Correlation

Correlation measures the strength and direction of the linear relationship between two variables. There are several types of correlation coefficients, each with its own characteristics and applications. Here are the most commonly used types:

1. Pearson Correlation Coefficient

- **Formula**:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

- **Characteristics**:
  - Measures the strength and direction of the linear relationship between two continuous variables.
  - Ranges from -1 to 1.
  - Positive values indicate a positive linear relationship, while negative values indicate a negative linear relationship.
  - A value of 0 indicates no linear relationship.
- **Use Case**: Used when both variables are continuous and the relationship is linear.

2. Spearman's Rank Correlation Coefficient

- **Formula**:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the ranks of corresponding variables, and n is the number of data points.

- **Characteristics**:
  - Measures the strength and direction of the monotonic relationship between two variables.
  - Ranges from -1 to 1.
  - Positive values indicate a positive monotonic relationship, while negative values indicate a negative monotonic relationship.
  - A value of 0 indicates no monotonic relationship.
- **Use Case**: Used when data are ordinal or when the relationship is not linear but monotonic.

3. Kendall's Tau

- **Formula**:

$$\tau = \frac{(P - Q)}{\sqrt{(P + Q + T_1)(P + Q + T_2)}}$$

where P is the number of concordant pairs, Q is the number of discordant pairs, and T_1 and T_2 are the number of ties in the variables.

- **Characteristics**:
  - Measures the strength and direction of the association between two variables.
  - Ranges from -1 to 1.
  - Positive values indicate a positive association, while negative values indicate a negative association.
  - Less sensitive to ties than Spearman's rank.
- **Use Case**: Used for ordinal data or when data may contain a large number of tied ranks.

4. Point-Biserial Correlation Coefficient

- **Formula**:

$$r_{pb} = \frac{M_1 - M_0}{s} \cdot \sqrt{\frac{n_1 \cdot n_0}{n^2}}$$

where M_1 and M_0 are the means of the continuous variable for the two groups defined by the binary variable, s is the standard deviation of the continuous variable, and n_1 and n_0 are the number of observations in each group.

- **Characteristics**:
  - Measures the strength and direction of the association between a binary variable and a continuous variable.
  - Ranges from -1 to 1.
- **Use Case**: Used when one variable is binary (dichotomous) and the other is continuous.

5. Phi Coefficient

- **Formula**:

  ϕ=ad−bc(a+b)(c+d)(a+c)(b+d)\phi = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}ϕ=(a+b)(c+d)(a+c)(b+d)ad−bc

  where aaa, bbb, ccc, and ddd are the frequencies in a 2x2 contingency table.

- **Characteristics**:
  - Measures the association between two binary variables.
  - Ranges from -1 to 1.

- **Use Case**: Used in the context of a 2x2 contingency table where both variables are binary.

6. Cramér's V

- **Formula**:

  $V = \sqrt{\frac{\chi^2}{n \cdot \min(k - 1, r - 1)}}$

  where χ^2 is the chi-square statistic, n is the total number of observations, k is the number of categories in one variable, and r is the number of categories in the other variable.

- **Characteristics**:
  - Measures the strength of association between two nominal (categorical) variables.
  - Ranges from 0 to 1.
- **Use Case**: Used for nominal data to determine the strength of association between categorical variables.

Summary

- **Pearson**: Best for linear relationships between continuous variables.
- **Spearman**: Useful for monotonic relationships and ordinal data.
- **Kendall's Tau**: Suitable for ordinal data and ties.
- **Point-Biserial**: For associations between a binary and a continuous variable.
- **Phi Coefficient**: For associations between two binary variables.
- **Cramér's V**: For association strength between nominal variables.

Choosing the right correlation coefficient depends on the type of data you have and the nature of the relationship you wish to measure.

1. Pearson Correlation Coefficient

- **Effect of Changing Origin**:
  - **No Effect**: Shifting the data along the y-axis (changing the origin) does not affect the Pearson correlation coefficient. This is because Pearson correlation measures the strength and direction of the linear relationship, not the absolute values of the variables.
- **Effect of Changing Scale**:
  - **No Effect**: Multiplying the data by a constant (scaling) also does not affect the Pearson correlation coefficient. This is because Pearson correlation is a standardized measure, and scaling both variables by the same factor cancels out in the calculation.

2. Spearman's Rank Correlation Coefficient

- **Effect of Changing Origin**:
  - **No Effect**: Adding a constant to all values does not change the ranks of the data. Spearman's correlation is based on the ranks of the data, not their actual values.
- **Effect of Changing Scale**:

    - **No Effect**: Multiplying all values by a constant does not affect the ranks and therefore does not change Spearman's correlation. It remains unaffected by changes in the scale of the data.

3. Kendall's Tau

- **Effect of Changing Origin**:
    - **No Effect**: Like Spearman's, Kendall's Tau is based on the ordering of the data (concordant and discordant pairs), so shifting the origin does not affect the correlation.
- **Effect of Changing Scale**:
    - **No Effect**: Multiplying the data by a constant does not change the relative order of the data, so Kendall's Tau remains unchanged by scaling.

4. Point-Biserial Correlation Coefficient

- **Effect of Changing Origin**:
    - **No Effect**: Similar to Pearson correlation, shifting the origin does not affect the point-biserial correlation coefficient.
- **Effect of Changing Scale**:
    - **No Effect**: Scaling the continuous variable by a constant does not change the point-biserial correlation coefficient because it is a measure of the relationship between a continuous variable and a binary variable.