Estadísticos básicos

▼ Varianza.

En estadística, la varianza es una medida que describe la dispersión o la propagación de un conjunto de datos. En otras palabras, la varianza indica qué tan dispersos están los valores de un conjunto de datos con respecto a su media aritmética. Se calcula como la media de los cuadrados de las diferencias entre cada valor y la media.

La fórmula para calcular la varianza muestral s^2 es la siguiente:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

donde:

- s^2 es la varianza muestral,
- n es el número de observaciones en el conjunto de datos,
- x_i son los valores individuales en el conjunto de datos,
- $\bar{x}$ es la media aritmética de los valores en el conjunto de datos.

Es importante destacar que en el denominador se utiliza (n-1) en lugar de n cuando se trata de la varianza muestral. Esto se hace para corregir el sesgo muestral y proporcionar una estimación no sesgada de la varianza poblacional.

La varianza puede ser una medida útil para comprender la dispersión de los datos, pero también tiene algunas limitaciones, especialmente cuando se trabaja con conjuntos de datos que tienen valores extremos. En tales casos, la desviación estándar (que es simplemente la raíz cuadrada de la varianza) a menudo se prefiere, ya que está en la misma escala que los datos originales y es más fácil de interpretar.

▼ Covarianza.

La covarianza es una medida estadística que describe la relación entre dos variables aleatorias. Más específicamente, la covarianza indica cómo dos variables cambian juntas. Si la covarianza es positiva, significa que las dos variables tienden a aumentar o disminuir juntas. Si es negativa, significa que una variable tiende a aumentar cuando la otra disminuye, y viceversa. Si la covarianza es cercana a cero, indica que no hay una relación lineal fuerte entre las dos variables.

La fórmula para calcular la covarianza muestral $(\operatorname{cov}(X, Y))$ entre dos conjuntos de datos, X e Y, con n observaciones es:

$$\operatorname{cov}(X, Y)=\frac{\sum_{i=1}^{n}(x_i-\bar{X})\cdot(y_i-\bar{Y})}{n-1}$$

donde:

- $\operatorname{cov}(X, Y)$ es la covarianza muestral entre X e Y,
- x_i e y_i son las observaciones individuales en los conjuntos de datos,
- $\bar{X}$ y $\bar{Y}$ son las medias de X e Y, respectivamente,
- n es el número de observaciones.

La covarianza puede ayudar a comprender la dirección de la relación entre dos variables, pero tiene la limitación de no ser fácilmente interpretable en términos de la fuerza de la relación debido a su dependencia de las escalas de las variables. Para superar esta limitación, se utiliza la correlación, que es la covarianza normalizada por el producto de las desviaciones estándar de las dos variables. La correlación varía entre -1 y 1, donde -1 indica una relación negativamente perfecta, 1 indica una relación positivamente perfecta y 0 indica la ausencia de una relación lineal.

▼ Correlación.

La correlación es una medida estadística que describe la fuerza y la dirección de una relación lineal entre dos variables. A diferencia de la covarianza, la correlación está normalizada, lo que significa que siempre toma valores entre -1 y 1. Una correlación de 1 indica una relación lineal positiva perfecta, -1 indica una relación lineal negativa perfecta, y 0 indica la ausencia de una relación lineal.

La fórmula para calcular la correlación muestral (r) entre dos conjuntos de datos, X e Y , con n observaciones es:

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{X})\cdot(y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{X})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{Y})^2}}$$

donde:

- r es la correlación muestral entre X e Y,
- x_i e y_i son las observaciones individuales en los conjuntos de datos,
- $\bar{X}$ y $\bar{Y}$ son las medias de X e Y, respectivamente.

La correlación es una medida útil porque no está afectada por las unidades de medida de las variables y proporciona información sobre la dirección y la fuerza de la relación entre ellas. Sin embargo, es importante señalar que la correlación no implica causalidad. Dos variables pueden estar correlacionadas, pero eso no significa que una variable cause la otra; puede haber otras variables o factores en juego. Además, la correlación solo mide relaciones lineales y puede no capturar relaciones no lineales entre las variables.