Nota estadística

Asociación entre variables cuantitativas: análisis de correlación

Jorge Camacho-Sandoval

En la investigación clínica frecuentemente se miden numerosas variables en los individuos incluidos en el estudio. Muchas veces interesa determinar si existe relación entre algunas de esas variables, o predecir el valor de una de ellas conociendo el valor de otras. En ocasiones interesa determinar si distintos instrumentos, métodos o personas obtienen valores similares cuando se mide una variable en las mismas unidades experimentales. Esos tres objetivos requieren métodos de análisis distintos. La presente nota se refiere al primer objetivo: determinar si existe asociación entre variables.

En el gráfico de dispersión del índice de tabaquismo y el índice de mortalidad por cáncer de pulmón adjunto, se puede observar que conforme se incrementa el índice de tabaquismo, se incrementa de forma lineal, el índice de mortalidad. Es decir, se puede representar la asociación entre esas variables, con una línea recta. En el segundo gráfico de dispersión, entre el número de cigarrillos fumados y la mortalidad por cáncer de riñón, se observa una relación que no es lineal, sino curvilínea. En ambos casos las variables están relacionadas, pero la forma de la relación es distinta.

El método más común de determinar si existe asociación **lineal** entre dos variables cuantitativas continuas es el Análisis de Correlación de Pearson. Con este método se obtiene el Coeficiente de Correlación de Pearson, usualmente representado por la letra R. Como suele utilizarse una muestra, lo que se obtiene en realidad es un estimado del coeficiente de correlación poblacional, r.

Dos aspectos importantes del coeficiente de correlación son su magnitud y su signo. La magnitud refleja la intensidad de la asociación entre las dos variables; el valor absoluto de la magnitud puede variar entre cero y uno. Valores cercanos a cero indican que las variables no están asociadas, es decir, que el valor de una variable es independiente del valor de la otra.

El signo, por su parte, refleja cómo están asociados los valores de ambas variables. Si el signo es positivo indica que a valores altos de una variable corresponden valores altos de la otra, o a valores bajos de una variable corresponden valores bajos de la otra. Si el signo es negativo, indica que a valores altos de una variable corresponden valores bajos de la otra. Es decir, el sigo positivo indica que los valores de ambas variables cambian en el mismo sentido, mientras que el signo negativo indica que cambian en sentido contrario. En la fórmula se observa que las unidades de ambas variables aparecen en el numerador y denominador, por lo tanto, se anulan aritméticamente, por lo que el coeficiente de correlación no tiene unidades de medición.

El cálculo del coeficiente de correlación es muy sencillo. Si se supone que se tienen dos variables cuantitativas continuas, por ejemplo, el número promedio de cigarros consumidos en cientos por persona (X), y la tasa de mortalidad por cáncer de pulmón en 15 localidades, en muertes por cien mil habitantes (Y), como se muestra en el Cuadro 1, una de las formas de cálculo es la siguiente (Zar, 1999):

$$r = \frac{\sum_{i=1}^{n} XY - \frac{\sum_{i=1}^{n} X \sum_{i=1}^{n} Y}{n}}{\sqrt{\left(\sum_{i=1}^{n} X^2 - \frac{\left(\sum_{i=1}^{n} X\right)^2}{n}\right)\left(\sum_{i=1}^{n} Y^2 - \frac{\left(\sum_{i=1}^{n} Y\right)^2}{n}\right)}} = \frac{8188.63 - \frac{(387.46)(306.39)}{15}}{\sqrt{\left(10530.73 - \frac{(387.46)^2}{15}\right)\left(6484.80 - \frac{(306.39)^2}{15}\right)}} = 0.80$$

Profesor, Maestría en Epidemiología, Postgrado en Ciencias Veterinarias, UNA.
Correspondencia:
Correo electrónico:
jcamacho@ice.co.cr

ISSN 0001-6002/2008/50/2/94-96
Acta Médica Costarricense, ©2008
Colegio de Médicos y Cirujanos

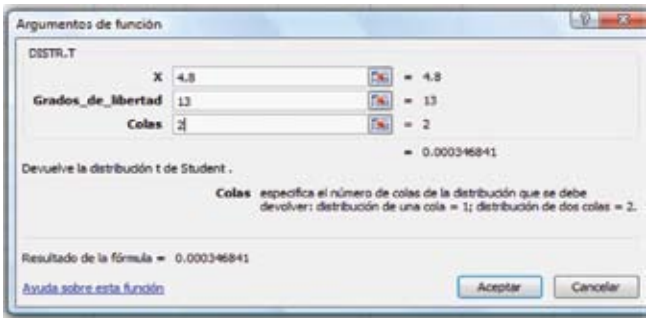


Figura 1. Probabilidad de la distribución t en MS Excel

En el ejemplo se encontró una alta correlación positiva entre las variables, con un coeficiente de correlación de 0.80. Para obtener el estimado del coeficiente de correlación no es necesario conocer la distribución de probabilidad de las variables; sin embargo, como se obtiene a partir de una muestra es preciso obtener indicadores de la variabilidad del estimado, como su error estándar o un intervalo de confianza. También es posible realizar pruebas de hipótesis, por ejemplo, para determinar si el coeficiente es estadísticamente diferente de cero. Para todo ello se requiere que las variables cumplan ciertos supuestos, específicamente, que tengan una distribución normal bivariada.

El error estándar de la correlación se calcula de la siguiente manera (Zar, 1999):

$$s_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-(0.80)^2}{15-2}} = 0.1664$$

Una prueba de hipótesis sobre el coeficiente de correlación se puede establecer en los términos siguientes: hipótesis nula H_o: r=0; hipótesis alternativa H_1: r≠0; estadístico de prueba: $t = \frac{r}{s_r}$ con n-2 grados de libertad y se rechaza la hipótesis nula sí $P(t)\langle 0.05$. En el caso del ejemplo $t = \frac{0.80}{0.1664} = 4.8$ con una probabilidad de 0.0003 (Figura 1), por lo tanto, se rechaza la hipótesis nula y se concluye que el coeficiente de correlación es significativamente distinto de cero.

Cuadro 1. Cigarrillos consumidos (cientos por persona) y mortalidad por cáncer de pulmón (muertes/100000 habitantes) en 15 localidades.

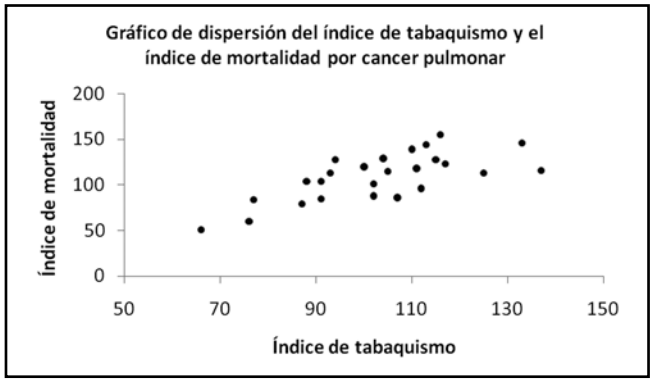
Localidad	Cigarrillos (X)	Mortalidad (Y)	XY	X^2	Y^2
1	18.20	17.05	310.31	331.24	290.70
2	25.82	19.80	511.24	666.67	392.04
3	18.24	15.98	291.48	332.70	255.36
4	28.60	22.07	631.20	817.96	487.08
5	31.10	22.83	710.01	967.21	521.21
6	33.60	24.55	824.88	1128.96	602.70
7	40.46	27.27	1103.34	1637.01	743.65
8	28.27	23.57	666.32	799.19	555.54
9	20.10	13.58	272.96	404.01	184.42
10	27.91	22.80	636.35	778.97	519.84
11	26.18	20.30	531.45	685.39	412.09
12	22.12	16.59	366.97	489.29	275.23
13	21.84	16.84	367.79	476.99	283.59
14	23.44	17.71	415.12	549.43	313.64
15	21.58	25.45	549.21	465.70	647.70
Suma	387.46	306.39	8188.63	10530.73	6484.80

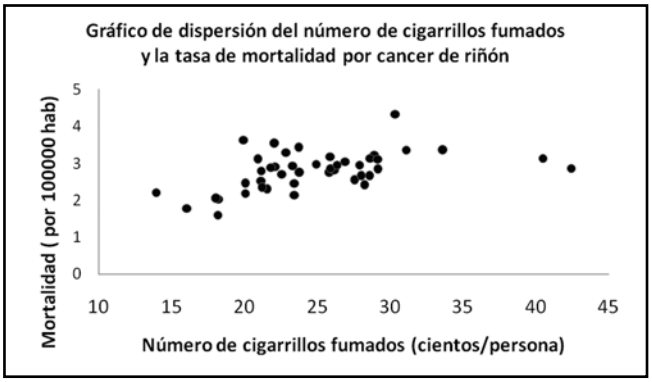
Fuente: (http://lib.stat.cmu.edu/DASL/Datafiles/cigcancerdat.html)

Si se prefiere, se puede construir un intervalo de confianza para el coeficiente de correlación. Como este no se distribuye normalmente, se debe realizar una transformación, de manera que el coeficiente de correlación transformado sí lo haga. La transformación se obtiene como:

$$z = 0.5\left(Ln\left[\frac{1+r}{1-r}\right]\right) = 0.5\left(Ln\left[\frac{1+0.8}{1-0.8}\right]\right) = 1.1$$

y su error estándar es: $s_z = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{15-3}} = 0.29$





En donde L_n se refiere al logaritmo natural o neperiano. El intervalo de confianza del coeficiente **transformado** se obtiene de forma convencional (Camacho, 2007). En el presente caso el intervalo de confianza del 95% se consigue de la siguiente manera:

Límite inferior:

$$z_1 = z - \frac{1.96}{\sqrt{n-3}} = 1.1 - \frac{1.96}{\sqrt{15-3}} = 0.53$$;

Límite superior:

$$z_2 = z + \frac{1.96}{\sqrt{n-3}} = 1.1 + \frac{1.96}{\sqrt{15-3}} = 1.67$$

El valor de los límites se refiere al coeficiente de correlación **transformado (z)**, por lo que se debe realizar el proceso inverso para obtener el intervalo de confianza de r. Los límites inferior y superior se obtienen como:

$$L_i = \frac{e^{2z_1} - 1}{e^{2z_1} + 1} = \frac{2.72^{(2\times 0.53)} - 1}{2.72^{(2\times 0.53)} + 1} = 0.49$$, y

$$Ls = \frac{e^{2z_2} - 1}{e^{2z_2} + 1} = \frac{2.72^{(2\times 1.67)} - 1}{2.72^{(2\times 1.67)} + 1} = 0.93$$

Es decir, se tiene un 95% de confianza de que el coeficiente de correlación en la población esté entre 0.49 y 0.93. La letra e representa la base de los logaritmos neperianos o naturales (2.72).

En una próxima nota se considerará el caso de variables que no cumplen el requisito de tener distribución normal bivariada.

Bibliografía

1. Camacho, J. 2007. ¿Hay diferencias significativas entre tratamientos? Primera parte. Acta Médica Costarricense 49(2): 81-82.
2. Zar, J. 1999. Biostatistical Analysis. 4th Ed. Prentice Hall, New Jersey. 663 pp.