Una Propuesta de Presentación del Tema de Correlación Simple

Por Prof. Carlos Araújo[1]

Departamento de Estadística, Facultad de Matemáticas, PUC

Las opiniones vertidas en este documento son de exclusiva responsabilidad del autor. El Departamento de Estadística de la PUC no necesariamente concuerda con dichas opiniones.
Todo comentario u observación puede ser enviado a: araujo@mat.puc.cl

Introducción

El concepto de Asociación o Correlación Estadística entre dos variables es, en general, otro de los temas de Estadística Descriptiva que no está correctamente presentado en los textos básicos de Estadística. Por ello es frecuente encontrar falsas y absurdas conclusiones supuestamente basadas en el estudio de correlaciones estadísticas como ejemplos para desprestigiar los métodos estadísticos (ver "La Correlación no Implica Relación Causa Efecto")

Insistimos que esta situación sólo refleja la incultura estadística de quienes proponen tales conclusiones y que dicha incultura es de responsabilidad de los que estamos a cargo de la enseñanza de esta disciplina.

Una Conceptualización de Correlación Estadística

Como se indicaba en la página 4 del Artículo 05, una adecuada interpretación del concepto de asociación o correlación en Estadística es fundamental independizar su definición de cualquier sugerencia sobre la relación causa-efecto.

Para ello se debe precisar en qué sentido se dice que dos variables cuantitativas están asociadas o correlacionadas estadísticamente. Al respecto se sugiere utilizar la siguiente caracterización:

En Estadística Descriptiva se dice que dos variables cuantitativas "están asociadas", "son dependientes", o "están correlacionadas" si cuando se aumentan los valores de una variable, los valores de la otra ***tienden*** *a:*

i) o bien a aumentar (y se dice que la asociación o dependencia es directa o que la correlación es positiva)

ii) o bien a disminuir (y se dice que la asociación o dependencia es inversa o que la correlación es negativa)

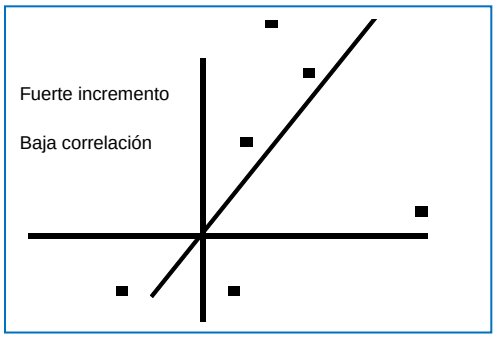
Cuando no se presenta esta tendencia se dice que las variables no están asociadas o no son dependientes o no están correlacionadas.

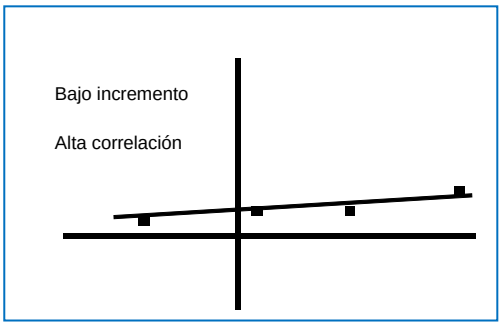
Para completar la conceptualización de la Correlación Estadística es necesario destacar que la "tendencia" se refiere a la seguridad o grado de certeza respecto del hecho de que *cuando una variable aumenta, la otra cumple o bien i) o bien ii).* Si se tiene una alta expectativa sobre esta ocurrencia, la Correlación es "alta"; si la expectativa de ocurrencia es "baja" la Correlación es baja.

No se debe confundir la calificación de "alto" o "bajo" con la magnitud del incremento o disminución. Los siguientes dibujos ilustran mejor esta situación.

---

[1] ***Prof. Carlos A. Araújo Ayesta*** fue Profesor del Centro Interamericano de Enseñanza de Estadística - CIENES (1967-1997), Asistente General del Director del CIENES (1974-1994) y Secretario Técnico de la Conferencia Interamericana de Estadística (CIE) de la OEA. A partir de enero de 2005 es Profesor en la Pontificia Universidad Católica de Chile - PUC.





La Correlación no Implica Relación Causa-Efecto

La asociación, correlación o dependencia en Estadística Descriptiva, **no implica relación causa-efecto.** En otras palabras, si cuando una variable aumenta la otra tiende a aumentar (o a disminuir) no es posible afirmar que esta última aumenta (o disminuye) **porque** la primera variable aumenta. Veamos un ejemplo. Sean:

X = número de iglesias en una ciudad
Y = número de bares en una ciudad

Si se consideran, por ejemplo las 10 principales ciudades de un país (las que tiene mayor número de habitantes) se obtiene la Matriz de Datos Bivariada

Unidad Inf. (ciudad)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
X = Nº Iglesias	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
Y = Nº Bares	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}

Si se representa gráficamente esta información en una nube de puntos seguramente mostrará un comportamiento positivo o directo en el sentido de que cuando una variable aumenta la otra tiende moderada o fuertemente a aumentar.

¿Esto significa entonces que para mantener o incrementar la fe religiosa se deberían promover el establecimiento de bares?, o bien que ¿para disminuir la ingesta de alcohol se debería cerrar iglesias? Obviamente esto no es así.

Si bien es cierto que en las ciudades con más iglesias existen más bares la explicación de **porqué** ocurre esta situación no se encuentra en el campo de la Estadística Descriptiva sino que dependerá de los especialistas del área de la cual provienen los datos.

Aunque que en este ejemplo no se requiere ser un especialista para darse cuenta de que existe una tercera variable Z = "nº de habitantes de la ciudad" que está asociada positivamente tanto con X como con Y y por lo tanto, cuando aumenta los bares, aumentan las iglesias y cualquier otro servicio (hospitales, cines, restaurantes, taxis, etc.) **porque** hay más habitantes.

Visualización Gráfica de la Correlación

Gráfico de Nube de Puntos

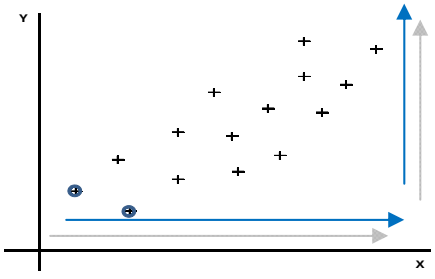


Fig. Nº 1 Correlación directa o positiva

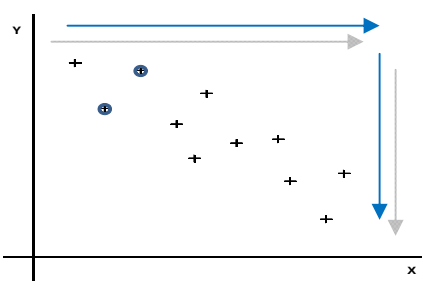


Fig. Nº 2 Correlación inversa o negativa

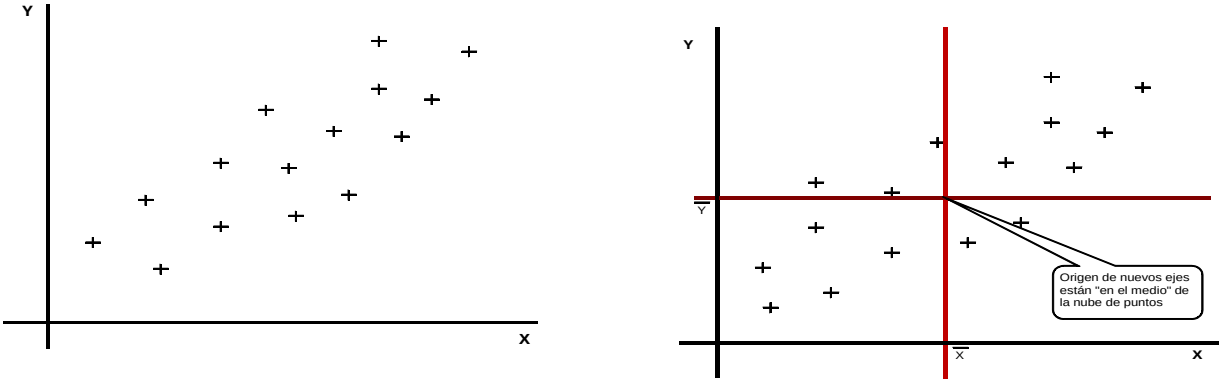
Es fácil "ver" que en la Figura Nº1 la "tendencia" es positiva como lo indican las flechas (en la medida en que se aumenta el valor de una variable se "esperan" valores "mayores" de la otra

variable aunque esta vinculación no ocurre siempre como lo muestran los puntos destacados de la Fig- Nº 1.

De igual forma, en la Figura Nº2 la "tendencia" es negativa como lo indican las flecas (en la medida en que se aumenta el valor de una variable se "esperan valores "menores" de la otra variable aunque los puntos destacados de la Fig- Nº 2 muestran que, en ese caso particular, al aumentar una variable la otra disminuye.

Comportamiento de los puntos en los Cuadrantes

Consideremos por ejemplo la situación gráfica de la Fig. Nº 1 y realicemos un cambio de coordenadas de tal forma que el nuevo origen sea el punto medio de las coordenada originales.



Este nuevo origen está determinado por el punto $\left(\overline{X},\overline{Y}\right)$.

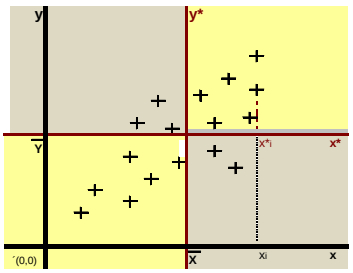
De esta forma el nuevo origen se encontrará situado "en el medio" de la nube de puntos i por lo tanto aparecerán distribuidos entre los cuatro cuadrantes definidos por estos nuevos ejes y podremos afirmar que:

Si "son más importantes" los puntos del Primer y Tercer Cuadrante que los puntos del Segundo o Cuarto existirá asociación directa o positiva

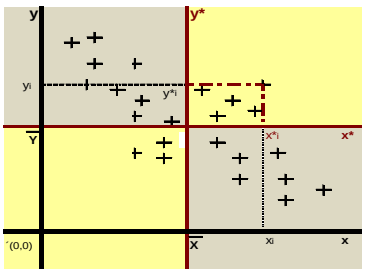
Si "son más importantes" los puntos del Segundo y Cuarto Cuadrante que los pontos del Primero y Tercero, debe existir una asociación inversa o negativa

Si los puntos "son igualmente importantes" o se distribuyen "por igual" entre los cuatro Cuadrantes es señal de que no hay asociación o dependencia.

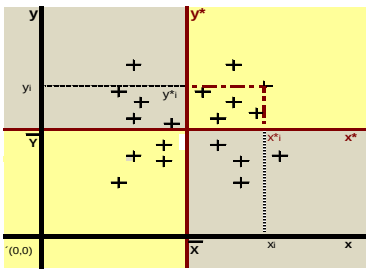
En general el gráfico corresponderá a una de las siguientes situaciones:



Correlación positiva



Correlación negativa



No hay Correlación

La importancia de los puntos en un cuadrante se "miden" por la cantidad de punto y por su distancia a los ejes (cuanto más distante, más importante)

Un Indicador de Asociación: la Covarianza

Intentaremos construir un indicador sobre el comportamiento de los puntos en los cuadrantes definidos en el punto anterior de forma tal que nos informe no sólo sobre si la correlación es positiva o negativa, sino además sobre la "importancia" de esta correlación, es decir sobre si la "tendencia" que muestra la asociación es "alta", "mediana", "baja" o inexistente.

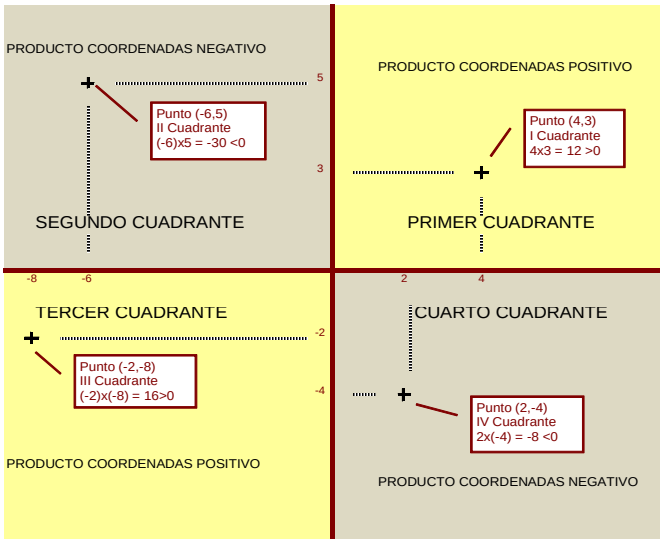
Usaremos las coordenadas definidas por el cambio de ejes, es decir si (x_i, y_i) es el punto en los ejes originales, entonces (x_i^*, y_i^*) donde : $x_i^* = x_i - \overline{X}$; $y_i^* = y_i - \overline{Y}$ serán las coordenadas del mismo punto en los nuevos ejes.

El producto de estas coordenadas nos informa si qué cuadrante se encuentra el punto y además nos indica se está cerca de los ejes, es decir

Si $x^* \times y^* > 0$ entonces el punto (x^*, y^*) está en el Primer Cuadrante o en el Tercer Cuadrante.

Si, en cambio, $x^* \times y^* < 0$ entonces (x^*, y^*) está en el Segundo o en el Cuarto Cuadrante.

La situación se muestra en la figura de al lado



Entonces, si $\sum_{i=1}^{N} x^* \times y^* > 0$ es porque los productos positivos "son importantes" más que los negativos y muestra entonces que los puntos del Primer y Tercer Cuadrante son "más importantes" que los del Segundo y Cuarto Cuadrante. En consecuencia, si $\sum_{i=1}^{N} x^* \times y^* > 0$ es señal de asociación o correlación directa o positiva.

Si en cambio $\sum_{i=1}^{N} x^* \times y^* < 0$ es porque los productos negativos "pesan" más que los positivos y muestra entonces que los puntos del Segundo y Cuarto Cuadrante son "más importantes" que del Primer y Tercer Cuadrante. En consecuencia, si $\sum_{i=1}^{N} x^* \times y^* < 0$ es señal de asociación o correlación inversa o negativa.

En consecuencia, si $P = \{ prod_i = (x_i \times y_i) \mid i = 1,2,...,n \}$ es el conjunto de los productos de las coordenadas, se elije como indicador de la asociación o correlación al "mejor" representante de estos productos el cual, por el principio de mínimos cuadrados, corresponde a la Media de dichos productos, es decir: $\overline{prod} = \frac{1}{n}\sum_{i=1}^{n} prod_i = \frac{1}{n}\sum_{i=1}^{n} x_i^* \times y_i^*$

Por lo tanto se define la Covarianza entre X e Y mediante

$$Cov(X,Y) = \frac{\sum_{i=1}^{n} x_i^* \times y_i^*}{n} = \frac{\sum_{i=1}^{n} (x_i - \overline{X})(y_i - \overline{Y})}{n} = \frac{\sum_{i=1}^{n} x_i \times y_i - n\overline{X}\,\overline{Y}}{n}$$

Entonces

Si $Cov(X,Y) > 0$ la asociación o correlación es directa o positiva.

Si $Cov(X,Y)<0$ la asociación o correlación es inversa o negativa.

Si $Cov(X,Y)\approx 0$ no hay asociación o correlación.

El problema es ¿qué valores de la Covarianza se consideran cerca del cero? Tal cual está construido el indicador resulta imposible dar una respuesta porque la Covarianza depende de la unidad de medida de las variables y entonces para "los mismos" datos (difieren sólo en la unidad de medida) el valor de la Covarianza su valor puede ser más grande o más chico.

Será necesario construir otro indicador que nos informe sobre si el valor de la Covarianza es alto o bajo y para ello encontraremos primeramente una cota para la Covarianza.

Para ello construiremos otro indicador llamado Coeficiente de Correlación Simple o Coeficiente de Correlación Lineal.

Otro Indicador de Asociación: El Coeficiente de Correlación

La Desigualdad de Cauchy - Schwart

Sean $a_i \in R, b_i \in R \;\; i=1,2,...,N$. Entonces: $\left(\sum_{i=1}^{N} a_i b_i\right)^2 \le \left(\sum_{i=1}^{N} a_i^2\right)\times\left(\sum_{i=1}^{N} b_i^2\right)$

Además la igualdad se verifica si y sólo si: $a_i = kb_i \;\; \forall\, i=1,2,...,N$

La demostración de esta desigualdad no forma parte de un curso de Estadística. En el caso de que esta demostración no sea conocida por los alumnos, mediante algunos ejemplos presentados en una hoja electrónica, el alumno puede entender rápidamente su significado y "aceptar "fácilmente su veracidad.

Si aplicamos la desigualdad reemplazando a_i por x_i^* y b_i por y_i^* se tiene que

$\left(\sum_{i=1}^{N} x_i^* y_i^*\right)^2 \le \left(\sum_{i=1}^{N}\left(x_i^*\right)^2\right)\times\left(\sum_{i=1}^{N}\left(y_i^*\right)^2\right)$. Dividiendo la desigualdad por N^2 se tiene que:

$$\left(\frac{\sum_{i=1}^{N} x_i^* y_i^*}{N}\right)^2 \le \left(\frac{\sum_{i=1}^{N}\left(x_i^*\right)^2}{N}\right)\times\left(\frac{\sum_{i=1}^{N}\left(y_i^*\right)^2}{N}\right) \Rightarrow \left(Cov(X,Y)\right)^2 \le V(X)\times V(Y)$$

Además, la igualdad es válida si y sólo si

$$y_i^* = kx_i^* \;\; \forall\, i=1,2,...,n \Leftrightarrow Y_i - \overline{Y} = kX_i - k\overline{X} \;\; \forall\, i=1,2,...,n \Leftrightarrow Y_i = kX_i + \overbrace{\left(\overline{Y} - k\overline{X}\right)}^{m} \;\; \forall\, i=1,2,...,n$$

Es decir la igualdad es válida si y sólo si, Y_i es una combinación lineal de X_i cualquiera sea el índice "i". Por ser "sólo si" tendremos entonces que si Y_i no es una combinación lineal de X_i entonces $\left(Cov(X,Y)\right)^2$ será estrictamente menor que $V(X)\times V(Y)$.

En consecuencia, puesto que $\left(Cov(X,Y)\right)^2 \le V(X)\times V(Y)$ se tiene que: $\left|Cov(X,Y)\right| \le \sqrt{V(X)V(Y)}$ por lo que $-\sqrt{V(X)V(Y)} \le Cov(X,Y) \le \sqrt{V(X)V(Y)}$

y dividiendo estas desigualdades por $\sqrt{V(X)V(Y)}$ se tiene que:

$-1 \leq \dfrac{Cov(X,Y)}{\sqrt{V(X)V(Y)}} \leq 1$. La igualdad a 1 (o a -1) ocurre si y sólo si $Y = a + bX$

El número $\dfrac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$ se define como **el Coeficiente de Correlación Simple o Lineal entre X e Y** y se designa por $\rho_{X,Y}$. Es decir: $\rho_{X,Y} = \dfrac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$

Si $\rho_{X,Y} = 1$ la correlación es la máxima correlación positiva o directa; si $\rho_{X,Y} = -1$ la correlación es la máxima correlación negativa o inversa y si $\rho_{X,Y} \approx 0$ no existe correlación o dependencia.

Conviene notar que si existe una dependencia funcional perfecta pero no lineal entre X e Y (por ejemplo $Y = X^2$ o bien $Y = \ln X$) la asociación o dependencia funcional sería la máxima, aunque el coeficiente de correlación NO es igual a 1 porque dicha dependencia no es lineal (es igual a 1 si y **solo si** la relación es lineal. Por esta razón al coeficiente de correlación se le llama coeficiente de correlación lineal.

Si bien no existe una regla general para decir si una correlación es alta media o baja, en general se podría proponer adoptar el siguiente criterio: