1. Definition of Covariance and Formula

Covariance measures the extent to which two random variables change together. It is calculated as the average of the product of the deviations of each pair of variables from their respective means. The formula for covariance between two variables \(X\) and \(Y\) is:

\[ \text{Cov}(X, Y) = \frac{1}{n} \sum\_{i=1}^{n} (X\_i - \bar{X})(Y\_i - \bar{Y}) \]

where \(X\_i\) and \(Y\_i\) are the individual data points, \(\bar{X}\) and \(\bar{Y}\) are the means of \(X\) and \(Y\), and \(n\) is the number of data points.

2. What Makes Correlations Better than Covariance?

Correlations are considered better than covariance because they provide a standardized measure of the relationship between two variables, which is independent of their units. While covariance can show the direction of a relationship, its magnitude is influenced by the scale of the variables, making it difficult to interpret. Correlation, on the other hand, normalizes covariance by dividing it by the product of the standard deviations of the variables, yielding a dimensionless value between -1 and 1, which makes comparisons easier.

3. Pearson and Spearman Correlation

\*\*Pearson Correlation\*\* measures the linear relationship between two continuous variables. It is computed by dividing the covariance of the variables by the product of their standard deviations. The formula is:

\[ r = \frac{\text{Cov}(X, Y)}{\sigma\_X \sigma\_Y} \]

Spearman Correlation assesses the monotonic relationship between two variables by using rank-order data. It calculates the Pearson correlation coefficient on the ranks of the data rather than the raw data itself. The formula for Spearman's rank correlation coefficient is:

\[ \rho = 1 - \frac{6 \sum d\_i^2}{n(n^2 - 1)} \]

where \(d\_i\) is the difference between the ranks of corresponding values and \(n\) is the number of data points.

4. Advantages of Spearman Correlation over Pearson Correlation

Spearman Correlation has several advantages over Pearson Correlation. It does not assume that the relationship between variables is linear, making it more robust for non-linear relationships. It also reduces the impact of outliers and skewed data because it uses ranks rather than raw values. Therefore, Spearman Correlation is more suitable for ordinal data or when the assumptions of Pearson Correlation (such as normality and homoscedasticity) are not met.

signm5. Central Limit Theorem

The Central Limit Theorem (CLT) states that the sampling distribution of the sample mean will approach a normal distribution as the sample size becomes large, regardless of the population's distribution shape. This is a fundamental principle in statistics because it justifies the use of normal distribution approximations in hypothesis testing and confidence interval estimation, especially when dealing with large samples. Essentially, the CLT allows for making inferences about population parameters even when the population distribution is unknown or not normal.