ReaR
ISNN 1989 4090
Revista electrónica de AnestesiaR
Abril 2021

ORIGINAL

Perro ladrador, poco mordedor. Correlación.

Molina Arias M.

Hospital Infantil Universitario La Paz, Madrid.

Resumen

Se define el concepto de correlación como la fuerza y sentido de asociación entre dos variables aleatorias. Se describen algunos de los coeficientes de correlación más utilizados, como el coeficiente de correlación de Pearson, el coeficiente de correlación de Spearman y el coeficiente tau de Kendall.

Introducción

Se define el concepto de correlación como la fuerza y sentido de asociación entre dos variables aleatorias. Se describen algunos de los coeficientes de correlación más utilizados, como el coeficiente de correlación de Pearson, el coeficiente de correlación de Spearman y el coeficiente tau de Kendall.

Tengo unos vecinos que tienen un perro que está todo el puñetero día ladrando. Es el típico perro que no abulta dos palmos, pero que los rentabiliza de una forma increíble ladrando a unos volúmenes desaforados, por no hablar de lo desagradable de su timbre de "voz".

Con estos perros enanos suele ocurrir, te ladran como poseídos por el demonio en cuanto te ven, pero, según dice la sabiduría popular, puedes estar tranquilo porque, cuanto más te ladran, menos probable es que te muerdan. Ahora que lo pienso, casi diría que hay una correlación inversa entre lo ladrador y lo mordedor que es uno de estos animalillos.

Y ya que hemos mencionado el término correlación, vamos a hablar de este concepto y de cómo medirlo.

¿Qué significa correlación?

Dice el diccionario de la lengua española que correlación es la correspondencia o relación recíproca entre dos o más cosas o series de cosas. Otra fuente de sabiduría, la Wikipedia, dice que, en lo referente a probabilidad y estadística, la correlación indica la fuerza y la dirección de una relación lineal y proporcionalidad entre dos variables estadísticas.

¿Qué quiere decir, entonces, que dos variables están correlacionadas? Pues

una cosa bastante más sencilla de lo que pueda parecer: que los valores de una de las variables cambian en un sentido determinado de forma sistemática cuando se producen cambios en la otra. Dicho de forma más sencilla, dadas dos variables A y B, siempre que el valor de A cambia en un sentido determinado, los de B cambiarán también en un determinado sentido, que puede ser el mismo o el contrario.

Y eso es lo que significa correlación. Solo eso, cómo cambia una con los cambios de la otra. Esto no quiere decir, para nada, que haya una relación de causalidad entre las dos variables, que es una asunción, generalmente errónea, que se hace con cierta frecuencia. Tan frecuente es esta falacia que tiene hasta un simpático nombre en latín, *cum hoc ergo propter hoc*, que puede resumirse para las mentes menos cultivadas como "correlación no implica causalidad". Porque dos cosas varíen juntas no significa que una sea causa de la otra.

Otro error frecuente es el de confundir correlación con regresión. En realidad, son dos términos que están muy relacionados. Mientras que el primero, correlación, solo nos dice si existe relación entre las dos variables, el análisis de regresión va un paso más allá y pretende encontrar un modelo que nos permita predecir el valor de una de las variables (la llamada dependiente) en función del valor que tome la otra variable (a la que llamaremos independiente o explicativa). En muchas ocasiones, estudiar si existe correlación es el paso previo antes de generar el modelo de regresión.

Coeficientes de correlación

Pues bien, de todos es conocida la afición del ser humano de medir y cuantificar, así que a nadie puede extrañarle que se inventasen los llamados coeficientes de correlación, de los que hay una familia más o menos numerosa.

Para calcular el coeficiente de correlación necesitamos, pues, un parámetro que nos permita cuantificar esta relación. Para ello podemos disponer de la covarianza, que indica el grado de variación conjunta de dos variables aleatorias.

El problema de la covarianza es que su valor depende de las escalas de medición de las variables, lo que nos impide realizar comparaciones directas entre distintos pares de variables. Para evitar este problema, recurrimos a una solución que ya nos es conocida y que no es otra que la estandarización. El producto de la estandarización de la covarianza serán los coeficientes de correlación.

Todos estos coeficientes tienen algo en común: su valor oscila desde -1 a 1. Cuánto más se aleje el valor de 0, mayor será la fuerza de la relación, que será prácticamente perfecta cuando alcance -1 o 1. En el 0, que es el valor nulo, en principio no existirá correlación entre las dos variables.

El signo del valor del coeficiente de correlación nos indicará la otra cualidad de la relación entre las dos variables: el sentido. Cuando el signo sea positivo significará que la correlación es directa: cuando una aumenta o disminuye, la otra lo hace también en el mismo sentido. Si el signo es negativo, la correlación será inversa: al cambiar una variable, la otra lo hará en el sentido opuesto (si una aumenta, la otra disminuye, y viceversa).

Hemos visto hasta aquí dos de las características de la correlación entre dos variables: la fuerza y el sentido. Existe una tercera, la forma, que depende del tipo de línea que defina el mejor modelo de ajuste. En esta entrada

nos vamos a quedar con la forma más sencilla, que no es otra que la correlación lineal, en la que la línea de ajuste es una recta, pero que sepáis que hay otros ajustes no lineales.

Coeficiente de correlación de Pearson

Ya hemos dicho que existe toda una serie de coeficientes de correlación que podremos calcular en función del tipo de variables que queramos estudiar y de la distribución de probabilidad que sigan en la población de la que procede la muestra.

El coeficiente de correlación de Pearson, también llamado coeficiente de correlación lineal producto-momento, es, sin duda, el más célebre de toda esta familia.

Como ya hemos dicho, no es más que la covarianza estandarizada. Hay varias formas de calcularlo, pero todos los caminos conducen a Roma, así que no me voy a resistir a poner la fórmula:

$r = S_{XY} / S_X S_Y$

Como vemos, la covarianza (en el numerador) se estandariza dividiéndola por el producto de las varianzas de las dos variables (en el denominador).

Para poder utilizar el coeficiente de correlación de Pearson, ambas variables tienen que ser cuantitativas, estar correlacionadas de forma lineal, distribuirse de forma normal en la población y cumplirse el supuesto de homocedasticidad, que quiere decir que la varianza de la variable Y debe ser constante a lo largo de los valores de la variable X. Una forma sencilla de comprobar este último supuesto es dibujar el diagrama de dispersión y ver si la nube se dispersa de forma similar a lo largo de los valores de la variable X.

Un factor a tener en cuenta que puede sesgar el valor de este coeficiente es la presencia de valores extremos (los *outliers*, para los amantes del inglés).

Coeficiente de correlación de Spearman

El equivalente no paramétrico del coeficiente de Pearson es el coeficiente de correlación de Spearman. Este, como ocurre con las técnicas no paramétricas, no emplea los datos directos para su cálculo, sino que utiliza su transformación en rangos.

Así, se utiliza cuando las variables son ordinales o cuando son cuantitativas, pero no cumplen el supuesto de normalidad y pueden transformarse en rangos.

Por lo demás, su interpretación es similar a la del resto de los coeficientes. Decir, además, que al calcularse con rangos es menos sensible que el coeficiente de Pearson ante la existencia de valores extremos.

Otra ventaja frente al de Spearman es que solo precisa que la correlación entre las dos variables sea monótona, que quiere decir que cuando una variable aumenta la otra lo hace también (igual cuando disminuyen) con una tendencia constante. Esto permite utilizarlo no solo cuando la relación es lineal, sino también en casos de relación logística y exponencial.

Coeficiente tau de Kendall

Otro coeficiente que utiliza también el rango de la variable es el coeficiente τ de Kendall. Al ser un coeficiente no paramétrico, es también una alternativa al coeficiente de Pearson cuando no se cumple el supuesto de normalidad, siendo más aconsejable que el de Spearman cuando la muestra es pequeña

y cuando hay mucha ligadura de rangos, que quiere decir que muchos datos ocupan la misma posición en los rangos.

Pero aún hay más…

Aunque hay algunos más, solo me voy a referir de forma específica a tres de ellos, útiles para estudiar variables cuantitativas:

1. Coeficiente de correlación parcial. Estudia la relación entre dos variables, pero tiene en cuenta y elimina la influencia de otras variables.

El caso más sencillo es el de estudiar dos variables X1 y X2, eliminando el efecto de una tercera variable X3. En este caso, si las correlaciones entre X1 y X3 y entre X2 y X3 son iguales a cero, se obtiene el mismo valor para el coeficiente de correlación parcial que si calculamos el coeficiente de Pearson entre X1 y X2.

En el caso de que queramos controlar más variables, la fórmula, que no pienso escribir, se vuelve más compleja, pero lo mejor es dejar que un programa de estadística lo calcule. Si el valor del coeficiente parcial es inferior al del coeficiente de Pearson querrá decir que la correlación entre ambas variables se debe parcialmente a las otras variables que controlamos. Cuando el coeficiente parcial es mayor que el de Pearson, las variables que se controlan enmascaran la relación entre las dos variables de interés.

2. Coeficiente de correlación semiparcial. Es similar al anterior, pero este semiparcial permite evaluar la asociación entre dos variables controlando el efecto de una tercera sobre una de las dos variables de interés (no sobre la dos, como el coeficiente parcial).

3. Coeficiente de correlación múltiple. Este permite conocer la correlación entre una variable y un conjunto de dos o más variables, todas ellas cuantitativas.

Y creo que con esto tenemos suficiente, por ahora. Hay algunos coeficientes más que son útiles para situaciones especiales. El que tenga curiosidad, que busque un libro de estadística gordo que seguro que los encuentra.

Significación e interpretación

Ya dijimos al comienzo que el valor de estos coeficientes podía oscilar de -1 a 1, siendo -1 la correlación negativa perfecta y 1 la correlación positiva perfecta.

Podemos hacer un paralelismo entre el valor del coeficiente y la fuerza de la asociación, que no es otra cosa que el tamaño del efecto. Así, valores de 0 indican asociación nula, de 0,1 asociación pequeña, de 0,3 mediana, 0,5 moderada, 0,7 alta y 0,9 asociación muy alta.

Para ir acabando, hay que decir que, para poder dar valor al coeficiente, este tiene que ser estadísticamente significativo. Ya sabéis que siempre trabajamos con muestras, pero lo que nos interesa es inferir el valor en la población, así que tenemos que calcular el intervalo de confianza del coeficiente que hayamos empleado. Si este intervalo incluye el valor nulo (el cero) o si el programa nos calcula el valor de p y es mayor de 0,05, no tendrá sentido valorar el coeficiente, aunque esté próximo a -1 o 1.

Nos vamos…

Y aquí lo dejamos por hoy. No hemos hablado nada del uso del coeficiente de correlación de Pearson para comparar la precisión de una prueba diagnóstica. Y

no hemos dicho nada porque no debemos utilizar este coeficiente con este fin. El coeficiente de Pearson depende mucho de la variabilidad intrasujetos y puede dar un valor muy alto cuando una de las mediciones sea sistemáticamente mayor que la otra, aunque no haya buena concordancia entre las dos. Para esto es mucho más adecuado utilizar el coeficiente de correlación intraclase, mejor estimador de la concordancia de medidas repetidas. Pero esa es otra historia…

Bibliografía

– Solanas A, Peró M. Coeficientes de correlación. Decisión estadística En: Peró Cebollero M, Leiva Ureña D, Guàrdia Olmos J, Solanas Pérez A, eds. Estadística aplicada a las ciencias sociales mediante R y R-Commander. Ibergarceta Publicaciones SL. Madrid, 2012; 405-31. (HTML)

– Sánchez-Villegas A, Martín-Calvo N, Martínez-González MA. Correlación y regresión lineal simple. En: Martínez MA, Sánchez-Villegas A, Toledo EA, Faulin A, eds. Bioestadística amigable, 3ª ed. Elsevier España SL. Barcelona, 2014; 269-326. (PDF)

– Amat Rodrigo J. Correlación lineal y regresión lineal simple. En: Estadística con R. Disponible en:

https://github.com/JoaquinAmatRodrigo/Estadistica-con-R/blob/master/PDF_format/24_Correlaci%C3%B3n_y_Regresi%C3%B3n_lineal.pdf. Consultado el 18/7/2020.

Correspondencia al autor

Manuel Molina Arias
mma1961@gmail.com
Servicio de Gastroenterología.
Hospital Infantil Universitario La Paz.
Madrid. España.

Aceptado para el blog en septiembre de 2020