Determinación del tamaño muestral para calcular la significación del coeficiente de correlación lineal

Autores: **Pértegas Díaz, S.** spertega@canalejo.org, **Pita Fernández, S.** spita@canalejo.org
Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Juan Canalejo. A Coruña.
Cad Aten Primaria 2001; 2002; 9: 209-211. Actualización 18/11/2002.

El coeficiente de correlación lineal de Pearson

En el análisis de estudios clínico-epidemiológicos con frecuencia interesa estudiar, a partir de los datos de un grupo de individuos, la posible asociación entre dos variables. En el caso de datos cuantitativos ello implica conocer si los valores de una de las variables tienden a ser mayores (o menores) a medida que aumentan los valores de la otra, o si no tienen nada que ver entre sí. La *correlación* es el método de análisis adecuado cuando se precisa conocer la posible relación entre dos variables de este tipo. Así, el grado de asociación entre dos variables numéricas puede cuantificarse mediante el cálculo de un *coeficiente de correlación*[1-5]. Debe entenderse, no obstante, que el coeficiente de correlación no proporciona necesariamente una medida de la causalidad entre ambas variables sino tan sólo del grado de relación entre las mismas[6].

La medida más habitualmente utilizada para el estudio de la correlación es el *coeficiente de correlación lineal de Pearson*. El coeficiente de Pearson mide el grado de asociación lineal entre dos variables cualesquiera, y puede calcularse dividiendo la covarianza de ambas entre el producto de las desviaciones típicas de las dos variables[1]. Para un conjunto de datos, el valor *r* de este coeficiente puede tomar cualquier valor entre –1 y +1. El valor de *r* será positivo si existe una relación directa entre ambas variables, esto es, si las dos aumentan al mismo tiempo. Será negativo si la relación es inversa, es decir, cuando una variable disminuye a medida que la otra aumenta. Un valor de +1 ó –1 indicará una relación lineal perfecta entre ambas variables, mientras que un valor 0 indicará que no existe relación lineal entre ellas. Hay que tener en consideración que un valor de cero no indica necesariamente que no exista correlación, ya que las variables pueden presentar una relación no lineal.

Para un conjunto de datos cualquiera, y una vez calculado el coeficiente de correlación entre un par de variables X e Y, puede realizarse un sencillo test de hipótesis, basado en la distribución t de Student, para valorar la significación del coeficiente de correlación y confirmar si existe o no una asociación estadísticamente significativa entre ambas características. Estudiar la significación estadística del coeficiente de correlación es, en definitiva, determinar si *r* es estadísticamente diferente de cero. Así mismo, puede obtenerse un intervalo de confianza para el coeficiente de correlación en la población. Sin embargo, mientras que el valor del coeficiente de correlación de Pearson puede ser calculado para cualquier conjunto de datos, la validez del test de hipótesis sobre la correlación entre las variables requiere que al menos una de ellas tenga una distribución normal en la población de la cual procede la muestra. Para el cálculo del intervalo de confianza, se requiere además que ambas variables presenten una distribución normal. Aún bajo esta suposición, la distribución del coeficiente de correlación no será normal, pero puede transformarse para conseguir un valor de *z* que siga una distribución normal y calcular a partir de él su correspondiente intervalo de confianza[2].

Cálculo del tamaño muestral para calcular la significación del coeficiente de correlación lineal de Pearson.

Supongamos que se quiere llevar a cabo un estudio con el fin de determinar si existe o no una relación significativa entre dos variables numéricas X e Y. Para llevar a cabo la investigación, se recoge una muestra de individuos en donde de cada uno de ellos se determina el valor que toma cada una de las dos variables. A continuación se muestra cómo calcular el tamaño de muestra necesario para contrastar la hipótesis de que el correspondiente coeficiente de correlación sea significativamente diferente de 0.

Como se dijo anteriormente, la distribución muestral del coeficiente de Pearson no es normal, pero bajo la suposición de que las dos variables de estudio presentan una distribución gaussiana, el coeficiente de Pearson puede transformarse para conseguir un valor de *z* que sigue una distribución normal. Se suele considerar la transformación de Fisher:

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

Siendo el error estándar de *z* aproximadamente igual a $1/\sqrt{n-3}$.

Utilizando esta aproximación, se obtiene fácilmente una fórmula para el cálculo del número de sujetos necesarios en esta situación. Para su cómputo, se precisará conocer:

- La magnitud de la correlación que se desea detectar (*r*). Esto es, se precisa tener una idea, a partir de publicaciones o estudios previos, del valor aproximado del coeficiente de correlación existente entre las dos variables a estudio.
- La seguridad con la que se desea trabajar, $1-\alpha$, o riesgo de cometer un error de tipo I. Generalmente se trabaja con una seguridad del 95% ($\alpha = 0{,}05$).
- El poder estadístico, $1-\beta$, que se quiere para el estudio, o riesgo de cometer un error de tipo II. Es habitual tomar $\beta = 0{,}2$ o, equivalentemente, un poder estadístico del 80%.

Se debe precisar además si el contraste de hipótesis se va a realizar con un planteamiento unilateral (el *r* calculado es mayor o menor de cero) o bilateral (el *r* calculado es diferente de cero) .

Si se dispone de los datos anteriores, el cálculo del tamaño muestral con un planteamiento bilateral puede realizarse mediante la expresión[7]:

$$n = \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)}\right)^2 + 3$$

donde los valores $z_{1-\alpha/2}$ y $z_{1-\beta}$ se obtienen de la distribución normal estándar en función de la seguridad y el poder elegidos para el estudio. En particular, para una seguridad del 95% y un poder estadístico del 80% se tiene que $z_{1-\alpha/2} = 1{,}96$ y $z_{1-\beta} = 0{,}84$. En las Tablas 1 y 2 se muestran los valores de estos parámetros utilizados con mayor frecuencia en el cálculo del tamaño muestral, en función de la seguridad y el poder con los que se trabaje.

Para un planteamiento unilateral, el razonamiento es análogo, llegando a la siguiente fórmula para el cálculo del tamaño muestral:

$$n = \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)}\right)^2 + 3$$

donde ahora el valor $z_{1-\alpha}$ se obtiene igualmente de la distribución normal estándar, siendo para una seguridad del 95% igual a $z_{1-\alpha} = 1{,}645$. La Tabla 1 muestra los valores más frecuentemente utilizados en función de la seguridad elegida cuando se trabaja con un planteamiento unilateral.

Como resulta habitual, las fórmulas anteriores pueden modificarse con el fin de ajustar el tamaño muestral previsto para el estudio a posibles pérdidas de información que se produzcan durante el desarrollo del mismo. Así, asumiendo un porcentaje de pérdidas L, el tamaño de la muestra a estudiar vendrá dado por:

$$n' = \frac{n}{1-L}$$

donde n denota el valor del tamaño muestral calculado por cualquiera de las dos fórmulas anteriores según el caso.

Ejemplo del cálculo del tamaño muestral para el cálculo del coeficiente de correlación entre dos variables

Supongamos que se desea estudiar la asociación entre la edad y el nivel de colesterol entre los pacientes que acuden a consulta en un determinado centro de salud. Para ello se diseña un estudio en el que se determinará mediante una analítica los valores de colesterol en una muestra aleatoria de los pacientes atendidos en ese centro durante un periodo de tiempo prefijado, de los que también se registrará su edad. Se cree que el valor del coeficiente de correlación lineal de Pearson entre los valores de la edad y el colesterol puede oscilar alrededor de r=0,4. Aplicando las fórmulas anteriores, con un planteamiento bilateral, una seguridad del 95% y un poder estadístico del 80%, se obtiene:

$$n = \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)} \right)^2 + 3 = \left(\frac{1{,}96 + 0{,}84}{\frac{1}{2}\ln\left(\frac{1+0{,}4}{1-0{,}4}\right)} \right)^2 + 3 \approx 47$$

Es decir, se necesitará estudiar a una muestra de 47 pacientes para detectar como significativo un valor del coeficiente de correlación de r=0,4.

Como resulta habitual, si el tamaño del efecto a detectar es menor, asumiendo ahora que el valor del coeficiente de correlación es aproximadamente igual a r=0,3, se obtiene:

$$n = \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)} \right)^2 + 3 = \left(\frac{1{,}96 + 0{,}84}{\frac{1}{2}\ln\left(\frac{1+0{,}3}{1-0{,}3}\right)} \right)^2 + 3 \approx 85$$

En este caso, se necesitaría incluir a 85 pacientes para llevar a cabo el estudio. Si, además, en este último caso se prevé un 20% de posibles pérdidas de información durante la ejecución del estudio, el tamaño muestral debe recalcularse según la siguiente expresión:

$$n' = \frac{n}{1-L} = \frac{85}{1-0{,}2} = 106{,}25 \approx 107$$

Es decir, se necesitaría una muestra de 107 pacientes para llevar a cabo la investigación.

No debe olvidarse que el precisar convenientemente el tamaño de muestra necesario para la ejecución de un estudio permite al investigador conocer el número mínimo de pacientes a estudiar para detectar como significativos efectos de una magnitud determinada. El no hacerlo podría llevar a realizar el estudio con un número insuficiente de casos y a cometer un error de tipo II, es decir, a no detectar una correlación significativa entre las dos variables cuando realmente la hay.

TABLA 1. Valores de $z_{1-\alpha}$ y $z_{1-\alpha/2}$ utilizados con mayor frecuencia en el cálculo del tamaño muestral en función de la seguridad $1-\alpha$ elegida para el estudio.

Seguridad	α	Prueba bilateral $z_{1-\alpha/2}$	Prueba unilateral $z_{1-\alpha}$
80 %	0,200	1,282	0,842
85 %	0,150	1,440	1,036
90 %	0,100	1,645	1,282
95 %	0,050	1,960	1,645
97,5 %	0,025	2,240	1,960
99 %	0,010	2,576	2,326

TABLA 2. Valores de $z_{1-\beta}$ utilizados con mayor frecuencia en el cálculo del tamaño muestral en función de el poder estadístico $1-\beta$ elegido para el estudio.

Poder estadístico	β	$z_{1-\beta}$
99 %	0,01	2,326
95 %	0,05	1,645
90 %	0,1	1,282
85 %	0,15	1,036
80 %	0,2	0,842
75 %	0,25	0,674
70 %	0,3	0,524
65 %	0,35	0,385
60 %	0,4	0,253
55 %	0,45	0,126
80 %	0,5	0,000

Bibliografía

1. Pita Fernández S. Relación entre variables cuantitativas. Cad Aten Primaria 1997; 4: 141-144. [Texto completo]

2. Altman D.G. Practical Statistics for Medical Research. London: Chapman&Hall, 1991.

3. Dawson-Saunders B, Trapp RG. Bioestadística Médica. 2ª ed. México: Editorial el Manual Moderno; 1996.

4. Milton JS, Tsokos JO. Estadística para biología y ciencias de la salud. Madrid: Interamericana McGraw Hill; 2001.

5. Armitage P, Berry G. Estadística para la investigación biomédica. Barcelona: Doyma; 1992.

6. Pita Fernández S. Correlación frente a causalidad. JANO 1996; (1774): 59-60.

7. Argimon Pallás J.M., Jiménez Villa J. Métodos de Investigación Clínica y Epidemiológica. 2ª ed. Madrid: Ediciones Harcourt, 2000.