Correlación

En probabilidad y estadística, la **correlación** indica la fuerza y la dirección de una relación lineal y proporcionalidad entre dos variables estadísticas. Se considera que dos variables cuantitativas están correlacionadas cuando los valores de una de ellas varían sistemáticamente con respecto a los valores homónimos de la otra: si tenemos dos variables (A y B) existe correlación si al aumentar los valores de A lo hacen también los de B y viceversa. La correlación entre dos variables no implica, por sí misma, ninguna relación de causalidad (Véase *cum hoc ergo propter hoc*).

Fuerza, sentido y forma de la correlación

La relación entre dos variables cuantitativas queda representada mediante la línea de mejor ajuste, trazada a partir de la nube de puntos. Los principales componentes elementales de una línea de ajuste y, por lo tanto, de una correlación, son la fuerza, el sentido y la forma:

- La **fuerza** extrema según el caso, mide el grado en que la línea representa a la nube de puntos: si la nube es estrecha y alargada, se representa por una línea recta, lo que indica que la relación es *fuerte*; si la nube de puntos tiene una tendencia elíptica o circular, la relación es *débil*.
- El **sentido** mide la variación de los valores de B con respecto a A: si al crecer los valores de A lo hacen los de B, la relación es *positiva*; si al crecer los valores de A disminuyen los de B, la relación es *negativa*.
- La **forma** establece el tipo de línea que define el mejor ajuste: la línea recta, la curva monotónica o la curva no monotónica.

Coeficientes de correlación

Existen diversos coeficientes que miden el grado de correlación, adaptados a la naturaleza de los datos. El más conocido es el coeficiente de correlación de Pearson (introducido en realidad por Francis Galton), que se obtiene dividiendo la covarianza de dos variables por el producto de sus desviaciones estándar. Otros coeficientes son:

- Coeficiente de correlación de Spearman
- Correlación canónica
- Coeficiente de Correlación Intraclase

Interpretación geométrica

Dados los valores muestrales de dos variables aleatorias $X(x_1, \ldots, x_n)$e $Y(y_1, \ldots, y_n)$, que pueden ser consideradas como vectores en un espacio a *n* dimensiones, puden construirse los "vectores centrados" como:

$X(x_1 - \bar{x}, \ldots, x_n - \bar{x})$e $Y(y_1 - \bar{y}, \ldots, y_n - \bar{y})$.

El coseno del ángulo alfa entre estos vectores es dada por la fórmula siguiente:

$$\cos(\alpha) = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

Pues $\cos(\alpha)$es el coeficiente de correlación muestral de Pearson. El coeficiente de correlación es el coseno entre ambos vectores centrados:

- Si *r* = 1, el ángulo $\alpha = 0°$, ambos vectores son colineales (paralelos).
- Si *r* = 0, el ángulo $\alpha = 90°$, ambos vectores son ortogonales.
- Si *r* =-1, el ángulo $\alpha = 180°$, ambos vectores son colineales de dirección opuesto.

Más generalmente: $\alpha = \arccos(r)$.

Por supuesto, del punto vista geométrica, no hablamos de *correlación lineal*: el coeficiente de correlación tiene siempre un sentido, cualquiera si que sea su valor entre -1 y 1. Nos informa de modo preciso, no tanto sobre el grado de dependencia entre las variables, que sobre su distancia angular en la hiperesfera a *n* dimensiones.

La Iconografía de las correlaciones es un método de análisis multidimensional que reposa en esta idea. La correlacion lineal se da cuando en una nube de puntos estos se encuentran o se distribuyen alrededor de una recta.

Distribución del coeficiente de correlación

El coeficiente de correlación muestral de una muestra es de hecho una varible aleatoria, eso significa que si repetimos un experimento o consideramos diferentes muestras se obtendrán valores diferentes y por tanto el coeficiente de correlación muestral calculado a partir de ellas tendrá valores ligeramente diferentes. Para muestras grandes la variación en dicho coeficiente será menor que para muestras pequeñas. R. A. Fisher fue el primero en determinar la distribución de probabilidad para el coeficiente de correlación.

Si las dos variables aleatorias que trata de relacionarse proceden de una distribución gaussiana bivariante entonces el coeficiente de correlación *r* sigue una distribución de probabilidad dada por:[1][2]

$$f(r) = \frac{(n-2)\,\Gamma(n-1)\,(1-\rho^2)^{\frac{n-1}{2}}\,(1-r^2)^{\frac{n-4}{2}}}{\sqrt{2\pi}\,\Gamma\left(n-\frac{1}{2}\right)(1-\rho r)^{n-\frac{3}{2}}}\;{}_2\mathbf{F}_1\left(\frac{1}{2},\frac{1}{2};\frac{2n-1}{2};\frac{\rho r+1}{2}\right)$$

donde:

$\mathbf{\Gamma}$ es la distribución gamma

${}_2\mathbf{F}_1(a,b;c;z)$ es la función gaussiana hipergeométrica.

Nótese que $E(r) = \rho - \frac{\rho(1-\rho^2)}{2(n-1)} + \cdots$, por tanto *r* es estimador sesgado de ρ.

Puede obtenerse un estimador aproximado no sesgado resolviendo la ecuación:

$$r = E(r) = \rho - \frac{\rho(1-\rho^2)}{2(n-1)} \text{for } \rho$$

Aunque, la solucón:

$$\breve{\rho} = r\left[1 + \frac{1-r^2}{2(n-1)}\right]$$

es subóptima. Se puede obtener un estimador sesgado con mínima varianza para grandes valores de *n*, con sesgo de orden $\frac{1}{n-1}$ buscando el máximo de la expresión:

$$\log f(r), \text{ i.e. } \hat{\rho} = r\left[1 - \frac{1-r^2}{2(n-1)}\right]$$

En el caso especial de que $\rho = 0$, la distribución original puede ser reescrita como:

$$f(r) = \frac{(1-r^2)^{\frac{n-4}{2}}}{\mathbf{B}\left(\frac{1}{2},\frac{n-2}{2}\right)}$$

donde $\mathbf{B}$ es la función beta.

Referencias

[1] Kenney, J. F. and Keeping, E. S., *Mathematics of Statistics*, Pt. 2, 2nd ed. Princeton, NJ: Van Nostrand, 1951.

[2] Correlation Coefficient - Bivariate Normal Distribution (http://mathworld.wolfram.com/CorrelationCoefficientBivariateNormalDistribution.html)

Enlaces externos

- Diccionario Estadístico - Divestadística (http://www.divestadistica.es/es/diccionario_estadistico.html#C) (en castellano)
- (http://cajael.com/mestadisticos/T1EDescriptiva/node20.php) Simulación de la correlación entre dos variables discretas con R (lenguaje de programación)

Fuentes y contribuyentes del artículo

Correlación *Fuente*: http://es.wikipedia.org/w/index.php?oldid=61001651 *Contribuyentes*: Acratta, Alhen, Bucho, Camilo, Davius, Diegusjaimes, Egozcue, El Quinche, Grillitus, Humberto, Jkbw, Juan Mayordomo, Lauranrg, Matdrodes, Mxcatania, Posesso, Redeyes, Technopat, Tirithel, 52 ediciones anónimas

Licencia