Estadística Descriptiva Bivariada (I: cont.): Otros índices de correlación

Eduardo López
Dpto. MIDE. UCM

Introducción

Dado que en este tema se van a exponer distintos índices de correlación, en esta introducción lo pertinente es proporcionar ideas o elementos claros para elegir con la mayor seguridad posible el adecuado coeficiente de correlación a los datos o información que se proporcione de dos (o más) variables.

Ya se habló de algunos supuestos que subyacen al uso de la correlación de Pearson. Se dijo que la distribución implicada era la normal bivariada. Esto significaba que las puntuaciones de Y_{ij} se distribuían normalmente para cada valor de X_{ij}. Igualmente, los valores de X_{ij} se debían distribuir normalmente para cada valor de Y_{ij}. El incumplimiento de este supuesto tiene poca influencia sobre la validez de la prueba cuando los grados de libertad (tamaño de la muestra menos dos unidades) eran más de 25 ó 30 unidades, o sea, para muestras por encima de 27 a 32 sujetos.

Se decía, además, que la relación entre X_{ij} e Y_{ij} debía ser lineal, porque si entre las dos variables existía una relación curvilínea, la utilización del coeficiente de correlación de Pearson haría inútil la detección de la relación. Además, aspecto a que se aludió pero que de los ejemplos escogidos se deducía, las variables debían referirse a características claramente cuantificables, como la edad, peso, inteligencia, rendimiento, ...

Es consecuencia de lo anterior el que, además del coeficiente de correlación de Pearson, existen otros índices que sirven a idénticos fines de calcular numéricamente la relación entre variables. Más aún, muchos de esos índices son extensión del coeficiente de correlación de Pearson, a los que se llega mediante la aplicación de dicho coeficiente a los datos.

Glass y Stanley (1974, 156/8), antes del estudio de los diversos índices de correlación y después de haber estudiado previamente el de Pearson, se detienen en el análisis de las dos variables que se ponen en relación en los índices bivariados. Analizan cuatro tipos de medidas de variables, la medida nominal dicotómica, la medida nominal dicotómica con una distribución normal subyacente, la medida ordinal y la medida de intervalo o de razón. Después en un cuadro de doble entrada con cuatro categorías cada una (medida nominal dicotómica, medida nominal dicotómica con una distribución normal subyacente, medida ordinal y medida de intervalo/razón) pone en relación la medida de la variable X y la medida de la variable Y, de cuya relación surgen los diversos índices de correlación.

Veamos cada una de ellas y, aunque en los temas del inicio ya se habló de las variables y de las escalas de medida, es pertinente recordarlo en este momento y adaptarlo a las actuales necesidades. De cualquier forma nos permitimos variar la presentación cuando las circunstancias lo exijan para una mayor adecuación a los propósitos del presente tema.

- **Medida nominal dicotómica.** Expresa mera categoría. Ya se dijo que existen algunas *variables*, que, aunque ordinariamente se denominen así, no son propiamente variables porque no admiten valores en un continuo, y que bien pueden llamarse *atributos o características* propiamente **dicotómicas**. La única operación de que es susceptible es la distinción o equivalencia.

Cuando se está hablando de la variable *titularidad de un centro (*variable *A)*, nos estamos refiriendo a si es de titularidad estatal (A_1) o de titularidad no estatal (A_2). Esto significa que lo único que se puede hacer con un determinado centro, el centro 1, es ubicarlo en una de dichas categorías, p. e. en la categoría A_1; el centro 2, p.e. en la categoría A_2 y el centro 3 en la categoría A_1, con lo que las únicas funciones y operaciones posibles son las de distinción/equivalencia. El centro 1 es *equivalente* al centro 3 en cuanto a titularidad, no distinguiéndose en algo más. A su vez el centro 2 es *distinto* en titularidad del centro 1 y del 3.

Dado que se está hablando de relación *bivariada,* se está mencionando dicotomía, pero cuando la relación es entre variables de más de dos categorías, se puede hablar de politomía y extender las categorías de las variables a más de dos. Son muchas las variables con las que se trabaja en educación. Piénsese en variables como sexo, asistencia a preescolar, éxito en selectividad, éxito en una materia, nivel de escolaridad, adscripción a una universidad, nacionalidad, etnia, raza, país de origen, elección de carrera, y muchas más.

- **Medida nominal dicotómica con una distribución normal subyacente**. Es una categoría con orden subyacente. La medida que estos autores llaman así no es otra que una variable de naturaleza continua, que ha sido dicotomizada. Por eso se distingue entre *dicotómica* y *dicotomizada*. Este tipo de variables se verán en los índices de correlación tetracórico y biserial.

Este tipo de variables se pueden extender a la educación, y en cualquier otros campos, hasta donde se desee. No es infrecuente que se hable de alumno de rendimiento bajo, alumno torpe y chico lento, no inteligente, antisocial, no motivado, ... -¿es preciso seguir?-, lo cual no son sino expresiones de unas variables, la primera, llamada *rendimiento,* aplicada a un alumno, previa dicotomización y asignación a una de las categorías, en este caso la baja. Otro tanto ocurre con las variables aprovechamiento en el estudio, inteligencia, sociabilidad, motivación, ... Por eso algunos índices, como el tetracórico, son tan usuales en educación, el cual trata con ambas variables de naturaleza continua pero *dicotomizadas*.

¿Cómo se contabilizan estos valores o categorías de las variables dicotomizadas? Cuando se trabaja con atributos (dicotomía) en una tabla de doble entrada se pone, p.e. masculino/femenino o femenino/masculino como categorias del sexo, por ser dicotómica pura la *variable* y se puede asignar un 0/1 ó 1/0 a las categorías, con la única salvedad de saber a quién se han dado el 0 ó el 1 para poder adecuadamente interprtetar los datos. Contrariamente, cuando se trabaja con variables dicotomizadas, en efecto, se asignan 0 y 1 a las categorías, pero no es arbitrario a quién se asigna el 0 y a quién el 1. Si son variables subyacentemente continuas, p.e. en el caso de la inteligencia, se dará el 1 al inteligente y el 0 a *torpe*. Este es un requisito derivado de la misma naturaleza de la variable y exigido por la interpretación del signo de la correlación.

- **Medida ordinal.** Las puntuaciones de los sujetos según este cómputo sólo llegan a asignar un rango en un grupo, rango que puede ser reflejo de varias situaciones, que se explican en el coeficiente de correlación de Spearman, pero que se resumen aquí.

Puede ocurrir que las posibilidades de expresión en un momento dado de una variable no alcancen más allá del rango; piénsese en el ámbito escolar en calificaciones de orden de los alumnos en variables tales como inteligencia y rendimiento en matemáticas, realizadas por los profesores o jueces, los cuales a veces no tienen información precisa, fruto de una medida. *Y* piénses en instrumentos de medida deficientes, por poco fiables y válidos, en cuyo caso lo pertinente sería convertir las puntuaciones de ambas variables a rangos. *Y*, finalmente, considérense variables de orden afectivo, actitudinal, no cognoscitivo, ... En estos casos la forma de presentación de los datos no puede ser otra que el rango, el cual expresa dos cosas, distinción (en lo que es común con las categorías de los atributos) y seriación individual (a diferencia de la anterior, que era seriación por bloques o grupos).

- **Medida de intervalo o de razón.** Además de distinción y seriación individual, la asignación de numerales a los sujetos, como expresión de variables, suele ser fruto de medida o evidencia. Esta escala, además de distinción (en lo que es común con las categorías de los atributos) y seriación individual (a diferencia de la anterior, que era seriación por bloques o grupos), añade el que entre las puntuaciones de cualquier número real existe una diferencia conocida en la cantidad de la característica analizada. *Y,* aunque no siempre sea así, se considera que las puntuaciones en este tipo de escalas se distribuyen normalmente.

Esta característica es la que ha llevado a muchos críticos a poner en duda que variables tan usuales en educación, como rendimiento, actitud, inteligencia, motivación, voluntad para aprender, ... pueden no ya ser expresadas en escala de razón, cosa que no puede ser, sino ni siquiera en escala de intervalo. A lo más que se puede llegar, dicen, es a una escala de orden. En este caso, de ser así, afirman que los índices estadísticos más apropiados son los que expresan orden y distinción, y nada más. Esto es lo que dará lugar a las pruebas no paramétricas. Por ello, proponen pasar habitualmente de puntuaciones en intervalo a orden, porque toda puntuación expresada en escala de razón o de intervalo se puede a su vez expresar en las escalas de orden inferior, rango o nombre (nominal).

Con estos cuatro tipos de medidas en las variables *X* e *Y* los autores mencionados construyen esta tabla en cuyas casillas insertan los diversos índices de correlación, aunque tal vez fuera perfectible o su contribución no es significativa.

	Medida de la variable X			
Medida de la variable Y	Dicotómica (pura)	Dicotómica, con distribución normal subyacente-**dicotomizada**	Ordinal	De razón o de intervalo
Dicotómica (pura)	Coeficiente Phi-ϕ (extensión de Pearson)	(No se conoce)	(No desarrollado)	(No desarrollado)
Dicotómica, con distribución normal subyacente-dicotomizada	(No desarrollado)	Coeficiente de correlación tetracórico: r_t (extensión de Pearson)	(No desarrollado)	(No desarrollado)
Ordinal	(No desarrollado)	(No desarrollado)	Coeficiente de correlación de Spearman: r_s	(No desarrollado)
De razón o de intervalo	Coeficiente de correlación biserial puntual: r_{bp}	Coeficiente de correlación biserial : r_b	Coeficiente de Spearman: r_s previa conversión de la variable Y ó X a orden	Coeficiente de correlación de Pearson: r_{xy}

De cualquier forma, en este tema se inserta en un *Anexo* una tabla en la que se presentan las condiciones o requisitos para el uso de los distintos coeficientes de relación (correlación y asociación) entre dos o más de dos conjuntos de datos.

I. Coeficientes de correlación

1. Coeficiente de correlación por rangos ordenados de Spearman-r_s

"Francis Galton fue el primero en utilizar esta técnica bastante obvia y Charles Spearman, el psicólogo británico, amplió su uso y fue recompensado por ello, ya que el coeficiente lleva su nombre, aunque parece ser que Karl Pearson opinaba que el mérito se le debía a Francis Galton" (Glass y Stanley, 1974, 172).

Este coeficiente se adapta a dos variables de naturaleza continua pero que vienen expresadas en escala de orden o rangos. Los rangos son habituales en el campo de la educación. En muchas ocasiones no es posible en un determinado momento medir a los alumnos en alguna variable y lo más que se puede conseguir es una seriación de los alumnos; en otras no es ni posible en un determinado ni en momentos siguientes, es decir, hay serios reparos en poder medir a los alumnos. Variables o situaciones de este tipo son frecuentes: Se recurre a los puestos de una clase, el orden en las calificaciones de un curso cuando no se tiene o dispone de información precisa, la ordenación en algún rasgo de personalidad, en alguna variable actitudinal, ... Sin embargo, esto no quiere decir que las variables vengan directa y necesariamente en rangos en una situación o problema. Se puede llegar al coeficiente de correlación de Spearman a través de las siguientes situaciones:

1) La situación obvia y más directa es la mencionada, es decir, aquélla en la que las dos variables, de naturaleza continua, vengan expresadas en escala de orden. Si esto ocurre así no hay cuestión y se ha de recurrir a este coeficiente. Puede pensarse en el ámbito escolar en calificaciones de orden de los alumnos en dos variables, p.e. en inteligencia y rendimiento en matemáticas, realizadas por los profesores o jueces, los cuales no suelen tener información precisa, que sea fruto de una medida precisa. Son muchos otros los casos que se pueden dar.

2) Otra situación, tal vez más frecuente, es la dos variables de naturaleza continua, una de las cuales es fruto de una medida precisa, y por tanto expresada en escala de intervalo, y la otra puede ser fruto de una ordenación por parte de alguien que asigna rangos a los mismos sujetos medidos en la otra variable. Tal vez esta situación sea más frecuente. Lo que se precisa en estos casos es convertir a rangos la variable que viene en escala de intervalo, con lo cual ambas ya están en formato de rangos.

3) Finalmente, dos variables de naturaleza continua y en escala de intervalo, para las cuales aparentemente el coeficiente natural de cálculo sería el coeficiente de correlación de Pearson, en algunos casos pueden demandar su conversión a rangos. ¿Cuáles son estas situaciones? Se me ocurren, al menos, dos: Cuando los instrumentos de medida son deficientes en diversos aspectos, p.e. son poco fiables y poco válidos, en cuyo caso lo pertinente sería convertir las puntuaciones de ambas variables a rangos; y cuando la naturaleza de las variables sea de orden afectivo, actitudinal, no cognoscitivo o físico. En ocasiones no es posible puntuar a un sujeto más allá del mero rango en alguna variable, p.e. ganas o voluntad de aprender, actitud hacia una materia, compañerismo, afición hacia un determinado campo, ...

Por tanto, como afirman Glass y Stanley (1974, 172) en forma de resumen, se puede decir que el coeficiente de Spearman se usa "cuando no se dispone, no se necesitan, o no son convenientes medidas más refinadas", en cuyo caso se pueden correlacionar las puntuaciones de rangos de unos mismos sujetos en dos variables.

La fórmula más conocida es la siguiente:

$$\mathbf{r_s = 1 - 6 \Sigma d^2_i /(N^3 - N)}$$

Pongamos un *ejemplo*. Dos profesores, uno de matemáticas y otro de lengua, califican a 8 mismos alumnos por rangos en aprovechamiento escolar.

Estos son los resultados:

Alumnos	**Rango en matemáticas : X**	**Rango en lengua: Y**	**d_i**	**d^2_i**	**X . Y**	**$r_s = 1 - 6 \Sigma d^2_i /(N^3 - N) = 1 - \{ 6 (28)/ (8^3 - 8)\} = 0.67$**
Ana	1	3	-2	4	3	
Carlos	2	1	+1	1	2	
Bartolomé	3	2	+1	1	6	
Horacio	4	5	-1	1	20	
Estela	5	7	-2	4	35	
Flavia	6	8	-2	4	48	
Donato	7	4	+3	9	28	
Gisela	8	6	+2	4	48	
				$^2_i = 28$	**Σ X . Y= 190**	

Glass y Stanley (1974, 174) proponen una fórmula alternativa a la anterior de cálculo más rápido y fácil, sobre todo cuando las diferencias de rangos y las potencias pueden generar perturbaciones:

$$r_s = 3/(n - 1) \{4\Sigma XY/n(n + 1) - (n + 1)\} =$$
$$3/(8 - 1) \{4 (190) / 8(8 + 1) - (8 + 1)\} = 0.67.$$

Interpretación. La *interpretación* del coeficiente de correlación de rangos de Spearman, por ser una extensión del coeficiente de correlación producto-momento de Pearson, no tiene otra interpretación que la de cualquier otro coeficiente de correlación, es decir, varía de 0 a 1. Solamente alcanzará el valor 1 si todos los individuos conservan en las dos variables idéntica posición, puesto que en la fórmula el valor de d^2, al ser 0, toda la fórmula queda como: 1 - 0 = 1.

Posiciones iguales en rangos. Puede surgir un problema cuando haya posiciones o *rangos iguales*. En efecto, la igualdad en medición es cosa corriente. Cuando se aplica una prueba de rendimiento o personalidad a los alumnos es habitual que los alumnos obtengan puntuaciones iguales; cuando un profesor ordena los alumnos en alguna variable es relativamente frecuente que asigne rangos iguales a algunos de ellos. ¿Qué rangos asignar a estos alumnos con rangos iguales? La respuesta no puede ser otra que asignar a todos ellos el promedio de los rangos que ocupen o se les habrían asignado si no hubiesen ocurrido coincidencias, de tal modo que siempre se cumpla que el rango final asignado coincida con el número total de unidades o alumnos del grupo.

¿Qué importancia o repercusión tiene este problema en los resultados del coeficiente de correlación y en el análisis de dicho coeficiente? Siegel (1976, 238) afirma que "si la proporción de ligas (coincidencias) no es grande, su efecto sobre r_s es insignificante y la fórmula aún puede usarse para el cálculo. Pero, si es grande, entonces hay que incorporar un factor de corrección".

En estos casos la fórmula es ésta:

$$\mathbf{r_s = (\Sigma x^2 + \Sigma y^2 - \Sigma d_i^2) \,/\, 2 \sqrt{\Sigma x^2 \,.\, \Sigma y^2}}$$

donde: Σx^2 : $(N^3 - N)/12 - \Sigma T_x$
Σy^2 : $(N^3 - N)/12 - \Sigma T_y$

y donde: T_x : $(t^3 - t\,)/\,12$, siendo t el número de observaciones coincidentes en un rango dado en X. En el ejemplo, existen dos coincidencias de 2 sujetos cada una. Por tanto: $\Sigma T_x = (2^3 - 2\,)/\,12 + (2^3 - 2\,)/\,12 = 1$;
T_y : $(t^3 - t\,)/\,12$, siendo t el número de observaciones coincidentes en un rango dado en Y. En el ejemplo existen dos coincidencias, la primera de 3 y la segunda de 2. Por tanto: $\Sigma T_y = (3^3 - 3\,)/\,12 + (2^3 - 2\,)/\,12 = 2.5$.

Supongamos el ejemplo anterior en el que algunos sujetos son calificados del mismo modo por sus respctivos profesores en las materias:

Alumnos	*Rango en mats. X*	*Rango en lengua: Y*	d_i	d^2_i	$\mathbf{r_s = (\Sigma x^2 + \Sigma y^2 - \Sigma d_i^2) / 2 \sqrt{\Sigma x^2 . \Sigma y^2} = (41 + 3}$ **/** $\mathbf{2\sqrt{41(39.5)} = 0.814}$
Ana	1	2	-1	1	
Carlos	2.5	2	+0.5	0.2	Correción en *X*: $\Sigma x^2 = (N^3 - N)/12 - \Sigma T_x =$
Bartolom	2.5	2	+0.5	5	$(8^3 - 8)/12 - \{(2^3 - 2\,)/\,12 + (2^3 - 2\,)/\,12\} =$
é	4	4	0	0.2	$42 - \{1\} = 41$.
Horacio	5	7.5	-2.5	5	
Estela	6.5	7.5	-1	0	Corrección en *Y*: $\Sigma y^2 = (N^3 - N)/12 - \Sigma T_y = (8^3 - 8)$
Flavia	6.5	5	+1.5	6.2	$- 3\,)/\,12 + (2^3 - 2\,)/\,12\} = 42 - 2.5 = 39.5$
Donato	8	6	+2	5	
Gisela				1	
				2.2	
				5	
				4	
			$\mathbf{\Sigma d_i^2 = 15}$		

Por tanto, la correlación por rangos de Spearman, al existir coincidencias en los rangos, es:

$$\mathbf{r_s = (\Sigma x^2 + \Sigma y^2 - \Sigma d_i^2) \,/\, 2 \sqrt{\Sigma x^2 \,.\, \Sigma y^2} = (41 + 39.5 - 15) \,/\, 2\sqrt{41(39.5)} = 0.814.}$$

Esto significa que la correlación se infraestima cuando no se tienen en cuenta las coincidencias en los rangos.

Resumen del procedimiento. Pueden indicarse los pasos siguientes en el cálculo del coeficiente de correlación por rangos de Spearman (Cfr. Siegel, 1976, 244/5):

1) Si los datos fueran dados en puntuaciones directas, es preciso ordenar las puntuaciones en las respectiva variables, *X e Y.*
2) Se asignan rangos a las observaciones o puntuaciones en la variable *X* de 1 a *N*, de tal modo que los rangos coincidan con el tamaño de N. Si hubiera puntuaciones coincidentes, es preciso asignar a todos ellos el promedio de los rangos que ocuparían o se les habrían asignado si no hubiesen ocurrido coincidencias, de tal modo que siempre se cumpla que el rango final asignado coincida con el número total de unidades o alumnos del grupo. Otro tanto es preciso hacer con la variable Y.
3) Una vez que cada sujeto en las variables *X* e *Y* tiene asignado su rango, se calcula la diferencia -d_i - entre los pares de rangos asignados a cada sujeto.
4) A continuación se eleva al cuadrado el valor d_i -d_i^2- para cada pareja de puntuaciones y a continuación se suman $\Sigma \mathbf{d_i}^2$.
5) Si es grande la proporción de coincidencias en los rangos, se calcula el valor de correlación recurriendo a una corrección, tal como se ha indicado.
6) Finalmente, el valor obtenido se interpreta como un valor de correlación de Pearson.

2. Coeficiente de correlación tetracórica-r_t

¿De qué naturaleza han de ser las variables y en qué escala han de venir para que sea pertinente aplicar el coeficiente de correlación tetracórico? Este es un índice que se adecua a variables -ambas- de naturaleza continua y que vienen dicotomizadas. Es un índice no inusual aunque, como se apunta más abajo, ha dejado de usarse tan frecuentemente debido al uso de la informática en el análisis de datos. Sin embargo, al menos en el campo de la educación es muy pertinente si se mira por el lenguaje que se usa.

En efecto, no son infrecuentes expresiones tales como "este alumno es de inteligencia baja"; "es un chico extrovertido"; "Fulanito es un chico de rendimiento alto", "Zutanito es de inteligencia alta mientras que Fulanito es de baja"; ... *Y* muchas otras expresiones que evidencian que variables de naturaleza continua han sido divididas en dos, han sido dicotomizadas.

Para seguir con un *ejemplo* a efectos de cálculo, veamos un caso, que es frecuente en el caso de la medida en educación. Cuando se analiza una prueba objetiva de rendimiento en orden a mejorarla, se analizan los items o preguntas de la prueba en orden a comprobar si realmente esos items son homogéneos. Un item será homogéneo si mide lo mismo que la prueba. La forma métrica o estadística de expresar la homogeneidad entre el item y la prueba consiste en calcular la correlación entre ambos elementos, item y prueba.

Supóngase que cada pregunta permite dar respuestas que van desde 0 a 5 puntos, sea porque se permiten aproximaciones a la correcta o por cualquier otra razón. *Y* por otra parte, cada alumno tiene una puntuación en la prueba total, exceptuada de esa puntuación el peso que pueda tener el item, que se quiere analizar. Una forma de abordar la relación entre el item y la prueba puede consistir en correlacionar la puntuación de cada sujeto en la prueba con la del item, siendo éstas unas puntuaciones que van de 0 a 5, es decir, se puede tratar como una variable continua, que por tener varios niveles de respuesta se puede expresar en escala de intervalo.

Para el ejemplo anterior el coeficiente de correlación tetracórico es pertinente. Dicen Glass y Stanley (1974, 165) que este coeficiente era muy usado antes de la existencia de ordenadores, pues facilitaba los cálculos y no era muy inexacto cuando el tamaño de las muestras era grande. En efecto, en la actualidad ha caído en desuso, lo cual no quiere decir que no tenga su interés. Supongamos el ejemplo anterior en el que se aplica a un grupo de 100 alumnos una prueba y se quiere analizar si un item, p.e. el 23, mide lo mismo que lo que mide la prueba, esto es, si es fiable. La fiabilidad del item se expresará a partir de la correlación que éste arroje con la prueba. La lógica pide que, si es fiable, los alumnos que puntúen alto en la prueba es lógico que también puntúen alto en el item.

Los datos que arrojan son los siguientes: De los 100 alumnos que son examinados, 31 responden bien (1) en la prueba y de ellos son 25 los que responden también correctamente (1) en el item 23; por el contrario, de los 69 que responden mal en la prueba (0) son nada menos que 64 los que responden incorrectamente (0) en el item 23. Estos y el resto de los datos están en la tabla.

Una observación es preciso hacer en la disposición de los datos. Así como en otros casos de tablas 2 *X* 2 no se ha recomendado una especial disposición de los datos, el requisito del índice de que sean variables de naturaleza continua las que sean objeto de análisis, aunque se hayan dicotomizado, implica que se deben disponer conforme al modelo de un eje de coordenadas con un punto cero de encuentro de los ejes cerca del cual se encuentran los valores bajos (en este caso los valores o frecuencias 0) y más lejos los valores superiores (valores 1). Esta es la razón por la que los valores de la variable *X* discurren de 0 a 1 hacia la derecha y los de la variable *Y* suban desde 0 a 1.

Item 23 (Y)	**Prueba objetiva (X) -menos item 23**		***Sub-totales***
	Mal: 0	**Bien: 1**	
Bien: 1	**5**	**25**	30
Mal: 0	**64**	**6**	70
Sub-totales	**69**	**31**	100

Son varias las formas de resolución del problema. Se puede resolver recurriendo a una ecuación de segundo grado, mediante una función trigonométrica (Cfr. Pérez Juste, 1985, 178/180) y mediante el uso de unas tablas de conversión de un cociente a valor de r_t. Este último procedimiento de uso de una *tabla de conversión de un cociente de productos de frecuencias a coeficiente de correlación,* que por otra parte se recomienda, es el que se va a abordar ahora, para el que se precisa seguir unos pocos pasos y atender una recomendación para su interpretación:

1) Cálculo del *cociente ad/bc* = AD/BC = 5 (6) / 25 (64) = 30/1600 = 0.0187.

Lo primero que es preciso advertir es que la tabla está calculada para valores del cociente no menores de 1. Por tanto, *si ad/bc es menor de 1, como es el caso, es preciso invertir la ubicación de los productos para que bc/ad sea mayor que 1 y entonces se lee directamente en la tabla el valor de r_t . Si ad/bc es mayor que 1, se busca el coeficiente de correlación bajo los valores de ad/bc y el valor de r_t resulta negativo*. De estas palabras se deduce, primero, que el coeficiente de correlación tetracórico presente es de signo positivo y se interpreta como reflejo de la existencia de relación directamente proporcional entre puntuación en el test y puntuación en el item. Dicho de otro modo significa que los alumnos que puntúan alto en la prueba (valor 1) vienen a ser los que puntúan correctamente en el item (valor 1), y al revés.

2) Por tanto, este segundo paso puede no ser necesario si es que en el primero el cociente AD/BC es igual o superior a uno. En segundo lugar, por tanto, para que este *cociente ad/bc* sea interpretable es preciso hacer que sea igual o superior a uno y, por tanto, es preciso convertirlo en BC/AD: *cociente ad/bc* = BC/AD = 25 (64) /5 (6) = 1600/30 = 53.33.

3) Finalmente, se ha de acudir a la tabla de conversión de los cocientes *cociente ad/bc* a valores de r_t . Para ello, lo primero que se ha de hacer es identificar en qué intervalo de valores de entre las tres columnas de BC/AD se encuentra el valor del *cociente ad/bc*, es decir, el valor en este caso de 53.33. Consultando la tabla se tiene que la correlación tetracórica buscada es 0.93:

Intervalo del cociente ad/bc en la columna de BC/AD de la tabla	**Valor correspondiente de r_t al del intervalo en el que está el valor del cociente**
49.84 - 58.79	**0.93**

Precauciones en el uso de r_t. No se insiste con suficiencia en una condición que hace de dicha correlación que sea un buen estimador de la correlación o no. En efecto, r_t será una estimación adecuada de r_{xy} cuanto más cerca de la mediana se encuentre el punto de dicotomización de las variables y, por tanto, cuanto los valores de *p* y *q* sean menos extremos y más próximos a 0.50. Dicho de otro modo, *este método no debe emplearse cuando las probabilidades de p_y y de p_x se alejen considerablemente de 0.50.* En la tabla se tienen las probabilidades de ambas variables en las categorías dicotomizadas que, como se ve, son desiguales: 0.30 vs. 0.70 y 0.31 vs. 0.69.

Item 23 (Y)	**Prueba objetiva (X)** –**menos item 23**		**Sub-totales**	
	Mal: 0	**Bien: 1**		
Bien: 1	5-**A**	25-**B**	30	$p_y = (a + b)/n = 30/100 = 0.30$
Mal: 0	64-**C**	6-**D**	70	$q_y = 0.70$
Sub-totales	69	31	100	
	$q_x = 69/100 = 0.69$	$p_x = (b + d)/n = 31/100 = 0.31$		

Si las proporciones se desvían de 0.50 los valores de la tabla son sobreestimaciones del valor verdadero de la correlación. Glass y Stanley (1974, 167) dan la regla de que si los valores de p_y *y de* p_x son mayores que 0.70 o menores que 0.30 no debería emplearse tal índice de correlación. En nuestro caso, por tanto, es dudoso que sea una buena opción el uso de este índice de correlación porque puede sobreestimar el verdadero valor de la correlación.

Ordinariamente, al usar este índice lo que se viene haciendo es hacer que el corte o dicotomización de las variables se haga, por ser continuas, en la mediana, la cual tiene la propiedad de dividir a la muestra de *p* y *q* en dos partes iguales y por tanto *p* y *q* serán igual a 0.50. Se podrían disponer los datos, aunque no con tanta perfección, conforme a la tabla cambiando el nombre de las variables:

Inteligencia (Y) –**corte en la mediana**	**Rendimiento (X)** –**corte en la mediana**		**Sub-totales**
	Bajo: 0	**Alto: 1**	
Alta: 1	5 **A**	45 **B**	50
Baja: 0	45 **C**	5 **D**	50
Sub-totales	50	50	100

Se supone también, finalmente, que las distribuciones subyacentes sean normales.

3. Coeficiente de correlación biserial puntual-r_{bp}

Este índice de correlación es una extensión del coeficiente de correlación producto-momento, cuya fórmula y nombre se deben a Pearson. A veces se utiliza la expresión correlación *producto-momento biserial.* Es apropiada para dos variables, unas de las cuales es de naturaleza continua y se presenta en escala de intervalo y la otra es discreta y se presenta en formato dicotómico y escala, naturalmente, nominal. "El término *biserial* se refiere al hecho de que existen *dos series* de observaciones en *X*: las puntuaciones en *Y* de **cero** o **uno**" (Glass y Stanley, 1974, 163).

Es bastante usual en educación. Permite poner en relación variables tales como sexo, asistencia a preescolar, dos niveles de escolaridad, superación de una asignatura, éxito en selectividad, permanencia en un curso, ... que se presentan en formato dicotómico, con cualquiera de las variables más recurrentes en educación, como el rendimiento, la inteligencia, rasgos de personalidad, interés por el estudio, motivación al aprendizaje, ...

Un campo de especial utilización es el de la medida en educación. En efecto, cuando se quiere analizar un item de una prueba, aquél suele venir en formato dicotómico, es decir, acierto o error; las puntuaciones en la prueba es lógico pensar que vengan en formato de escala de intervalo proveniente de una variable de naturaleza continua. Correlacionando el item con la prueba se comprueba el grado de homogeneidad del item, de tal modo que si la correlación es alta, el item es homogéneo con la prueba, de lo contrario, no mide lo mismo que la prueba. Asimismo, utilizando el mismo formato pero variando la prueba, en este caso siendo distinta pero relacionada con la prueba primera, se viene utilizando el coeficiente de correlación biserial puntual como índice de validez de un item. En efecto, si un item correlaciona alto con la puntuación en la prueba relacionada que sirve para validarlo, entonces el item es válido, de lo contrario, no.

Existen varias fórmulas para el cálculo del coeficiente de correlación biserial puntual, pero todas son variaciones de la producto-momento de Pearson.

En efecto,

$$\mathbf{r_{bp} = \{ (media\ de\ X_{.1} - media\ de\ X_{.0})/s_x\} \sqrt{n_1 n_0/n(n-1)}}$$

donde (se indican notaciones de otras fórmulas):

media de $X_{.1}$: Es la *media* en *X* de las puntuaciones 1 en *Y*;
media de $X_{.0}$: es la *media* en *X* de las puntuaciones 0 en *Y*;
s_x: Es la desviación típica de las *n* puntuaciones en *X*;

n_1: Es el número de puntuaciones con valor 1 en *Y*;
n_0: Es el número de puntuaciones con valor 0 en *Y*;
n:es igual a $n_1 + n_0$
media de $X_.$: *Media* total de *X*.

Veamos un ejemplo (página siguiente). Los datos se pueden interpretar afirmando que la relación entre permanencia en un curso y motivación hacia el aprendizaje es moderada, en principio. En promedio los que permanecen en el curso (los 1) son los que obtienen altas puntuaciones en motivación hacia el aprendizaje (64.25), mientras que los que abandonan (los 0) obtienen una puntuación promedio sensiblemente menor (61.14).

Alumno	Permanencia en el curso: 1-permanencia; 0-abandono	Motivación hacia el aprendizaje (escala de estimación)	Cálculos de r_{bp}
Ana	1	59	1. $r_{bp} = \{(\textit{media} \text{ de } X_{.1} - \textit{media} \text{ de } X_{.0})/s_x\} \sqrt{n_1 n_0 / n(n-1)} = \{(64.25 - 61.14)/3.91\} \sqrt{8(7)/15(15-1)} = 0.41$ 2. $r_{bp} = \{(\textit{media} \text{ de } X_{.1} - \textit{media} \text{ de } X_{.})/s_x\} \sqrt{n_1 n / n_0(n-1)} = \{(64.25 - 62.80)/3.91\} \sqrt{8(15)/7(15-1)} = 0.41$ 3. $r_{bp} = \{(\textit{media} \text{ de } X_{.} - \textit{media} \text{ de } X_{.0})/s_x\} \sqrt{n_0 n / n_1(n-1)} = \{(62.80 - 61.14)/3.91\} \sqrt{7(15)/8(15-1)} = 0.41$
Beatriz	0	67	
Carla	1	63	
David	1	65	
Eduardo	0	55	
Fernanda	1	72	
Gerardo	0	62	
Honorio	0	60	
Ignacio	1	64	
Javier	1	66	
Kika	1	63	
Luisa	0	61	
María	1	62	
Nicolás	0	63	
Ovidio	0	60	
$n_1 = 8$; $n_1 = 7$ $n = 15$	*media* de $X_{.1}$= 64.25 *media* de $X_{.0}$ = 61.14	*media* de $X_{.}$= 62.80 s_x= *3.91*	

Es como si los datos estuvieran dispuestos así:

Permanencia (1)	X_1	(0) No permanencia	X_0
Ana	59	Beatriz	67
Carla	63	Eduardo	55
David	65	Gerardo	62
Fernanda	72	Honorio	60
Ignacio	64	Luisa	61
Javier	66	Nicolás	63
Kika	63	Ovidio	60
María	62		

En el fondo, este índice es una medida de la diferencia entre las puntuaciones promedio en motivación al aprendizaje-*X* de los que permanecieron (obtuvieron 1) y los que no permanecieron (obtuvieron 0). Cuando se hable de la prueba *t*, que es una prueba de diferencia de medias entre dos muestras, el estadístico *t* se puede convertir en índice de correlación biserial puntual (Cfr. Welkowitz et al., 1981, 244/6).

Como acaba de verse, o se deduce de los comentarios sobre la derivación de este índice del coeficiente de correlación producto-momento de Pearson, este coeficiente tiene un recorrido de 0 a 1, admitiendo ambos signos.

4. Coeficiente de correlación biserial-r_b

Ál comenzar véanse dos ejemplos adecuados a la correlación de que se habla. En efecto, Glass y Stanley (1974, 168) citan el *ejemplo* del tiempo que los estudiantes tardan en equilibrar reacciones químicas (variable *X*) y la habilidad para equilibrarlas (variable *Y*). Esta última variable, aunque puede ser medida a través de una prueba de rendimiento de varios items con resultado supuestamente normal, va a ser medida con un solo item, en cuyo caso la habilidad para equilibrar reacciones químicas resultará de *cero* y *uno*, aunque bajo esas calificaciones subyace rendimiento distribuído normalmente.

Por su parte, Welkowitz *et al.* (1981, 247) citan el ejemplo de la relación entre la aptitud para el trabajo y el éxito en el trabajo (*Y*). Lo que *s*e hace es aplicar un test de habilidad (*X*) a los solicitantes de empleo a la entrada; son admitidos todos y, pasado medio año, el responsable, que ha visto trabajar a los admitidos, los califica como *competentes* e *incompotentes*. Esta realmente no es una verdadera dicotomía, puesto que los empleados despliegan su competencia a lo largo de un continuo que, además, se distribuye normalmente.

Aplicación. ¿En qué situaciones se aplica la correlación biserial? Afirman Glass y Stanley (1974, 168) que el

> "coeficiente de correlación biserial es una estimación de la correlación producto-momento entre *X* y las puntuaciones normalmente distribuídasY, que se supone, subyacen a la dicotomía de "ceros" y "unos". Esta situación es similar a la que se presentó cuando se hubo de recurir a r_t , exceptuando que entonces *ambas* variables eran dicotómicas con distribuciones normales subyacentes".

Por tanto, las dos variables de las que trata la correlación biserial son variables de naturaleza cuantitativa continua, una de las cuales es presentada en escala de intervalo y la otra ha sido *dicotomizada,* y debajo de las cuales subyace una distribución normal.

Cálculo a través de un ejemplo. Veamos el ejemplo de Glass y Stanley (1974, 169), que se ilustra gráficamente. Como se dijo anteriormente, se trata de poner en relación el tiempo dedicado al estudio del equilibrio de reacciones químicas (*X*) y la puntuación en un item de una prueba que mide habilidad para equilibrar las reaciones químicas (*Y*). Esta puntuación se presenta como 0 (no acierto) y 1 (acierto), aunque en ella subyace una continuidad.

Son varias las fórmulas para el cálculo de la correlación biserial, que son coincidentes:

- r_b = (*Media* de 1 - *media* total)/s_{tx} (p_1/y)
- r_b = (*Media* de 1 - *media* de 0)/s_t {(p_1 p_0/y)}
- r_b = (*Media* de 1 - *media* de 0)/s_x {n_1 (n_0)/u *n* $\sqrt{}$(n^2 -*n*)}

donde:

p_1 : probabilidad equivalente a 1 en el total: n_1 / *n*

p_0 : probabilidad equivalente a 0 en el total: n_0 / *n*

n : tamaño total de la muestra;

s_{tx} : desviación típica en *X* de todos los sujetos de la muestra;

Media de 1: *media* de las puntuaciones en *X* de los sujetos que obtuvieron 1 en *Y*;

media total : *media* de las puntuaciones en *X* de todos los sujetos,

media de 0 : *media* de las puntuaciones en *X* de los sujetos que obtuvieron 0 en *Y*;

n_1 : tamaño de la muestra que obtuvo 0 en la variable dicotomizada;

Los valores de *y* ó *u* precisan una mayor aclaración.

Para poner en relación las dos variables, dado que la variable cuantitativa continua multicategorial *X* debe distribuirse normalmente, es preciso situar a los sujetos del grupo 1 en la variable *Y* en la parte superior de las puntuaciones en *X* y los del grupo 0 de la variable *Y* en la pñarte inferior. Dicha ubicación ha de hacerse en términos de proporciones bajo la curva normal -proporción de sujetos 1 y proporción de sujetos 0-, para identificar en la curva la altura - *y* - que divide los grupos 1 y 0.

Si los grupos 1 y 0 fueran iguales en tamaño, la proporción de sujetos 1 y 0 sería de 0.50, la ubicación sería el centro de la curva y la puntuación típica de *X* sería 0, con lo cual en la curva normal el valor de la altura de la ordenada *y* sería la que indica la tabla de áreas y ordenadas de la curva normal, es decir, 0.3989. Esto es así porque el área de la parte mayor y la menor es de 0.50.

Es decir, calculada la proporción de 1 y 0 en la variable Y, se escoge la que es proporción menor, porque el valor menor es el que suele encontrarse en los valores de la tabla de la curva normal. Identificada la proporción menor de 1 ó 0, se busca este valor en la tabla y en la misma línea horizontal se busca en la columna de valores de *y* el que le corresponde.

Esto se entenderá mejor con el ejemplo de la tabla y gráfico. Como en el problema la proporción de 1 es: 11/18 = 0.61 y la de 0 es: 1 - 0.61 = 0.39 = 7/18, representadas estas proporciones y consideradas como áreas bajo la curva normal, a un área de 0.39 le corresponde un valor de *y* de 0.3836. Este es el valor de *y ó u.*

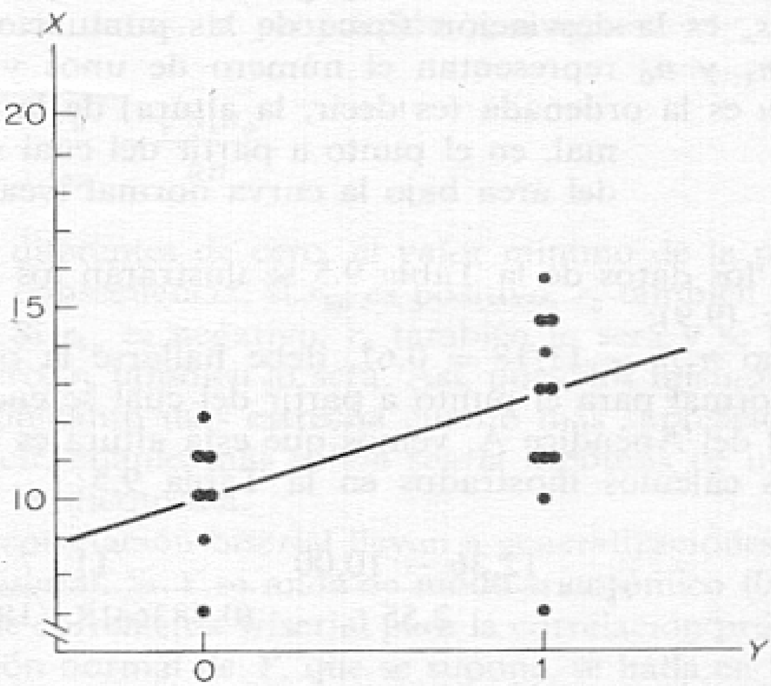
Con este valor calculado, se puede calcular el valor de r_b según las tres fórmulas anteriores:

- r_b = (*Media* de 1 - *media* total)/s_{tx} (p_1/y) = (12.36 - 11.44)/2.55 (0.61/0.3836) = 0.60;
- r_b = (*Media* de 1 - *media* de 0)/s_t {p_1 p_0/y} = (12.36 - 10.00)/2.55 {0.61(0.39)/0.3836} = 0.60;
- r_b = (*Media* de 1 - *media* de 0)/s_x {n_1 (n_0)/u *n* $\sqrt{(n^2 - n)}$} = (12.36 - 10.00)/2.55 {11 (7)/0.3836 (18) $\sqrt{(18^2 - 18)}$} = 0.60.

ILUSTRACION DE LOS CALCULOS DEL COEFICIENTE DE CORRELACION BISERIAL

Tiempo gastado en estudio X	*Puntuación en el ítem de la prueba* Y
16	1
12	0
11	0
7	1
15	1
14	1
10	0
11	0
15	1
9	0
13	1
7	0
13	1
11	1
10	0
11	1
10	1
11	1

Diagrama de dispersión



Resumen de los estadígrafos

$n_1 = 11, \quad n_0 = 7, \quad n = 18$

$\bar{X}_{.1} = 12{,}36, \quad \bar{X}_{.0} = 10{,}00, \quad \bar{X}_{.} = 11{,}44$

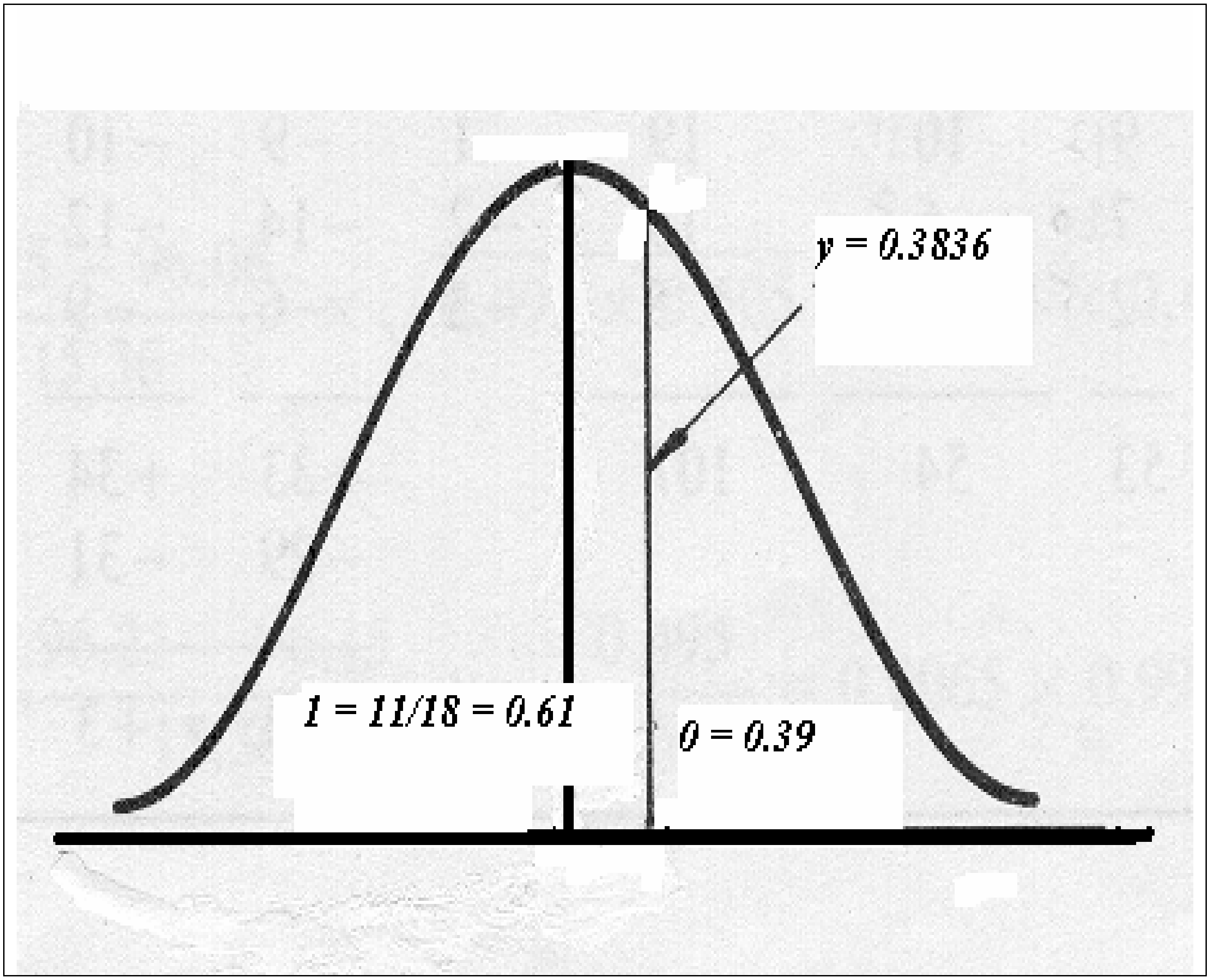
$s_x = 2{,}55, \quad u = 0{,}3836$

$$r_{bis} = \frac{\bar{X}_{.1} - \bar{X}_{.0}}{s_x} \cdot \frac{n_1 n_0}{un\sqrt{n^2 - n}}$$

$$= \frac{12{,}36 - 10{,}00}{2{,}55} \cdot \frac{11 \cdot 7}{(0{,}3836)18\sqrt{18^2 - 18}} = 0{,}60$$

Supuestos e interpretación. Este índice de correlación, a diferencia de otros, arroja en ocasiones unos resultados un tanto desconcertantes al dar resultados que van más allá de ± 1.00, lo cual repercute en la interpretación, situación ésta que confunde cuando sucede, sobre todo si no se estaba advertido. Este problema está relacionado con el incumplimiento de ciertos supuestos exigidos para su utilización, en especial el de la normalidad de la variable continua en escala de intervalo.

Cuando se obtienen valores extremos de r_b puede significar que las puntuaciones en la variable continua en escala de intervalo o multicategorizada (*X*) no se distribuían normalmente o que, dado que la muestra es pequeña, existen fluctuaciones de muestreo produciendo una distribución claramente platicúrtica. De hecho, para una *n* menor en la muestra de 15 la correlación biserial puede tomar valores superiores a +1.25. Por otra parte, del mismo modo que los límites de r_b pueden ser mayores de ± 1.00, también pueden ser menores, siendo el caso de una distribución de *X* leptocúrtica. Así, pues, se ha de tener cuidado en el cumplimiento de los supuestos, en especial el que la variable continua multicategorial o en escala de intervalo sea claramente continua y el tamaño de las muestras no sea pequeño.



II. Coeficientes de asociación

1. Coeficiente cuádruplo ó Phi-ϕ

Con frecuencia se dice (Cfr. Glass y Stanley, 1974, 158) que "tanto *X* como *Y* se han dicotomizado". Es una confusión en la que a veces no se cae en la cuenta. Propiamente, tal como se incluye en la tabla, en vez de decir se la ha *dicotomizado* debe decirse **es dicotómica**. Welkowitz *et al.* (1981, 247) lo dicen claramente: "se desea determinar la relación entre dos variables y *ambas* son genuinamente dicotómicas". En efecto, si nos fijamos en las *variables* asistencia a preescolar, éxito en selectividad, éxito en una materia, y muchas otras que no se mencionan en este momento, no son variables dicotomizadas sino dicotómicas. Es decir, el investigador no ha *dicotomizado* las variables; ellas *son* dicotómicas. En *asistir a preescolar* no cabe sino asistir/no asistir; en *éxito en selectividad* no cabe sino aprobar/suspender (nótese que no de dice *nota en selectividad*); en *éxito en una materia* no cabe igualmente sino pasarla/no pasarla.

Pues bien, el coeficiente *Phi* es un índice que se adapta a *variables* o atributos en formato dicotómico, que ordinariamente se presentan, dependiendo del atributo, como 0/1, sí/no, +/-, o de alguna otra forma. Welkowitz *et al.* (1981, 248) afirman que este indice se comprende adecuadamente, por estar estrechamente relacionado, cuando se comenta el estadístico *Ji cuadrado* (χ^2).

Así, pues, los datos se dispondrán conforme al modelo siguiente, si se tratara de un caso de poner en relación en 12 estudiantes universitarios el sexo (*X*) con el éxito en selectividad (*Y*) (Cfr. Glass y Stanely, 1974, 158):

Estudiante	X: Sexo (var: 1; hemr: 0)	Y: Selectividad (apt: 1; no apt: 0)	Cálculos

<table>
<tr><td>1</td><td>0</td><td>0</td><td rowspan="12">

- la proporción de puntuaciones 1 en *X* - p_x - es: 5/12 0.4167;
- la proporción de puntuaciones 0 en *X* - q_x - es: 7/12 = 1 0.4167 = 0.5833;
- la proporción de puntuaciones 1 en *Y* - p_y - es: 6/12 0.50;
- la proporción de puntuaciones 0 en *Y* - q_y - es: 6/12 0.50;
- la proporción de 1 tanto en *X* como en *Y* - p_{xy} - es: 4/12 0.3333.

$\phi = (p_{xy} - p_x p_y)/\sqrt{} \; p_x q_x p_y q_y$ = **0.507**

</td></tr>
<tr><td>2</td><td>1</td><td>1</td></tr>
<tr><td>3</td><td>0</td><td>1</td></tr>
<tr><td>4</td><td>0</td><td>0</td></tr>
<tr><td>5</td><td>1</td><td>1</td></tr>
<tr><td>6</td><td>1</td><td>0</td></tr>
<tr><td>7</td><td>0</td><td>0</td></tr>
<tr><td>8</td><td>1</td><td>1</td></tr>
<tr><td>9</td><td>0</td><td>0</td></tr>
<tr><td>10</td><td>0</td><td>1</td></tr>
<tr><td>11</td><td>0</td><td>0</td></tr>
<tr><td>12</td><td>1</td><td>1</td></tr>
</table>

Estos datos se disponen en una tabla de doble entrada conforme al modelo siguiente en orden a los cálculos:

Variable Y: Éxito en selectividad	**Variable X: Sexo**		**Totales**
	▪ **Mujer: 0**	▪ **Hombre: 1**	
▪ **Aprobado: 1**	a: 2	b: 4	a+b: **6**
▪ **Suspenso: 0**	c: **5**	d: 1	c+d: 6
Totales	a+c: 7	b+d:5	a+b+c+d = 12

Sin embargo, por razones que luego se verán al utilizar el estadístico de *Chi Cuadrado,* se va a utilizar idéntico esquema pero con unas frecuencias muy superiores a las contempladas en la tabla anterior:

Variable Y: Éxito en selectividad	**Variable X: Sexo**		**Totales**
	▪ **Mujer: 0**	▪ **Hombre: 1**	
▪ **Aprobado: 1**	a: 9	b: 27	a+b: **36**
▪ **Suspenso: 0**	c: **44**	d: **4**	c+d: **48**
Totales	a+c: **53**	b+d:31	a+b+c+d = 84

El coeficiente de correlación producto-momento de Pearson es una medida de la relación entre *X* e *Y* que, cuando se calcula a partir de datos monimales dicotómicos, se conoce como coeficiente *phi*. En Glass y Stanley (1974, 159/160) se demuestra cómo se puede pasar desde la fórmula de Pearson, si fueran los datos dicotómicos, a coeficiente *Phi*. La fórmula final de *Phi* es la siguiente:

$$\phi = (p_{xy} - p_x p_y)/\sqrt{} \; p_x q_x p_y q_y.$$

Para el cálculo de *Phi* se precisa conocer la proporción de los distintos sucesos en *X* e Y. Así:

- la proporción de puntuaciones 1 en *X* - p_x - es: 31/84 = 0.369;
- la proporción de puntuaciones 0 en *X* - q_x - es: 53/84 = 1 - 0.369 = 0.631;
- la proporción de puntuaciones 1 en *Y* - p_y - es: 36/84 = 0.429;
- la proporción de puntuaciones 0 en *Y* - q_y - es: 48/84 = 0.571;
- la proporción de 1 tanto en *X* como en *Y* - p_{xy} - es: 27/84 = 0.321.

Sustituyendo en la fórmula: $\phi = \{0.321 - (0.369)\ 0.631\}/\sqrt{}\ 0.369\ (0.631)\ 0.429\ (0.571) = \{0.321 - 0.1583\}/\sqrt{}\ 0.057 = 0.703$.

Cuando no es importante conocer las propiedades de p_x y p_y y se considera más conveniente tabular los datos dicotómicos tal como se encuentran en una tabla de contingencia, puede calcularse *Phi* mediante la fórmula: $\phi = (bc - ac) / \sqrt{}(a+c)\ (b+d)\ (a+b)\ (c+d)$. Sustituyendo: $\phi = (1188 - 36)/ \sqrt{}53(31)36(48) = 0.683$. Es un valor próximo a 0.70, que se acepta.

Propiedades de Phi. ¿Cómo se puede interpretar *Phi*? Se ha dicho que *Phi* es una extensión del coeficiente de correlación de Pearson al caso de variables dicotómicas. Pero la interpretación de ambas no es idéntica. No es fácil suponer que las dos variables dicotómicas se distribuyan normalmente como ocurre en Pearson. Por tanto, no puede interpretarse *Phi* como se interpreta Pearson; es decir, no es fácil interpretar *Phi* de 0.703 como si fuera equivalente a una r_{xy} de 0.703 en Pearson.

La razón de lo anterior es muy sencilla: *Phi* no puede tomar el valor de +1 sino solamente cuando ($a + b$) y ($b + d$) sean iguales y, por consiguiente: $a = d$ en una tabla de contingencia de dos por dos. Es decir, sólo cuando la proporción de 1 es la misma en *X* e Y. *Y,* si esto es así, cuando $a + c = c + d$ y, por consiguiente: $a = d$.

Otra condición es que $a = d = 0$. *Y*, finalmente, que los cuadrantes estén vacíos. Es decir, el valor de *Phi* será la unidad cuando las frecuencias de cualquiera de las dos casillas diagonales sean nulas. En los demás casos se da un valor máximo de *Phi* -***Phi* máximo**- que es preciso calcular para poder interpretar su valor empírico. Por tanto, y sustituyendo en el problema:

$$\phi_{max} = \sqrt{}(f_{1-m}/f_m)\ (f_p/f_{1-p}) = \sqrt{}(31/53)\ (48/36) = 0.883,$$

donde:
f_m : frecuencia marginal mayor de las cuatro, esto es: 53;
f_{1-m}: frecuencia marginal complementaria a la anterior, esto es: 31;
f_p : frecuencia marginal mayor de la otra fila o columna, esto es: 48;
f_{1-p} : frecuencia marginal complementaria a la anterior, esto es: 36.

Es decir, el valor de 0.70 se ha de interpretar a partir de un valor máximo -0.883- que es el valor que para la disposición de las frecuencias es el máximo que los datos pueden arrojar en el supuesto de que la relación fuera perfecta. Es decir, 0.70 es una correlación muy alta para el valor máximo de 0.883. Sin embargo, ¿no es posible llegar a calcular una correlación interpretable en términos de correlación de Pearson, es decir, de 0 a 1? Es posible.

Volvamos a partir de los datos. Existe una prueba estadística, que se estudiará más adelante como prueba, entre otras cosas, de independiencia o relación entre variables y como modelo de distribución muestral, llamada *Chi Cuadrado*-χ^2 de Pearson, que, a partir de ella, se permite obtener un valor de *Phi* interpretable como un coeficiente de correlación de Pearson, es decir, de 0 a 1.

La fórmula de *Chi Cuadrado* es ésta: $\chi^2 = N(AD-BC)^2 / (A+B)(C+D)(A+C)(B+D)$. Sustituyendo los datos del problema se tiene: $\chi^2 = 84\ (36 - 1188)^2 / 36\ (48)\ 53\ (31) = 39.26$.

Una vez obtenido el valor de *Chi Cuadrado*, Welkowitz *et al.* (1981, 323) introducen una fórmula de *Phi* que sirve para convertir *Phi* a índice de correlación recurriendo al valor de *Chi Cuadrado*. Esta es la fórmula:

$$\phi = \sqrt{\chi^2 / N} = \sqrt{39.26 / 84} = 0.68.$$

Este índice se interpreta como un coeficiente de correlación de Pearson.

2. Coeficiente de Contingencia-C

Afirma Siegel (1976, 227) que el coeficiente *C* es una medida del grado de asociación o relación entre dos conjuntos de atributos. Por tanto es especialmente útil cuando la información es clasificatoria, que viene en consecuencia en escala nominal. Se aplica a dos atributos cada uno de los cuales puede presentar dos o más de dos categorías.

Responde a la fórmula: $C = \sqrt{\chi^2 / (N + \chi^2)}$.

Como el presente caso contempla dos atributos constando cada uno de dos categorías, el valor de *Chi Cuadrado* se puede calcular en un ejemplo concreto y proceder al cálculo de *C*, sabiendo que *N* es el tamaño total de la muestra.

De cualquier forma, dada la posibilidad de transformar *Chi Cuadrado* a índice de corrrelación, adaptándose igualmente a datos de idéntica naturaleza, es preferible *Chi Cuadrado* a l coeficiente *C*, que ha caído en desuso.

Conclusión

Una de las conclusiones más claras que se derivan de los coeficientes de asociación es que, dado que son en sí ininterpretables, los autores han buscado fórmulas para convertirlos a índices de correlación, puesto que las correlaciones son interpretables de 0 a 1 y no precisan calcular el valor teórico máximo para interpretar el empírico.

Existen otros índices de correlación, tal como se indica en las tablas, que no se desarrollan en las presentes páginas. Se insertan algunas referencias para el lector pueda ampliar.

Glosario

Asociación	Escala de la variable
Coeficiente cuádruplo	Escala de orden
Coeficiente de Contingencia	Escala de razón
Coeficiente de correlación biserial	Escala nominal
Coeficiente de correlación biserial puntual	Naturaleza de la variable
Coeficiente de correlación de Spearman	Razón de correlación
Coeficiente de correlación tetracórico	Relación curvilínea
Coeficiente *eta*	Relación lineal
Coeficiente *Phi*	Suposiciones paramétricas
Coeficiente *tau* de Kendall	Variable continua
Correlación	Variable dicotómica
Correlación múltiple	Variable dicotomizada
Escala de intervalo	Variable discreta

Referencias

Glass, G.V. y Stanley, J.C. (1974): *Métodos estadísticos aplicados a las ciencias sociales*, Cap. 9 (Medidas adicionales de correlación)

Fernández Díaz, M.J. *et al.* (1990): *Resolución de problemas de estadística aplicada a las ciencias sociales. Guía práctica para profesores y alumnos*, 3. Medidas correlacionales, pp. 58-72.

Pérez Juste, R. (1985): *Estadística Descriptiva*, Tema IX (Coeficientes de correlación relacionados con el *r* de Pearson) y Tema *X* (La correlación en el caso de variables nominales y ordinales).

Siegel, S. (1976): *Estadística no paramétrica aplicada a las ciencias de la conducta*, Cap. 9 (Las medidas de correlación y sus pruebas de significación)

Welkowitz, J. *et al.* (1981): *Estadística aplicada a las Ciencias de la Educación*, caps. 12 (Otras técnicas de correlación) y 16 (Ji cuadrado).