Correlaciones que no llegan a 1

Gabriel Cavada Ch.[1,2]
[1]*Facultad de Medicina, Universidad de los Andes.*
[2]*Escuela de Salud Pública, Universidad de Chile.*

Correlations that do not reach 1

Como es sabido, la medida de asociación más utilizada entre dos variables continuas, es el conocido **coeficiente de correlación de pearson**, ρ, también es ampliamente sabido que dicha medida es un número real que varía entre -1 y 1 y que en la medida que éste se acerque a -1 o a 1, estamos en presencia de una asociación lineal que alcanza la perfección, sea esta directa o inversa.

Muchas veces ρ se utiliza incluso para sugerir la causalidad de una variable Y debida a X, que expresamos mediante la relación: $Y = \alpha + \beta X$.

En la práctica o conocimiento de la evidencia, manejamos valores de correlación cercanos a 1 (o a -1) que somos capaces de explicar muy razonablemente desde el punto de vista biológico, o incluso imaginarnos correlaciones altas que darían origen a muy plausibles y prometedoras preguntas o hipótesis de investigación. Esta situación, muy deseable por cierto, nos llevará a un diseño de investigación que obligadamente nos pondrá en la situación de tener que calcular un tamaño de muestra, que con cierta significación y potencia, nos permitan estimar a ρ; cuando este momento crucial llega, la pregunta es ¿qué nivel de correlación queremos detectar?, esto es por ejemplo ρ = 0,5; 0,6; 0,8 etc.

Así: si asumiéramos la relación lineal a través de la cual se digiera que el índice de masa corporal es capaz de explicar el nivel de triglicéridos en población de adultos obesos, ¿qué nivel de correlación asumiríamos para calcular el tamaño de muestra para diseñar el estudio?, la respuesta más inmediata nos la daría la evidencia publicada, pero está claro que la literatura nos entregará un rango de valores, uno en cada estudio y con suerte con su intervalo de confianza, y pues ¿cuál tomar? La respuesta más correcta, sólo puede obtenerse si se conocen las distribuciones de probabilidad de las variables involucradas, pues si esta información está disponible podemos calcular los valores mínimo y máximo del coeficiente de correlación que podrían alcanzar las variables que serán estudiadas.

En efecto, el Doctor Carles Cuadras de la Universidad de Barcelona, en un curso doctoral de probabilidades, a modo de ejercicio, nos propuso demostrar que si dos variables aleatorias estaban estandarizadas (con valor medio de 0 y con varianza 1), las correlaciones mínimas y máximas estaban dadas por las siguientes expresiones:

$$\rho^- = \int_0^1 F^{-1}(t)G^{-1}(1-t)dt$$

$$\rho^+ = \int_0^1 F^{-1}(t)G^{-1}(t)dt$$

Es decir, que el coeficiente de correlación de Pearson, en términos generales estaba comprendido entre el valor de estas dos integrales, donde F y G denotan las funciones de probabilidad acumuladas de las respectivas variables, en consecuencia las integrales contienen las respectivas funciones inversas. Así se demuestra que el coeficiente de correlación de Pearson varía según la desigualdad:

$$\rho^- \leq \rho(X,Y) \leq \rho^+$$

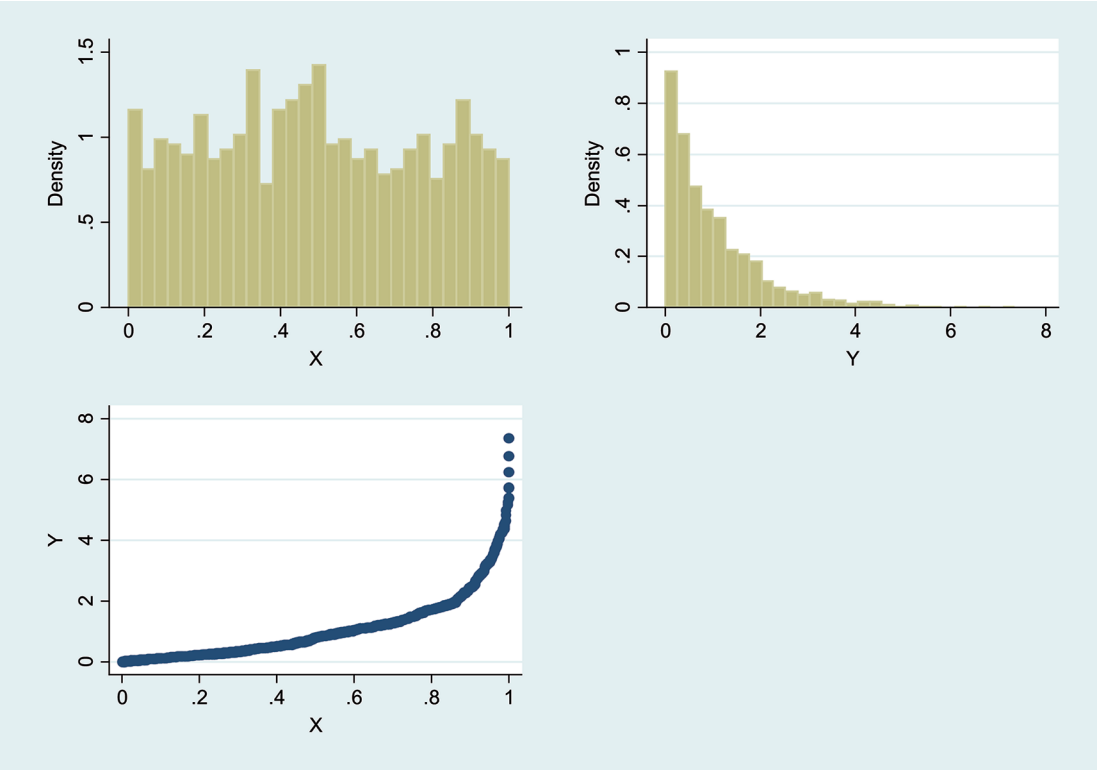
Esta desigualdad sólo tiene límites de -1 y 1, si las variables aleatorias provienen de las mismas distribuciones de probabilidades (ambas normales, ambas exponenciales, ambas uniformes, etc…).

Volviendo al ejemplo de asociar IMC con triglicéridos en sujetos obesos, antes de intentar dar un valor para la correlación de Pearson, deberíamos despejar el problema de si ellas tienen la misma distribución de probabilidades, pues si esto no es evidente, no es plausible proponer valores de correlación cercanos a 1, y se deberían calcular las integrales propuestas para afinar un valor: ¡gran desafío para el bioestadístico del equipo de investigación!

Como este artículo no pretende "profundidades matemáticas" sino ilustrar la idea, simularemos mil de pares de observaciones y calcularemos cuál es la correlación máxima a la que podríamos aspirar en una propuesta de investigación, si es que el par de variables en juego tuvie-

Comentarios de Bioestadística

ran distintas distribuciones de probabilidad, por ejemplo X sea uniforme en el intervalo [0,1] e Y sea exponencial con $\alpha = 1$. Para este ejemplo, la correlación máxima a la que se puede aspirar es $\rho = 0{,}9$ aproximadamente, el gráfico siguiente muestra los histogramas de X e Y, y, el gráfico de correlación entre X e Y:



La correlación muestral alcanzada es de $r_{XY} = 0{,}877$. Es decir, si alguien pide calcular un tamaño de muestra para detectar una correlación de 95%, debe saber que nunca logrará evidencia que avale su propósito.