J. Rodríguez Cano

ASOCIACION

DE

VARIABLES

INDICE

INTRODUCCION

En Sociología predominan variables de nivel nominal y ordinal (llamadas genéricamente **categóricas** o **cualitativas**) debido a la naturaleza misma del objeto de estudio y a la falta de instrumentos estandarizados que permitan cuantificar.

La relación o **asociación** de variables cualitativas se puede describir numéricamente. Se han desarrollado muchas medidas o **coeficientes** para hacerlo, presentando los datos en las llamadas 'tablas de contingencia' o 'de doble entrada'.

Los coeficientes tienen fórmulas distintas, miden aspectos diversos de una tabla, o bien, si miden los mismos aspectos, los ponderan de manera desigual. Se puede decir que cada uno operacionaliza diferentemente el concepto teórico 'asociación de variables.'

La consecuencia es que coeficientes que supuestamente miden lo mismo, entregan resultados diferentes y comúnmente no traducibles, lo que sería aceptable si cada coeficiente pudiera ser considerado una dimensión distinta y se creara un índice a partir de ellos. Pero no es así: cada uno es visto como una medida real de asociación. Esta situación afecta la interpretación y comparación de resultados de investigaciones empíricas.

En sus comienzos, el concepto de asociación no es estadístico. Aparece inicialmente en escritos de lógicos ingleses del siglo XIX (Mill, Boole, Jevons).

A partir del concepto lógico, varios investigadores empiezan a crear formas de medir asociación empíricamente. El estadístico belga Quetelet presenta en 1832 un coeficiente de 'influencia'. Al empezar este siglo, G. Udny Yule desarrolla bases algebraicas y medidas para lo que denomina 'asociación' y Karl Pearson crea coeficientes paramétricos de 'variabilidad dependiente' y de 'contingencia'.

La falta de acuerdo inicial sobre qué es y cómo se mide asociación hace que empiecen a aparecer coeficientes que sólo tienen en común que sus valores no exceden de 1. Hasta hoy se usan coeficientes de todo tipo: derivados de 'chi-cuadrado', de la correlación 'producto-momento' (r) de Pearson, de Q de Yule, de Goodman y Kruskal o basados en ellos, y otros.

Este trabajo examina el concepto de asociación desde su origen, su desarrollo teórico y la formación de coeficientes. Se ven distintos enfoques y se presentan dos modelos que permiten reducir los coeficientes a una base común haciéndolos comparables.

Se consideran tablas de 4 celdas (2x2) y tablas mayores (>2x2), pero preferentemente las primeras. Muchos pro-

blemas de tablas mayores pueden ser vistos como meras extensiones de los de tablas de 2x2.

En este trabajo los conjuntos de casos son tomados como 'colectivo' (o población o universo). "Ésta parece ser la manera de empezar en la construcción de medidas racionales de asociación." (Goodman y Kruskal 1954, p.733). Por lo tanto, no se examinan distribuciones muestrales ni problemas de significación.

El término 'explicación' y sinónimos son usados aquí sólo en sentido estadístico. Los problemas de causalidad son de gran importancia metodológica, pero exceden el marco de este trabajo.

El Capítulo 1 trata del tipo de razonamiento en asociación de variables.

El Capítulo 2 analiza escritos de lógicos ingleses del siglo XIX en que se presenta el concepto de asociación.

El Capítulo 3 trata de la obra de G. Udny Yule, que sienta las bases de la teoría moderna de asociación de variables, se analizan coeficientes para tablas de 2x2, y se examina la controversia de Yule y Pearson.

El Capítulo 4 estudia el aporte de Goodman y Kruskal, analiza medidas posteriores, y presenta dos modelos para comparar fórmulas y valores numéricos de coeficientes.

CAPITULO 1

RAZONAMIENTO BASICO

La distribución de una variable en un grupo puede ser observada al establecer cuántos individuos corresponden a cada categoría. Se puede decir entonces que el grupo ha sido medido o sus individuos clasificados de acuerdo a ella.

El paso siguiente puede ser medir al grupo en una segunda variable. Al especificar el número de individuos en cada combinación de categorías, se tiene una medición simultánea del grupo en ambas variables.

Estos datos pueden presentarse más claramente en una tabla de doble entrada, como la siguiente de 2x2:

TABLA 1

	y1	y2	
x1	a	b	a+b
x2	c	d	c+d
	a+c	b+d	n

En esta tabla se muestran simultáneamente tres mediciones:

(1) Variable **y**:

Valores	Frecuencias (número de individuos)
y1	a+c
y2	b+d
	n

(2) Variable **x**:

Valores	Frecuencias (número de individuos)
x1	a+b
x2	c+d
	n

(3) Variables **x** e **y**:

Valores	Frecuencias (número de individuos)
x1y1	a
x1y2	b
x2y1	c
x2y2	d
	n

Se puede decir que **x** explica a **y** si el conocimiento del valor de la primera ayuda a la predicción del valor de la segunda. El siguiente ejemplo puede ilustrar esto:

Hacemos correr a diez niños una distancia corta y clasificamos a cinco como rápidos y a cinco como lentos. Para explicar las diferencias de velocidad podemos observar la distribución de una segunda variable, por ejemplo 'sexo'.

Si hay tres niñas rápidas y tres lentas, y dos niños rápidos y dos lentos (véase la Tabla 2), podemos decir que para este grupo el sexo no explica la velocidad: ambos sexos tienen igual proporción de rápidos (.5). En este grupo, el saber de qué sexo es un niño no sirve para predecir su velocidad.

TABLA 2

		Velocidad		
		Rápido	Lento	
Sexo	Niña	3	3	6
	Niño	2	2	4
		5	5	10

Examinemos otra variable, 'edad', con categorías mayor y menor. Supongamos que tenemos cinco niños mayores, de los cuales tres son rápidos, y que de los cinco menores sólo dos lo son (véase la Tabla 3).

TABLA 3

		Velocidad		
		Rápido	Lento	
Edad	Mayor	3	2	5
	Menor	2	3	5
		5	5	10

En este grupo hay una proporción mayor de niños rápidos entre los de más edad que entre los menores. Si queremos predecir la velocidad de un niño sin conocer su edad y pronosticamos que es rápido, la probabilidad de acertar es .5. Pero si sabemos que es mayor, la probabilidad de que sea rápido es .6. Si es menor, predeciremos que es lento. Saber la edad de un niño de este grupo ayuda a conocer su velocidad.

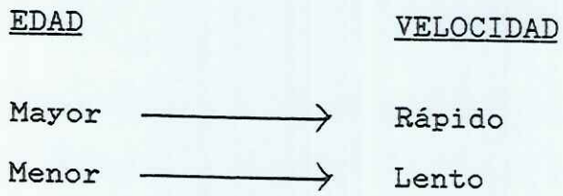
Queremos después saber si el peso de los niños incide en su velocidad y los clasificamos en pesados y livianos. Si observamos que los primeros tienden a ser lentos y los segundos rápidos (véase Tabla 4), podemos concluir que el peso de los niños influye en su velocidad pues permite predecirla.

TABLA 4

		Velocidad		
		Rápido	Lento	
Peso	Pesado	1	4	5
	Liviano	4	1	5
		5	5	10

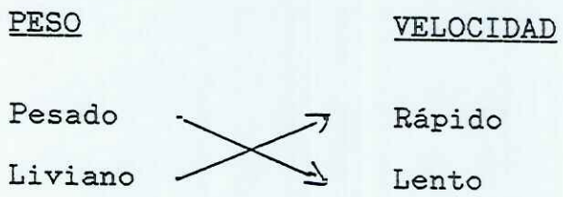
Si volvemos a la Tabla 3 notamos que cada categoría de edad se puede asociar con una categoría de velocidad: mayor con rápido y menor con lento. Decimos que las variables están asociadas a través de una relación funcional: velocidad como función de edad. (Véase el Gráfico 1.)

GRAFICO 1



Con respecto a la Tabla 4, se ve en ella que se puede asociar liviano con rápido y pesado con lento. Velocidad es función de peso. (Véase el Gráfico 2.)

GRAFICO 2



Por el contrario, en la Tabla 2 no hay asociación entre las categorías, no puede relacionarse un sexo con rapidez o con lentitud. Al no estar asociadas las categorías, tampoco lo están las variables: las diferencias de velocidad no son explicadas por diferencias de sexo. Se puede decir que estas variables son **independientes**.

Asociación e **independencia** son, entonces, situaciones opuestas. Asociación es la ausencia de independencia o, dicho de otra manera, asociación es **dependencia.**

Independencia es un estado extremo; asociación no lo es. Lo primero puede verse en la Tabla 2: cualquier cambio en los números dentro de la tabla transformaría la situación de independencia en dependencia (o asociación). Por otra parte, los números de una tabla en que hay asociación pueden ser cambiados una y otra vez sin volver a la situación de independencia. Las Tablas 3 y 4 son ejemplos de esto: los números dentro de la tabla cambian y persiste la asociación.

La situación de independencia es una sola para cada tabla y no tiene sentido calificarla de mayor o menor. Pero puede haber más de una situación de dependencia en una tabla (dados los mismos marginales) y hablaremos de una asociación mayor o menor.

Hay que hacer notar una diferencia importante entre los Gráficos 1 y 2. En el primero, las flechas unen las categorías mayores de las variables (mayor edad con mayor velocidad) y también las menores (menor edad con menor velocidad). En el segundo, las flechas ligan la categoría mayor de cada variable con la menor de la otra: más peso con menos velocidad, y menos peso con más velocidad. La asociación del primer gráfico es llamada **positiva**; la del segundo, **negativa**.

El que la asociación sea positiva o negativa se refiere a la llamada **dirección** de la relación, pero no afecta a la fuerza de la misma. Esto es fácil de observar al 'invertir' una variable, como por ejemplo si 'velocidad' es transformada en 'lentitud': la categoría de mayor velocidad pasa a ser la de menor lentitud y viceversa, pero las frecuencias no cambian.

Las relaciones, determinadas por estas frecuencias, se mantienen constantes. Lo que cambia es su dirección: la asociación positiva 'edad-velocidad' es ahora la negativa 'edad-lentitud.'

Volviendo a los datos del ejemplo (Tablas 3 y 4): ¿es más **fuerte** la asociación de edad o la de peso con velocidad? La asociación sería de fuerza máxima si todos los niños de igual velocidad correspondieran a una sola categoría de la otra variable. Por ejemplo, si todos los niños livianos fueran rápidos y los pesados fueran lentos, como lo muestra la Tabla 5. Esta asociación perfecta entre categorías (y entre variables) permitiría una predicción sin error pues bastaría conocer el peso del niño para saber si es veloz o no lo es.

TABLA 5

		Velocidad		
		Rápido	Lento	
Peso	Pesado	0	5	5
	Liviano	5	0	5
		5	5	10

La Tabla 4 no muestra asociación perfecta como la Tabla 5: dos niños (liviano-lento y pesado-rápido) difieren del patrón ideal. Esto es peor en la Tabla 3: cuatro niños no corresponden a ese patrón (dos menores-rápidos y dos mayores-lentos). La Tabla 4 se acerca más que la 3 al estado ideal de la 5. Se puede afirmar que en la primera hay una asociación mayor, más fuerte, que en la segunda.

El procedimiento usado en este ejemplo es posible porque los datos son muy simples. Las tablas de contingencia con datos verdaderos requieren métodos más complejos para medir asociación. Los coeficientes de asociación son las fórmulas para hacerlo.

El establecimiento de una ley empírica se puede verificar a través de "la frecuencia de su ocurrencia. No, sin embargo, su frecuencia absoluta. La pregunta no es si ocurre a menudo o rara vez, en el significado común de estas palabras, sino si ocurre más a menudo de lo que explica el azar, más que lo que racionalmente podría esperarse si la co-incidencia fuera casual. Debemos decidir, por lo tanto, qué grado de frecuencia explica el azar en una co-incidencia." (p.346)

"Tenemos que deducir de la frecuencia observada de co-incidencias el efecto del azar, y si algo queda, ello es el hecho residual que prueba existencia de la ley." (p.347) Si lo observado es más que lo esperado -o menos, diríamos hoy, pues la diferencia puede ser negativa- se concluye que hay una relación.

Mill especifica frecuencia esperada en caso de independencia: "Si encontramos en nuestras observaciones que A aparece en un caso de cada dos, y B en uno de cada tres, entonces, si no hay conexión ni repulsión entre ellos ni entre cualquiera de sus causas, las instancias en que ambos han de existir, o, mejor dicho, han de co-existir, será un caso de cada seis. Porque A estará en tres casos de cada seis; y B, apareciendo en un caso de cada tres sin importar la presencia o ausencia de A, estará en un caso de esos tres." (p.347)

CAPITULO 2

ORIGEN DEL CONCEPTO DE ASOCIACION

El concepto de asociación de variables no es creado por estadísticos ni aparece por necesidades de investigación empírica sino en trabajos de John Stuart Mill y W. Stanley Jevons. Este último basa su sistema en escritos de Augustus de Morgan y George Boole.

La primera fuente de la teoría de asociación es la obra de Mill "A System of Logic" (1843), donde reconoce la influencia del azar en las 'uniformidades' (aparición conjunta de fenómenos) y la necesidad de eliminarlo al formular leyes empíricas. Intenta determinar cuántas confirmaciones son necesarias para considerar una uniformidad como establecida, una ley empírica.

El problema básico que aborda Mill es que la aparición conjunta de dos fenómenos no es prueba de que no coexistan sólo por azar. "Una co-incidencia puede ocurrir una y otra vez, y ser sólo casual. Sería inconsistente con lo que sabemos del orden de la naturaleza el dudar de que toda co-incidencia casual se ha de repetir tarde o temprano." (p.346)

Mill no se preocupa de cuánta es la diferencia entre observado y esperado. Si algo queda, sea mucho o poco, es prueba de ley empírica. Importa la diferencia como tal, no su tamaño. "Si A ocurre en una mayor proporción de casos cuando existe B que cuando no existe, entonces B también ha de ocurrir en una mayor proporción cuando existe A que cuando no existe, y hay, por lo tanto, alguna conexión entre A y B." (p.348) Esto se puede expresar como en la ecuación (1) [usando la notación de la Tabla 6]:

$$\text{Si} \quad \frac{a}{a+c} > \frac{b}{b+d}, \quad \text{entonces} \quad \frac{a}{a+b} > \frac{c}{c+d} \qquad (1)$$

TABLA 6

	B	no B	
A	a	b	a+b
no A	c	d	c+d
	a+c	b+d	n

No hay números en la exposición de Mill. Las cantidades son 'todos', 'algunos' y 'nada'; las desviaciones son importantes cualitativa y no cuantitativamente.

El uso de cantidades numéricas en lógica es iniciado por Augustus de Morgan (1847). Basado en esto, George Boole (1854) idea un método para determinar los límites mayor y menor del número de individuos en una categoría lógica. Los desarrollos de Morgan y Boole son base de la obra de W. Stanley Jevons.

Jevons (1870) crea un método que usa ecuaciones para determinar número de objetos en categorías. Cosiste en dividir clases en todas sus posibles subclases, procedimiento ideado por Boole. Si existen las clases (o cualidades) A y B, las posibles subclases son AB, Ab, aB y ab (indicando con minúsculas la ausencia de la cualidad: 'a' = 'no A'). La Tabla 7 muestra la notación de Jevons en una tabla de contingencia actual:

TABLA 7

	B	b	
A	AB	Ab	(A)
a	aB	ab	(a)
	(B)	(b)	

El siguiente es un ejemplo que da Jevons para mostrar su método. En 100 sujetos de la clase A hay 45 B y 53 C. Existe la condición de que B=BC, o sea, que todos los sujetos B sean también C. Son subclases posibles ABC, ABc, AbC y Abc. Jevons aplica números a diversas ecuaciones:

$$BC + Bc = (B)$$
$$bC + bc = (b)$$
$$BC + bC = (C)$$
$$Bc + bc = (c)$$
$$(C) + (c) = (A)$$
$$(B) + (b) = (A)$$

y determina así el número de elementos en clases y subclases. (pp.181-182).

Con una tabla de contingencia, el problema se habría resuelto fácilmente. De acuerdo al ejemplo:

TABLA 8

	C	c	
B	45		45
b			
	53		100

y con sólo llenar las celdas vacías (ver Tabla 9):

TABLA 9

	C	c	
B	45	0	45
b	8	47	55
	53	47	100

Es así como una tabla de contingencia puede ser considerada una presentación compacta del conjunto de ecuaciones que surgen al combinar dos variables. El desarrollo de subclases de Boole y Jevons es realizado sin notarlo por el usuario actual de una tabla de contingencia: al llenar celdas vacías está resolviendo ecuaciones.

La mayor contribución de Jevons a la teoría de asociación es la solución del siguiente problema general: "Dados los números de tres clases de objetos, A, B y C, determinar qué circunstancias o condiciones exigirá la existencia de una clase ABC." (p.184) Eliminando la constante A, la solución de Jevons es:

$$(BC) = (B) + (C) - (A) + (bc) \qquad (2)$$

Como $bc \geq 0$, el límite inferior de BC es el exceso de la suma de B y C sobre A. Usando notación actual (Tabla 1,

página 5), la ecuación (2) sería:

$$a = (a+b) + (a+c) - n + d \qquad (3)$$

y, dado que **d** no puede ser negativo, **a** no puede ser menor que la suma de sus marginales (línea y columna) menos el total de casos. También, si **d** = 0, los marginales de **a** no pueden sumar menos que el número de casos total **n**. Ambas conclusiones son aplicables a cualquier celda de la tabla.

Finalmente, la transformación:

$$a - d = (a+b) + (a+c) - n \qquad (4)$$

muestra que la diferencia entre dos celdas diagonales es igual a la diferencia entre la suma de los marginales de la primera celda y el número total de casos.

La asociación de variables es para Jevons "las combinaciones de ciertas cualidades o clases de las cosas." (p.180)

Su método es la base del trabajo de Yule y, como consecuencia, de la moderna teoría de asociación de variables.

Las principales contribuciones de Jevons son la expansión de grupos a subclases, determinación de sus frecuencias, y establecimiento de relaciones entre ellas.

Desarrolla su método como una "ayuda para todos aquellos problemas matemáticos que incluyan consideraciones lógicas" (p.195), pero no lo transforma en un método aplicable empíricamente. Escribe: "No debemos estimar el valor de una teoría por sus resultados prácticos inmediatos." (p.195)

CAPITULO 3

YULE: TEORIA DE LA ASOCIACION

La teoría de asociación de variables, tal como se conoce hasta hoy, fue establecida por George Udny Yule (1871-1951) en varios artículos escritos entre 1900 y 1912 y en su libro 'An Introduction to the Theory of Statistics' (1911).

El artículo fundamental es 'On the Association of Attributes in Statistics' (1900), en que Yule introduce el término 'asociación' para lo que Mill llamaba 'conexión', Jevons 'combinación', Quetelet 'influencia', y Pearson habría de llamar 'variabilidad dependiente' o 'contingencia'.

TABLA 10

	Y	no Y	
X	a	b	a+b
no X	c	d	c+d
	a+c	b+d	n

Siguiendo a Mill, Yule define asociación como dependencia, o alejamiento de la independencia, siendo este último el concepto del que se debe partir: "Dos cualidades o atributos se

definen como independientes si la probabilidad de encontrarlas juntas es el producto de las probabilidades de encontrar a cada una separadamente." (p.270). Con la nomenclatura de la Tabla 10:

$$\frac{a}{n} = \frac{a+b}{n} \cdot \frac{a+c}{n} \qquad (5)$$

"[(5) y (6) son] la única prueba legítima de dependencia o independencia -asociación o no asociación". (p.270)

$$an = (a+b)(a+c) \qquad (6)$$

(6) es transformable en (7), más útil conceptualmente: la igualdad de las frecuencias observadas y esperadas.

$$a = \frac{(a+b)(a+c)}{n} \qquad (7)$$

Es fácil ver que las ecuaciones (8) a (11), que definen independencia a partir de cada celda, se implican mutuamente.

$$\frac{a}{n} = \frac{a+b}{n} \cdot \frac{a+c}{n} \qquad (8)$$

$$\frac{b}{n} = \frac{a+b}{n} \cdot \frac{b+d}{n} \qquad (9)$$

$$\frac{c}{n} = \frac{a+c}{n} \cdot \frac{c+d}{n} \qquad (10)$$

$$\frac{d}{n} = \frac{b+d}{n} \cdot \frac{c+d}{n} \qquad (11)$$

Si se multiplican los lados derechos de (8) y (11), el producto es igual al de los lados derechos de (9) y (10), por lo que los productos de los lados izquierdos deben ser iguales:

$$\frac{a}{n} \cdot \frac{d}{n} = \frac{b}{n} \cdot \frac{c}{n},$$

lo que resulta en:

$$ad=bc \qquad (12)$$

que Yule llama "los productos de las frecuencias contrarias de

segundo orden" (p.270) y es la fórmula más práctica para determinar independencia en tablas de cuatro celdas. Por lo tanto,

$$ad-bc \qquad (13)$$

es 0 en caso de independencia y diferente de 0 en caso de asociación; si ésta es positiva, (13) > 0, y si es negativa, (13) < 0. Como (13) no tiene valores límites no sirve como coeficiente de asociación; es, sin embargo, la base de la mayor parte de ellos.

Cuando no hay independencia, las ecuaciones (5) a (12) se transforman en desigualdades, y (13) $\neq$ 0. Para medir el grado de dependencia se necesita "algún tipo de coeficiente de asociación que debería reemplazar al coeficiente de correlación de variables continuas y ser una medida de cercanía de la asociación, por un lado hacia independencia completa y por otro hacia asociación completa." (p.271)

Yule fija reglas que debe seguir tal coeficiente:

1, ser 0 cuando variables son independientes, y sólo cuando lo son;

2, ser +1 sólo cuando variables están completamente asociadas en forma positiva y sólo en ese caso: cuando b=0 ó c=0; y

3, ser -1 sólo cuando variables están completamente asociadas en forma negativa y sólo en ese caso: cuando a=0 ó d=0. (p.271)

Estas reglas son aún exigibles para todo coeficiente de asociación. Si las variables son nominales, el signo no tiene sentido.

Las reglas 2 y 3 requieren una celda vacía para que un coeficiente llegue a los valores extremos +1 ó -1. Pero puede ser útil a veces que para llegar a esos valores ambas celdas deban estar vacías. Este problema es reconocido por Yule en una nota agregada posteriormente al artículo: [notación de Tabla 1]

"En varias ocasiones me ha parecido posible que me he limitado mucho en esta sección al definir el caso de asociación completa simplemente como equivalente al caso lógico. Se podría haber obtenido un coeficiente de asociación de mayor conveniencia analítica al definir a x e y como completamente asociados sólo cuando todos los X fueran no X y todos los Y fueran no Y. Distinguir el caso lógico mediante un valor determinado de la asociación tiene, sin embargo, obvias ventajas." (p.271)

La asociación completa puede ser expresada lógicamente en las siguientes hipótesis:

1. 'X es condición suficiente pero no necesaria de Y.' Significa que X puede determinar Y, pero hay otras variables (por ejemplo, Z) que también pueden hacerlo:

GRAFICO 3

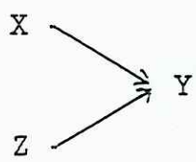


TABLA 11

	Y	no Y
X		0
no X		

2. 'X es condición necesaria pero no suficiente de Y.' Significa que sólo X puede determinar Y, pero requiere otra variable, Z, para hacerlo:

GRAFICO 4

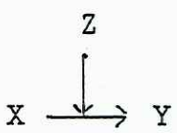


TABLA 12

	Y	no Y
X		
no X	0	

3. 'X es condición necesaria y suficiente de Y.' Significa que sólo X puede determinar Y, y no requiere otra variable para hacerlo. (Ver Kang 1972)

GRAFICO 5

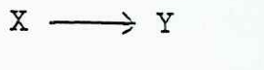


TABLA 13

	Y	no Y
X		0
no X	0	

La hipótesis 1 puede expresarse simplemente como 'si X entonces Y'; la hipótesis 2, 'si Y entonces X'; y la hipótesis 3, 'si X entonces Y, y si Y entonces X'. El 'caso lógico' corresponde a las dos primeras hipótesis.

Las hipótesis 1 y 2 se refieren a lo que Boudon llama 'implicación simple', y la tercera a lo que llama 'implicación recíproca'. (Boudon 1974) Un coeficiente puede identificar asociación completa (valor absoluto = 1) con implicación simple o con implicación recíproca.

Yule (1900) presenta lo que puede ser llamado el primer coeficiente de asociación: Q, llamado así en honor de Adolphe Quetelet (1769-1874), un estadístico belga que había creado un complicado coeficiente de 'grado de influencia' con el símbolo 'phi' (Quetelet 1832; Yule 1900, pp. 280-282):

$$\phi = \frac{\frac{a}{a+c} - \frac{a+b}{n}}{\frac{a+b}{n}}$$

La fórmula de Q es:

$$Q = \frac{ad - bc}{ad + bc}$$

que cumple las reglas indicadas por el mismo Yule. Es de implicación simple pues con una celda vacía alcanza valores extremos: +1 si 'b' o 'c' = 0, y -1 si 'a' o 'd' = 0.

Q es simple y fácil de calcular. Si se aplica a las Tablas 2 a 5, su valor es respectivamente 0, .38, -.88 y -1, mostrando independencia en la Tabla 2, asociación positiva en la Tabla 3, y negativas crecientes en las Tablas 4 y 5. En ésta, que tiene dos celdas vacías, habría bastado una para dar el valor -1.

Q es independiente del número de casos. Por ejemplo, si se duplican las frecuencias de la Tabla 3, como muestra la Tabla 14, Q mantiene su valor .38:

TABLA 14

		Velocidad		
		Rápido	Lento	
Edad	Mayor	6	4	10
	Menor	4	6	10
		10	10	20

También Q es independiente de las razones entre marginales siempre que las razones entre celdas permanezcan constantes. Por ejemplo, si se duplica el número de niños de más peso en la Tabla 4, manteniendo 1 a 4 la razón de rápidos a lentos, tal como aparece en la Tabla 15, el valor de Q sigue siendo -.88. Esta propiedad permite aplicar Q a proporciones y porcentajes.

TABLA 15

		Velocidad		
		Rápido	Lento	
Peso	Pesado	2	8	10
	Liviano	4	1	5
		6	9	15

En Yule (1911) se presenta un nuevo criterio para independencia en tablas de 2x2: la **igualdad de proporciones**, que toma las siguientes cuatro formas, cada una de las cuales implica a las demás:

$$\frac{a}{a+b} = \frac{c}{c+d} \qquad (14)$$

$$\frac{b}{a+b} = \frac{d}{c+d} \qquad (15)$$

$$\frac{a}{a+c} = \frac{b}{b+d} \qquad (16)$$

y

$$\frac{c}{a+c} = \frac{d}{b+d} . \qquad (17)$$

La igualdad de proporciones está relacionada algebraicamente con (7) y (12), y desde este punto de vista no parece ser un avance. Lo es como **diferencia de proporciones**, coeficiente simbolizado habitualmente por la letra **D** y a veces presentado como diferencia porcentual al multiplicarse por cien.

Yule hace notar que el resultado numérico de una diferencia de proporciones 'por filas',

$$\frac{a}{a+b} - \frac{c}{c+d} , \qquad (18)$$

suele ser diferente del resultado 'por columnas',

$$\frac{a}{a+c} - \frac{b}{b+d} . \qquad (19)$$

Indica que la diferencia que hay que usar es la que "conteste directamente la pregunta específica que tenemos en mente" (Yule y Kendall 1968, p.26): si x es la variable independiente (Dyx), es preferible (18); si la independiente es y (Dxy), hay que usar (19).

D es, por lo tanto, asimétrico: exige definir una variable como independiente y otra como dependiente. Es también de implicación recíproca ya que define asociación perfecta en forma restrictiva, alcanzando el valor absoluto 1 sólo si dos celdas diagonales están vacías.

Q, Dyx y Dxy tienen valor 0 en caso de que x e y sean independientes, y valor absoluto 1 en caso de asociación perfecta con dos celdas vacías. En toda otra circunstancia el valor absoluto de Q es mayor que el de cualquiera de los D. Los tres coeficientes tienen igual signo (+,-).

Yule (1912) aplica el coeficiente D a datos sobre recuperaciones y muertes entre pacientes vacunados y no vacunados en varios distritos y hospitales de Leicester (Inglaterra) durante una epidemia de viruela (Tabla 16). Yule muestra que el valor máximo que puede alcanzar D es a veces engañosamente bajo.

Entre los no vacunados .88 se recuperaron; .99 entre los vacunados. Dyx es sólo .11. Aunque se hubieran recuperado

todos los vacunados, Dyx habría llegado apenas a .12. La razón es que la proporción de muertos es pequeña (21/357 < 6%) y la consecuente gran diferencia entre marginales (de recuperados y muertos) limita el rango de Dyx. En esta tabla, Q es .86. Yule usa este ejemplo para justificar su preferencia de Q sobre D.

TABLA 16. **Relación de Vacunación (x) y Recuperación (y) en Leicester, en Epidemia de Viruela de 1892-3**

	Recuperados	Muertos	
Vacunados	197	2	199
No Vacunados	139	19	158
	336	21	357

Yule limita el concepto de asociación a tablas de 2x2. Para tablas mayores usa el término de Pearson 'contingencia', y no crea coeficientes para ellas.

Karl Pearson (1857-1936) critica la teoría de asociación de variables de Yule en varios artículos a partir de 1901.

Para Pearson, el supuesto de normalidad es esencial. Concibe la tabla de 2x2 como una superficie de correlación normal dividida en cuatro por una paralela a cada eje. Escribe: "Todo nuestro razonamiento en este artículo está basado en normalidad de frecuencia." (1901, p. 17)

Reconoce que "es difícil encontrar datos que sigan esa ley dentro del límite de errores probables" (p.17); sin embargo, "queremos una función que no difiera mucho de la correlación (ya que) ésta no es considerada como algo propio de la distribución normal sino como algo significativo para todas las posibles distribuciones." (p.18) Debe ser "una función que desaparezca y alcance la unidad con 'r' y que sea igual a 'r' si se divide la tabla por las medianas." (p.15)

Para medir el grado de 'variabilidad dependiente', que así llama a la asociación, sugiere cuatro coeficientes de implicación simple, Q1, Q3, Q4 y Q5 (pp.15-16). Son variaciones de Q de Yule -al que Pearson llama Q2-, tienen fórmulas muy complejas y no asumen valores negativos.

Tras aplicar los cinco coeficientes a quince series

de datos seleccionados para cubrir un gran rango de valores, Pearson concluye que Q5 es preferible pues se acerca más a los valores de r. La fórmula de Q5 es:

$$Q5 = \operatorname{sen} \frac{\pi}{2} \frac{1}{\sqrt{1 + k^2}},$$

$$\text{en que} \quad k^2 = \frac{4\,abcd\,n^2}{(ad - bc)^2\,(a+d)\,(b+c)}.$$

En un artículo publicado tres años después, Pearson (1904) presenta el concepto de **contingencia** para variables nominales. Distingue contingencia de asociación, restringiendo ésta a relación de variables dicotómicas, quizás porque Q de Yule sólo se aplica a tales variables.

"El concepto de contingencia nos permite generalizar la noción de asociación de dos atributos desarrollada por el Sr. Yule. Podemos clasificar a individuos no sólo en dos grupos alternativos sino en cuantos grupos de atributos exclusivos queramos." (p.474)

Pearson define contingencia como diferencia de frecuencias observadas y esperadas. "Considero cualquier medida del

alejamiento de la clasificación con respecto a la probabilidad independiente como una medida de su contingencia." (p.445)

Entonces, a partir de su coeficiente chi-cuadrado, χ^2, presentado en Pearson (1900), introduce el 'promedio del cuadrado de la contingencia' ('mean square contingency') usando como símbolo la letra 'phi' al cuadrado, ϕ^2 (1904, p.446):

$$\phi^2 = \frac{\chi^2}{n}$$

Aunque chi-cuadrado no tiene límites, ϕ^2 no puede ser superior a 1 en tablas de 2x2, ya que en éstas el valor de chi-cuadrado no puede ser mayor que el número de casos (n).

Chi-cuadrado puede ser mayor que n en tablas >2x2 por lo que Pearson limita ϕ^2 a tablas de 2x2. Desarrolla luego otra fórmula (p.461):

$$\phi^2 = \frac{(ad - bc)^2}{(a+b)\,(a+c)\,(b+d)\,(c+d)} \qquad (20)$$

Al extraer raíz en (20), resulta el actual coeficiente Ø (Pearson y Heron 1913, p.167), cuya fórmula, pero sin el símbolo Ø, había presentado años antes Pearson (1901, p. 12):

$$Ø = \frac{ad - bc}{\sqrt{(a+b)\,(a+c)\,(b+d)\,(c+d)}}, \qquad (21)$$

con rango de -1 a +1, aplicable a variables nominales y ordinales. Ø tiene igual numerador que Q, habitual en futuros coeficientes.

Ø es r aplicado a tablas de 2x2. Por lo tanto, $Ø^2$ se puede interpretar operacionalmente como r^2: la proporción de variación de cada variable explicada por variación de la otra. r^2 es el llamado 'coeficiente de determinación'.

Ø es de implicación recíproca: alcanza 1 sólo si b y c están vacíos, y -1 sólo si a y d lo están. Para que ambos valores extremos sean posibles en una tabla, los cuatro marginales deben ser iguales.

Al igual que Q, Ø es independiente del número de casos, pero, al revés de Q, es afectado por variaciones en la razón de los marginales aunque la razón de las celdas permanezca constante.

Ø es la media geométrica de las 'comparaciones de proporciones':

$$Ø = \sqrt{Dyx \cdot Dxy} \qquad (22)$$

por lo que su valor está entre los dos, salvo cuando coinciden en valores extremos. Q, Ø y los D tienen valor 0 en caso de independencia y valor absoluto 1 en caso de asociación perfecta con dos celdas vacías. En toda otra circunstancia el valor absoluto de Q es mayor. Los cuatro coeficientes tienen igual signo.

Pearson presenta otros coeficientes para tablas de 2x2. El 'coeficiente de contingencia', C :

$$C = \sqrt{\frac{Ø^2}{1 + Ø^2}} = \sqrt{\frac{\chi^2}{n + \chi^2}} \,,$$

que requiere gran número de categorías con frecuencias muy pequeñas y parecidas para acercarse al valor 1, aunque sin alcanzarlo. (1904, p.449)

El 'coeficiente tetracórico' (Pearson 1901) requiere variables continuas con distribución normal bivariada, en

cuyo caso es equivalente a r. Yule lo llama el 'coeficiente normal'. (1912, p.613) La fórmula es muy compleja (ver Walker y Lev 1953, p.274), y existen gráficos en que interpolando se puede obtener el valor.

La dura controversia entre las posiciones de Pearson y Yule puede ejemplificarse con un artículo que David Heron, colega de Pearson, publica sobre el entonces reciente libro de Yule (1911). "Hay que lamentar mucho que en el reciente libro de texto del Sr. G. Udny Yule no haya habido más cuidado para asegurarse de que los procesos ahí descritos tengan base teórica firme." (1911, p.109)

Heron critica que Q y Ø difieran en resultados: "Es muy claro que (Yule) no está consciente de las limitaciones de los procesos que recomienda y que no ha empleado la prueba más obvia de su validez, que es comparar resultados obtenidos por ambos métodos al aplicarlos a los mismos datos." (p.111)

El criterio de validez para Heron es el coeficiente tetracórico, que supone que las tablas de contingencia son superficies bivariadas normales, denotando la idea de continuidad, no aplicable a atributos y poco probablemente a variables categóricas.

El último párrafo del artículo de Heron indica la creciente popularidad de la posición de Yule: "Esta discusión del peligro de las fórmulas de correlación del Sr. Yule no es ociosa ya que el coeficiente de asociación ha sido recientemente adoptado por algunos autores italianos. De esta manera, mucho daño se hará a la ciencia estadística, debido a la postulación de conclusiones falsas que resulten del uso de fórmulas inadecuadas." (p.122)

En parte de su artículo de 1912, Yule responde a Heron. A las objeciones de éste sobre diferencias en valores numéricos de coeficientes, replica que ambos coeficientes tienen propiedades esencialmente distintas, y, dado que esto ocurre con todos los coeficientes, ninguno puede ser usado como prueba de validez de los demás.

Ante la crítica de Heron de que Q toma diferentes valores al cambiar el punto de corte de las dicotomías, Yule, mediante una tabla de 7x7 usada por Pearson, muestra que el coeficiente tetracórico varía más con tales cambios que Q y Ø. Agrega que ésta no es una prueba de invalidez, ya que "es imposible la estabilidad de un coeficiente de asociación ante divisiones en una tabla." (p.634)

Al final de su artículo, Yule elogia el trabajo de Pearson en este campo como "un trabajo matemático notable (pero)

el valor del método sugerido depende enteramente de la existencia empírica de los supuestos. Estos nunca han sido puestos a prueba adecuadamente y las pocas pruebas que he aplicado han bastado para demostrar que son, para decir lo menos, de validez muy dudosa." (p.640)

CAPITULO 4

INTERPRETACION OPERACIONAL DE LOS COEFICIENTES

En 1954, Leo Goodman y William Kruskal publican un trascendental artículo en el que sugieren usar coeficientes cuyos valores numéricos sean expresables en palabras (como probabilidades u otra forma); por ejemplo, qué significa que un coeficiente sea .40. Llaman a esto una **interpretación operacional**.

La única interpretación operacional existente hasta entonces era la del coeficiente de correlación r, que consistía en interpretar r^2 como proporción de variación de una variable explicada por variación de la otra.

Había también una tradición de construcción de coeficientes sin interpretación operacional basados en chi-cuadrado: contingencia de Pearson, T de Tschuprow, y C de Cramér. (Ver fórmulas de T y C en Loether y McTavish 1974.)

Sobre ésta escriben Goodman y Kruskal (1954, p.740): "Que un excelente test de independencia esté basado en chi-cuadrado no quiere decir que esta medida o alguna función suya

sea medida apropiada del grado de asociación. No hemos visto una defensa convincente de coeficientes basados en chi-cuadrado como medidas de asociación."

Bishop, Fienberg y Holland (1975): "Las medidas basadas en chi-cuadrado tienen la ventaja de estar basadas en una prueba de significación. La mayor dificultad al usarlas es que no tienen interpretación clara." (p.393)

Finalmente, Agresti y Agresti (1979): "Las magnitudes de estas medidas no se pueden interpretar fácilmente o compararse para tablas de diferentes dimensiones, por lo que sugerimos que no sean usadas." (p.223)

En Goodman y Kruskal (1954) aparecen las bases de dos modelos de interpretación operacional: 'de pares de casos' y 'reducción proporcional del error'. La primera se aplica a coeficientes de variables ordinales y consiste en transformar sus fórmulas para poder compararlas. La segunda, a coeficientes de cualquier nivel de medición, cuyos valores expresa como mejoría de predicción.

Goodman y Kruskal aplican cada interpretación operacional sólo a un coeficiente. Autores posteriores las han ampliado a otros coeficientes. El presente trabajo intenta analizar y generalizar estos modelos.

4.1: INTERPRETACION OPERACIONAL 'DE PARES DE CASOS'

La base de este modelo es tomar los casos de una tabla de contingencia y formar pares. El número de pares es igual a la 'combinación de 2 en n', que se simboliza 'nC2' y se calcula con el producto ½n(n-1).

Los pares de casos pueden ser de diferentes tipos. Al formular coeficientes usando tipos de pares, se pueden interpretar en términos de éstos y compararlos.

Para entender la noción de 'tipos de pares', supóngase que de entre los casos de la Tabla 17 tomamos dos casos cualesquiera, los que podemos llamar casos I y II.

TABLA 17

	y+	y-	
x+	a	b	a+b
x-	c	d	c+d
	a+c	b+d	n

Si el caso I supera al caso II en ambas variables, o sea, si I es x+ e y+, y II es x- e y-, se dice que es un **par concordante**, que sugiere asociación positiva entre x e y. En la tabla, el caso I está en la celda a y el caso II en d. El

producto ad indica cuántos pares concordantes hay en la tabla.

Si el caso I supera al caso II en una variable, pero es superado en la otra, se tiene una discordancia, un **par discordante**, sugiriendo asociación negativa entre las variables. En la tabla, un caso está en b y otro en c. El producto bc muestra cuántos pares discordantes hay en la tabla.

Si ad > bc, o sea, si predominan pares concordantes sobre discordantes, la asociación es positiva. Si ad < bc, la asociación es negativa. Si ad=bc, se tiene la fórmula (10), una manera de determinar independencia.

Los pares concordantes difieren en ambas variables, al igual que los discordantes. Pero hay pares que difieren en una sola variable. Puede ser que los casos I y II tengan valor igual en x, y diferente en y; sería un **par empatado sólo en x**, que se da cuando hay un caso en la celda a y otro en b, o uno en c y otro en d. El número de estos pares es ab + cd.

Un **par empatado sólo en y** está formado por un caso en a y otro en c, o por uno en b y otro en d. El número total es ac + bd.

Si la relación de variables es asimétrica, ab + cd

es el número de pares de casos empatados en la variable independiente (x) pero no en la dependiente (y), y ac + bd el de pares empatados en la dependiente pero no en la independiente.

Finalmente, hay **pares empatados en ambas variables**; cada uno de estos pares está formado por dos casos que están en la misma celda. El número de estos pares es igual a la suma de la combinaciones de 2 en a, 2 en b, 2 en c y 2 en d.

El número total de pares de una tabla de contingencia es la suma de los cinco tipos: concordantes, discordantes, empatados en x, empatados en y, y empatados en ambas variables. La Tabla 18 muestra esta distribución de pares.

TABLA 18: Distribución de Pares de Casos de la Tabla 17

Tipo de Par	Número de Pares	Símbolo
Concordante	ad	C
Discordante	bc	D
Empatado sólo en x	ab + cd	EX
Empatado sólo en y	ac + bd	EY
Empatado en x e y	aC2 + bC2 + cC2 + dC2	Z
Total	nC2	

Los pares C, D y EX están formados por casos que son diferentes en la variable dependiente: 'DY'; los pares C, D y EY, por casos diferentes en la variable independiente: 'DX'. Por lo tanto,

$$\begin{aligned} DY &= C + D + EX \\ &= ad + bc + ab + cd \\ &= a(b+d) + c(b+d) \\ &= (a+c)\,(b+d) \end{aligned}$$

y también

$$\begin{aligned} DX &= C + D + EY \\ &= ad + bc + ac + bd \\ &= a(c+d) + b(c+d) \\ &= (a+b)\,(c+d) \end{aligned}$$

La interpretación 'de pares de casos' hace comparables coeficientes al expresar sus fórmulas en tipos de pares. Para tablas de 2x2 se consideran aquí los coeficientes Q, Dyx y ø.

$$Q = \frac{ad - bc}{ad + bc} = \frac{C - D}{C + D}. \qquad (23)$$

$$Dyx = \frac{a}{a+b} - \frac{c}{c+d}$$

$$= \frac{a\,(c+d) - c\,(a+b)}{(a+b)\,(c+d)}$$

$$= \frac{ac + ad - ac - bc}{ac + ad + bc + bd}$$

$$= \frac{ad - bc}{ad + bc + ac + bd}$$

$$= \frac{C - D}{C + D + EY} \qquad (24)$$

o también $$\frac{C - D}{DX} \cdot \qquad (25)$$

$$\emptyset = \frac{ad - bc}{\sqrt{(a+b)\,(a+c)\,(b+d)\,(c+d)}}$$

$$= \frac{ad - bc}{\sqrt{[(a+b)\,(c+d)]\,[(a+c)\,(b+d)]}}$$

$$= \frac{ad - bc}{\sqrt{(ad+bc+ac+bd)\,(ad+bc+ab+cd)}}$$

$$= \frac{C - D}{\sqrt{(C + D + EY)\,(C + D + EX)}} \qquad (26)$$

$$= \frac{C - D}{\sqrt{(DX)\,(DY)}} \; . \qquad (27)$$

Al tener el mismo numerador, los coeficientes tienen igual signo. Sus diferencias numéricas se deben a que los denominadores no consideran los mismos tipos de pares.

La Tabla 19 (igual a la Tabla 15 del Cap. 3) relaciona peso y velocidad de 15 niños:

TABLA 19

		Velocidad		
		Rápido	Lento	
Peso	Pesado	2	8	10
	Liviano	4	1	5
		6	9	15

Los valores de los coeficientes son:

según (23): $$Q = \frac{2 - 32}{2 + 32} = -.88 \qquad (28)$$

según (24): $$Dyx = \frac{2 - 32}{2 + 32 + 16} = -.60 \qquad (29)$$

según (25): $$Dyx = \frac{2 - 32}{10 \cdot 5} = -.60 \qquad (30)$$

según (26): $$\emptyset = \frac{2 - 32}{\sqrt{(2 + 32 + 16)\,(2 + 32 + 20)}} = -.58 \qquad (31)$$

según (27): $$\emptyset = \frac{2 - 32}{\sqrt{(10 \cdot 5)\,(6 \cdot 9)}} = -.58 \qquad (32)$$

Q es mayor porque considera menos pares en el denominador. El valor Q=-.88 quiere decir que pares discordantes superan a concordantes en 88% del total de ambos tipos de pares.

Dyx incluye pares empatados en la variable dependiente, aumentando el denominador. Disminuye la fuerza de la relación si, además de concordantes y discordantes, hay pares de

niños de diferente peso que corren a igual velocidad.

Dyx considera todos los pares de niños de diferente peso (x). Si el de más peso corre más, la asociación es más positiva; si corre menos, es más negativa; si corren a la misma velocidad, la asociación no es más positiva ni más negativa sino menor.

El denominador de Dyx (29) incluye pares diferentes en ambas variables o que empatan sólo en la dependiente. Todos ellos difieren en la independiente, resultando la fórmula (30). El valor Dyx=-.60 indica que pares discordantes superan a concordantes en 60% del total de pares diferentes en la variable independiente.

Ø sigue el razonamiento de Dyx con respecto a pares empatados, pero como es coeficiente simétrico, incluye empatados sólo en x además de empatados sólo en y.

Esta interpretación operacional es directamente aplicable a coeficientes de tablas >2x2. Se consideran los coeficientes gamma, d de Somers y τb (tau-b).

TABLA 20

	y1	y2	y3	
x1	a	b	c	a+b+c
x2	d	e	f	d+e+f
x3	g	h	i	g+h+i
	a+d+g	b+e+h	c+f+i	n

Con la nomenclatura de la Tabla 20, los pares son:

Concordantes: a(e+f+h+i) + b(f+i) + d(h+i) + ei

Discordantes: c(d+e+g+h) + b(d+g) + f(g+h) + eg

Empatados en x: a(b+c) + bc + d(e+f) + ef + g(h+i) + hi

Empatados en y: a(d+g) + dg + b(e+h) + eh + c(f+i) + fi

Empatados en x e y: aC2 + bC2 + ... + hC2 + iC2

Total de pares: nC2

Gamma es igual a Q en tablas de 2x2. Creado por Goodman y Kruskal (1954, pp.747-754) "es lo mismo que el coeficiente de asociación Q de Yule" (p.750). La fórmula de gamma es (23).

d, creado por Somers (1962), es igual a D en tablas de 2x2. Por lo tanto, es asimétrico, siendo **dyx** la versión más usada (con x simbolizando la variable independiente). Su fórmula es (24) o (25).

τ (tau), creado en 1948 por Kendall (1975, capítulo 3) mide la asociación de variables expresadas en rangos. Kendall desarrolla tres coeficientes: τa (tau-a) para datos sin empates en rangos, y τb y τc para datos con empates. Estos dos son aplicables por lo tanto a tablas de contingencia con variables ordinales.

τb es aplicable a tablas cuadradas (sólo en ellas llega a ± 1). τc a cualquier tipo de tablas; pero es "difícil de interpretar y a este respecto menos satisfactorio que τb." (Blalock 1972, p.423)

τb es igual a $\emptyset$ en tablas de 2x2. Es también simétrico. Su fórmula es (26) ó (27).

En cada línea de la Tabla 21 se presenta un coeficiente para tabla mayor, su equivalente en tabla 2x2, y la fórmula de ambos:

TABLA 21: Equivalencias y Fórmulas de Coeficientes

Tablas de 2x2	Tablas > 2x2	Fórmula
Q	gamma	$\frac{C - D}{C + D}$
Dyx	dyx	$\frac{C - D}{C + D + EY}$
Ø	τb ...	$\frac{C - D}{\sqrt{(C + D + EY)\,(C + D + EX)}}$

Si las variables son independientes o hay asociación perfecta de implicación recíproca, los seis coeficientes tienen igual valor. Fuera de estos casos extremos, Q (gamma) es mayor en valores absolutos que los demás. Ø (τb) puede ser mayor o menor que Dyx (dyx).

Tal como ocurre con Ø y D (22), τb es la media geométrica de dyx y dxy. (Recuérdese que r lo es de los coeficientes de regresión byx y bxy.)

Los coeficientes dan valores menores cuantos más tipos de pares considera el denominador. El caso extremo es τa, que los toma a todos en cuenta. Su fórmula es:

$$\tau a = \frac{C - D}{C + D + EX + EY + Z} = \frac{C - D}{nC2}$$

Para que el valor de τa pueda llegar a ±1, puede haber como máximo un caso por celda. Sólo así Z=0. Por lo tanto, τa es inaplicable en tablas de contingencia.

CONCLUSIONES.

El objetivo del modelo es igualar numeradores de coeficientes (C-D) y radicar sus diferencias en los denominadores, mostrando qué tipos de pares considera cada coeficiente y qué ponderación les da.

Al hacer comparables las fórmulas, se aclara cómo operacionaliza cada coeficiente el concepto de asociación.

Q (gamma) lo operacionaliza como las variaciones (aumento o disminución) de una variable ante variaciones de la otra.

Dyx (dyx), como las variaciones (o no variación) de la dependiente ante variaciones de la independiente. Puede ser que aquélla no varíe, y tal respuesta es importante porque indica una asociación de menor fuerza. Si varía, interesa si aumenta o disminuye.

Ø (τb) lo hace de la misma manera que Dyx, pero siendo simétrico, toma a una variable como independiente y después a la otra. Compensa la multiplicación de ambas situaciones extrayendo raíz.

El modelo 'de pares de casos' expresa las fórmulas de los coeficientes en pares de casos, con numerador C-D. Un coeficiente es interpretable operacionalmente por este modelo si su fórmula es transformable a ese formato.

4.2: INTERPRETACION OPERACIONAL DE 'REDUCCION PROPORCIONAL DEL ERROR'

La 'reducción proporcional del error' (RPE) interpreta asociación como mejoría en la predicción de una variable dependiente al conocer la independiente. Por ejemplo, si predecimos mejor la posición política de una persona al conocer su nivel socioeconómico, podemos decir que ambas variables están asociadas.

RPE supone dos situaciones en que se predice alguna característica de una variable dependiente: (I) sin más información; y (II) conociendo su relación con una variable independiente. Estas situaciones se asocian en la fórmula:

$$RPE = \frac{\text{Errores en (I)} - \text{Errores en (II)}}{\text{Errores en (I)}}, \qquad (33)$$

o sea, qué proporción de los errores cometidos en la situación I se reduce en la situación II. Esta proporción es el grado de asociación de las variables.

La naturaleza de las predicciones -¿qué se predice?- y la definición de 'error' varían en general con el nivel

de medición de las variables y en particular con cada coeficiente.

Además de los ya vistos, se han de analizar como RPE dos coeficientes presentados por Goodman y Kruskal (1954) para variables nominales. Uno es 'lambda', y el otro tiene el mismo símbolo 'τb' del coeficiente de Kendall, por lo que se ha de llamar aquí 'τGK'.

Lambda (Goodman y Kruskal 1954, pp.740-747) mide asociación de variables nominales en tablas > 2x2, como la Tabla 22 en que 1200 personas son clasificadas de acuerdo a lugar de nacimiento y área de trabajo.

TABLA 22

		Area de Trabajo			
		Area A	Area B	Area C	
Lugar de Nacimiento	Santiago	130	220	180	530
	Norte	140	70	60	270
	Sur	90	160	150	400
		360	450	390	1200

Situación I: se predice área de trabajo de una persona sin conocer su lugar de nacimiento. Mejor predicción es la categoría modal: Area B. Al hacerlo para los 1200 casos habrá 450 aciertos (la frecuencia modal) y 1200-450=750 errores.

Situación II: se predice conociendo lugar de nacimiento. Si la persona es de Santiago, mejor predicción es que trabaja en el Area B; habrá 220 aciertos. Si es del norte se predice Area A con 140 aciertos. Si es del sur, Area B con 160 aciertos. El total de aciertos es 520, y los errores son 1200-520=680.

De acuerdo a (33),

$$\text{Lambda} = \frac{750 - 680}{750} = \frac{70}{750} = .09$$

Al predecir area de trabajo conociendo lugar de nacimiento se eliminaron 70 de 750 errores cometidos al no conocerlo. La proporción eliminada es .09, valor de lambda.

En la situación I se predice la moda (o modo) de la variable dependiente, y error es todo caso fuera de la clase modal.

En la situación II, se predice la moda de la variable dependiente para cada categoría de la variable independiente, y error es todo caso fuera de la clase modal de la categoría.

La proporción de reducción de errores no es igual a la de aumento de aciertos, que es

$$\frac{520 - 450}{450} = \frac{70}{450} = .16,$$

proporción en que II aumenta los aciertos de I.

Al principio de esta sección se dice que RPE indica mejoría en la predicción. Esta mejoría se determina por reducción de errores, no por aumento de aciertos.

τGK (Goodman y Kruskal 1954, pp.759-760) es también asimétrico. Difiere de lambda en que éste predice la moda y τGK predice ubicación de casos en la variable dependiente.

Usando la Tabla 22, situación I es asignar al azar los 1200 casos a categorías de la variable dependiente, de modo que haya 360 personas en el Area A, 450 en B y 390 en C.

Error es asignar una persona a una categoría equivocada. Habiendo 360 personas en el Area A, 840 no pertenecen a ella. La probabilidad de asignar a alguien equivocadamente a esa área es 840/1200, y como son 360 las personas asignadas, se pueden predecir (840/1200)·360 = 252 errores.

Los errores para el Area B son (750/1200)·450 = 281.25; y para C, (810/1200)·390 = 263.25. Por lo tanto, en la situación I hay 252 + 281.25 + 263.25 = 796.5 errores.

En la situación II se repite la asignación, pero dentro de cada categoría de la variable independiente. Las 530 personas de Santiago son asignadas al azar a las tres áreas; para A se pueden esperar (400/530)·130 = 98.11 errores; para B, (310/530)·220 = 128.68; y para C, (350/530)·180 = 118.87.

El proceso se repite para Norte y Sur, y se suman los errores esperados. Hay 771.09 en total. De acuerdo a (33),

$$\tau GK = \frac{796.5 - 771.09}{796.5} = \frac{25.41}{796.5} = .03$$

Este valor de τGK se puede interpretar operacionalmente diciendo que la ubicación de estas personas en sus áreas de trabajo mejora 3% cuando se sabe el lugar de nacimiento.

En tablas de 2x2, el valor de τGK es igual a ϕ^2 y, por lo tanto, a τb^2.

Con **variables ordinales**, se toman nuevamente pares de casos. Se predice si son concordantes o discordantes.

En la situación I se predice al azar, y a la larga se puede esperar que la mitad de las veces haya aciertos y la otra mitad errores.

En la situación II se sabe si la relación de las variables es positiva o negativa. Si es positiva, cuando un caso supera al otro en la variable independiente se predice que lo supera también en la dependiente (concordante); si es negativa, se predice que es superado en la dependiente (discordante). Se definen como errores pares discordantes si la relación es positiva o concordantes si es negativa.

La Tabla 23 (igual a Tabla 19) puede servir para interpretar operacionalmente como RPE coeficientes para variables ordinales.

TABLA 23

		Velocidad: Rápido	Velocidad: Lento	
Peso	Pesado	2	8	10
	Liviano	4	1	5
		6	9	15

Pares concordantes y discordantes suman 34. En la situación I se pueden esperar 17 aciertos y 17 errores.

En II se predice 'discordante' ya que las variables están relacionadas negativamente: Q=-.88 [fórmula (28)]. Hay 2 pares concordantes y 32 discordantes, o sea, 2 errores. De acuerdo a (33),

$$Q\ (\text{gamma}) = \frac{17 - 2}{17} = \frac{15}{17} = .88 \qquad (34)$$

Se puede interpretar Q (gamma) como mejoría al predecir velocidad conociendo peso. La mejoría es de .88, pues 15 de 17 errores de la situación I han sido reducidos en la situación II.

Si no aparece el signo menos en (34) es porque lo interpretado operacionalmente es el valor numérico del coeficien-

te. El signo sirve para predecir 'concordante' o 'discordante' en la situación II.

Al interpretar Dyx (dyx) como RPE (Somers 1968) se definen aciertos y errores de otra manera. Si es una relación positiva, son errores los pares discordantes; si es negativa, son errores los concordantes; pero un par que empata en la variable dependiente no se toma como concordante o discordante, sino como algo intermedio. Se define, por lo tanto, como ½ acierto y ½ error. Entonces,

$$Dyx = dyx = \frac{C - D}{C + D + EY} = \frac{(C + \tfrac{1}{2}EY) - (D + \tfrac{1}{2}EY)}{(C + \tfrac{1}{2}EY) + (D + \tfrac{1}{2}EY)} . \qquad (35)$$

De acuerdo a la Tabla 23,

$$Dyx = dyx = \frac{2 - 32}{2 + 32 + 16} = \frac{(2 + \tfrac{1}{2}16) - (32 + \tfrac{1}{2}16)}{(2 + \tfrac{1}{2}16) + (32 + \tfrac{1}{2}16)}$$

$$= \frac{10 - 40}{10 + 40} = \frac{-30}{50} = -.60$$

Hay 10 pares concordantes y 40 discordantes. La situación I es predecir al azar concordancia o discordancia en los 50 pares de casos; pueden esperarse 25 errores. En II, sabiendo que hay asociación negativa, se predicen pares discordantes; los 10 pares concordantes son errores.

La interpretación RPE, de acuerdo a (33), es:

$$Dyx = dyx = \frac{25 - 10}{25} = \frac{15}{25} = .60 \qquad (36)$$

Se puede interpretar operacionalmente Dyx (dyx) como mejoría al predecir velocidad conociendo peso. La mejoría es .60 pues 15 de los 25 errores de I han sido reducidos en II.

La diferencia numérica con Q (gamma) se debe a que Dyx (dyx) define pares concordantes y discordantes agregándoles pares empatados en la variable dependiente.

La situación de τb y ϕ como RPE es diferente. El primero no ha podido ser interpretado operacionalmente de esta manera (que sepa este autor), salvo para tablas de 2x2 a través de ϕ.

En el artículo considerado básico del enfoque RPE, Costner (1965) dice: "No he podido diseñar reglas o definiciones para ninguno de los τ de Kendall que permitan interpretarlos como RPE cuando hay empates. Aunque fuera posible formularlas, serían indudablemente tan complejas que la aplicación de la interpretación RPE sería limitada." (pp.347-348)

Wilson (1969) presenta una interpretación para τb que tiene dos defectos: su aplicación es muy complicada, y en ciertos casos la situación II produce más error que la I, resultando RPE negativa. Agrega al final del artículo: "Por esto, la noción de predicción para variables ordinales parece ser considerablemente menos útil que para variables de intervalo, y alguna otra perspectiva puede ser más apropiada." (p.342)

Dicen Loether y McTavish (1974): "Infortunadamente, cuanto más complicado es el denominador, más difícil se hace expresar una clara definición operacional como RPE, y esto sucede con τb." (p.290)

Reynolds (1977, p.85) menciona varios artículos que han tratado el problema del RPE negativo, y repite la crítica de Wilson de que sigue vigente.

$\emptyset$ es derivable algebraicamente de r, es la correlación de dos dicotomías. Por ello, como se indica en el capítulo

anterior, ϕ^2 se puede interpretar operacionalmente como r^2: la proporción de variación de una variable explicada por variación de la otra.

Como se verá más adelante, ϕ también puede interpretarse a través de r como PRE.

Otra interpretación operacional de ϕ como RPE se debe a la mencionada igualdad de ϕ^2 y τGK.

A través de ϕ, τb tiene las dos interpretaciones RPE, pero sólo para tablas de 2x2. En éstas, los valores absolutos de τGK, ϕ^2 y τb^2 son iguales.

Se ha indicado antes que RPE es aplicable a cualquier nivel de medición. Se muestra ahora la interpretación de r, aunque brevemente pues el coeficiente está fuera del marco de este trabajo.

En la situación I se predice un valor de y; mejor estimador es el promedio. En la situación II, conociendo el valor en x, se predice y mediante la ecuación de regresión.

Error en I es la varianza de y; en II es la varianza con respecto a la recta de regresión, o sea, el cuadrado del llamado 'error standard de la estimación'.

Por lo tanto, r^2 puede ser interpretado operacionalmente como reducción proporcional de varianza de y al conocer su relacion con x. También puede indicar cómo mejora predicción de y al conocer x.

La interpretación RPE ha dominado el campo de coeficientes de asociación.

Costner (1965) propone adoptar el criterio de que "las medidas de asociación en investigación social deben ser interpretables operacionalmente en términos de reducción proporcional del error". (p.342) Espera que esto lleve a eliminar coeficientes que no sean interpretables como RPE, sirva para elegir entre medidas alternativas, y ayude a especificar un marco general para desarrollar coeficientes de asociación con fines especiales. (p.342)

RPE es para Agresti y Agresti (1979) "un concepto que unifica la medición de asociación para variables nominales,

ordinales y de intervalo." (p.217)

Reynolds (1984) señala que "la alternativa más popular es la lógica de RPE." (p.49)

Dos enfoques interesantes son los de Leik y Gove, y Kim. Leik y Gove (1971) proponen que todas las interpretaciones RPE, en cualquier nivel de medición, se basen en la fórmula 'de pares de casos'. "A medida que los datos aumentan sus propiedades matemáticas de nominal a ordinal a igualdad de intervalos, las medidas de asociación deben agregar las nuevas propiedades al formato usado en niveles más bajos." (p.279) Los autores diseñan varios coeficientes; algunos son variaciones de medidas ya existentes.

Kim (1984) presenta, basado en la lógica de RPE, medidas de asociación que llama 'de reducción de incertidumbre' (en inglés: PRU) "que provean estadígrafos descriptivos para análisis de tablas de contingencia, de la misma manera que las correlaciones simple, parcial y múltiple lo hacen para el análisis de regresión múltiple." (p.4) Su enfoque excede el marco de este artículo.

CONCLUSIONES.

El modelo de RPE consiste en dos situaciones de predicción, una definición de error, y una fórmula general para todos los coeficientes: (33).

Los tres primeros elementos varían entre niveles de medición y entre coeficientes. La fórmula (33) es común.

Un coeficiente es interpretable de acuerdo a este modelo si puede definir situaciones de predicción y error, y si es expresable como (33).

La interpretación 'de pares de casos' reduce coeficientes a un formato comparable, pero no interpreta fácilmente valores numéricos. RPE los interpreta mediante (33), fórmula que, al ser igual para todo coeficiente, no permite comparación.

Se puede concluir que la superioridad de la interpretación 'de pares de casos' radica en que permite comparar fórmulas de coeficientes, y la de RPE en que permite interpretar sus valores numéricos.

REFERENCIAS

Agresti, A., y Agresti, B.F. (1979) Statistical Methods for the Social Sciences. Dellen.

Bishop, Y.M.M., Fienberg, S.E., y Holland, P.W. (1975) Discrete Multivariate Analysis. MIT Press.

Blalock, H.M. (1972) Social Statistics. 2ª ed. McGraw-Hill.

Boole, G. (1854) An Investigation of the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities. Londres. [Reeditado por Dover, N.York, 1958.]

Boudon, R. (1974) The Logic of Sociological Explanation. Penguin.

Costner, H.L. (1965) "Criteria for Measures of Association." American Sociological Review 30, pp.341-353.

Goodman, L.A., y Kruskal, W.H. (1954) "Measures of Association for Cross Classifications." Journal of the American Statistical Association 49, pp.732-764.

Heron, D. (1911) "The Danger of Certain Formulae Suggested as Substitutes for the Correlation Coefficient." Biometrika 8, pp.109-122.

Jevons, W.S. (1870) "On a General System of Numerically Definite Reasoning." Memoirs of the Manchester Literary and Philosophical Society. [Las citas son de Pure Logic and Other Minor Works, editado por R. Adamson y H.A. Jevons, Lenox Hill, N.York, 1971.]

Kang, T.S. (1972) "Linking Form of Hypothesis to Type of Statistic: An Application of Goodman's Z." American Sociological Review 37, pp.357-365.

Kendall, M. (1975) Rank Correlation Methods. 4ª ed. Griffin.

Kim, J.O. (1984) "PRU Measures of Association for Contingency Table Analysis." Sociological Methods and Research 13, pp.3-44.

Leik, R.K., y Gove, W.R. (1971) "Integrated Approach to Measuring Association." Sociological Methodology 1971, pp.279-301.

Loether, H.J., y McTavish, D.G. (1974) Descriptive Statistics for Sociologists. Allyn and Bacon.

Mill, J.S. (1843) A System of Logic, Ratiocinative and Inductive, Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation. Londres. [Las citas son de la edición de Longmans, Londres, 1884.]

Morgan, A. (1847) Formal Logic, or The Calculus of Inference, Necessary and Probable. Londres.

Pearson, K. (1900) "On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it Can Be Reasonably Supposed to Have Arisen from Random Sampling." Philosophical Magazine, Fifth Series L, pp.157-175.

Pearson, K. (1901) "On the Correlation of Characters not Quantitatively Measurable." Philosophical Transactions of the Royal Society of London, Series A, Vol.195, pp.1-47.

Pearson, K. (1904) "On the Theory of Contingency and its Relation to Association and Normal Correlation" Draper's Company Research Memoirs, Biometric Series I. [Las citas son de Karl Pearson's Early Statistical Papers, Cambridge University Press, 1956; primera edición: 1948.]

Pearson, K., y Heron, D. (1913) "On Theories of Association." Biometrika 9, pp.159-315.

Quetelet, A. (1832) Sur la Possibilite de Mesurer l'Influence des Causes qui Modifient les Elements Sociaux. Bélgica.

Reynolds, H.T. (1977) The Analysis of Cross-Classifications. Free Press.

Reynolds, H.T. (1984) Analysis of Nominal Data. 2ª ed. Sage.

Somers, R.H. (1962) "A New Asymmetric Measure of Association for Ordinal Variables." American Sociological Review 27, pp.799-811.

Somers, R.H. (1968) "On the Measurement of Association." American Sociological Review 33, pp.291-292.

Walker, H.M., y Lev, J. (1953) Statistical Inference. Holt, Rinehart and Winston.

Wilson, T.P. (1969) "A Proportional-Reduction-in-Error Interpretation for Kendall's Tau-b." Social Forces 47, pp.340-342.

Yule, G.U. (1900) "On the Association of Attributes in Statistics." Philosophical Transactions of the Royal Society of London, Series A, Vol.194, pp.257-319.

Yule, G.U. (1911) An Introduction to the Theory of Statistics. Griffin.

Yule, G.U. (1912) "On the Methods of Measuring Association between Two Attributes." Journal of the Royal Statistical Society 75, pp.579-642.

Yule, G.U., y Kendall, M. (1968) An Introduction to the Theory of Statistics. 14ª ed. Hafner. [Una edición más de Yule (1911). Kendall aparece desde la 11ª edición en 1937 y continúa tras la muerte de Yule en 1951.]
