Revista FABICIB, Año 2004, Volumen 8, págs. 255-259

# Notas Estadísticas 2: El coeficiente de correlación lineal

Carrera, E.; Vaira, S.; Contini, L.

Dpto. de Matemática, Area Estadística. Facultad de Bioquímica y Ciencias Biológicas. Universidad Nacional del Litoral. CC. 242. Ciudad Universitaria, Paraje El Pozo. (3000) Santa Fe.

**RESUMEN:** Ante el uso indiscriminado del coeficiente de correlación lineal, el objetivo de este trabajo fue analizar qué mide y qué no mide este coeficiente. Se mencionan además el coeficiente de correlación de Spearman y otros alternativos.

**Palabras claves:** correlación - correlación lineal - diagramas de dispersión .

**SUMMARY: Statistical notes 2: The linear correlation coefficient. Carrera, E.; Vaira, S.; Contini, L.**. Since the linear correlation coefficient is indiscriminately used the goal of this paper was to analyze what this coefficient can or cannot measure. Spearman's correlation coefficient and other alternative methods are also mentioned.

**Key words:** correlation – linear correlation – scatter diagram.

*** Correspondencia:**
Tel.: 0342-4575210
e-mail: ecarrera@unl.edu.ar

Recibido: 1-07-04
Aceptado: 23-09-04

## El coeficiente de correlación de Pearson

Uno de los objetivos más frecuentes de la investigación en ciencias experimentales en general, es hallar la asociación entre dos variables. Para ello se utiliza frecuentemente el coeficiente de correlación de Pearson, $\rho$. Ahora bien, en realidad se hace uso y abuso de este coeficiente. Por ello el objetivo de esta nota es señalar en parte qué es y qué no es, así como qué mide y qué no mide este coeficiente de correlación.

El coeficiente de correlación es una medida bivariada de asociación o de relación entre dos variables aleatorias cualesquiera y se define:

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

donde Cov(X,Y) es la covarianza de las dos variables aleatorias X e Y o sea la expresión que *verdaderamente mide* la variación simultánea de X con Y y que depende de las unidades en que están medidas. Por ello hoy está de moda el Análisis de la Covarianza (ANCOVA), no sólo el Análisis de la Varianza (ANOVA), pero esto es tema de otro artículo. La división de esta covarianza por los desvíos estándar de X y de Y, $\sigma_x$ y $\sigma_y$ respectivamente, propuesto por Karl Pearson (1857-1936), permitió, siempre que sean positivos, obtener una medida adimensional de variabilidad conjunta: $\rho$, y su valor es un simple número.

## El coeficiente de correlación lineal de Pearson

Si ambas variables son *continuas* y tienen una distribución normal conjunta o al menos cada una de ellas tiene una distribución normal, puede probarse y hay un teorema que lo demuestra, que: $-1 \leq \rho \leq 1$ y que cuando $|\rho| = 1$, existe una función lineal, $Y = a + bX$, que indica que la relación entre las variables es *exactamente lineal*. Si $b > 0$ y $\rho = 1$, entonces la relación es lineal directa y si $\rho = -1$ y por ende $b < 0$ la relación lineal es inversa (1-3).

Por ello el coeficiente de correlación pasa a denominarse *coeficiente de correlación lineal de Pearson*.

Pero esto sucede en la población, el investigador trabaja sobre una muestra de tamaño **n** de la misma.

## El coeficiente de correlación lineal muestral

El valor del estadístico, o sea la cantidad calculada con los datos muestrales, es simbolizado por **r** o **R** en algunas publicaciones, se calcula a través de la siguiente expresión:

$$r = \frac{S_{XY}}{S_X S_Y}$$

, donde $S_x$ y $S_y$ son los desvíos estándar en la muestra de cada variable y $S_{xy}$ es la covarianza muestral de X e Y, que se calcula como:

$$S_{XY} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Este coeficiente de correlación muestral mide nada más ni nada menos que la *asociación lineal* entre las variables X e Y que son de interés del investigador y cuyos valores ha obtenido durante su trabajo. También varía, como su correspondiente parámetro poblacional, entre –1 y 1. Si entre las variables hay otro tipo de asociación, el coeficiente no lo detecta (3-6).

Sería erróneo interpretar que un $\rho = 0$ indica no relación entre las variables o, es más, independencia entre ellas. Un coeficiente de correlación igual a cero sólo está indicando que ambas variables no están relacionadas linealmente pero puede haber entre ellas otro tipo de relación (2, 3, 7). En cualquier calculadora que tenga funciones estadísticas simples se puede calcular el valor de $\rho$. No es necesario un software estadístico para hacerlo.

Si se necesitan además plantear pruebas de hipótesis estadísticas acerca de la fuerza de la asociación lineal y su significancia estadística, es necesario hacer supuestos sobre la distribución conjunta de probabilidades de X e Y. La distribución de interés según se señaló es la normal bivariada y $\rho$ es uno de los parámetros característicos de la misma, luego los pares de puntos debieron ser extraídos de una distribución normal bivariada (1-3).

La mejor forma de interpretar el coeficiente de correlación lineal y saber si es la medida de asociación adecuada, es realizar primeramente un gráfico de dispersión donde se encuentren todos los pares (x,y) que se han medido.

## Importancia del diagrama de dispersión

No se desconoce que una parte importante de la estadística inferencial se basa en un buen análisis exploratorio de los datos, que consiste en un conjunto de técnicas gráficas y cálculo de medidas resúmenes que permiten describir una o varias variables, buscando compararlas o relacionarlas. En este tema de buscar asociaciones no es menos importante graficar, a través de los llamados diagrama de dispersión, los pares de mediciones (8-10). La búsqueda de patrones así como poder establecer el tipo de asociación entre variables cuantitativas puede realizarse observando estos diagramas. De igual manera que el histograma, el diagrama de cajas o el de tallo-hoja permiten visualizar y detectar comportamientos «normales» o "no normales", la nube de puntos (scatter diagram) permitirá detectar la relación entre ellas. Si existe relación lineal aportará además la dirección y la fuerza de la misma, así como, si no fuese lineal sugerirá, algún otro tipo de relación.

Si la nube de puntos tiene una forma de elipse o de cigarro, **r** representa la medida adecuada. Cuanto más aplastada sea la elipse, mayor será la relación lineal (Figura 1).

## Correlación versus Concordancia

En ocasiones pueden existir diferentes métodos de medida de un mismo parámetro y posiblemente uno de estos sea el patrón de referencia (conocido como *gold standard*) y el otro, uno que el investigador quiere emplear como método alternativo. Interesa a éste determinar si las mediciones obtenidas con uno y otro método *concuerdan*, si es así pueden reemplazarse estos métodos entre sí. Esto difiere de la calibración donde las mediciones con el nuevo método son comparadas con el verdadero valor o con mediciones realizadas con un método altamente exacto.

Existe la tendencia a utilizar el coeficiente de correlación muestral de Pearson como medida descriptiva de la concordancia entre ambos métodos, pero éste no es el procedimiento adecuado; se fundamenta en que puede ocurrir que el método utilizado por el investigador tenga un error sistemático, por ejemplo, cada medición hecha por el método a comparar sea 3 unidades más alta que las correspondientes al de referencia, con lo cual el valor de la correlación dará 1 o cercana a 1 y no obstante existir una relación lineal casi perfecta entre ambos métodos, no se puede concluir que concuerden o que haya concordancia (4, 7-9).

## Otras medidas de asociación

Cuando alguna de las variables que se analizan conjuntamente no es continua, hay medidas de asociación a tener en cuenta que dependen del tipo y lo que se quiere medir entre ellas, no se dará una descripción detallada de las mismas, pero se mencionan algunas de uso más frecuente en la literatura científica como ser el coeficiente de correlación de Spearman, simbolizado por $\rho_s$ (poblacional) y por $r_s$ (muestral). Es utilizado cuando las dos variables son ordinales, además de ser la alternativa no paramétrica, en reemplazo del coeficiente de correlación lineal de Pearson, cuando no puede evaluarse o verificarse el supuesto de normalidad bivariada o al menos de normalidad de cada una (5, 7, 10, 11).

Otros coeficientes que deben ser usados cuando las variables son categóricas ordinales o nominales, una continua y la otra no son el Coeficiente Punto Biserial ($r_{pb}$), que es un caso especial de la correlación de Pearson cuando una variable es continua y la otra es nominal y dicotómica; el coeficiente Phi ($\varphi$), la $\tau$ de Kendall, entre otros (12, 13).

## Conclusión

Tiene el coeficiente de correlación lineal características que lo hacen la medida de asociación más usada al momento de describir el comportamiento conjunto entre dos variables continuas, como resultado de la adecuada interpretación y uso del mismo se resumen sus propiedades más importantes a manera de conclusión:

– El valor del coeficiente de correlación es independiente de cualquier unidad de medida.

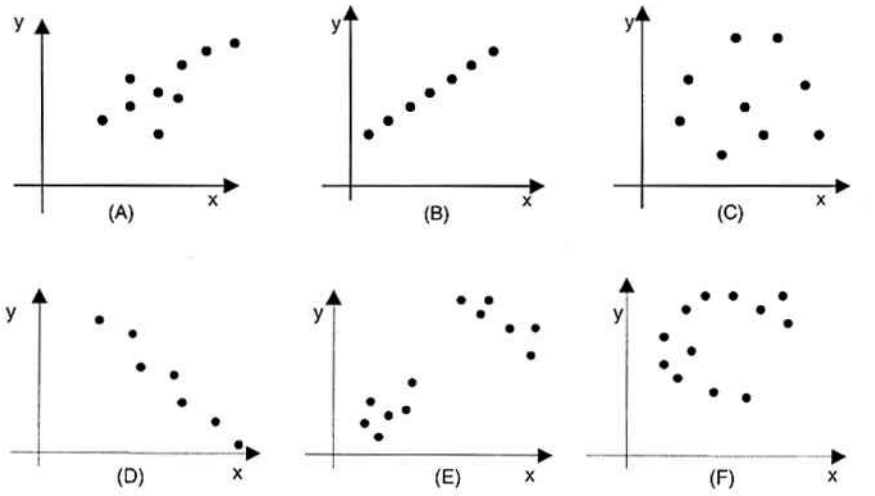
– El valor del coeficiente de correlación lineal se ve influenciado de forma importante ante la presencia de un valor extremo, como sucede con la desviación estándar. Se sugiere en este caso una transformación de los datos que cambie la escala de

medida y de esta manera se modere el efecto de los valores extremos.

– El coeficiente de correlación lineal no se debe extrapolar más allá del rango de valores observados de las variables de estudio, ya que la relación entre X e Y puede cambiar fuera de dicho rango.

– El coeficiente de correlación lineal de Pearson sólo mide asociación lineal entre variables y su empleo debe estar apoyado en la observación de diagrama de dispersión.

**Figura 1**: Diagramas de dispersión que ilustran diferentes tipos de relaciones entre variables




(A) **r** = 0,70. (B) **r** = 1, relación lineal directa perfecta. (C) **r** = 0, no correlacionadas linealmente e independientes. (D) **r** = -0,85. (E) **r** = 0,70, dispersión de puntos en grupos, posible presencia de dos poblaciones. (F) **r** = 0,85, correlación alta y no lineal.

## Bibliografía

**1.** Hogg, R.; Craig, A., 1975. "Introduction to mathematical statistics". The Macmillan Company, (New York). 50 - 68, 360 - 370.

**2.** Bickel, P.; Doksum, K., 1977. "Mathematical Statistics". Prentice Hall. (New Jersey). 20 - 25, 200 - 230, 450 - 460.

**3.** Casella, G; Berger, R., 2002. "Statistical Inference". Duxbury Advanced Series, Thomson Learning. (Australia). 160 - 180, 260 - 266.

**4.** Freund, J.; Walpole, R., 1980. "Mathematical Statistics". Prentice hall, Inc. (London). 120 - 130, 419 - 440.

**5.** Myers, R. H., 2000. "Classical and modern regression with application". PWS Kent Publishing Company. (Boston). 425 – 449.

**6.** Minami, M.; Shimizu, K.1998. Estimation for a common correlation coefficient in bivariate normal distributions with missing observations. Biometrics. **54**: 1136 – 1146.

**7.** Altman, Douglas G., 1997. "Practical statistics for medical research". Chapman and Hall. (London). 7ma reimpresión. 339 – 358, 403 - 411.

**8.** Altman, D.; Bland, J., 1983. Measurement in medicine: the analysis of method comparasion studies. The Statistician, **32**, 307 - 317.

**9.** Bland J.; Altman D., 1986. Statistical method for assessing agreement between two methods of clinical measurement. The Lancet, I, 307-310.

**10.** Fukuda, H; Ohashi, Y., 1997. A guideline for reporting results of statistical analysis in japanese journal of clinical oncloIgy. Jpn J Clin Oncol; **27**,3: 121-127.

**11.** Conover, W. J., 1998. "Practical Nonparametric Statistic". John Willey & Sons. (New York). 213 – 229.

**12.** Agresti, A., 1996. "An introduction to categorical data analysis". John Willey & Sons. (New York).

**13.** Armitage, P.; Berry, G. 1997. "Estadística para la investigación biomédica". Harcourt Brace. (Madrid). 145-165, 419-440.