Relación entre variables cuantitativas

Pita Fernández S., Pértega Díaz S.
Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Juan Canalejo. A Coruña.
Cad Aten Primaria 1997; 4: 141-144. Actualización 30/03/2001.

En el análisis de los estudios clínico-epidemiológicos surge muy frecuentemente la necesidad de determinar la relación entre dos variables cuantitativas en un grupo de sujetos. Los objetivos de dicho análisis suelen ser:

a. Determinar si las dos variables están correlacionadas, es decir si los valores de una variable tienden a ser más altos o más bajos para valores más altos o más bajos de la otra variable.
b. Poder predecir el valor de una variable dado un valor determinado de la otra variable.
c. Valorar el nivel de concordancia entre los valores de las dos variables.

Correlación

En este artículo trataremos de valorar la asociación entre dos variables cuantitativas estudiando el método conocido como correlación. Dicho cálculo es el primer paso para determinar la relación entre las variables. La predicción de una variable. La predicción de una variable dado un valor determinado de la otra precisa de la regresión lineal que abordaremos en otro artículo.

La cuantificación de la fuerza de la relación lineal entre dos variables cuantitativas, se estudia por medio del cálculo del coeficiente de correlación de Pearson (1-3). Dicho coeficiente oscila entre –1 y +1. Un valor de –1 indica una relación lineal o línea recta positiva perfecta. Una correlación próxima a cero indica que no hay relación lineal entre las dos variables.

El realizar la representación gráfica de los datos para demostrar la relación entre el valor del coeficiente de correlación y la forma de la gráfica es fundamental ya que existen relaciones no lineales.

El coeficiente de correlación posee las siguientes características (4):

a. El valor del coeficiente de correlación es independiente de cualquier unidad usada para medir las variables.
b. El valor del coeficiente de correlación se altera de forma importante ante la presencia de un valor extremo, como sucede con la desviación típica. Ante estas situaciones conviene realizar una transformación de datos que cambia la escala de medición y modera el efecto de valores extremos (como la transformación logarítmica).
c. El coeficiente de correlación mide solo la relación con una línea recta. Dos variables pueden tener una relación curvilínea fuerte, a pesar de que su correlación sea pequeña. Por tanto cuando analicemos las relaciones entre dos variables debemos representarlas gráficamente y posteriormente calcular el coeficiente de correlación.
d. El coeficiente de correlación no se debe extrapolar más allá del rango de valores observado de las variables a estudio ya que la relación existente entre X e Y puede cambiar fuera de dicho rango.
e. La correlación no implica causalidad. La causalidad es un juicio de valor que requiere más información que un simple valor cuantitativo de un coeficiente de correlación (5).

El coeficiente de correlación de Pearson (r) puede calcularse en cualquier grupo de datos, sin embargo la validez del test de hipótesis sobre la correlación entre las variables requiere en sentido estricto (4): a) que las dos variables procedan de una muestra aleatoria de individuos. b) que al menos una de las variables tenga una distribución normal en la población de la cual la muestra procede. Para el cálculo válido de un intervalo de confianza del coeficiente de correlación de r ambas variables deben tener una distribución normal. Si los datos no tienen una distribución normal, una o ambas variables se pueden transformar (transformación logarítmica) o si no se calcularía un coeficiente de correlación no paramétrico (coeficiente de correlación de Spearman) que tiene el mismo significado que el coeficiente de correlación de Pearson y se calcula utilizando el rango de las observaciones.

El cálculo del coeficiente de correlación (r) entre peso y talla de 20 niños varones se muestra en la tabla 1. La covarianza, que en este ejemplo es el producto de peso (kg) por talla (cm), para que no tenga dimensión y sea un coeficiente, se divide por la desviación típica de X (talla) y por la desviación típica de Y (peso) con lo que obtenemos el coeficiente de correlación de Pearson que en este caso es de 0.885 e indica una importante correlación entre las dos variables. Es evidente que el hecho de que la correlación sea fuerte no implica causalidad. Si elevamos al cuadrado el coeficiente de correlación obtendremos el coeficiente de determinación (r^2=0.783) que nos indica que el 78.3% de la variabilidad en el peso se explica por la talla del niño. Por lo tanto existen otras variables que modifican y explican la variabilidad del peso de estos niños. La introducción de más variable con técnicas de análisis multivariado nos permitirá identificar la importancia de que otras variables pueden tener sobre el peso.

Tabla 1. Cálculo del Coeficiente de correlación de Pearson entre las variables talla y peso de 20 niños varones

Y Peso (Kg)	X Talla (cm)	$X-\overline{X}$	$Y-\overline{Y}$	$(X-\overline{X})*(Y-\overline{Y})$
9	72	5.65	1.4	7.91
10	76	9.65	2.4	23.16
6	59	-7.35	-1.6	11.76
8	68	1.65	0.4	0.66
10	60	-6.35	2.4	-15.24
5	58	-8.35	-2.6	21.71
8	70	3.65	0.4	1.46
7	65	-1.35	-0.6	0.81
4	54	-12.35	-3.6	44.46
11	83	16.65	3.4	56.61
7	64	-2.35	-0.6	1.41
7	66	-0.35	-0.6	0.21
6	61	-5.35	-1.6	8.56
8	66	-0.35	0.4	-0.14
5	57	-9.35	-2.6	24.31
11	81	14.65	3.4	49.81
5	59	-7.35	-2.6	19.11
9	71	4.65	1.4	6.51
6	62	-4.35	-1.6	6.96
10	75	8.65	2.4	20.76
				$\sum$ 290.8

$$X\left(Media\ de\ \overline{X}=66.35\right)$$

$$Y\left(Media\ de\ \overline{Y}=7.6\right)$$

$$Covarianza=\frac{\sum\left(\overline{X}-X\right)*\left(\overline{Y}-Y\right)}{n-1}=\frac{290.8}{19}=15.30$$

$$r=\frac{covarianza}{S_x*S_y}=\frac{15.30}{8.087*2.137}=0.885$$

S_x = Desviación típica x = 8.087

S_y = Desviación típica y = 2.137

Test de hipótesis de r

Tras realizar el cálculo del coeficiente de correlación de Pearson (r) debemos determinar si dicho coeficiente es estadísticamente diferente de cero. Para dicho calculo se aplica un test basado en la distribución de la t de student.

$$Error\ estandard\ de\ r = \sqrt{\frac{1-r^2}{n-2}}$$

Si el valor del r calculado (en el ejemplo previo r = 0.885) supera al valor del error estándar multiplicado por la t de Student con n-2 grados de libertad, diremos que el coeficiente de correlación es significativo.

El nivel de significación viene dado por la decisión que adoptemos al buscar el valor en la tabla de la t de Student.

En el ejemplo previo con 20 niños, los grados de libertad son 18 y el valor de la tabla de la t de student para una seguridad del 95% es de 2.10 y para un 99% de seguridad el valor es 2.88. (Tabla 2)

$$Error\ estandard\ de\ r = \sqrt{\frac{1-0.885^2}{20-2}} = 0.109$$

Como quiera que r = 0.885 > a 2.10 * 0.109 = 2.30 podemos asegurar que el coeficiente de correlación es significativo (p<0.05). Si aplicamos el valor obtenido en la tabla de la t de Student para una seguridad del 99% (t = 2.88) observamos que como r = 0.885 sigue siendo > 2.88 * 0.109 = 0.313 podemos a su vez asegurar que el coeficiente es significativo (p<0.001). Este proceso de razonamiento es válido tanto para muestras pequeñas como para muestras grandes. En esta última situación podemos comprobar en la tabla de la t de student que para una seguridad del 95% el valor es 1.96 y para una seguridad del 99% el valor es 2.58.

Intervalo de confianza del coeficiente de correlación.

La distribución del coeficiente de correlación de Pearson no es normal pero no se puede transformar r para conseguir un valor z que sigue una distribución normal (transformación de Fisher) y calcular a partir del valor z el intervalo de confianza.

La transformación es:

$$z = 1/2 L_n \frac{1+r}{1-r}$$

L_n representa el logaritmo neperiano en la base e

$$El\ error\ standard\ de\ z\ es = \frac{1}{\sqrt{n-3}}$$

donde n representa el tamaño muestral. El 95% intervalo de confianza de z se calcula de la siguiente forma:

$$z_1(limite\ \inf\ erior) = z - 1.96/\sqrt{n-3}$$

$$z_2(limite\ \sup erior) = z + 1.96/\sqrt{n-3}$$

Tras calcular los intervalos de confianza con el valor z debemos volver a realizar el proceso inverso para calcular los intervalos del coeficiente r

$$\frac{e^{2z1}-1}{e^{2z1}+1} \quad a \quad \frac{e^{2z2}-1}{e^{2z2}+1}$$

Utilizando el ejemplo de la Tabla 1, obtenemos r = 0.885

$$z = 1/2L_n \frac{1+0.885}{1-0.885} = 1.398$$

95% intervalo de confianza de z

$$z_1 = 1.398 - 1.96/\sqrt{20-3} = 0.922$$

$$z_2 = 1.398 + 1.96/\sqrt{20-3} = 1.873$$

Tras calcular los intervalos de confianza de z debemos proceder a hacer el cálculo inverso para obtener los intervalos de confianza de coeficiente de correlación r que era lo que buscábamos en un principio antes de la transformación logarítmica.

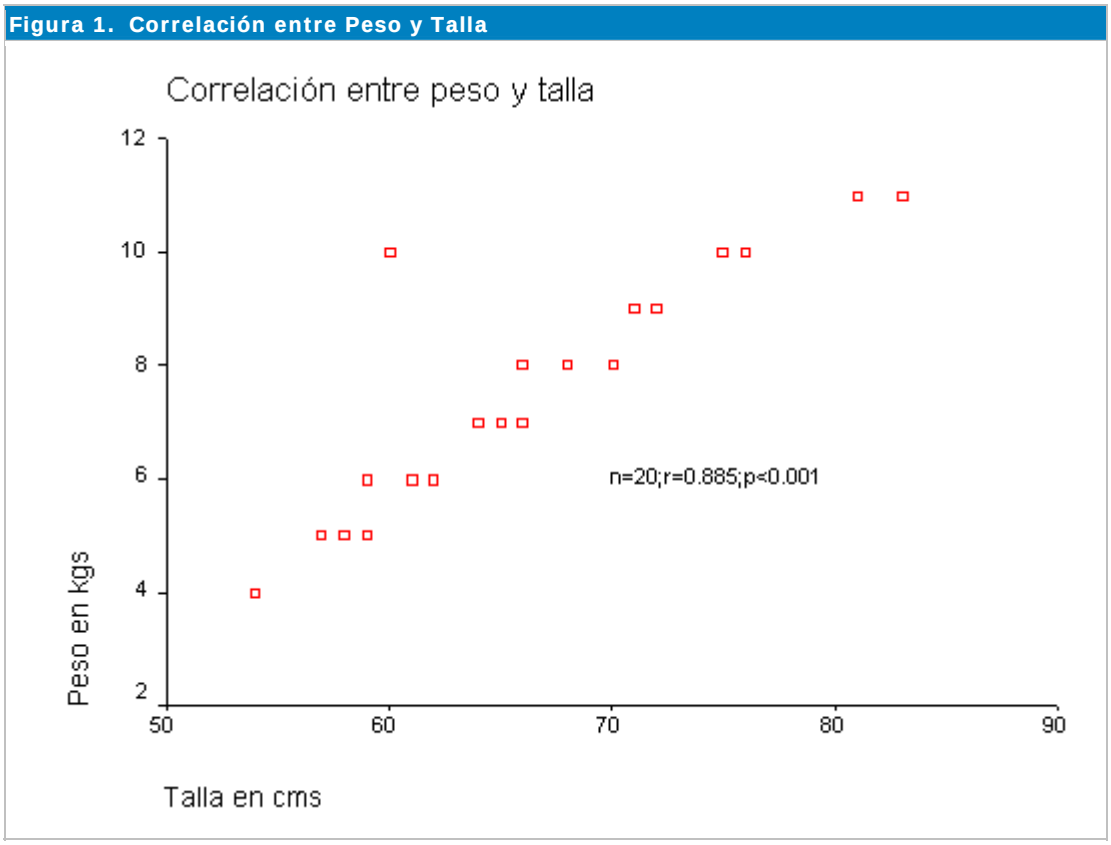
$$\frac{e^{2*0.922}-1}{e^{2*0.922}+1} \quad a \quad \frac{e^{2*1.873}-1}{e^{2*1.873}+1}$$

0.726 a 0.953 son los intervalos de confianza (95%) de r.

Presentación de la correlación

Se debe mostrar siempre que sea posible la gráfica que correlaciona las dos variables de estudio (Fig 1). El valor de r se debe mostrar con dos decimales junto con el valor de la p si el test de hipótesis se realizó para demostrar que r es estadísticamente diferente de cero. El número de observaciones debe a su vez estar indicado.

Figura 1. Correlación entre Peso y Talla



Interpretación de la correlación

El coeficiente de correlación como previamente se indicó oscila entre –1 y +1 encontrándose en medio el valor 0 que indica que no existe asociación lineal entre las dos variables a estudio. Un coeficiente de valor reducido no indica necesariamente que no exista correlación ya que las variables pueden presentar una relación no lineal como puede ser el peso del recién nacido y el tiempo de gestación. En este caso el r infraestima la asociación al medirse linealmente. Los métodos no paramétrico estarían mejor utilizados en este caso para mostrar si las variables tienden a elevarse conjuntamente o a moverse en direcciones diferentes.

La significancia estadística de un coeficiente debe tenerse en cuenta conjuntamente con la relevancia clínica del fenómeno que estudiamos ya que coeficientes de 0.5 a 0.7 tienden ya a ser significativos como muestras pequeñas (6). Es por ello muy útil calcular el intervalo de confianza del r ya que en muestras pequeñas tenderá a ser amplio.

La estimación del coeficiente de determinación (r^2) nos muestra el porcentaje de la variabilidad de los datos que se explica por la asociación entre las dos variables.

Como previamente se indicó la correlación elevada y estadísticamente significativa no tiene que asociarse a causalidad. Cuando objetivamos que dos variables están correlacionadas diversas razones pueden ser la causa de dicha correlación: a) pude que X influencie o cause Y, b) puede que influencie o cause X, c) X e Y pueden estar influenciadas por terceras variables que hace que se modifiquen ambas a la vez.

El coeficiente de correlación no debe utilizarse para comparar dos métodos que intentan medir el mismo evento, como por ejemplo dos instrumentos que miden la tensión arterial. El coeficiente de correlación mide el grado de asociación entre dos cantidades pero no mira el nivel de acuerdo o concordancia. Si los instrumentos de medida miden sistemáticamente cantidades diferentes uno del otro, la correlación puede ser 1 y su concordancia ser nula (7).

Coeficiente de correlación de los rangos de Spearman

Este coeficiente es una medida de asociación lineal que utiliza los rangos, números de orden, de cada grupo de sujetos y compara dichos rangos. Existen dos métodos para calcular el coeficiente de correlación de los rangos uno señalado por Spearman y otro por Kendall (8). El r de Spearman llamado también rho de Spearman es más fácil de calcular que el de Kendall. El coeficiente de correlación de Spearman es exactamente el mismo que el coeficiente de correlación de Pearson calculado sobre el rango de observaciones. En definitiva la correlación estimada entre X e Y se halla calculado el coeficiente de correlación de Pearson para el conjunto de rangos apareados. El coeficiente de correlación de Spearman es recomendable utilizarlo cuando los datos presentan valores externos ya que dichos valores afectan mucho el coeficiente de correlación de Pearson, o ante distribuciones no normales.

El cálculo del coeficiente viene dado por:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

en donde $d_i = r_{xi} - r_{yi}$ es la diferencia entre los rangos de X e Y.

Los valores de los rangos se colocan según el orden numérico de los datos de la variable.

Ejemplo: Se realiza un estudio para determinar la asociación entre la concentración de nicotina en sangre de un individuo y el contenido en nicotina de un cigarrillo (los valores de los rangos están entre paréntesis) (2).

X	Y
Concentración de Nicotina en sangre (nmol/litro)	**Contenido de Nicotina por cigarrillo (mg)**
185.7 (2)	1.51 (8)
197.3 (5)	0.96 (3)
204.2 (8)	1.21 (6)
199.9 (7)	1.66 (10)
199.1 (6)	1.11 (4)
192.8 (6)	0.84 (2)
207.4 (9)	1.14 (5)
183.0 (1)	1.28 (7)
234.1 (10)	1.53 (9)
196.5 (4)	0.76 (1)

Si existiesen valores coincidentes se pondría el promedio de los rangos que hubiesen sido asignado si no hubiese coincidencias. Por ejemplo si en una de las variables X tenemos:

X (edad)	(Los rangos serían)
23	1.5
23	1.5
27	3.5
27	3.5
39	5
41	6
45	7
…	…

Para el cálculo del ejemplo anterior de nicotina (2) obtendríamos el siguiente resultado:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6\left[(2-8)^2 + (5-3)^2 + (8-6)^2 + \ldots + (4-1)^2\right]}{10(10^2-1)} = 1 - \frac{6(120)}{10(99)} = 0.27$$

Si utilizamos la fórmula para calcular el coeficiente de correlación de Pearson de los rangos obtendríamos el mismo resultado

$$r_s = \frac{n\sum r_x r_y - \sum r_x \sum r_y}{\sqrt{\left[n\sum r_x^2 - \left(\sum r_x\right)^2\right]\left[n\sum r_y^2 - \left(\sum r_y\right)^2\right]}}$$

$$\sum r_x = \sum r_y = 55 \qquad \sum r_x^2 = \sum r_y^2 = 385$$

$$\sum r_x r_y = 2\,(8) + 5\,(3) + 8\,(6) + \ldots + 4\,(1) = 325$$

$$r_s = \frac{10(325) - 55(55)}{\sqrt{\left[10\,(385) - 55^2\right]\left[10(385) - 55^2\right]}} = 0.27$$

La interpretación del coeficiente r_s de Spearman es similar a la Pearson. Valores próximos a 1 indican una correlación fuerte y positiva. Valores próximos a –1 indican una correlación fuerte y negativa. Valores próximos a cero indican que no hay correlación lineal. Así mismo el r_s^2 tiene el mismo significado que el coeficiente de determinación de r^2.

La distribución de r_s es similar a la r por tanto el calculo de los intervalos de confianza de r_s se pueden realizar utilizando la misma metodología previamente explicada para el coeficiente de correlación de Pearson.

Bibliografía

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

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

3- Martín Andrés A, Luna del Castillo JD. Bioestadística para las ciencias de la salud. 4ª ed. Madrid: ORMA; 1993.

4- Altman DA. Practical statistics for medical research. 1th ed., repr. 1997. London: Chapman & Hall; 1997.

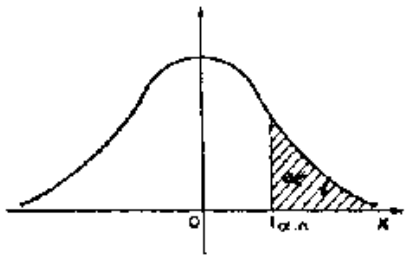
5- Pita Fernández S. Correlación frente a causalidad JANO 1996; (1174): 59-60.

6- Feintein AR. Tempest in a P-pot?. (Editorial). Hypertension 1985; 7: 313-318. [Medline]

7- Bland JM, Altman DG. Statistical methods for assesing agreement between two methods of clinical measurement. Lancet 1986; 1: 307-310. [Medline]

8- Conover WJ. Practical nonparametric statistics. 3rd . ed. New York: John Wiley & Sons; 1998.

Tabla 2. Distribución t de Student



gl \ α/2	0,40	0,30	0,20	0,10	0,050	0,025	0,010	0,005	0,001	0,0005
1	0,325	0,727	1,376	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,22	12,94
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032	5,893	6,859
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,405
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,611	3,922
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,767
24	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	0,255	0,529	0,851	1,303	1,648	2,021	2,423	2,704	3,307	3,551
50	0,255	0,528	0,849	1,298	1,676	2,009	2,403	2,678	3,262	3,495
60	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	3,232	3,460
80	0,254	0,527	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,415
100	0,254	0,256	0,845	1,290	1,660	1,984	2,365	2,626	3,174	3,389
200	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,339
500	0,253	0,525	0,842	1,283	1,648	1,965	2,334	2,586	3,106	3,310
∞	0,233	0,524	0,842	1,282	1,645	1,960	2,326	2,576	3,090	3,291