

Cálculo práctico de la covarianza

Para calcular la covarianza de una distribución estadística bidimensional usando la definición es necesario utilizar todas las desviaciones de los datos, que se basan en el cálculo previo de las medias. Esto presenta algunos problemas prácticos:

- * Las medias fácilmente pueden ser un número con muchas cifras decimales; en ese caso, las operaciones se complican y, además, se van acumulando errores.
- * Si cambia alguno de los datos, cambian las medias, y por tanto hay que recalcular todas las desviaciones.

Por estos motivos, se utiliza para calcular la covarianza otro método más sencillo desde el punto de vista práctico, que se basa en una propiedad de la covarianza.

Propiedad de la covarianza

La covarianza de un conjunto de datos bidimensional es igual a la media de los productos de los datos menos el producto de las medias.

Expresión simbólica

Consideramos los valores $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$; llamamos \bar{x} e \bar{y} a las medias de cada variable y σ_{xy} a la covarianza. Entonces, se verifica:

$$\sigma_{xy} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n} = \frac{\sum x_i y_i}{n} - \bar{x} \bar{y}$$

Demostración

Para facilitarte la comprensión de la demostración general, comenzamos por la demostración en un caso más sencillo y la usamos de guía para la generalización.

Demostración para dos datos

Consideramos los datos, (x_1, y_1) y (x_2, y_2) ; las medias: « $\bar{x} = (x_1 + x_2) : 2$ » y « $\bar{y} = (y_1 + y_2) : 2$ ».

Calculamos la covarianza como la media de los productos de las dos desviaciones:

$$\begin{aligned} \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y})}{2} &= \frac{x_1 y_1 - x_1 \bar{y} - \bar{x} y_1 + \bar{x} \bar{y} + x_2 y_2 - x_2 \bar{y} - \bar{x} y_2 + \bar{x} \bar{y}}{2} = \\ &= \frac{x_1 y_1 + x_2 y_2 - (x_1 + x_2) \bar{y} - (y_1 + y_2) \bar{x} + 2 \bar{x} \bar{y}}{2} = \frac{x_1 y_1 + x_2 y_2}{2} - \frac{(x_1 + x_2)}{2} \bar{y} - \frac{(y_1 + y_2)}{2} \bar{x} + \bar{x} \bar{y} = \\ &= \frac{x_1 y_1 + x_2 y_2}{2} - \bar{x} \bar{y} - \bar{y} \bar{x} + \bar{x} \bar{y} = \frac{x_1 y_1 + x_2 y_2}{2} - \bar{x} \bar{y} \end{aligned}$$

Vemos que hemos obtenido la media de los productos de los datos menos el producto de las medias.

Demostración general

Sabemos que « $\bar{x} = (\sum x_i) : n$ » y « $\bar{y} = (\sum y_i) : n$ »

Calculamos la covarianza como la media de los productos de las desviaciones:

$$\begin{aligned} \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n} &= \frac{\sum (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y})}{n} = \frac{\sum x_i y_i}{n} - \frac{\sum x_i}{n} \bar{y} - \bar{x} \frac{\sum y_i}{n} + \frac{n \bar{x} \bar{y}}{n} = \\ &= \frac{\sum x_i y_i}{n} - \bar{x} \bar{y} - \bar{x} \bar{y} + \bar{x} \bar{y} = \frac{\sum x_i y_i}{n} - \bar{x} \bar{y} \end{aligned}$$

Vemos que hemos obtenido la media de los productos de los datos menos el producto de las medias.