Correlation & Covariance

Correlation and covariance are measures of the relationship between two variables.

- Covariance: Measures the direction of the relationship (positive or negative).
- Correlation: Measures both the direction and the strength of the relationship on a standardized scale (-1 to +1).

These measures are fundamental in statistics and data science, especially for exploratory data analysis and feature selection in machine learning.

1. Covariance

Covariance indicates whether two variables increase or decrease together.

- **Formula**

Population Covariance

$$Cov(X,Y) = \frac{\Sigma\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{N}$$

Sample Covariance

$$Cov(X,Y) = \frac{\Sigma\left(x_i-\bar{x}\right)\left(y_i-y\right)}{N-1}$$

These are the formula for finding Population and Sample Covariance.

where,

- x_i = data value of x
- y_i = data value of y
- $\bar{x}$ = mean of x
- $\bar{y}$ = mean of y
- N = number of data values.

- **Interpretation**
  - Cov(X, Y) > 0: Positive relationship (when X increases, Y tends to increase).
  - Cov(X, Y) < 0: Negative relationship (when X increases, Y tends to decrease).
  - Cov(X, Y) ≈ 0: No linear relationship.

- **Limitations**
  - Units depend on the variables → not standardized.
  - Difficult to compare across datasets.

- **Example**
  - Temperature vs Ice cream sales → Positive covariance.

2. Correlation

Correlation standardizes covariance, giving a measure between -1 and +1.

- **Formula**

$$PXY = \frac{COV_{XY}}{\sigma_X \sigma_Y}$$

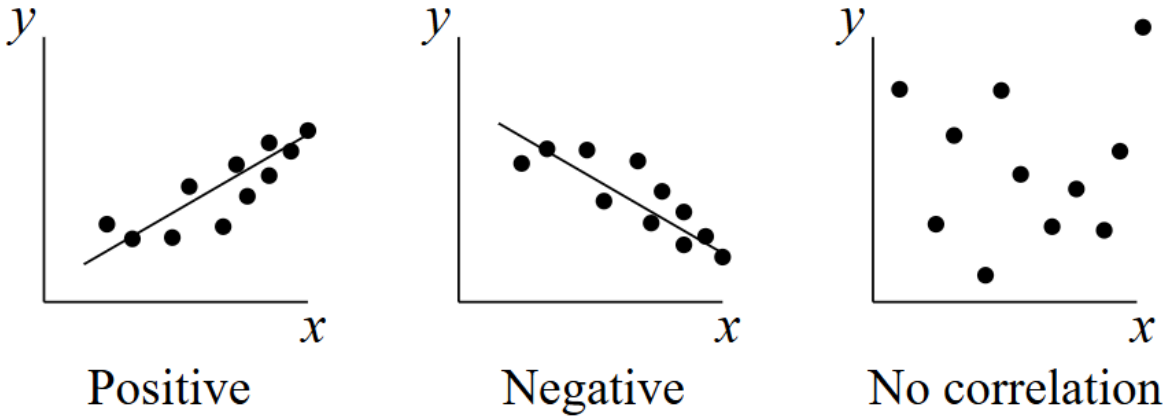
$$r = \frac{N * \sum xy - (\sum x)(\sum y)}{\sqrt{\left[N * \sum x^2 - (\sum x)^2\right] * \left[N * \sum y^2 - (\sum y)^2\right]}}$$

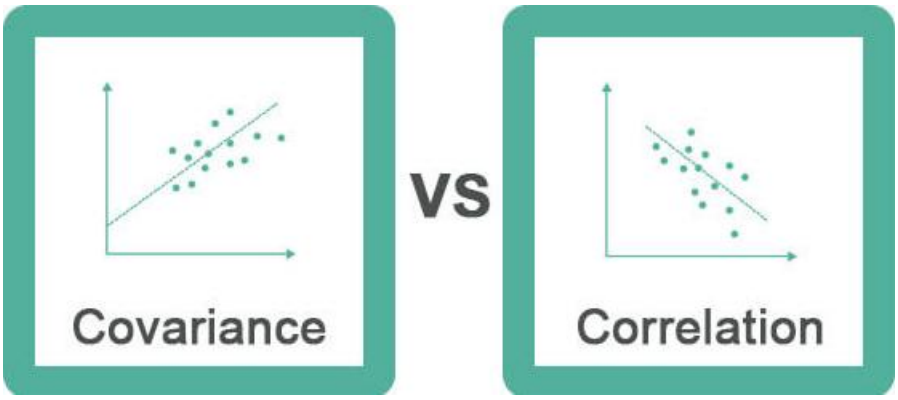
- **Interpretation**
  - r = +1 → Perfect positive linear relationship.
  - r = -1 → Perfect negative linear relationship.
  - r = 0 → No linear relationship.

- **Types of Correlation**
  - **Pearson Correlation** — Measures linear relationships.
  - **Spearman Correlation** — Rank-based, used for monotonic relationships.
  - **Kendall's Tau** — Non-parametric correlation.

- **Example**
  - Height and weight → Positive correlation.
  - Study hours and number of errors → Negative correlation.

3. Visual Examples





4. Applications

- **Exploratory Data Analysis (EDA)** → Discover relationships between features.
- **Feature Selection** → Remove redundant variables in machine learning.
- **Multicollinearity Detection** → Check correlation between predictors in regression models.

5. Common Mistakes

- **Correlation ≠ Causation →** Just because variables are correlated doesn't mean one causes the other.
- Using **Pearson correlation** on non-linear data.
- Ignoring the impact of **outliers**, which can distort correlation.