REVISTA ELECTRÓNICA DE PSICOLOGÍA
Vol. 4, No. 2, Julio 2000
ISSN 1137-8492

Tutorial sobre coeficientes de correlación con una o dos variables categóricas

Palmer, A., Jiménez, R. y Montaño, J.J.
Area de Metodología de las Ciencias del Comportamiento.
Facultad de Psicología.
Universitat de les Illes Balears.
e-Mail: alfonso.palmer@uib.es

1.- Introducción
2.- Coeficientes de correlación entre una variable continua y otra categórica
2.1.- Variable categórica dicotómica
2.1.1.- Coeficiente de correlación biserial puntual
2.1.2.- Coeficiente de correlación biserial
2.2.- Variable categórica politómica
2.2.1.- Correlación eta
3.- Correlación con una medida ordinal: coeficiente de correlación biserial por rangos
4.- Coeficientes de correlación en tablas de contingencia bivariantes
4.1.- Coeficiente de correlación de Pearson
4.2.- Coeficiente phi
4.3.- Correlación tetracórica
4.4.- Correlación de Spearman
4.5.- Correlación de Kendall
4.6.- Correlación eta

1. Introducción

En un tutorial anterior a éste se habló del estudio de la correlación entre dos variables continuas (Palmer, Jiménez y Montaño, 2000), concretamente se analizó el coeficiente de correlación lineal de Pearson y derivados (correlación parcial, semiparcial y múltiple), así como el coeficiente de correlación ordinal de Spearman. El aspecto más destacado de dicho tutorial son las direcciones de Internet que se ofrecen al usuario para poder acceder a información y al cálculo de los coeficientes de correlación tratados.

En el tutorial que aquí presentamos se realiza un análisis de los distintos coeficientes de correlación que se pueden aplicar a dos variables, pero para el caso concreto en el que al menos una de estas variables sea categórica. Se comenzará hablando de coeficientes de correlación entre una variable continua y otra categórica, iniciando dicha andadura por aquellos casos en los que la variable categórica sea dicotómica, pasando a continuación a hablar de la relación entre una variable continua y otra politómica. También se considerarán aquellos casos en los que tengamos una variable dicotómica y otra continua que sea tratada de forma ordinal a través de los rangos asociados a sus valores. Finalmente, en la última sección analizaremos los distintos coeficientes de correlación que se pueden aplicar cuando las dos variables a relacionar sean categóricas (tablas de contingencia).

El objetivo de este tutorial pasa por pretender dar una base teórica para la interpretación de los distintos coeficientes de correlación así como ejemplificar el cálculo de éstos, aunque el aspecto más destacado es la posibilidad que se brinda al usuario de acceder a través de Internet al cálculo de algunos de los coeficientes de correlación tratados en este tutorial.

2.- Coeficientes de correlación entre una variable continua y otra categórica

2.1.- Variable categórica dicotómica

2.1.1.- Coeficiente de correlación biserial puntual (r_{bp})

Este coeficiente es muy utilizado en el ámbito de la Psicometría para evaluar la discriminación de los ítems que componen un test. La Psicometría es una disciplina que según Muñiz (1992) puede definirse en términos generales como el conjunto de métodos, técnicas y teorías implicadas en la medición de las variables psicológicas. Dentro de este conjunto de métodos de medición encontramos los tests, instrumentos de medida regidos por un procedimiento estándar y que son utilizados para obtener valores de los indicadores de un determinado constructo. El test está formado por una serie de ítems que representan variables que operativizan el constructo a medir (una habilidad, un rasgo de personalidad, etc.).

La discriminación de un ítem es la capacidad de éste para distinguir entre individuos con diferente situación en el rasgo o atributo medido por el test. Un ítem discriminará bien si las respuestas y puntuaciones de cada uno de estos sujetos al ítem están relacionadas con sus puntuaciones globales del test, de modo que cada ítem se convierte en un pequeño test que facilita la puntuación que describe la capacidad, actitud, etc. de la persona evaluada.

Una forma de calcular el índice de discriminación de un ítem *i*, se da mediante el índice de correlación biserial puntual entre el ítem *i* (variable dicotómica: respuesta correcta ó respuesta incorrecta) y la puntuación total *X* del test (variable continua), de la que el ítem forma parte como elemento. El coeficiente de correlación biserial puntual oscila como la r de Pearson entre ±1, interpretándose de igual modo.

Si los ítems están bien enunciados, es de esperar que exista correlación entre la respuesta dada al ítem y la puntuación del sujeto en la prueba.

En general, la correlación biserial puntual se establece como una correlación de Pearson entre dos variables, con la particularidad de que una de esas variables es de tipo continuo y la otra es una variable dicotómica (no dicotomizada, como ocurre en el caso de la correlación biserial). En el caso

concreto que se expone, en el que la discriminación va a ser calculada mediante un coeficiente de correlación ítem-test, se considera a cada uno de los ítems *i* como una variable dicotómica, puesto que lo que se tiene en cuenta es si el sujeto contesta o no correctamente al ítem. A la respuesta correcta se le puede asignar el valor uno y a la respuesta incorrecta el valor cero, de manera que cualquier sujeto obtendrá, como vector de respuesta a la prueba un conjunto de unos y de ceros. Las puntuaciones globales de los sujetos en la prueba, si las hay en número suficiente, pueden considerarse como valores de una variable continua. (En ítems no acertados incluimos ítems no alcanzados y omisiones).

Sea p la proporción de respuestas correctas dadas a un cierto ítem y sea q = 1-p la de incorrectas. Designamos mediante m_x, m_p, m_q y S^2_x, respectivamente, a la media de las puntuaciones obtenidas en la prueba por el conjunto de todos los sujetos, la media de las puntuaciones obtenidas en la prueba por aquellos que dieron respuesta correcta a ese ítem, la media de las puntuaciones en la prueba por aquellos que no respondieron correctamente al ítem y la variancia de las puntuaciones en la prueba.

El coeficiente de correlación biserial puntual se obtiene mediante la expresión:

$$r_{bp} = \frac{m_p - m_q}{S_X}\sqrt{p_i q_i}$$

(La diferencia m_p-m_q será tanto mayor cuanto más difieran en puntuación los que contestan correctamente, de aquellos que lo hacen incorrectamente. Para $m_p > m_q$, entonces r_{bp} tenderá a crecer con m_p-m_q, o sea, con la corrección entre el test y la respuesta correcta al ítem).

O bien, podemos decir que, para un cierto ítem *i*, el coeficiente de correlación biserial puntual puede obtenerse mediante:

$$r_{bp} = \frac{m_p - m_x}{S_x}\sqrt{\frac{p_i}{q_i}}$$

donde m_p representa la media de las puntuaciones de los sujetos que responden correctamente al ítem *i*, m_x indica la media de las puntuaciones de los examinados en el test, S_x es la desviación estándar de las puntuaciones obtenidas por todos los examinados en el test y p_i expresa la dificultad del ítem.

Para realizar los cálculos indicados anteriormente, a la puntuación total del test (X) hay que descontarle el ítem cuyo índice de discriminación se pretende hallar (X-i), de lo contrario una de las variables (el ítem) estaría impropiamente incluida en la otra (el test).

Si al calcular la correlación ítem-test no se descontase de éste las puntuaciones correspondientes al ítem, se estaría elevando impropiamente la correlación, pues estrictamente no se estaría correlacionando el ítem con el resto de los ítems (test), sino con un test que incluiría también al ítem en cuestión. En suma, se estaría correlacionando una variable (test) con parte de ella (ítem). Cuando el test consta de un número elevado de ítems este efecto puede ser de poca relevancia empírica, pero ello no legitima su incorrección.

Lo más sencillo es calcular la corrección descontando el ítem. Sin embargo, si por cualquier razón se tiene la correlación ítem-test sin descontar los efectos del ítem, puede utilizarse la siguiente fórmula de corrección para obtener la correlación pertinente:

$$r_{i(x-i)} = \frac{r_{ix}S_x - S_i}{\sqrt{S_i^2 + S_x^2 - 2r_{ix}S_iS_x}}$$

donde $r_{i(x-i)}$ es la correlación entre el ítem *j* y el test tras descontar el ítem *(x-i)*, r_{ix} es la correlación ítem-test cuando el ítem está incluido en el test, S_x es desviación estándar del test y S_i es la desviación estándar del ítem.

A continuación se proporciona la tabla de respuestas de 10 sujetos a 5 ítems, donde el valor 1 indica «respuesta correcta» y el valor 0 «respuesta incorrecta»:

	I T E M S					
	1	2	3	4	5	PUNTUACIONES
SUJETO 1	1	1	0	1	0	3
SUJETO 2	1	1	1	1	1	5
SUJETO 3	1	0	1	0	0	2
SUJETO 4	0	1	0	1	0	2
SUJETO 5	1	1	0	1	1	4
SUJETO 6	1	0	1	0	0	2
SUJETO 7	0	1	1	1	1	4
SUJETO 8	1	1	0	0	0	2
SUJETO 9	1	0	1	0	0	2
SUJETO 10	1	1	1	0	1	4
Total Item	8	7	6	5	4	30
p_i	0.8	0.7	0.6	0.5	0.4	
q_i	0.2	0.3	0.4	0.5	0.6	

La media m_x del test viene dada por : $m_x = \Sigma\, p_i$. Este valor coincide con el hallado a partir de las puntuaciones totales del test:

$$m_x = \frac{3+5+2+2+4+2+4+2+2+4}{10} = \frac{30}{10} = 3$$

$$m_x = 0.8 + 0.7 + 0.6 + 0.5 + 0.4 = 3$$

La variancia del test es una *variancia poblacional*. Así pues, su divisor es *n* y no *n-1*.

$$S_x^2 = \frac{\Sigma(x_i - m_x)^2}{n} = \frac{12}{10} = 1.2$$
$$S_x = \sqrt{1.2} = 1.095$$

Cálculo del coeficiente de correlación biserial puntual para el ítem 3:

La puntuación media de los sujetos que han contestado correctamente al ítem viene dada por:

$$M_p = \frac{5+2+2+4+2+4}{6} = \frac{19}{6} = 3.1\hat{6}$$

La media de las puntuaciones totales al test, la desviación estándar, así como el índice de dificultad del ítem se ha calculado con anterioridad, lo cual nos permite hallar el coeficiente de correlación biserial puntual.

$$r_{bp} = \frac{m_p - m_x}{S_x}\sqrt{\frac{p_i}{q_i}} = \frac{3.1\hat{6} - 3}{1.095}\sqrt{\frac{0.6}{0.4}} = 0.1864$$

Podemos acceder a Internet para el cálculo del coeficiente de correlación biserial puntual. *Jeff Kandra*. La aplicación presenta el siguiente aspecto (han sido introducidos los valores del ítem 3):

Compute
- Allows you to compute the Point-Biserial Correlation using the data you have entered below.

Sample Data 1
- Retrieves some sample data for you to use.
Sample Data 2
- Retrieves more sample data for you to use.

Examinee (subject)	Score on Item (enter 0 or 1)	Score on Test	Examinee (subject)	Score on Item (enter 0 or 1)	Score on Test
1	0	3	2	1	5
3	1	2	4	0	2
5	0	4	6	1	2
7	1	4	8	0	2
9	1	2	10	1	4
11			12		
13			14		
15			16		
17			18		
19			20		
21			22		
23			24		
25			26		
27			28		
29			30		

Enter data in the boxes. You do not need to enter data in all of the rows of boxes. When entering the scores, do not type commas, periods, or other special characters. After you have entered the data, press the (COMPUTE) button. A report will then be generated and displayed on your computer. The data that you enter will be saved automatically so that you will not have to enter the data more than once.

Se accederá a la siguiente pantalla cuando sea pulsado el botón «Compute»:

Item Analysis: Correlational Methods: Point-Biserial

Examinee (subject)	Score on Item	Score on Test
1	0	3
2	1	5
3	1	2
4	0	2
5	0	4
6	1	2
7	1	4
8	0	2
9	1	2
10	1	4

X1 = (19 / 6) = 3.16666666666667 = Average test score for subjects that scored 1 on the item.

X2 = (11 / 4) = 2.75 = Average test score for subjects that scored 0 on the item.

Xt = (30 / 10) = 3.0000 = Average test score for all subjects.

p = 0.6000 = Portion of subjects scoring 1 on the item.

q = 0.4000 = Portion of subjects scoring 0 on the item.

Sx = SQRT((10(102) - (30)^2) / (10)(10)) = 1.0954 = Variance in test scores based on sum of squares.

rpb = ((3.16666666666667 - 2.75)SQRT(0.6000(0.4000))) / 1.0954 = 0.1863 = Point-Biserial correlation coefficient.

Como se puede observar, da información sobre la puntuación media de los sujetos que han contestado correctamente al ítem (*X1*), la puntuación media de los sujetos que no han contestado correctamente al ítem (*X2*), la media de puntuación del test para todos los sujetos *(Xt*), la proporción de sujetos que han contestado correctamente al ítem (*p*) y los que han contestado de forma incorrecta (*q*), la raiz cuadrada de la variancia poblacional del test (*Sx*) y, finalmente, el valor del coeficiente de correlación biserial puntual (*rbp*). Podemos observar que calcula el coeficiente de correlación biserial puntual a partir de la siguiente expresión:

$$r_{bp} = \frac{m_p - m_q}{S_X}\sqrt{p_i q_i}$$

cuyos elementos ya han sido definidos anteriormente.

Si descontamos a las puntuaciones totales del test, el valor del ítem, podemos obtener el **coeficiente de correlación biserial puntual corregido**. Debemos pues, hallar la puntuación media de los sujetos que aciertan el ítem, así como la media y la desviación estándar de las puntuaciones totales corregidas:

PUNTUACIONES TOTALES CORREGIDAS	
3 4 1 2 4 1 3 2 1 3	$m_p = \frac{4+1+1+3+1+3}{6} = \frac{13}{6} = 2.1\hat{6}$ $m_x = \frac{3+4+1+2+4+1+3+2+1+3}{10} = \frac{24}{10} = 2.4$ $S_x^2 = \frac{12.4}{10} = 1.24$ $S_x = \sqrt{1.24} = 1.113$

Así, el coeficiente de correlación biserial puntual corregido vendrá dado por:

$$r_{bp} = \frac{m_p - m_x}{S_x}\sqrt{\frac{p_i}{q_i}} = \frac{2.1\hat{6} - 2.4}{1.113}\sqrt{\frac{0.6}{0.4}} = -0.2567$$

Podemos observar cómo este resultado coincide con el que obtenemos al aplicar la fórmula de corrección del coeficiente:

$$r_{i(x-i)} = \frac{r_{ix}S_x - S_i}{\sqrt{S_i^2 + S_x^2 - 2r_{ix}S_iS_x}} =$$

$$\frac{(0.1864)(1.095) - (0.48989)}{\sqrt{(0.24) + (1.2) - 2(0.1864)(0.48989)(1.095)}} = -0.2567$$

A la luz de los resultados, se puede comprobar el sesgo que se produce en el valor final al incluir la puntuación obtenida en el ítem. La discriminación de un ítem se verá pues, muy afectada, por la inclusión impropia del mismo en el procedimiento de cálculo. Este sesgo disminuye a medida que aumenta el número de ítems que componen la puntuación total del test.

A continuación se ofrece el valor del coeficiente de correlación biserial puntual para cada uno de los ítems que constituyen el test anterior, así como el de dicha correlación corregida.

Vector de coeficientes de correlación biserial puntual

ITEM	1	2	3	4	5
$\mathbf{r_{bp}}$	0	0.5978	0.1864	0.5479	0.93207

Vector de coeficientes de correlación biserial puntual corregida

ITEM	1	2	3	4	5
$\mathbf{r_{bp}}$	-0.3429	0.2182	-0.2567	0.1104	0.80

2.1.2.- Coeficiente de correlación biserial (r_b)

El coeficiente de correlación biserial se utiliza también cuando el ítem a analizar se presenta de forma dicotómica. Sin embargo, el ítem no es dicotómico por naturaleza, sino que se trata de una variable continua, distribuida según la curva normal y que el analizador decide dicotomizar.

Evidentemente, el paso de una escala continua a una dicotómica implica una pérdida de información considerable, lo que hace desaconsejable éste método.

Aún así, en la práctica podemos encontrarnos con escalas de gradación, que por distintas razones tienden a dicotomizarse (formatos de respuesta dicotomizada). En tal caso, para el cálculo de la discriminación del ítem, utilizamos el coeficiente de correlación biserial que se obtiene a partir de:

$$r_b = \frac{m_p - m_x}{S_x}\frac{p}{y}$$

donde m_p, m_x, S_x y p representan los índices ya conocidos (apartado anterior), y el valor de y representa el valor de la ordenada correspondiente al valor de la puntuación estándar, en la curva normal, que deja por debajo un área igual a p, valor éste que puede ser encontrado en tablas (Palmer, 1995a).

Si la distribución del ítem no se ajusta a la curva normal, especialmente si es platicúrtica o bimodal (Muñiz,1992), la r_b puede tomar valores superiores a uno. Esto hace aconsejable el uso del coeficiente de correlación biserial puntual (apartado anterior), cuya relación con r_b se expresa mediante:

$$r_{bp} = \frac{r_b y}{\sqrt{pq}}$$

Vamos a suponer que los 5 ítems de la siguiente tabla, que se han administrado a 10 sujetos, fueron dicotomizados de forma artificial (el valor 1 indica «respuesta correcta» y el valor 0 «respuesta incorrecta»):

	I T E M S					
	1	2	3	4	5	PUNTUACIONES
SUJETO 1	1	1	0	1	0	3
SUJETO 2	1	1	1	1	1	5
SUJETO 3	1	0	1	0	0	2
SUJETO 4	0	1	0	1	0	2
SUJETO 5	1	1	0	1	1	4
SUJETO 6	1	0	1	0	0	2
SUJETO 7	0	1	1	1	1	4
SUJETO 8	1	1	0	0	0	2
SUJETO 9	1	0	1	0	0	2
SUJETO 10	1	1	1	0	1	4
Total Item	8	7	6	5	4	30
p_i	0.8	0.7	0.6	0.5	0.4	
q_i	0.2	0.3	0.4	0.5	0.6	

Cálculo de la correlación biserial para el ítem 3:

$$r_b = \frac{m_p - m_x}{S_x}\frac{p}{y} = \frac{3.1\widehat{6} - 3}{1.095}\frac{(0.6)}{(0.3863)} = 0.2363$$

Podemos comprobar la correspondencia entre el coeficiente de correlación biserial y el biserial puntual:

$$r_{bp} = \frac{r_b y}{\sqrt{pq}} = \frac{(0.2363)(0.3863)}{\sqrt{0.24}} = 0.1864$$

El resultado se corresponde con el hallado para el coeficiente de correlación biserial puntual no corregido, del ítem 3 (apartado anterior).

El valor del coeficiente de correlación biserial para cada ítem se recoge a continuación:

Vector de coeficientes de correlación biserial

ITEM	1	2	3	4	5
$\mathbf{r_b}$	0	0.7878	0.2363	0.6866	1.18

Observe que el coeficiente para el ítem 5 supera el valor 1, hecho que se podría explicar por lo expuesto anteriormente sobre el efecto de la no normalidad de la distribución de un ítem.

2.2. Variable categórica politómica

2.2.1. Correlación eta

Eta es un índice de asociación que se utiliza cuando la variable independiente está medida en una escala nominal (variable categórica) y la variable dependiente está medida en una escala de intervalo o de razón. Por este motivo el índice eta siempre es asimétrico. El valor de eta cuadrado (también llamado *razón de correlación*) se interpreta como la proporción de variancia de la variable dependiente que es explicada por la variable independiente (Palmer, 1996a).

Se muestra a continuación cómo calcular eta cuadrado (η^2):

$$\eta^2 = \frac{SC_{ENTRE}}{SC_{ENTRE} + SC_{RESIDUAL}}$$

donde:

$$SC_{ENTRE} = \sum_{i=1}^{k} n_i (m_i - G)^2$$

$$SC_{RESIDUAL} = \sum_{i=1}^{k} (n_i - 1) s_i^2$$

Ejemplo

Dada la variable dependiente *tiempo de reacción* y la variable independiente categórica *dosis de un fármaco* con cuatro condiciones experimentales, se pretende buscar la proporción de variabilidad de la variable dependiente que es explicada por la variable independiente. A continuación se proporcionan los datos básicos para realizar el análisis:

	DOSIS			
	0cc	5cc	10cc	15cc
	129	142	176	198
	135	167	178	198
	138	157	177	197
	149	153	173	191
	146	153	187	182
	141	160	191	191
	138	149	171	187
	138	145	185	190
	133	165	188	194
	155	156	176	190
	143	150	180	193
	149	154	177	183
	143	162	198	183
	141	134	188	190
	149	144	178	202
m_i	141.8	152.73	181.53	191.27
s^2	48.74	81.92	57.12	35.35
n	15	15	15	15
				N=60
				G=166.83

El valor de SC_{ENTRE} es el siguiente:

$$SC_{ENTRE} = \sum_{i=1}^{k} n_i (m_i - G)^2 =$$
$$= 15\left[(141.8 - 166.83)^2 + (152.73 - 166.83)^2 + (181.53 - 166.83)^2 + (191.27 - 166.83)^2\right] =$$
$$= 24578.33$$

Cálculo de la $SC_{RESIDUAL}$:

$$SC_{RESIDUAL} = \sum_{i=1}^{k} (n_i - 1)s_i^2 = 14\left[48.74 + 81.92 + 57.12 + 35.35\right] = 3124$$

(Nota: los valores han sido calculados utilizando 12 decimales)

Finalmente, obtenemos el valor de eta cuadrado:

$$\eta^2 = \frac{SC_{ENTRE}}{SC_{ENTRE} + SC_{RESIDUAL}} = \frac{24578.33}{24578.33 + 3124} = 0.887$$

Vemos como este valor coincide con el proporcionado por el SPSS:

Medidas de asociación

	Eta	Eta cuadrado
Tiempo de reacción * DOSIS	,942	,887

Podemos concluir, por tanto, que la variable independiente *dosis* permite explicar el 88,7% de la variabilidad de la variable dependiente *tiempo de reacción*.

En el contexto del ANOVA, donde una variable dependiente continua se relaciona con una variable independiente nominal, el índice eta cuadrado también puede ser expresado en función del valor de la prueba F hallado como resultado del análisis de la variancia:

$$\eta^2 = \frac{(k-1)F}{(k-1)F + (N-k)}$$

donde:

$$F = \frac{\frac{SC_{ENTRE}}{k-1}}{\frac{SC_{RESIDUAL}}{N-k}}$$

y *k* es el número de condiciones de la variable independiente.

3.- Correlación con una medida ordinal: coeficiente de correlación biserial por rangos (r_{bRANK})

Curenton (1956) y Glass (1966) proponen un coeficiente de correlación biserial por rangos para aplicar en situaciones en las que las puntuaciones se presentan como un rango u ordenación de grupo.

Este coeficiente, conocido como el coeficiente de Glass se calcula mediante la expresión:

$$r_{b(RANK)} = \frac{2}{N}(\overline{R}_A - \overline{R}_B)$$

donde N es el número total de sujetos, $\overline{R}_A$ es el rango medio en el criterio de los sujetos que aciertan el ítem, y $\overline{R}_B$ es el rango medio en el criterio de los sujetos que fallan el ítem.

Así, para la tabla de sujetos siguiente se asignan rangos a las puntuaciones totales de los sujetos. En caso de que se produzcan empates se asigna el rango medio:

X:	21	22	23	24	25	36	47	48	49	510
R:			3			6		8		10

	ITEMS						
	1	2	3	4	5	PUNTUACIONES	RANGOS

SUJETO 1	1	1	0	1	0	3	6
SUJETO 2	1	1	1	1	1	5	10
SUJETO 3	1	0	1	0	0	2	3
SUJETO 4	0	1	0	1	0	2	3
SUJETO 5	1	1	0	1	1	4	8
SUJETO 6	1	0	1	0	0	2	3
SUJETO 7	0	1	1	1	1	4	8
SUJETO 8	1	1	0	0	0	2	3
SUJETO 9	1	0	1	0	0	2	3
SUJETO 10	1	1	1	0	1	4	8

Coeficiente de correlación biserial por rangos para el ítem 3:

$$r_{b(RANK)} = \frac{2}{N}(\overline{R}_A - \overline{R}_B) = \frac{2}{10}\left(\frac{35}{6} - \frac{20}{4}\right) = 0.1\hat{6}$$

El valor del coeficiente de correlación biserial por rangos que obtenemos para cada ítem será:

Vector de coeficientes de correlación biserial por rangos

ITEM	1	2	3	4	5
$\mathbf{r_{b(RANK)}}$	0	0.7142	0.1666	0.6	1

Este índice es una estimación de la correlación por rangos de Spearman. Es importante señalar que, en caso de muchos empates, como es el que nos ocupa, este índice está muy sesgado.

4.- Coeficientes de correlación en tablas de contingencia bivariantes

Veremos a continuación distintos tipos de coeficientes de correlación que pueden ser aplicados a tablas de contingencia bivariantes, es decir, en aquellos casos en los que se relacionen dos variables categóricas. En una tabla de contingencia bivariante RxC (R categorías de la variable fila y C categorías de la variable columna) cada uno de los elementos de la muestra utilizada se clasifica en un nivel (categoría) concreto de cada una de las variables categóricas, y en cada una de las casillas de la tabla de las RxC existentes se representa la frecuencia (n_{ij}) de aquellos elementos de la muestra que pertenecen a un nivel concreto de la variable fila (*i* niveles) y a otro nivel concreto de la variable columna (*j* niveles).

4.1.- Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson **r** entre dos variables X e Y, viene dado por:

$$r = \frac{Cov(X, Y)}{\sqrt{S(X)\,S(Y)}}$$

Sean X e Y las dos variables, fila y columna, que forman una tabla de contingencia, los elementos para obtener el coeficiente de correlación de Pearson vienen dados por:

$$Cov(X, Y) = \sum X_i\, Y_j\, n_{ij} - \frac{\left(\sum_{i=1}^{r} X_i n_{i.}\right)\left(\sum_{j=1}^{c} Y_j n_{.j}\right)}{N}$$

$$S(X) = \sum_{i=1}^{r} X_i^2 n_{i.} - \frac{\left(\sum_{i=1}^{r} X_i n_{i.}\right)^2}{N}$$

$$S(Y) = \sum_{j=1}^{c} Y_j^2 n_{.j} - \frac{\left(\sum_{j=1}^{c} Y_j n_{.j}\right)^2}{N}$$

En el caso particular de una tabla de contingencia 2x2, cada una de las variables cualitativas puede ser codificada, por ejemplo, con los valores 0 y 1, de forma que X={0,1} y Y={0,1}, pudiéndose obtener el coeficiente de correlación de Pearson entre las dos variables.

En este caso, la expresión que permite obtener el valor del coeficiente de correlación, deducida de la expresión general de Pearson mostrada anteriormente, viene dada por:

$$r = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1.}n_{2.}n_{.1}n_{.2}}}$$

Este valor se moverá en el intervalo [-1,+1], y tomará el valor cero cuando las dos variables sean independientes o bien cuando no haya ningún tipo de relación lineal entre ellas.

Al igual que en el caso de variables cuantitativas, el valor de r^2 puede interpretarse como la variancia de la variable dependiente que es explicada por la variable independiente.

El valor que toma este índice en el ejemplo presentado más adelante vale:

$$r = \frac{130 \times 3 - 97 \times 20}{\sqrt{227 \times 23 \times 150 \times 100}} = -0.17515$$

Se puede comprobar que este valor es el proporcionado por el SPSS:

		Valor	Error típ. asint.	T aproximada	Sig. aproximada
Intervalo por intervalo	R de Pearson	-,17515	,04805	-2,80157	,00549

A continuación, el SPSS proporciona el valor ASE1 del error estándar (*Error típ. asint.*) supuesto que la hipótesis alternativa es cierta. Este valor viene dado por medio de los siguientes cálculos:

$$\overline{X} = \frac{\sum_{i=1}^{r} X_i n_{i.}}{N}$$

$$\overline{Y} = \frac{\sum_{j=1}^{c} Y_j n_{.j}}{N}$$

En primer lugar se obtienen las medias de las variables fila X y columna Y:

Se definen las dos cantidades siguientes:

$$S = Cov(X, Y) \quad ; \quad T = \sqrt{S(X) S(Y)}$$

donde los valores de Cov(X,Y), S(X) y S(Y) han sido descritos anteriormente, pero que pueden ser escritos asimismo en los siguientes términos:

$$Cov(X, Y) = \sum_i \sum_j n_{ij} (X_i - \overline{X})(Y_j - \overline{Y})$$

$$S(X) = \sum_i n_{i.} (X_i - \overline{X})^2$$

$$S(Y) = \sum_j n_{.j} (Y_j - \overline{Y})^2$$

A partir de estos valores se puede obtener, bajo el supuesto de la hipótesis alternativa, el valor de la variancia:

$$VAR\ 1 = \frac{1}{T^4} \sum_{ij} n_{ij} \left[T(X_i - \overline{X})(Y_j - \overline{Y}) - \frac{S}{2T} \left[(X_i - \overline{X})^2 S(Y) + (Y_j - \overline{Y})^2 S(X) \right] \right]^2$$

de forma que:

$$ASE1 = \sqrt{VAR1}$$

La variancia, bajo el supuesto de que la hipótesis nula es cierta, viene dada por:

$$VAR\ 0 = \frac{\sum_{ij} n_{ij} X_i^2 Y_j^2 - \frac{\left(\sum_{ij} n_{ij} X_i Y_j \right)}{N}}{\left(\sum n_{i.} X_i^2 \right) \left(\sum n_{.j} Y_j^2 \right)}$$

Por otra parte, el valor de t para probar la hipótesis nula de que el coeficiente de correlación es nulo, viene dado por:

$$t = \frac{r \sqrt{N-2}}{\sqrt{1 - r^2}}$$

cuya significación se realiza situando el valor de t en la distribución de Student definida por t(N-2).

Ejemplo

A efectos de obtener el valor del coeficiente de correlación y del error estándar ASE1, en una tabla 2x2, es más cómodo utilizar los valores 0 y 1 como códigos de las dos variables que forman la tabla, tal como se muestra a continuación:

Recuento

		VC		
		0	1	Total
VF	0	130	97	227
	1	20	3	23
Total		150	100	250

En primer lugar comprobaremos que el cambio de código numérico no afecta al valor del coeficiente de correlación. Calculamos los valores de Cov (X,Y), S(X) y S(Y):

$$\mathrm{Cov}(X,Y) = 3 - \frac{23 \times 100}{250} = -6.2$$

$$S(X) = 23 - \frac{(23)^2}{250} = 20.884$$

$$S(Y) = 100 - \frac{(100)^2}{250} = 60$$

A partir de ellos se obtiene el valor del coeficiente de correlación:

$$r = \frac{-6.2}{\sqrt{20.884 \times 60}} = -0.175149629$$

valor que coincide con el calculado anteriormente por la vía de los tamaños.

Para obtener el valor de ASE1 se obtienen, en primer lugar, las medias de X e Y:

$$\overline{X} = \frac{23}{250} = 0.092 \;\; ; \;\; \overline{Y} = \frac{100}{250} = 0.4$$

El valor de ASE1 viene dado por:

$$\mathrm{VAR}\,1 = \frac{1}{T^4} \sum_{ij} n_{ij} \left[T(X_i - \overline{X})(Y_j - \overline{Y}) - \frac{S}{2T} \left[(X_i - \overline{X})^2 S(Y) + (Y_j - \overline{Y})^2 S(X) \right] \right]^2$$

donde:

$$S = -6.2 \;\; ; \;\; T = \sqrt{1253.04}$$

Aplicando todos los datos se obtiene que:

$$\mathrm{VAR}\,1 = \frac{1}{(\sqrt{1253.04})^4} [130 \times (1.6397576)^2 + 97 \times (-1.2509252)^2 +$$

$$+ 20 \times (-8.2319014)^2 + 3 \times (24.275542)^2] = 0.00230845$$

$$\mathrm{ASE}\,1 = 0.0480463$$

Es interesante hacer notar que los programas SPSS y BMDP dan diferentes valores para la prueba t del coeficiente de correlación. A continuación se proporcionan los dos listados:

		Valor	Error típ. asint.	T aproximada	Sig. aproximada
Intervalo por intervalo	R de Pearson	-,17515	,04805	-2,80157	,00549

```
STATISTIC                      VALUE     ASE1 T-VALUE DEP.
-----------------------------------------------------------
CORRELATION COEFFICIENT       -0.175    0.048  -3.161
```

Así pues, mientras el SPSS proporciona el valor t=-2.80157, el BMDP proporciona el valor t=-3.161. La razón de esta diferencia radica en el hecho de que los dos programas utilizan métodos diferentes de cálculo.

El valor de la prueba t para contrastar la hipótesis nula, utilizada por el SPSS, es:

$$t = \frac{(-0.17515)\sqrt{248}}{\sqrt{1-(-0.17515)^2}} = -2.80$$

Si buscamos este valor en la distribución t con 248 grados de libertad, veremos que se encuentra entre los valores de α =0.01 y de α =0.002. Así pues, diríamos que se acepta la hipótesis alternativa con un grado de significación P<0.01. El listado del SPSS nos proporciona el grado de significación exacto P=0.00549.

El valor de la prueba t para contrastar la hipótesis nula, utilizada por el BMDP, se basa en el cálculo del error estándar del coeficiente de correlación bajo la hipótesis nula, que puede escribirse, asimismo, de la siguiente forma:

$$VAR\ 0 = \frac{1}{T^2}\left[\Sigma\Sigma n_{ij}(X_i - \overline{X})^2(Y_j - \overline{Y})^2 - \frac{S^2}{N}\right]$$

Así pues, el cálculo del error estándar ASE0 proporciona el valor:

$$VAR\ 0 = \frac{1}{(\sqrt{1253.04})^2}[(0.1760512) + (0.29556288) +$$

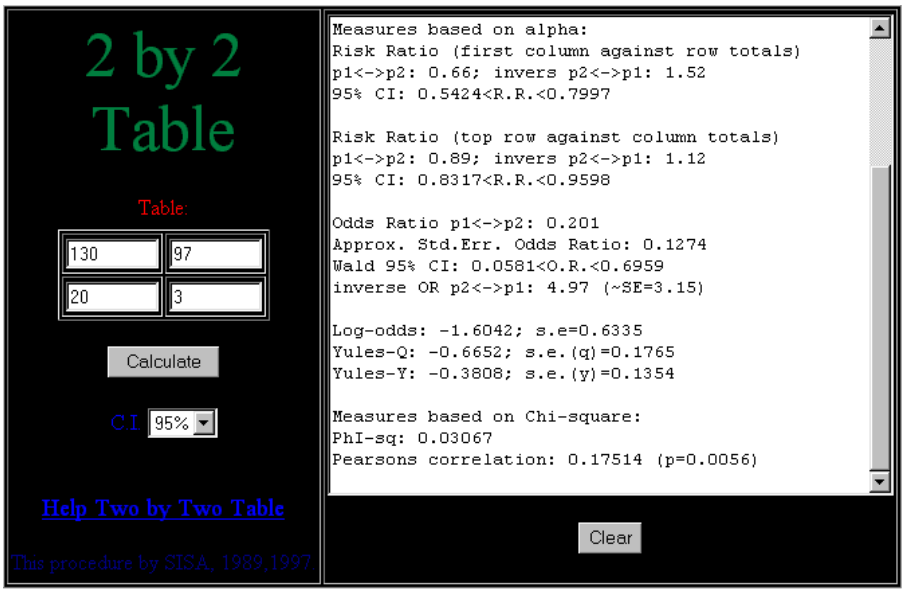
$$+ (2.6382848) + (0.89042112) - \frac{(-6.2)^2}{250}] = 0.00306978$$

$$ASE\ 0 = \sqrt{VAR\ 0} = 0.055405616$$

El valor de la prueba t viene dado, entonces, por:

$$t = \frac{r}{ASE\ 0} = \frac{-0.17515}{0.055405616} = -3.1612252$$

En la siguiente dirección se accede a un calculador de estadísticos e índices de asociación en tablas de contingencia 2x2, incluyendo el cálculo del coeficiente de correlación de Pearson: http://home.clara.net/sisa/twoby2.htm. Se debe introducir en dicho calculador los valores observados en las casillas de la tabla y pulsar el botón «Calculate». Entre los resultados obtenidos, el coeficiente de correlación de Pearson aparece bajo el nombre 'Pearsons correlation'. *SISA*. Ejemplo:



En la ultima línea aparece el índice que buscamos. Vemos como coincide su valor con el obtenido manualmente en el ejemplo anterior:

$$r = \frac{Cov(X,Y)}{\sqrt{S(X)S(Y)}} = 0.17514$$

4.2.- Coeficiente phi

El coeficiente phi, en tablas de contingencia 2x2, es equivalente al coeficiente de correlación de Pearson, por lo que ambos gozan de las mismas

propiedades. Su valor, una vez calculado el valor del ji-cuadrado de Pearson de la tabla, puede obtenerse, de forma sencilla, por medio de:

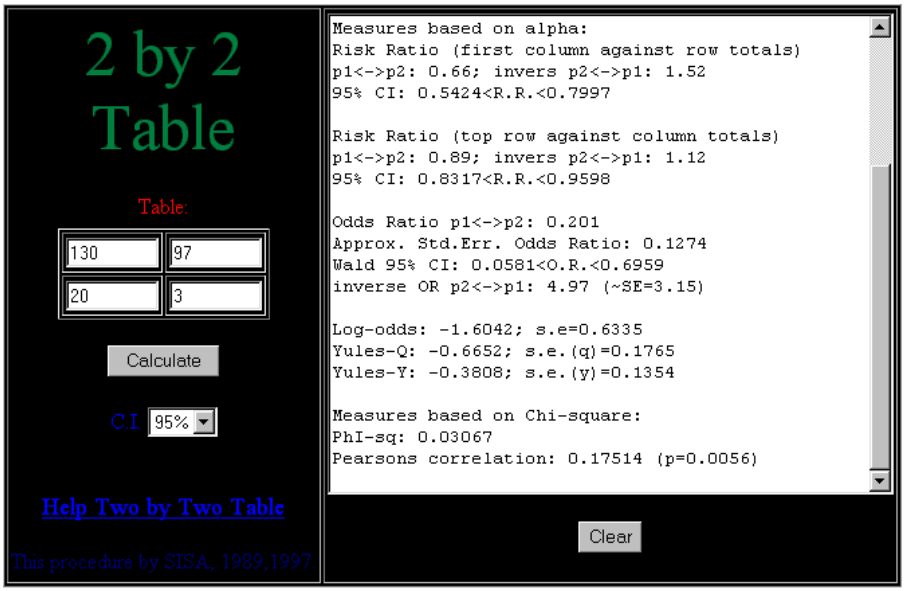
$$\phi = \sqrt{\frac{\chi^2}{N}}$$

En este caso, para el ejemplo visto en el apartado 4.1 se obtiene:

$$\phi = \sqrt{\frac{7.66937}{250}} = 0.175149$$

El inconveniente de este cálculo es que no sabemos qué signo presenta el coeficiente phi, ya que éste puede ser, en una tabla 2x2, positivo o negativo. Para ello, hemos de calcularlo tal como se calcula el coeficiente de correlación.

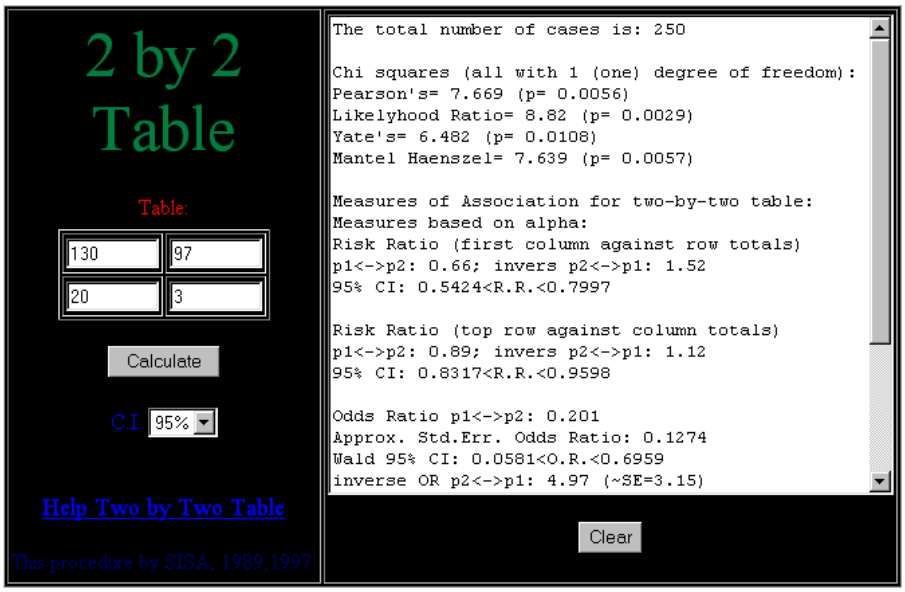
En la siguiente dirección se accede a un calculador de estadísticos e índices de asociación en tablas de contingencia 2x2, incluyendo el cálculo del coeficiente Phi cuadrado: http://home.clara.net/sisa/twoby2.htm. Se debe introducir en dicho calculador los valores observados en las casillas de la tabla y pulsar el botón «Calculate». Entre los resultados obtenidos, el coeficiente Phi cuadrado aparece bajo el nombre 'Phi-sq'. El coeficiente Phi se obtendría simplemente calculando la raíz cuadrada del coeficiente Phi cuadrado. *SISA*. Ejemplo:



En la penúltima línea aparece el índice que buscamos. Vemos como el valor de la raíz cuadrada de *Phi-sq* coincide con el valor de *Phi* calculado anteriormente:

$$\phi = \sqrt{\phi^2} = \sqrt{\frac{\chi^2}{N}} = \sqrt{0.03067} = 0.1751$$

El valor del ji-cuadrado de Pearson (X^2) es proporcionado también por el mismo *applet*, bajo el nombre *Pearson's*, tal como podemos ver al principio de la ventana de texto:



Valor máximo de phi

Fijados los valores totales de fila y de columna de la tabla 2x2, el valor máximo que puede tomar, en valor absoluto, el coeficiente phi viene dado (Dixon, 1988) por el siguiente esquema:

1. Si $n_{11}n_{22} < n_{12}n_{21}$ entonces el valor máximo de phi viene dado por el valor de la expresión siguiente cuyo valor absoluto sea menor que 1:

$$\phi_{max} = -\sqrt{\frac{n_{1.}n_{.1}}{n_{2.}n_{.2}}}$$

$$\phi_{max} = -\sqrt{\frac{n_{2.}n_{.2}}{n_{1.}n_{.1}}}$$

2. En caso contrario, el valor máximo de phi viene dado por el valor de la expresión siguiente que sea menor que 1:

$$\phi_{max} = \sqrt{\frac{n_{1.}n_{.2}}{n_{2.}n_{.1}}}$$

$$\phi_{max} = \sqrt{\frac{n_{2.}n_{.1}}{n_{1.}n_{.2}}}$$

Actualmente el SPSS proporciona el signo del coeficiente phi, pero no proporciona el valor máximo del coeficiente. El BMDP proporciona ambos resultados, tal como se puede comprobar en el siguiente listado (obtenido a través de la tabla de contigencia de ejemplo del apartado 4.1):

```
STATISTIC                      VALUE
  ------------------------------------
  PHI = CRAMER'S V              -0.175
  MAXIMUM VALUE FOR PHI         -0.260
```

El cálculo del valor máximo del coeficiente phi, supuesto fijos los totales marginales, viene dado, puesto que 130x3 < 97x20, por:

$$\phi_{max} = -\sqrt{\frac{227 \times 150}{23 \times 100}} = -3.8476$$

$$\phi_{max} = -\sqrt{\frac{23 \times 100}{227 \times 150}} = -0.259899$$

Así pues, ya que el primer valor es superior a 1, en valor absoluto, el valor máximo viene dado por el segundo valor.

4.3.- Correlación tetracórica

La *correlación de Pearson*, r, analiza la relación lineal entre dos variables cuantitativas; cuando una de las variables cuantitativas se dicotomiza, el cálculo de la relación entre ellas se obtiene por medio de la *correlación biserial*, r_b, mientras que cuando las dos variables cuantitativas se dicotomizan, el índice que permite estudiar la relación entre ellas es la *correlación tetracórica*, r_t.

Así pues, la correlación tetracórica presupone que las dos variables son esencialmente continuas y que además se distribuyen normalmente.

Sea la siguiente tabla de contingencia 2x2:

	1	0	Total
1	a	b	a + b
0	c	d	c + d
Total	a + c	b + d	N

El cálculo de la correlación tetracórica es algo complejo ya que debe ser obtenido por iteración a partir de una serie infinita en r_t, que viene dada (Amón, 1978) por medio de:

$$\frac{ad-cb}{N^2 yy'} = r_t + z'z\frac{r_t^2}{2} + (z^2-1)(z'^2-1)\frac{r_t^3}{6} + (z^3-3z)(z'^3-3z')\frac{r_t^4}{24} + \ldots$$

donde a,b,c,d son las frecuencias de la tabla de contingencia, N el tamaño de muestra total, z es la puntuación estándar que divide la distribución normal en dos áreas de valores (a+c)/N y (b+d)/N, z' es la puntuación estándar que divide la distribución normal en dos áreas de valores (c+d)/N y (a+b)/N. Los valores de y e y' son las ordenadas en la distribución normal correspondientes a los valores de z y z'.

Cuando los totales marginales de fila son similares, y los totales marginales de columna son similares, la siguiente expresión permite obtener el valor aproximado del coeficiente de correlación tetracórico:

$$r_t = \cos\left(\frac{180^\circ \sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right)$$

En el caso en el que el producto **bc** sea superior al producto **ad**, entonces la expresión a utilizar es:

$$r_t = \cos\left(\frac{180^\circ \sqrt{ad}}{\sqrt{ad}+\sqrt{bc}}\right)$$

Las expresiones anteriores nos permiten obtener la magnitud del valor. El signo de r_t es positivo si ad >bc, es decir si las concordancias superan a las discrepancias, y negativo si ad < bc. El signo de la correlación lo debe poner el experimentador en función del resultado obtenido.

Existen tablas que permiten obtener de forma fácil el valor aproximado de r_t conocido el valor del cociente ad/bc.

Ambas aproximaciones dan resultados útiles cuando los porcentajes marginales de fila y de columna se encuentran entre 0,40 y 0,60, y la muestra N es grande, siendo mejor a medida que N aumenta y los porcentajes se aproximan a 0,50 (caso en el que cada variable cuantitativa se dicotomiza por encima y por debajo de su mediana).

Ejemplo 1: Supongamos que en un ejemplo sobre vacunación y epidemia, la vacuna pueda ser una dosis y la infección pueda ser asimismo medida cuantitativamente. Al utilizar las dos variables dicotomizadas, se podría obtener la correlación tetracórica. Los datos observados son:

Recuento

	NO_VAC	VAC	Total
NO_INF	130	97	227
INF	20	3	23
Total	150	100	250

En primer lugar, puesto que 3x130 < 20x97 (ad<bc), el valor de la correlación será negativo. El valor aproximado viene dado por:

$$r_t = \cos\left(\frac{180^\circ \sqrt{3x130}}{\sqrt{3x130}+\sqrt{20x97}}\right) = -0.5632$$

Si buscamos en tablas, puesto que 1940/390 = 4.974, el valor hallado resulta ser 0.56, con signo negativo.

El valor exacto proporcionado por el BMDP es:

```
STATISTIC                    VALUE     ASE1 T-VALUE DEP.
-------------------------------------------------------
TETRACHORIC CORRELATION     -0.434    0.133  -3.089
```

Esto significa la existencia de una correlación negativa que, tal como están codificadas las variables, significa que a mayor dosis de vacuna menor infección.

Sin embargo, el valor exacto es bastante diferente del valor hallado de forma aproximada, ya que en este caso se puede comprobar que los porcentajes de fila son muy desiguales: 90.8% y 9.2%.

Ejemplo 2: Sean los datos correspondientes al estudio de la presencia o ausencia de un determinado síntoma a partir de dos instrumentos de medida, datos que se presentan a continuación, y supongamos que se han dicotomizado las dos variables fila y columna.

Recuento

	PRESENTE	AUSENTE	Total
PRESENTE	13	2	15
AUSENTE	7	8	15
Total	20	10	30

En primer lugar, puesto que 13x8 > 7x2 (ad>bc), el valor de la correlación será positivo.

El valor aproximado viene dado por:

$$r_t = \cos\left(\frac{180^\circ \sqrt{7 \times 2}}{\sqrt{7 \times 2} + \sqrt{13 \times 8}}\right) = 0.665$$

Si buscamos en tablas, puesto que 104/14 = 7.42857, el valor hallado resulta ser 0.66.

El valor exacto proporcionado por el BMDP es:

```
STATISTIC                      VALUE    ASE1 T-VALUE DEP.
-----------------------------------------------------------
TETRACHORIC CORRELATION        0.646   0.205   2.178
```

Así pues, podemos observar que, en este caso, al ser los porcentajes marginales bastante similares entre sí, los valores aproximados son valores bastante cercanos al valor exacto.

4.4.- Correlación de Spearman

A continuación vamos a exponer el cálculo del coeficiente de correlación de Spearman en una tabla de contingencia. El cálculo de esta correlación para dos variables cuantitativas se ha visto dentro de un artículo anterior creado por los mismos autores (Palmer, Jiménez y Montaño, 2000).

Para calcular dicho coeficiente en una tabla de contingencia es necesario que los niveles de cada una de las variables a relacionar esten codificados de forma ordinal.

El cálculo del coeficiente de correlación de Spearman viene dado por:

$$r_s = \frac{N^3 - N - 6\sum d^2 - 6(t' + u')}{\sqrt{N^3 - N - 12t'}\ \sqrt{N^3 - N - 12u'}}$$

donde t' y u' permite realizar la corrección por empates, lo que caracteriza a los datos de una tabla de contingencia, que se obtienen por medio de:

$$t' = \frac{\left(\sum t^3 - \sum t\right)}{12} \quad ; \quad u' = \frac{\left(\sum u^3 - \sum u\right)}{12}$$

El valor del sumatorio de d^2 viene dado por:

$$\sum d^2 = n_{ij}(R_i - R_j)^2$$

A efectos de práctica analizaremos los datos de la siguiente tabla de contingencia:

Recuento

		VC		
		POSITIVO	NEGATIVO	Total
VF	BAJA	10	35	45
	MEDIA	25	15	40
	ALTA	13	12	25
Total		48	62	110

Tenemos tres dosis de un determinado fármaco (1=Baja, 2=Media, 3=Alta) y su resultado (1=positivo, 2=negativo). Se trataría de saber qué asociación hay entre dosis y resultado y en qué dirección se da.

A efectos de obtener la correlación de Spearman, los datos originales deben ser sustituidos por su rango. Ahora bien, los 48 sujetos, de los 110, que han dado resultado positivo tienen todos ellos el mismo rango, por lo que a cada uno de ellos le correspondería el rango promedio, el cual se obtiene por:

(1+48)/2 = 24.5

A los 62 sujetos con resultado negativo le corresponden los rangos que van desde el 49 hasta el 110, por lo que el rango promedio resulta ser:

(49+110)/2 = 79.5

Estos dos valores constituyen los rangos R_j por columna.

Asimismo, para los 45 sujetos que han recibido dosis baja, le corresponden los rangos comprendidos entre 1 y 45, por lo que el rango promedio es:

(1+45)/2 = 23

Los 40 sujetos que recibieron dosis media, presentan un rango entre 46 y 85 con un rango promedio de valor:

(46+85)/2 = 65.5

Finalmente, los 25 sujetos que recibieron dosis alta se encuentran desde el rango 86 hasta el rango 110, por lo que su rango promedio será:

(86+110)/2 = 98

Estos tres valores constituyen los rangos R_i por fila.

Con los datos de la tabla original y los rangos promedios, se construye la siguiente tabla que nos permitirá obtener fácilmente el valor del sumatorio de los valores d^2:

VF	VC	n_{ij}	R_i	R_j	d^2
1	1	10	23	24.5	22.5
1	2	35	23	79.5	111728.75
2	1	25	65.5	24.5	42025
2	2	15	65.5	79.5	2940
3	1	13	98	24.5	70229.25
3	2	12	98	98	4107
					231052.5

Por otro lado, a continuación se realizan los cálculos correspondientes para la corrección por empates:

t	t^3	u	u^3
45	91125	48	110592
40	64000	62	23838
25	15625	110	348920
110	170750		
$t' = \frac{170750 - 110}{2}$		$u' = \frac{348920 - 110}{2}$	
$= 14220$		$= 29067.5$	

Con todos los cálculos anteriores puede obtenerse el valor del coeficiente de correlación de Spearman en la tabla de contingencia utilizada:

$$r_s = \frac{110^3 - 110 - 6(231052.5) - 6(14220 + 29067.5)}{\sqrt{110^3 - 110 - 12(14220)}\sqrt{110^3 - 110 - 12(29067.5)}} = -0.295235$$

Podemos probar la hipótesis nula de independencia a través de la estandarización de dicho valor bajo la distribución N(0,1):

$$Z = r_s\sqrt{n-1} = -0.295235\sqrt{110-1} = -3.08234$$

valor que bajo la distribución N(0,1) da lugar a una probabilidad bilateral inferior a 0.002, es decir, que la probabilidad de equivocarnos si rechazamos la hipótesis nula de independencia entre variables es inferior al 0,2%.

Los resultados proporcionados por el SPSS vienen dados a continuación:

		Valor	Error típ. asint.	T aproximada	Sig. aproximada
Ordinal por ordinal	Correlación de Spearman	-,29524	,09049	-3,21132	,00174

Los datos indican que existe una relación estadísticamente significativa entre la variable independiente 'dosis' y la respuesta (p=0.00174) y que además esta relación tiene un sentido negativo (r_s=-0.295235), es decir, que a mayor dosis disminuye la frecuencia de resultados negativos.

4.5.- Tau-a de Kendall

Este índice, al igual que el coeficiente de correlación de Spearman, está indicado en el caso en que queramos medir la asociación existente en dos variables X e Y categóricas, con categorías codificadas de forma ordinal. Sin embargo, mide dicha asociación de distinta manera a como lo hace el coeficiente de correlación de Spearman.

El índice Tau-a de Kendall está basado en la noción de concordancia (Palmer, 1996b), teniendo en cuenta que no se realiza corrección por empates. Su interpretación es la diferencia entre la proporción de pares concordantes y de pares discordantes.

Dados dos pares de observaciones (X_i,Y_i), (X_j,Y_j):

Los pares son concordantes si: $(X_i-X_j)(Y_i-Y_j)>0$

Los pares son discordantes si: $(X_i-X_j)(Y_i-Y_j)<0$

Los pares están ligados o empatados si: $(X_i-X_j)(Y_i-Y_j)=0$

El número total de pares de observaciones en una muestra de tamaño N viene dado por N(N-1)/2.

Para obtener el número de pares concordantes (C) en una tabla de contingencia, se multiplica la frecuencia de cada casilla por la suma de las frecuencias de las casillas que quedan por debajo y a la derecha de ella.

Para obtener el número de pares discordantes (D) en una tabla de contingencia, se multiplica la frecuencia de cada casilla por la suma de las frecuencias de las casillas que quedan por debajo y a la izquierda de ella.

Definimos S como el número de pares concordantes menos el número de pares discordantes:

$$S=C-D$$

La estimación de la Tau-a de Kendall viene dado por:

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

Este valor está comprendido entre -1 (todos los pares son discordantes, lo que indica una perfecta asociación negativa entre variables) y +1 (todos los pares son concordantes, lo que indica una perfecta asociación positiva entre variables), tomando el valor cero cuando X e Y son independientes.

Para probar la hipótesis nula $H_o:\tau=0$, se puede utilizar la aproximación:

$$z = \frac{\hat{\tau}_a}{SE(\hat{\tau}_a)}$$

siendo:

$$SE(\hat{\tau}_a) = \sqrt{\frac{2(2N+5)}{9N(N-1)}}$$

El índice Z sigue una distribución N(0,1). De hecho esta aproximación es buena para n > 10, mientras que para valores inferiores de n es mejor buscar las probabilidades exactas en la tabla correspondiente.

Ejemplo

Calcularemos Tau-a a partir de la misma tabla de contingencia que fue utilizada para ejemplificar el cálculo del coeficiente de correlación de Spearman (apartado anterior):

Recuento

		VC		
		POSITIVO	NEGATIVO	Total
VF	BAJA	10	35	45
	MEDIA	25	15	40
	ALTA	13	12	25
Total		48	62	110

Tenemos tres dosis de un determinado fármaco (1=Baja, 2=Media, 3=Alta) y su resultado (1=positivo, 2=negativo).

Cálculo del número de pares concordantes (C) y del número de pares discordantes (D):

$$C=10(15+12)+25(12)=570$$

Interpretación del cálculo de C: Los pares concordantes con los 10 (1,1) pares son aquellos con los siguientes rangos: (2,2) y (3,2). Los pares concordantes con los 25 (2,1) pares son aquellos con el siguiente rango: (3,2). En todos ellos se cumple que $(X_i-X_j)(Y_i-Y_j)>0$.

$$D=35(25+13)+15(13)=1525$$

Interpretación del cálculo de D: Los pares discordantes con los 35 (1,2) pares son aquellos con los siguientes rangos: (2,1) y (3,1). Los pares discordantes con los 15 (2,2) pares son aquellos con el siguiente rango: (3,1). En todos ellos se cumple que $(X_i-X_j)(Y_i-Y_j)<0$.

A continuación se calcula el valor de S:

$$S=C-D=-955$$

Así pues:

$$\tau_a = \frac{S}{\frac{N(N-1)}{2}} = \frac{-955}{\frac{110(109)}{2}} = -0.1593$$

Notar que si X e Y son independientes entonces $\tau = 0$, pero que si $\tau = 0$ no necesariamente significa que X e Y sean independientes.

Probaremos la hipótesis nula de Tau-a:

$$z = \frac{\tau_a}{SE(\tau_a)} = \frac{-0.1593}{\sqrt{\frac{2(2(110)+5)}{9(110)(109)}}} = -2.4668$$

valor que bajo la distribución N(0,1) nos permite concluir que existe relación entre la dosis del fármaco y el resultado, con una probabilidad de equivocarnos inferior a 0.0138 en modo bilateral.

Si las X representan puntuaciones de un individuo y las Y representan una ordenación objetiva conocida, entonces los empates no indican concordancia y Tau-a es la medida recomendada por Kendall.

El índice Tau-a es una correlación de rangos que, según Everitt (1977), no es aplicable a tablas de contingencia ya que este índice asume que no hay observaciones empatadas.

4.6.- Correlación eta

Ya se ha explicado en el apartado 2.2.1 que la correlación eta se puede utilizar en aquel caso en el que busquemos la relación existente entre una variable independiente medida en una escala nominal y una variable dependiente medida en una escala de intervalo o de razón. En este apartado calcularemos eta en el caso concreto de que la variable dependiente se halla previamente categorizado.

Cuando X (VF) es la variable dependiente:

$$\eta_X = \sqrt{1 - \frac{S_X}{S(X)}}$$

Cuando Y (VC) es la variable dependiente:

$$\eta_Y = \sqrt{1 - \frac{S_Y}{S(Y)}}$$

A continuación se muestra la siguiente tabla de contingencia que relaciona la terapia recibida por una serie de sujetos (variable independiente nominal) con el resultado obtenido (positivo o negativo), variable dependiente medida originalmente en una escala de intervalo (puntuaciones) y que ha sido dicotomizada:

Recuento

		VC		
		POSITIVO	NEGATIVO	Total
VF	FARMACOLOGICA	10	35	45
	CONDUCTUAL	25	15	40
	PSICOANALITICA	13	12	25
Total		48	62	110

Las expresiones para la obtención de S(X) y de S(Y), así como el cálculo de los valores de S(X)=66.363636 y S(Y)=27.054545 para el ejemplo utilizado, se ha visto y desarrollado en el cálculo del coeficiente de correlación de Pearson. Los valores de S_X y de S_Y se obtienen mediante las siguientes expresiones:

$$S_X = \sum X_i^2 n_{ij} - \sum_{j=1}^{c} \frac{1}{n_{.j}} \left(\sum_{i=1}^{r} X_i n_{ij} \right)^2$$

$$S_Y = \sum Y_j^2 n_{ij} - \sum_{i=1}^{r} \frac{1}{n_{i.}} \left(\sum_{j=1}^{c} Y_j n_{ij} \right)^2$$

Para los datos del ejemplo, en los que la variable fila toma valores de 1 a 3 y la variable columna toma los valores 1 y 2, los resultados del índice eta, ofrecidos por el SPSS, vienen dados por:

			Valor
Nominal por intervalo	Eta	VF dependiente	,27677
		VC dependiente	,36790

El cálculo manual se realiza de la siguiente forma. Se calculan en primer lugar los valores de S(X) y de S(Y), así como los valores de S_X y de S_Y:

$$S_X = 430\left(\frac{(10+50+39)^2}{48}+\frac{(35+30+36)^2}{62}\right)=61.280242$$

$$S_Y = 296\left(\frac{(10+70)^2}{45}+\frac{(25+30)^2}{40}+\frac{(13+24)^2}{25}\right)=23.392777$$

Se sustituyen los valores en la expresión correspondiente:

$$\eta_{VF}=\sqrt{1-\frac{61.280242}{66.363636}}=0.2767654$$

$$\eta_{VC}=\sqrt{1-\frac{23.392777}{27.054545}}=0.3678962$$

En este caso, como la VC es la variable dependiente, el valor de eta válido sería η_{vc}=0.3678962, valor que elevado al cuadrado nos indicaría que el porcentaje de variabilidad de la variable dependiente explicado por la variable independiente es del 13,53%.

BIBLIOGRAFIA

AMON, J. (1978). *Estadística para psicólogos 1. Estadística descriptiva*. Madrid: Pirámide.

BRUNING, J.L. y KINTZ, B.L. (1987). *Computational handbook of statistics*. 3rd edition. London: Scott, Foresman and Company.

CURENTON, E.E. (1957). Rank-biserial correlation. *Psychometrika, 21, 287-290*.

DIXON, W.J. (Ed.) (1988). *BMDP statistical software manual*. Berkeley, CA: University of California Press.

EVERITT, B.S. (1977). *The analysis of contingency tables*. London: Chapman and Hall.

GLASS, G.V. (1966). Note on rank biserial correlation. *Educational and Psychological Measurement, 26, 622-631*.

GIBBONS, J.D. (1993). *Nonparametric measures of association*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-091. Newbury Park, CA: Sage.

MUÑIZ, J. (1992). *Teoría clásica de los test*. Madrid: Pirámide.

PALMER, A. (1995a). *Tablas de estadística*. Palma de Mallorca: Servei de Publicacions i Intercanvi Científic. Col.lecció Materials Didàctics, 7. Sèrie de Metodologia. Universitat de les Illes Balears.

PALMER, A. (1995b). *Análisis del diseño completamente aleatorizado I: Comparación de dos medias*. Palma de Mallorca: Servei de Publicacions i Intercanvi Científic. Col.lecció Materials Didàctics, 8. Sèrie de Metodologia. Universitat de les Illes Balears.

PALMER, A. (1996a). *Análisis del diseño completamente aleatorizado II: Análsis de la variancia unifactorial*. Palma de Mallorca: Servei de Publicacions i Intercanvi Científic. Col.lecció Materials Didàctics, 9. Sèrie de Metodologia. Universitat de les Illes Balears.

PALMER, A. (1996b). *El análisis de tablas de contingencia bivariantes*. Palma de Mallorca: Servei de Publicacions i Intercanvi Científic. Col.lecció Materials Didàctics, 15. Sèrie de Metodologia. Universitat de les Illes Balears.

PALMER, A., JIMENEZ, R. Y MONTAÑO, J.J. (2000). *Tutorial sobre el coeficiente de correlación lineal de Pearson*.