de educación alcanzado por sus padres:

(Y) *Nivel de educación de los jóvenes*	(X) *Nivel de educación de los padres*			
	Bajo	*Medio*	*Alto*	*Total*
Alto	54	110	136	300
Medio	118	106	96	320
Bajo	142	74	60	276
Total	314	290	292	896

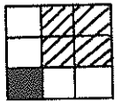
* Se dice que hay empate entre dos objetos o casos cuando ambos ocupan la misma posición, es decir, tienen el mismo valor ordinal.

Antes de proceder a calcular los diferentes tipos de pares de casos es preciso determinar qué diagonal es la «positiva», es decir, qué diagonal une las celdillas que contienen los valores «alto-alto» y «bajo-bajo», en ambas variables. En este ejemplo, la diagonal positiva es la que une el extremo inferior izquierdo con el extremo superior derecho de la tabla, mientras que la diagonal contraria es la negativa. Denominaremos con una *s* el final de la diagonal positiva y con una *d* el final de la diagonal negativa. De este modo, nos aseguraremos de que los pares N_s y N_d se calculan correctamente, y de que el signo del coeficiente final refleja la dirección de la asociación. Pasemos a calcular los diferentes tipos de pares.

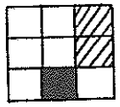
a) T=número total de pares diferentes:

$$T=\frac{N(N-1)}{2}=\frac{896\,(896-1)}{2}=400{,}960$$

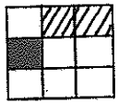
b) N_s=número de pares semejantes o concordantes. Se localiza en primer lugar la celdilla que corresponde al extremo *s* de la tabla, como se indica en el diagrama. La frecuencia de esta celdilla se multiplica por la suma de las frecuencias de las celdillas que se encuentran arriba y a la derecha (ya que la celdilla *s* se encuentra en el extremo izquierdo-inferior). A continuación se realiza el mismo procedimiento con el resto de las celdillas que se encuentran arriba y a la derecha de la celdilla *s*, tal como se indica en el diagrama, y se suman todos los productos:



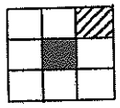
$142\times(110+136+106+96)=\ 63.616$



$74\times(136+96)\ =\ 17.168$



$118\times(110+136)\ =\ 29.028$

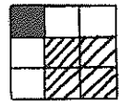


$106\times(136)\ =\ 14.416$

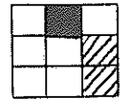
$N_s=124.228$

c) N_d=número de pares desemejantes o discordantes. Se calcula del mismo modo que el N_s, con la excepción de que la celdilla de partida

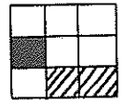
comienza en el extremo *d* de la tabla y procede a partir de ahí hacia abajo. De este modo, la frecuencia de la celdilla se multiplica por la suma de las frecuencias de las celdillas que se encuentran a la derecha y abajo:



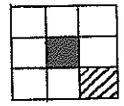
$54\times(106+96+74+60)\ =18.144$



$110\times(96+60)\ =17.160$



$118\times(74+60)\ =15.812$



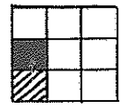
$106\times(60)\ =\ 6.360$

$N_d=57.476$

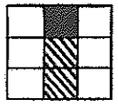
De la simple comparación de las magnitudes de N_s y N_d se deduce que el número de pares semejantes es mayor que el número de pares desemejantes, lo que revela la existencia de una asociación positiva.

d) T_x=número de pares «empatados» sólo en la variable independiente (X). Estos son los pares que se forman dentro de la misma categoría de la variable *x*, tal como se indica en el siguiente gráfico. Para su cálculo se elige una celdilla que encabeza una columna, se multiplica su frecuencia por la suma de las frecuencias de las celdillas que se encuentran debajo de la primera, y así sucesivamente:

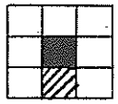
$54\times(118+142)\ =14.040$



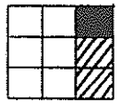
$118\times(142)\ =16.756$



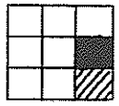
$110 \times (106+74)$ $= 19.800$



$106 \times (74)$ $= 7.844$



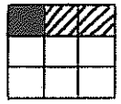
$136 \times (96+60)$ $= 21.216$



$96 \times (60)$ $= 5.760$

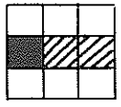
$T_x = 85.416$

e) T_y = número de pares «empatados» sólo en la variable dependiente (Y). Se calculan al igual que en el caso anterior, a excepción de que los productos se forman dentro de las categorías de la variable dependiente, es decir, a lo largo de las filas, como sigue:

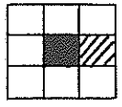


$54 \times (110+136)$ $= 13.284$

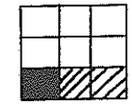
$110 \times (136)$ $= 14.960$



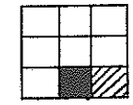
$118 \times (106+96)$ $= 23.836$



$106 \times (96)$ $= 10.176$



$142 \times (74+60)$ $= 19.028$



$74 \times (60)$ $= 4.440$

$T_y = 85.724$

f) T_{xy} = número de pares empatados simultáneamente en X e Y. Consiste en la suma de los pares que se pueden formar a partir de los casos que caen en la misma celdilla, esto es, que tienen idénticos valores en X e Y. Para cada celdilla se calculan a partir de la expresión:

$$f\,(f-1)/2$$

en donde f es la frecuencia de cada celdilla. Para el ejemplo anterior sería:

$$\begin{aligned}
54 \quad & (54-1)/2 = 1.431 \\
110 \quad & (110-1)/2 = 5.995 \\
136 \quad & (136-1)/2 = 9.180 \\
118 \quad & (118-1)/2 = 6.903 \\
106 \quad & (106-1)/2 = 5.565 \\
96 \quad & (96-1)/2 = 4.560 \\
142 \quad & (142-1)/2 = 10.011 \\
74 \quad & (74-1)/2 = 2.701 \\
60 \quad & (60-1)/2 = 1.770
\end{aligned}$$

$T_{xy} = 48.116$

En la actualidad, los programas de ordenador diseñados para el análisis estadístico de los datos sociológicos contienen el cálculo de los diversos números de pares, para cualquier tipo de tabla, con lo que el investigador se ve aliviado en su tedioso cómputo. Con todo, es importante conocer el detalle de su cálculo, para hacerse una idea más completa de los fundamentos lógicos de las medidas de asociación. Conocidos los valores de los diferentes tipos de pares, ya se está en condiciones de sustituirlos en las fórmulas que expresan las diferentes medidas de asociación que veremos a continuación, ya que todas ellas incluyen algunos de los valores que hemos calculado. En todos estos coeficientes que vamos a ver, el numerador es el mismo, $N_s - N_d$, cuya diferencia va a indicar el carácter positivo ($N_s > N_d$) o negativo ($N_s < N_d$) de la asociación. Se trata de medidas tipo *RPE* que indican la reducción proporcional en el error que se produce al utilizar la variable independiente como pre-

dictora de la distribución de la variable dependiente. En lo que se diferencian entre sí las diversas medidas de asociación es en la composición del denominador. El estudio de las relaciones que guardan entre sí las diferentes medidas de asociación fue realizado, para el caso de los datos sociológicos, por Robert H. Somers (1962), y básicamente vamos a seguir aquí el esquema que dicho autor ha desarrollado en su trabajo.

8.3.2. *Coeficiente Tau-a de Kendall*

Es el más intuitivo de todos los coeficientes que miden la asociación entre variables ordinales, siguiendo el criterio de la reducción proporcional en el error. Fue definido por Kendall como la diferencia entre los pares semejantes y desemejantes en relación al número total de pares diferentes:

$$t_a = \frac{N_s - N_d}{T} \qquad [8.4]$$

Volviendo a los datos del ejemplo anterior, el coeficiente Tau de Kendall sería:

$$t_a = \frac{124.228 - 57.476}{400.960} = 0{,}17$$

El coeficiente Tau de Kendall varía entre $-1{,}0$ y $+1{,}0$, indicando el valor cero la incapacidad de una variable por reducir los errores que cabría esperar al distribuir al azar los valores de la otra variable. Cuando la asociación es negativa, el coeficiente Tau va acompañado de un signo negativo, mientras que el signo positivo indica una asociación positiva. El valor de la unidad indica que todos los posibles pares son del mismo tipo (semejantes o desemejantes, según el signo del coeficiente). Es una medida simétrica, ya que no es preciso distinguir entre variable independiente y variable dependiente al calcular N_s, N_d y T, y no depende del tamaño de la tabla ni del número de rangos de las variables ordinales.

Sin embargo, el coeficiente Tau-a presenta un inconveniente, y es que cuando existen empates, como ocurre con frecuencia, el coeficiente no puede alcanzar el valor de 1,0, porque el denominador, cuando existen empates, siempre será mayor que N_s o N_d.

8.3.3. *Coeficiente Gamma de Goodman y Kruskal*

Cuando la muestra consta de un número amplio de casos y son muy pocos los valores ordinales que pueden alcanzar los casos, el número de empates será muy grande, con lo que no está recomendado utilizar el

coeficiente Tau-a de Kendall, ya que el máximo valor posible del coeficiente no alcanza la unidad. Una solución al problema de obtener un coeficiente igual a 1,0 cuando existen empates consiste, sencillamente, en la eliminación de los empates no sólo del numerador, como ocurre con el coeficiente Tau de Kendall, sino igualmente en eliminarlos del denominador.

El coeficiente Gamma (γ) de Goodman y Kruskal permite precisamente realizar dicha eliminación. Se trata de una medida simétrica de la asociación de dos variables ordinales que, a diferencia del coeficiente Tau-a, siempre puede alcanzar los valores límites de $-1{,}0$ a $+1{,}0$, independientemente del número de empates que presenten los datos. La fórmula para calcular el valor de Gamma es la siguiente:

$$\text{Gamma} = \frac{N_s - N_d}{N_s + N_d} \qquad [8.5]$$

Como se observa, el numerador es el mismo que para Tau-a, y el denominador es simplemente la suma de los pares que se encuentran ordenados de forma diferente en ambas variables. Tal como se ha dicho anteriormente, el valor de Gamma oscila entre $-1{,}0$ y $+1{,}0$. En efecto, si todos los pares no empatados son semejantes, en tal caso $N_d = 0$ y $\text{Gamma} = \frac{N_s - 0}{N_s + 0} = 1$; mientras que si todos los pares no empatados son desemejantes, en tal caso $N_s = 0$ y $\text{Gamma} = \frac{0 - N_d}{0 + N_d} = -1$. Cuando $N_s = N_d$, Gamma $= 0$. De cualquier forma, $N_s - N_d < N_s + N_d$, ya que N_s y N_d son números positivos y, en consecuencia, $N_s - N_d / N_s + N_d$ será (en valor absoluto) menor que 1.

Si utilizamos los datos calculados a partir de la tabla que relaciona el nivel de educación de los jóvenes con el nivel de educación de los padres, el valor de Gamma será el siguiente:

$$\text{Gamma} = \frac{124.228 - 57.476}{124.228 + 57.476} = \frac{66.752}{181.704} = 0{,}37$$

El valor de Gamma se puede interpretar como la reducción proporcional en el error cometido al predecir el ordenamiento de los casos en una variable mediante el conocimiento de la ordenación de los casos en otra variable, en lugar de realizar la predicción basándose en una ordenación aleatoria de los casos en las dos variables.

Resulta de interés destacar que, para el caso de una variable 2×2, el valor de Gamma es el mismo que se obtendría si en su lugar hubiéramos utilizado el coeficiente Q de Yule. Por esta razón se puede considerar que el coeficiente Gamma es una versión generalizada del coeficiente Q de Yule para tablas en las que el número de filas y columnas sea superior a dos.

8.3.4. *Coeficiente d de Somers*

Aparte de los coeficientes Gamma y Tau, tenemos dos medidas asimétricas, d_{yx} y d_{xy}, que han sido introducidas por Somers (1962), y que se definen como sigue:

$$d_{yx}=\frac{N_s-N_d}{N_s+N_d+T_y} \qquad [8.6]$$

$$d_{xy}=\frac{N_s-N_d}{N_s+N_d+T_x}$$

Al tratarse de una medida asimétrica de asociación se hace preciso distinguir entre la variable independiente y la variable dependiente. De este modo, si se pretendiera predecir la ordenación de los casos en una variable dependiente utilizando para ello una variable independiente o predictora, la predicción afectaría no sólo a los pares que se encuentran ordenados de forma diferente en cada variable (los pares N_s y N_d), sino que se realizaría también una predicción de los casos T que son diferentes en la variable predictiva, pero que se encuentran empatados en la variable dependiente. La diferencia en la variable independiente permite realizar una predicción incluso de los casos de empate en la variable dependiente. Es así como el denominador de la medida de asociación d contiene todos los pares para los que se puede formular una predicción, esto es, $N_s+N_d+T_y$ (o T_x), según que sea X o Y la variable que se considera dependiente. El numerador, como se observa en [8.6], es otra vez la diferencia entre los pares semejantes y los pares desemejantes, y sólo se incluyen los empates de la variable que se va a predecir, quedando excluidos del cómputo los empates de la variable predictora. Al igual que los coeficientes anteriores, el coeficiente d de Somers se puede interpretar como la reducción proporcional en los errores que se cometen al predecir el ordenamiento de los casos en la variable dependiente cuando se tiene en cuenta el ordenamiento de los casos en la variable independiente, en lugar de realizar la predicción del ordenamiento de los casos por medios aleatorios.

Al igual que vimos al estudiar otra medida de asociación asimétrica, el coeficiente Lambda, los dos valores que se pueden obtener de d a partir de una misma tabla (según que la variable que se tome como independiente sea X o Y) suelen ser también diferentes entre sí.

8.3.5. *Coeficiente Tau-b de Kendall*

Existe otro coeficiente Tau debido a Kendall, que permite estudiar otro tipo de asociación. Supongamos que deseamos encontrar una medida del grado de asociación que sea simétrica pero que, a diferencia del coeficiente Gamma, tenga en cuenta los empates que se producen en una u otra variable, pero no los empates que se forman en ambas, T_{xy}. Pues bien, en tal caso conviene utilizar el coeficiente Tau-b, que se puede considerar como un promedio de los dos coeficientes d de Somers que pueden calcularse a partir de una misma tabla. Dicho coeficiente se expresa, de hecho, como la raíz cuadrada del producto de los dos coeficientes d:

$$T_b=\sqrt{d_{xy}\cdot d_{yx}}$$

Pero la forma operativa de utilizar el coeficiente T_b es a partir directamente del número de cada tipo de pares, tal como sigue:

$$T_b=\frac{N_s-N_d}{\sqrt{(N_s+N_d+T_y)\,(N_s+N_d+T_x)}} \qquad [8.7]$$

Al igual que los coeficientes anteriores, T_b puede tomar valores que oscilan entre —1,0 y +1,0, según sea el sentido de la asociación, y su magnitud señala cuán fuerte es la asociación entre dos variables. Sin embargo, cuando la tabla no es cuadrada, es decir, el número de filas no es igual al de las columnas, el coeficiente Tau-b no puede llegar a valer la unidad, dado que cuando hay un número diferente de filas que de columnas existirán más pares empatados en una variable (la que tiene menos categorías) que en la otra variable. Con todo, se trata de una medida simétrica muy útil del grado de asociación entre dos variables ordinales, porque, a diferencia del coeficiente Tau-a, sólo tiene en cuenta para su cálculo los tipos de pares más relevantes para la asociación.

8.3.6. *Coeficiente rho de Spearman*

Uno de los coeficientes más utilizados para medir la asociación entre las variables sociológicas de tipo ordinal es el *rho* (r_s) de Spearman. La lógica que sigue este coeficiente para medir la dirección y la fuerza de la asociación es diferente de la que hemos visto hasta ahora. Su uso viene recomendado en aquellos casos en que se cuenta con el ordenamiento de todos los casos individuales en las dos variables, de tal modo que en cada variable los ordenamientos tienen un recorrido de 1 a N. En la tabla 8.3 se contiene un ejemplo de los ordenamientos de algunas regiones españolas según la evaluación que la población residente en ellas hace, en una escala del 1 al 10, de la actuación de los empresarios y de los obreros.

Un ordenamiento se refiere a las medias de la evaluación, en una escala del 1 al 10, de la actuación de los empresarios en general, y el segundo ordenamiento se refiere a la evaluación de la actuación de los obreros. Lo que se trata de saber es si la población, en una misma re-

TABLA 8.3

Medias y ordenamiento de la evaluación de la actuación de empresarios y obreros, en algunas regiones españolas

Región	Empresarios		Obreros		d	d^2
	Media	Orden	Media	Orden		
Cataluña	3,87	6	6,81	5	1	1
País Vasco	3,82	7	6,17	6	1	1
Andalucía	4,78	2	7,64	3	—1	1
Canarias	5,87	1	8,30	1	0	0
Madrid	4,57	5	7,18	4	1	1
Barcelona	4,65	4	6,06	7	—3	9
Galicia	4,78	3	8,08	2	1	1
					$d=0$	$d^2=14$

FUENTE: Banco de Datos, CIS, 1982.

gión, evalúa diferentemente o en el mismo sentido a los empresarios y a los obreros.

El coeficiente *rho* (r_s) de Spearman es una medida adecuada para el problema que hemos planteado, ya que mide el grado de asociación de dos variables ordinales, basándose en la diferencia entre rangos. Si no existe diferencia alguna es igual a cero. A efectos de cálculo se utiliza el sumatorio de los valores de las diferencias al cuadrado, porque la suma de los valores simples es siempre igual a cero. Cuando $\Sigma d^2 \neq 0$, sabemos que las dos variables no se ordenan idénticamente. Con el fin de interpretar el valor de tal diferencia se utiliza el coeficiente *rho* de Spearman, que se define del siguiente modo:

$$r_s = 1 - \frac{6\,\Sigma d^2}{n\,(n^2-1)} \qquad [8.8]$$

Para el caso de los datos que se contienen en la tabla 8.3, su valor es el siguiente:

$$r_s = 1 - \frac{6 \cdot 14}{7\,(7^2-1)} = 1 - \frac{84}{336} = 1 - 0{,}25 = 0{,}75$$

El valor de *rho* (r_s) varía entre $-1{,}0$ y $+1{,}0$, indicando el primer valor una ordenación opuesta de los casos en las variables, y el segundo valor un perfecto acoplamiento de las dos ordenaciones. Cuando $r_s=0$, significa que no existe una ordenación sistemática de ningún tipo entre las dos variables.

En realidad, la fórmula del coeficiente *rho* de Spearman es la del

coeficiente r de Pearson (una medida de asociación para variables de intervalo, que veremos en el próximo capítulo) aplicado a ordenamientos. La interpretación de r_s se hace no en términos de la reducción proporcional en el error, sino en términos de la fuerza de asociación o correlación entre variables. Su uso está muy indicado en la investigación sociológica, siempre que se desee conocer si la ordenación de una variable está o no asociada a la ordenación de otra variable para los mismos usos. Otro ejemplo, con datos hipotéticos, nos va a permitir comprobar las posibilidades del coeficiente de Spearman para el análisis sociológico.

Supongamos que en ocho provincias españolas se ha producido, al comparar los resultados de dos elecciones diferentes, un incremento de los votos emitidos a favor de un partido regionalista y una disminución de los votos emitidos a favor de un partido de ámbito nacional, y se pretende saber si el incremento de votos para un partido y la disminución de votos del segundo partido es un fenómeno político que se encuentra relacionado en las ocho provincias. A esta cuestión se puede responder ordenando las ocho provincias según el porcentaje de pérdidas y ganancias respectivo de votos de ambos partidos y calculando un coeficiente *rho* de Spearman, como se hace a continuación:

Número de orden de la provincia	1	2	3	4	5	6	7	8	
Rango por disminución del partido nacional	8	1	5	3	2	7	6	4	
Rango por incremento del partido regional	8	1	5	5	2	7	4	4	
d, diferencia de rango	0	0	0	−2	0	0	2	0;	$d = 0$
d^2	0	0	0	4	0	0	4	0;	$d^2 = 8$

Aplicando la fórmula [8.8]:

$$r_s = 1 - \frac{6 \cdot 8}{8\,(64-1)} = 1 - \frac{48}{504} = 1 - 0{,}09 = 0{,}91$$

Lo que revela una alta correlación entre ambos movimientos electorales en las ocho provincias consideradas. Con el conocimiento de este estadístico, la interpretación sociológica de los resultados electorales sería ahora más sencilla y significativa.

8.4. LA MATRIZ DE ASOCIACIONES

Con frecuencia, los investigadores sociales calculan simultáneamente un número de medidas similares de asociación, que sirven para poner de manifiesto el tipo de relación que existe entre todos los pares posibles

de un conjunto de variables. Al colocar en una misma matriz todos los resultados se obtiene una evidente ventaja comparativa, ya que de una sola ojeada es posible observar el modelo de asociaciones que configuran las diversas variables. Un ejemplo de una matriz de asociaciones se incluye en la tabla 8.4, utilizando coeficiente Gamma.

TABLA 8.4

Matriz de asociaciones utilizando coeficientes Gamma entre cuatro variables culturales

	Tradic.	Patern.	Racion.	Nepot.
Tradicionalismo ...	—	23	10	03
Paternalismo	—	—	00	12
Racionalismo	—	—	—	07
Nepotismo	—	—	—	—

FUENTE: Rafael LÓPEZ PINTOR, «Satisfacción en el trabajo...» *REOP*, 44, 1976, páginas 113 y 114.

En un estudio que se enmarca dentro de la sociología de las organizaciones, López Pintor (1976) pretende encontrar una explicación satisfactoria a ciertas actitudes y comportamientos de la organización burocrática. Para explicar la satisfacción en el trabajo utiliza tres tipos de variables: sociológicas, orientaciones de valor y variables específicamente de organización. Para estudiar la orientación cultural de los funcionarios de una organización burocrática utiliza cuatro medidas referentes a las siguientes variables: tradicionalismo, paternalismo, racionalismo y nepotismo. Medidas estas variables a través de los correspondientes indicadores, calcula el grado de asociación que existe entre las cuatro variables, tomadas dos a dos, mediante el cálculo del coeficiente Gamma. Los resultados obtenidos son los que se recogen en la tabla 8.4. Los coeficientes obtenidos presentan unos valores ciertamente bajos, lo que revela la inexistencia o debilidad de la asociación entre las cuatro variables culturales. López Pintor, apoyándose en la teoría del conflicto de valores, interpreta la ausencia de una fuerte asociación entre las cuatro variables culturales como la manifestación de un potencial conflictivo en el sentido de enfrentamiento, yuxtaposición o falta de valores.

Obsérvese que al ser Gamma una variable simétrica, sólo se necesita incluir los coeficientes en una sola mitad de la matriz, tal como aparece en la tabla 8.4, ya que los coeficientes que debieran aparecer en la segunda mitad son idénticos (simétricos) a los de la primera. Por eso, sólo se suelen dar en las matrices de coeficientes de asociación (cuando éstos son simétricos, tales como el Gamma o el *r* de Pearson) los resultados para una sola mitad.

8.5. TERMINOLOGÍA

Se recomienda la memorización y comprensión del significado de cada uno de los términos y conceptos siguientes:

- Reducción proporcional del error (*RPE*).
- Medida de asociación tipo *RPE*.
- Coeficiente Lambda.
- Coeficiente Tau-y de Goodman y Kruskal.
- Pares de observaciones:
  - Pares semejantes.
  - Pares desemejantes.
  - Pares empatados (en una sola variable o en ambas).
- Coeficiente Tau-a de Kendall.
- Coeficiente Tau-b de Kendall.
- Coeficiente Gamma de Goodman y Kruskal.
- Coeficiente *d* de Somers.
- Coeficiente *rho* de Spearman.
- Matriz de correlaciones.

EJERCICIOS

1. En una encuesta realizada entre la población juvenil, se obtuvo la siguiente distribución de la identificación religiosa de los jóvenes según el lugar de residencia:

	Lugar de residencia			
Religiosidad	*Rural*	*Semi-urbano*	*Urbano*	*Metro-politano*
Católico practicante	320	305	188	80
Católico no practicante	432	290	170	62
Indiferente	280	212	126	66
No creyente	60	35	20	3

Calcular el valor de la asociación entre ambas variables mediante el coeficiente Lambda, considerando el lugar de residencia como la variable independiente.