Nota estadística

Asociación entre variables: correlación no paramétrica

Jorge Camacho-Sandoval

Resumen

Se describe la forma de estimar el coeficiente de correlación de Spearman y las condiciones en que resulta apropiada su utilización. También se describe como realizar una prueba de hipótesis para determinar si el coeficiente estimado es significativamente distinto de cero.

Abstract

A description is made on how to estimate the Spearman correlation coefficient and in what conditions it is appropriate to use it. Also, how to make hypotheses tests to determine if the estimated coefficient is significantly different from zero is described.

En el número anterior (ACM 50: 94-96) se hizo referencia al análisis de correlación lineal entre dos variables cuantitativas, conocido como el análisis de correlación de Pearson. Ese método requiere, cuando se desea hacer pruebas de hipótesis o estimar intervalos de confianza del coeficiente de correlación, como es usual, que las variables tengan una distribución normal bivariada.

¿Qué pasa cuando las variables no tienen una distribución normal bivariada? Eso ocurre, por ejemplo, cuando una o ambas variables se miden en una escala ordinal o de intervalo. Para ese caso existen pruebas que no exigen ese requisito. Dos de ellas son las pruebas de Spearman y de Kendall; ambas utilizan, en vez de los valores de las variables, sus rangos, es decir, el número de orden del valor de cada observación de la variable dentro del conjunto de observaciones. La prueba de Spearman tiene la ventaja de ser muy sencilla de calcular.

Para estimar el coeficiente de correlación de Spearman, primero se deben obtener los rangos para cada una de las observaciones de ambas variables. Para ello se considera una variable y se asigna el rango 1 al valor más pequeño, 2 al siguiente valor más pequeño y así sucesivamente hasta llegar al rango n que le corresponde a la observación con el valor más alto. Luego se repite el procedimiento para la otra variable.

El coeficiente de correlación de Spearman, r_s, se puede obtener con la siguiente fórmula:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

En donde n es el número de casos o pacientes y d es la diferencia entre los rangos de las variables para cada paciente o unidad de observación. No obstante, esa fórmula supone que no hay valores repetidos, es decir que no hay 2 o más pacientes a los que les corresponda el mismo rango para una misma variable. Si existen pacientes con valores repetidos, se les asigna a esos pacientes el rango promedio y se usa una fórmula de cálculo alternativa.

Profesor del Programa de Maestría en Epidemiología, Posgrado en Ciencias Veterinarias, UNA.
Correspondencia:
Correo electrónico:
jcamacho@ice.co.cr

ISSN 0001-6002/2008/50/3/144-146
Acta Médica Costarricense, ©2008
Colegio de Médicos y Cirujanos

Cuadro 1. Puntuación de actividad motora en extremidades inferiores y balance en pacientes con accidentes cerebro vasculares

Paciente	Actividad Motora (AM)	Balance (B)	Rango AM	Rango B	$(R.\ AM)^2$	$(R.\ B)^2$	(R. AM) x (R. B)	(R. AM) - (R. B)	$((R.\ AM) - (R.\ B))^2$
1	30	12	17.5	15	306.25	225.00	262.50	2.5	6.25
2	11	1	4	1.5	16.00	2.25	6.00	2.5	6.25
3	30	14	17.5	18.5	306.25	342.25	323.75	-1	1
4	20	14	10	18.5	100.00	342.25	185.00	-8.5	72.25
5	30	6	17.5	7	306.25	49.00	122.50	10.5	110.25
6	30	14	17.5	18.5	306.25	342.25	323.75	-1	1
7	30	13	17.5	16	306.25	256.00	280.00	1.5	2.25
8	22	3	12	4.5	144.00	20.25	54.00	7.5	56.25
9	18	9	7	11.5	49.00	132.25	80.50	-4.5	20.25
10	20	9	10	11.5	100.00	132.25	115.00	-1.5	2.25
11	30	14	17.5	18.5	306.25	342.25	323.75	-1	1
12	3	2	2	3	4.00	9.00	6.00	-1	1
13	14	3	6	4.5	36.00	20.25	27.00	1.5	2.25
14	12	4	5	6	25.00	36.00	30.00	-1	1
15	6	1	3	1.5	9.00	2.25	4.50	1.5	2.25
16	24	10	13.5	14	182.25	196.00	189.00	-0.5	0.25
17	24	8	13.5	9	182.25	81.00	121.50	4.5	20.25
18	20	7	10	8	100.00	64.00	80.00	2	4
19	19	9	8	11.5	64.00	132.25	92.00	-3.5	12.25
20	2	9	1	11.5	1.00	132.25	11.50	-10.5	110.25
		Suma	210	210	2850.00	2859.00	2638.25		432.50

A manera de ejemplo se considera un estudio que se realizó para comparar herramientas de evaluación para determinar la recuperación de pacientes que sufrieron accidentes cerebro vasculares. Se evaluaron numerosas variables, entre ellas el balance y la función motora de los miembros inferiores. En ambos casos se asignó un puntaje, con un máximo de 14 puntos para el balance y de 30 para la actividad motora (http://www.statsci.org/data/oz/strokeass.html).

En el estudio, cuyos datos se muestran en el Cuadro 1, se observa que algunos pacientes tienen el mismo valor para las variables en consideración, es decir, existen valores repetidos. Por ejemplo, los pacientes 16 y 17 tienen un puntaje en actividad motora de 24 y les corresponde la posición o rango 13 y 14 si se ordenan los datos ascendentemente; el rango promedio de esas dos observaciones es 13.5, por lo tanto es el rango que se les asigna a ambos pacientes para la variable correspondiente.

Uno de los objetivos del estudio puede ser determinar si existe correlación entre ambas variables pero como éstas no siguen una distribución normal bivariada, no es posible utilizar la correlación de Pearson. Si se estima el coeficiente de correlación de Spearman, utilizando la fórmula anterior, se obtiene el siguiente valor:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} = 1 - \frac{6\,(432.50)}{20^3 - 20} = 0.675$$

Por otra parte, como existen valores repetidos en ambas variables y para varios pacientes, es necesario introducir una corrección o utilizar formas alternativas de cálculo. Existen varias opciones, una de ellas es aplicar la fórmula utilizada para obtener el coeficiente de correlación de Pearson, pero utilizando los rangos en vez de los valores o puntajes de las variables, como ya se mencionó.

De esa manera, el coeficiente de correlación de Spearman es:

$$r_S = \frac{\sum_{i=1}^{n} XY - \frac{\sum_{i=1}^{n} X \sum_{i=1}^{n} Y}{n}}{\sqrt{\left(\sum_{i=1}^{n} X^2 - \frac{\left(\sum_{i=1}^{n} X\right)^2}{n}\right)\left(\sum_{i=1}^{n} Y^2 - \frac{\left(\sum_{i=1}^{n} Y\right)^2}{n}\right)}} = \frac{2638.25 - \frac{(210)(210)}{20}}{\sqrt{\left(2850 - \frac{(210)^2}{20}\right)\left(2859 - \frac{(210)^2}{20}\right)}} = 0.667$$

Como se observa, los valores obtenidos por ambos métodos son muy similares y pueden usarse indiferentemente si el número de valores repetidos no es muy elevado. Si hay muchos valores repetidos, es mejor utilizar la última forma de cálculo.

Una vez estimado el coeficiente de correlación de Spearman, es conveniente realizar una prueba de hipótesis con la H_0: $r_S=0$; H_1: $r_S\neq0$. Para decidir si se rechaza o no la hipótesis nula (H_0), se utiliza un valor crítico correspondiente al nivel de significancia deseado. Esos valores se pueden obtener en muchos textos o en la red (http://www.atozee.co.uk/S151/spearman3.html). En el Cuadro 2 se muestran los valores críticos para distinto número de casos y niveles de significancia.

En el caso del ejemplo, para 20 pacientes y una prueba de hipótesis con un nivel de significancia del 5%, el valor crítico es 0.45. Como el valor estimado del coeficiente de correlación de Spearman, 0.67, es superior al valor crítico, se rechaza la hipótesis nula y se concluye que el coeficiente de correlación es significativamente distinto de cero, es decir que existe una asociación significativa entre la actividad motora de las extremidades inferiores y el balance, en pacientes que sufrieron accidentes cerebro vasculares.

Cuadro 2. Valores críticos para una prueba de dos colas del Coeficiente de Correlación de Spearman

Número de casos	Valor crítico para α:		
	P=0.1	P=0.05	P=0.01
7	0.714	0.786	0.929
8	0.643	0.738	0.881
9	0.600	0.683	0.833
10	0.564	0.648	0.794
12	0.506	0.591	0.777
14	0.456	0.544	0.715
16	0.425	0.506	0.665
18	0.399	0.475	0.625
20	0.377	0.450	0.591
22	0.359	0.428	0.562
24	0.343	0.409	0.537
26	0.329	0.392	0.515
28	0.317	0.377	0.496
30	0.306	0.364	0.478

Referencias

1. Camacho-Sandoval, J. Asociación entre variables cuantitativas: análisis de correlación. Acta Médic Costarric 2008; 50: 94-96.
2. Siegel, S. & Castellan, N. J. 1998. Estadística no paramétrica aplicada a las ciencias de la conducta. 4ª Ed. Trillas, México. 437 p.
3. Zar, J. 1999. Biostatistical Analysis. 4th Ed. Prentice Hall, New Jersey. 663 pp.