**When to Use What Statistical Analysis**

This document is a comprehensive guide designed to assist students and novice researchers in understanding and selecting appropriate statistical tests for their research projects. It covers a wide range of commonly used statistical analyses. Each section provides a simple explanation of the test, scenarios where it can be appropriately applied, and key assumptions necessary for its valid execution. Additionally, the document includes a 'Definition List' to clarify technical terms, making it an all-encompassing resource for those seeking to apply statistical methods in their research accurately and effectively. This guide is designed to be accessible, offering clear, concise descriptions that demystify complex statistical concepts, making them approachable for users with varying levels of statistical knowledge.

**Table of Contents**

3. T-tests

4. ANOVA (Analysis of Variance)

13. Chi-Square Test

14. Pearson Correlation Coefficient

15. Spearman’s Rank Correlation

16. Mann-Whitney U Test

17. Wilcoxon Signed-Rank Test

19. Linear Regression Analysis

20. Logistic Regression

23. Definition List

1. **T-tests**

* **Simple Explanation:** Compares the averages of two groups to see if they are significantly different from each other.
* **Types:**
  + **Independent Samples T-test:** Compares means of two independent groups.
  + **Paired Samples T-test:** Compares means from the same group at different times.
  + **Scenario:**
    - **Independent Samples:** Comparing the average blood pressure of patients using two different medications.
      * **Key assumptions:** The two groups being compared should be independent of each other. This means the participants in one group have no relation to the participants in the other group.
      * **Normality:** The data in each group should be approximately normally distributed. This assumption is less of a concern with larger sample sizes due to the Central Limit Theorem.
      * **Homogeneity of Variances:** The variances in the two groups should be approximately equal. This can be tested using Levene's test for equality of variances.
      * **Scale of Measurement:** The dependent variable should be measured at least at the interval scale (i.e., the difference between values is meaningful).
      * **Random Sampling:** Ideally, the data should be collected from a random sample from the population.
    - **Paired Samples:** Measuring student performance before and after a new teaching method is implemented.
      * **Key Assumptions:**
      * **Related or Paired Groups:** This test is used when the same subjects are used in both groups (e.g., measurements before and after a treatment on the same subjects) or in cases of matched subjects in both groups.
      * **Normality of Difference Scores:** The differences between the pairs should be approximately normally distributed. This is more crucial in smaller samples.
      * **Scale of Measurement:** The dependent variable should be measured at least at the interval scale.
      * **No Outliers:** Outliers can significantly affect the results, so it's important to check for outliers in the difference scores.
      * **Random Selection and Assignment:** Ideally, participants should be randomly selected and assigned to pairs or conditions.

1. **ANOVA (Analysis of Variance)**
   * **Simple** **Explanation**: ANOVA is a statistical method used to compare the means of three or more groups to determine if there's at least one significant difference among them. It assesses whether the average differences between groups are more than what would be expected by chance.
   * **One-way ANOVA (Single-factor ANOVA)**
     + **Explanation:** Used when there's only one independent variable (factor) and it compares the means of different levels (groups) of this variable.
     + **Scenario:** A researcher wants to compare the effectiveness of four different types of fertilizers on plant growth. Here, the types of fertilizer are the independent variable with four levels, and plant growth is the dependent variable.
     + **Key assumptions:**
       1. **Independence of Observations:** The data in each group must be collected independently. This means the scores of an individual in one group do not influence the scores of an individual in another group. This is usually ensured by random sampling.
       2. **Normality:** The data in each group should be approximately normally distributed. This assumption tends to be less critical with larger sample sizes due to the Central Limit Theorem, but it's important for smaller samples.
       3. **Homogeneity of Variances:** The variance among the groups should be approximately equal. This can be tested using Levene's test for equality of variances. If variances are not equal, the robustness of ANOVA is affected, and it might be necessary to use other approaches, like a Welch's ANOVA.
       4. **Scale of Measurement:** The dependent variable should be measured on an interval or ratio scale (i.e., a scale with meaningful numerical values and intervals).
       5. **Random Sampling:** Ideally, the data should be collected from a random sample from the population. This helps in generalizing the findings to the larger population.
   * **Two-way ANOVA**
     + **Explanation:** Examines the effect of two independent variables on a dependent variable. It can also explore the interaction effect between these two variables.
     + **Scenario:** A study aims to determine the impact of diet (vegetarian, non-vegetarian) and exercise routine (regular, irregular) on blood pressure. This test will not only look at each factor individually but also whether the combination of diet and exercise has a unique effect.
     + **Key assumptions:**
       1. **Independence of Observations:** Similar to the one-way ANOVA, the observations must be independent. This means the data in one group must not influence the data in another, and this is typically ensured through random sampling or random assignment.
       2. **Normality:** The distribution of the residuals (not the variables themselves) should be approximately normal. This is especially important for small sample sizes. For larger samples, the ANOVA is robust to deviations from normality.
       3. **Homogeneity of Variances (Homoscedasticity):** The variances across different groups (for each combination of categories of the two independent variables) should be approximately equal. This assumption can be tested using tests like Levene’s test or Hartley’s test.
       4. **No Significant Interactions (for the additivity assumption):** While Two-way ANOVA is used to test interactions between factors, it also assumes that the effects of one independent variable are the same at all levels of the other variable (additivity). Significant interaction effects may violate this assumption.
       5. **Random Sampling/Assignment:** The data should be collected from a randomly selected sample from the population or through a process of random assignment if it's an experiment.
       6. **Scale of Measurement:** The dependent variable should be measured at an interval or ratio scale.
       7. **Measurement of the Dependent Variable:** The dependent variable should be continuous (i.e., measured on an interval or ratio scale).
   * **Repeated Measures ANOVA**
     + **Explanation:** Used when the same subjects are measured multiple times under different conditions or over time. It's like a paired t-test, but for more than two measurements.
     + **Scenario:** A clinical trial tests the effectiveness of a new drug on reducing stress levels. The same group of participants' stress levels are measured at four different times: before treatment, after one month, after three months, and after six months of treatment.
     + **Key assumptions:** 
       1. **Sphericity**: This assumption pertains to the variances of the differences between all possible pairs of within-subject conditions, which should be approximately equal. If sphericity is violated, it can lead to inaccuracies in significance testing. Mauchly's test is commonly used to test this assumption. If sphericity is not met, adjustments like the Greenhouse-Geisser or Huynh-Feldt corrections are often applied.
       2. **Normality of Residuals**: The distribution of the residuals (differences between observed values and values predicted by the model) for each group should be approximately normal. This is particularly important for small sample sizes. For larger samples, the test is more robust to violations of normality.
       3. **No Outliers**: Outliers can have a disproportionately large effect on the results of a Repeated Measures ANOVA, so it's important to check for and address any outliers in the data.
       4. **Independence of Observations**: While the same subjects are measured multiple times in a Repeated Measures ANOVA, the assumption of independence still applies to the different cases or subjects in the study. The repeated measures (conditions or time points) on the same subject are not independent, but the different subjects should be.
       5. **Covariate Balance**: If covariates are used, they should be balanced across the repeated measures. Imbalance can lead to biased estimates and incorrect inferences.
       6. **Random Selection**: Ideally, participants should be randomly selected from the population. While this is not always feasible, especially in repeated measures designs, the random selection of subjects helps in generalizing the results.
   * **Multivariate ANOVA (MANOVA)**
     + **Explanation:** Similar to ANOVA, but it can handle multiple dependent variables at the same time.
     + **Scenario:** A study is conducted to see how different teaching methods (traditional, online, blended) impact students' grades in math and their reading comprehension scores. Here, both grades and comprehension scores are dependent variables.
     + **Key assumptions:**
       1. **Independence of Observations:** Similar to ANOVA, MANOVA assumes that the observations are independent of each other. This means that the results or responses of one participant do not influence the results of another.
       2. **Multivariate Normality:** The data should be multivariate normally distributed within each group or level of the independent variable(s). This is an extension of the normality assumption in ANOVA, applied to multiple dependent variables simultaneously.
       3. **Homogeneity of Covariance Matrices:** This assumption extends the homogeneity of variance assumption in ANOVA. MANOVA assumes that the covariance matrices of the dependent variables are equal across the groups. This can be tested with Box's M test.
       4. **No Multicollinearity:** The dependent variables should not be too highly correlated with each other. High correlation among dependent variables can inflate variance and affect the results of the MANOVA.
       5. **Linearity:** MANOVA assumes linear relationships among all pairs of dependent variables, all pairs of independent variables, and between each pair of dependent and independent variables.
       6. **Absence of Outliers:** Like ANOVA, outliers can have a significant impact on MANOVA results. It's important to check for and address outliers in the multivariate space of the dependent variables.
       7. **Adequate Sample Size:** MANOVA requires a larger sample size compared to ANOVA, particularly because it deals with multiple dependent variables. A common rule of thumb is to have more cases (observations) than the number of dependent variable combinations.
   * **Mixed-design ANOVA**
     + **Explanation:** Combines features of both repeated measures and two-way ANOVAs, used when there are both between-subjects and within-subjects variables.
     + **Scenario:** A research study evaluates the impact of therapy (two types: cognitive, behavioral) over time (before, after 6 months, after 12 months) on patient anxiety levels, with patients randomly assigned to therapy types.
     + **Key assumptions:**
       1. **Independence of Observations:** For the between-subjects factor, the groups must be independent of each other. This means participants in one group should not influence participants in another group.
       2. **Sphericity (for within-subjects factors):** Similar to repeated measures ANOVA, sphericity assumes that the variances of the differences between all combinations of related groups (within-subjects factor levels) are equal. If this assumption is violated, it can be corrected using adjustments like Greenhouse-Geisser or Huynh-Feldt.
       3. **Normality:** The dependent variable should be approximately normally distributed for each combination of groups of the between-subjects factor and levels of the within-subjects factor. This assumption is more critical in smaller samples.
       4. **Homogeneity of Variances (for between-subjects factors):** The variance among the groups for the between-subjects factor should be similar. This is assessed using Levene’s test of equality of variances.
       5. **No Significant Interactions (for the additivity assumption):** While Mixed-design ANOVA is used to examine interactions, it assumes that the effect of one factor is consistent at all levels of the other factor.
       6. **Random Sampling/Assignment:** The data should be collected from a randomly selected sample from the population, or subjects should be randomly assigned to groups if it's an experimental design.
       7. **Homogeneity of Covariance Matrices (for MANOVA only, if applicable):** In cases where multiple dependent variables are analyzed (as in MANOVA), the covariance matrices should be similar across different groups.
       8. **Absence of Outliers:** Outliers should be checked for and addressed, as they can have a significant impact on the results.
2. **Chi-Square Test**
   * **Simple Explanation:** Checks if there is a significant relationship between two categorical variables.
   * **Scenario:** Seeing if there is a relationship between gender (male/female) and preference for a certain kind of movie (action/romance/sci-fi).
   * **Key assumptions:**
     + **Categorical Data:** Both variables should be categorical (nominal or ordinal). The Chi-Square test is not suitable for continuous or highly granular data.
     + **Independence of Observations:** Each case or observation should be independent of each other. This means the outcome of one case should not influence or be influenced by the outcome of another.
     + **Sample Size:** A common guideline is that the total sample size should be sufficiently large, typically N > 20. However, more important than the total sample size is the expected frequency in each cell of the contingency table.
     + **Expected Frequencies:**
       1. **Chi-Square Goodness of Fit Test:** Each expected frequency should be 5 or more. This test is used to compare the observed distribution of data to an expected probability distribution.
       2. **Chi-Square Test of Independence:** Each cell of the contingency table should have an expected frequency of at least 5. If this condition is not met, the Fisher’s Exact Test might be a more appropriate choice, especially for 2x2 tables.
     + **No Overlapping Categories:** The categories for each variable should be mutually exclusive and exhaustive. This means that each observation fits into one and only one category for each variable.
     + **Random Sampling:** The data should be collected through a process that ensures random sampling. This is important for the generalizability of the results.
3. **Pearson Correlation Coefficient**
   * **Simple Explanation:** Measures how much two continuous variables move together (positively or negatively).
   * **Scenario:** Investigating if there is a relationship between hours studied and exam scores.
   * **Key assumptions:**
     + **Linearity:** The relationship between the two variables should be linear. This means that a change in one variable should result in a proportional change in the other variable. Non-linear relationships are not adequately captured by Pearson's r.
     + **Continuous Variables:** Both variables should be continuous. Continuous data can take any value within a range and are not restricted to categories (like categorical data).
     + **Bivariate Normal Distribution:** The pair of variables should be approximately normally distributed. More specifically, the joint distribution of these variables (considering them together) should be normal. This assumption is especially important for small sample sizes.
     + **Homoscedasticity:** This refers to the assumption that the variance within each variable should be constant at all levels of the other variable. In other words, the data should form a roughly circular cloud in a scatterplot. Heteroscedasticity, where the variance changes, can distort the correlation.
     + **Independence of Observations:** Each pair of observations should be independent of all other pairs. This means that the data for one pair should not influence or be influenced by any other pair.
     + **No Outliers:** Outliers can have a significant impact on the Pearson correlation coefficient, potentially leading to misleading results. It's important to identify and address outliers before conducting the analysis.
4. **Spearman’s Rank Correlation**
   * **Simple Explanation:** Similar to Pearson but for data that's ranked (like 1st, 2nd, 3rd) or not normally distributed.
   * **Scenario:** Looking at the relationship between the rank of employees in a company hierarchy and their job satisfaction.
   * **Key assumptions:**
     + **Ordinal or Continuous Data:** Spearman’s correlation can be used with either ordinal data (data that can be ranked or ordered) or continuous data. It is particularly useful when the data do not meet the assumptions necessary for Pearson’s correlation.
     + **Monotonic Relationship:** The relationship between the two variables should be monotonic. This means that as one variable increases, the other variable either consistently increases or decreases. Unlike Pearson’s r, the relationship does not need to be linear, but it should be monotonically increasing or decreasing.
     + **Independence of Observations:** Each pair of observations should be independent of all other pairs. This means the presence or value of one pair of data points should not influence or determine the presence or value of another pair.
     + **No Tied Ranks (or awareness of how to handle them):** Ideally, there should not be tied ranks in the data. Tied ranks occur when the same value appears more than once in the data set, leading to the same rank for different values. If ties are present, Spearman's rank correlation coefficient needs to be adjusted to account for this.
     + **Appropriate Sample Size:** While Spearman’s rank correlation can be used with small sample sizes, the reliability of the correlation estimate increases with larger samples. u
5. **Mann-Whitney U Test**
   * **Simple Explanation:** Compares the differences between two groups when the data is not normally distributed.
   * **Scenario:** Comparing customer satisfaction ratings (on a scale from 1-10) between two small cafes.
   * **Key assumptions:**
     + **Independence of Samples:** The two groups being compared must be independent, meaning the observations in one group are not related to the observations in the other group. This is typically the case in studies where different subjects are assigned to different groups.
     + **Ordinal or Continuous Data:** The dependent variable should be ordinal (can be ranked) or continuous.
     + **Similar Shape and Spread:** While the Mann-Whitney U test does not assume normal distribution of the data in the two groups, it does assume that the shapes of the distributions are similar. This means the general pattern and spread of the data should be similar across the groups, even if they have different medians.
     + **Random Sampling:** The data should be drawn from a random sample, meaning that each individual or observation in the population has an equal chance of being included in the sample.
     + **No Tied Ranks (or handling them appropriately):** The test includes adjustments for tied ranks (where two or more values are the same), but a large number of ties can affect the validity of the test.
6. **Wilcoxon Signed-Rank Test**
   * **Simple Explanation:** Compares two related groups to see if there are significant differences between them.
   * **Scenario:** Checking if a training program has improved employee performance before and after its implementation.
   * The Wilcoxon Signed-Rank Test is a non-parametric statistical test used to compare two related samples, matched samples, or repeated measurements on a single sample. It's an alternative to the paired t-test when the data do not meet the assumptions of the t-test. The key assumptions for conducting a Wilcoxon Signed-Rank Test are:
     + **Paired or Matched Data:** The test requires paired or matched samples. This means the data come from the same participants measured at two different times (e.g., before and after a treatment), or from matched subjects in two different conditions.
     + **Ordinal or Continuous Data:** The dependent variable should be continuous (measurable and able to have decimal values) or ordinal (data that can be ranked).
     + **Symmetry of the Distribution:** For the differences between the pairs, the distribution should be symmetric. While the test does not require the original data to be normally distributed, the differences between paired observations should be symmetrically distributed around the median.
     + **No Tied Differences for the Exact Test:** The test works best when there are no ties in the differences (i.e., when no two differences are the same). However, the Wilcoxon test has been adapted to handle ties.
     + **Random Selection and Assignment:** Ideally, the participants should be randomly selected from the population, and, in the case of an experiment, randomly assigned to conditions. This is important for the generalizability of the results.
7. **Linear Regression Analysis**
   * **Simple Explanation:** Predicts the value of a variable based on the value of another variable.
   * **Scenario:** Predicting a house's price based on its square footage.
   * **Key assumptions:**
     + **Linearity:** The relationship between the independent and dependent variables should be linear. This means that a change in the independent variable(s) should result in a proportional change in the expected value of the dependent variable.
     + **Independence of Residuals:** The residuals (the differences between the observed values and the values predicted by the model) should be independent. In other words, the residuals should not display any pattern when plotted against time or the independent variables. This is often assessed by examining a plot of residuals versus predicted values or time.
     + **Homoscedasticity:** The variance of residuals (errors) should be constant for all levels of the independent variable(s). This means the spread of the residuals should be approximately the same across all levels of the independent variables (homogeneity of variance).
     + **Normal Distribution of Residuals:** The residuals should be approximately normally distributed. While slight deviations from normality are not typically a major issue, especially with larger sample sizes, severe skewness or kurtosis can be problematic.
     + **No Multicollinearity (in multiple regression):** In cases of multiple regression, the independent variables should not be too highly correlated with each other. High multicollinearity can make it difficult to distinguish the individual effects of the independent variables.
     + **No Autocorrelation (in time series data):** If the data are time series, the residuals should not be correlated with each other (i.e., no autocorrelation). The Durbin-Watson test is commonly used to test for this.
     + **Adequate Sample Size:** Generally, a larger sample size provides more reliable results. A common rule of thumb is that there should be at least 10 observations per independent variable in the analysis.
8. **Logistic Regression**
   * **Simple Explanation:** Used when the outcome is binary (like yes/no) and predicts the probability of the outcome.
   * **Scenario:** Predicting whether a student will pass (1) or fail (0) an exam based on their study hours.
   * **Key assumptions:**
     + **Binary Outcome Variable:** The dependent variable should be binary (e.g., success/failure, yes/no, 1/0). Logistic regression can also be extended to categorical dependent variables (Multinomial Logistic Regression) or ordinal dependent variables (Ordinal Logistic Regression).
     + **Independence of Observations:** Each observation should be independent of the others. This means the outcome for one observation should not influence or be influenced by the outcome of another observation.
     + **No Multicollinearity:** The independent variables should not be too highly correlated with each other. High multicollinearity can lead to difficulties in determining the individual effect of each independent variable.
     + **Linear Relationship between the Logit of the Outcome and the Independent Variables:** In logistic regression, it is assumed that the logit transformation of the outcome variable has a linear relationship with the independent variables. The logit function is the natural log of the odds that the outcome variable equals a particular value.
     + **No Outliers or Highly Influential Points:** Outliers and influential data points can have a significant effect on the logistic regression model, potentially leading to overfitting or biased estimates.
     + **Large Sample Size:** Logistic regression requires a sufficiently large sample size to achieve reliable results. A common rule of thumb is having at least 10 cases with the least frequent outcome for each independent variable in the model.
     + **Goodness of Fit:** The model should fit the data well. Goodness of fit in logistic regression can be assessed using tests like the Hosmer-Lemeshow test or by looking at classification tables or ROC curves.
     + **Absence of Perfect Separation:** Perfect separation (or quasi-complete separation) happens when the outcome variable separates a predictor variable or a combination of predictor variables completely. This can lead to issues with model estimation and should be checked for.

**Definition List**

1. Variables:
   * Categorical Variables: Groups or labels like types of fruit or yes/no answers.
   * Continuous Variables: Numbers that can vary, like height or temperature.
   * Independent Variable: Something changed by the researcher to see its effect.
   * Dependent Variable: Something measured in response to changes in the independent variable.
2. Data Characteristics:
   * Normally Distributed/Skewed Data: Data that either fall in a normal bell-shaped pattern or are tilted to one side.
   * Binary Outcome: Results that only have two options, like pass/fail.
3. Testing and Measurement:
   * Non-Parametric Test/Nonparametric Methods: Tests used for data that don't fit a normal pattern or when there's not much data.
   * Repeated Measures: Taking multiple measurements of the same thing over time.
   * Multivariate Analysis: Looking at more than one outcome measure at a time.
4. ANOVA Specifics:
   * Levels: Different groups or conditions being compared.
   * Between-Subjects Variable: Differences between different groups of subjects.
   * Within-Subjects Variable: Differences within the same subjects under varying conditions.
   * Interaction Effect: When two variables together have a different effect than each on their own.
5. Error Types:
   * Type I Error: Incorrectly thinking something has an effect when it doesn't (false positive).
   * Type II Error: Not detecting an effect when there is one (false negative).
6. Inference Statistics:
   * P-Value: A measure that helps determine whether your results are likely due to chance.
   * Confidence Interval: A range that is likely to contain the true value you're trying to estimate.
   * Effect Size: The size or strength of an effect or relationship.
   * Power of a Test: The likelihood that your test will find an effect if there is one to be found.
7. Regression Analysis:
   * Regression Coefficient: A number showing the size of the effect one variable has on another.
   * Odds Ratio: A measure of association between an outcome and a variable, used in logistic regression.
   * Residuals: The differences between observed values and those predicted by a model.