Statistical Tools for Data Science

Composition: The process of combining two or more variables to create a new variable. This can involve simple arithmetic operations or more complex statistical techniques.

Example: Combining height and weight to calculate body mass index (BMI) is a common example of composition in statistics.

Distribution: The pattern of values that a variable takes on, including the frequency and spread of those values.

Example: The distribution of test scores in a classroom can be visualized using a histogram, which shows the frequency of scores across a range of values.

Relationship: The association between two or more variables, which can be positive, negative, or neutral.

Example: There may be a positive relationship between the number of hours studied and the grade received on an exam, indicating that as one variable increases, so does the other.

Comparison: The act of examining two or more variables or groups to identify similarities and differences.

Example: Comparing the performance of two different advertising campaigns to determine which one was more effective in reaching the target audience.

Null hypothesis: A statement that assumes there is no significant difference or relationship between variables.

Example: The null hypothesis for a study examining the effectiveness of a new medication might be that there is no significant difference in symptom relief between the medication and a placebo.

Alternative hypothesis: A statement that assumes there is a significant difference or relationship between variables.

Example: The alternative hypothesis for the medication study would be that there is a significant difference in symptom relief between the medication and a placebo.

Test statistic: A value calculated from sample data that is used to make inferences about population parameters.

Example: The t-value calculated from a sample of exam scores can be used to determine whether the mean score for the population is significantly different from a particular value.

Critical region: The range of values that would lead to the rejection of the null hypothesis.

Example: If the critical region for a particular test is between 1.96 and -1.96, a calculated test statistic falling outside of that range would lead to rejection of the null hypothesis.

Significance level: The probability of rejecting the null hypothesis when it is true.

Example: A significance level of 0.05 means that there is a 5% chance of rejecting the null hypothesis when it is true.

Critical value: The value that separates the critical region from the non-critical region.

Example: The critical value for a particular test might be 1.96, which would separate the critical region from the non-critical region for a two-tailed test.

Type 1 error: The incorrect rejection of the null hypothesis when it is true.

Example: Concluding that a new medication is effective when it is no more effective than a placebo would be a type 1 error.

Type 2 error: The failure to reject the null hypothesis when it is false.

Example: Concluding that a new medication is no more effective than a placebo when it is effective would be a type 2 error.

Parametric tests: Statistical tests that assume that the data being analyzed follows a specific distribution, such as the normal distribution.

Example: A t-test assumes that the data being analyzed follows a normal distribution.

Non-parametric tests: Statistical tests that do not assume that the data being analyzed follows a specific distribution.

Example: The Wilcoxon rank-sum test is a non-parametric alternative to the t-test.

Normal distribution: A bell-shaped probability distribution that is symmetric and characterized by its mean and standard deviation.

Example: The distribution of heights in a population may follow a normal distribution.

Normality test: A statistical test used to determine whether a set of data follows a normal distribution.

Example: The Shapiro-Wilk test is a commonly used normality test.

The Shapiro-Wilk test: A statistical test used to determine whether a set of data follows a normal distribution. The null hypothesis is that the data is normally distributed, while the alternative hypothesis is that it is not.

Theoretical distribution: A probability distribution that is derived from mathematical principles, rather than being based on actual data.

Kolmogorov-Smirnov test: A statistical test used to compare the distribution of a sample with a theoretical distribution. The test statistics are based on the maximum difference between the cumulative distribution functions of the sample and the theoretical distribution.

Homogeneity test: A statistical test used to determine whether two or more populations have the same distribution.

Four types of relation: Connection refers to a relationship between two variables where one variable directly affects the other. **Correlation** refers to a statistical relationship between two variables that may or may not be causal. **Causation** refers to a relationship where one variable directly causes a change in another variable. **Prediction** refers to the ability to forecast one variable based on the value of another variable.

Chi-squared test: A statistical test used to determine whether there is a significant difference between expected and observed frequencies in a contingency table.

Test of homogeneity: A statistical test used to determine whether two or more groups have the same distribution.

Test of independence: A statistical test used to determine whether there is a significant relationship between two categorical variables.

T-test: A statistical test used to compare the means of two groups of data.

T-test types: There are two main types of T-tests. **Independent samples t-test**, which is used when the two samples being compared are independent, and paired samples t-test, which is used when the two samples are dependent. **Paired sample T-test** is used to compare the means of two related samples.

ANOVA: Analysis of variance is a statistical test used to compare the means of three or more groups.

Correlation types: There are two main types of correlation: positive correlation, where two variables move in the same direction, and negative correlation, where two variables move in opposite directions.

Covariance: A measure of the extent to which two variables are linearly related.

The Pearson correlation: A measure of the strength and direction of the linear relationship between two variables.

Population attributes: Characteristics of a population, such as mean, variance, and standard deviation.

Sample attributes: Characteristics of a sample, such as mean, variance, and standard deviation.

Measure of dispersed: A measure of how spread out the data is, such as variance, range, standard deviation, and interquartile range.

Variance: A measure of how far a set of numbers is spread out from their average.

Range: The difference between the largest and smallest values in a set of data.

Standard deviation: A measure of how much the data deviates from the mean.

Interquartile range: The difference between the upper and lower quartiles in a set of data.

Mean absolute deviation: A measure of the average distance between each data point and the mean.

Interval estimation: A method of estimating a population parameter using an interval of values that is likely to contain the true value of the parameter.

Confidence interval: An interval of values calculated from sample data that is likely to contain the true value of a population parameter with a specified level of confidence.

Measure of central tendencies: A measure of the center of a set of data, such as mean, median, and mode.