1. Hypothesis Testing

Hypothesis testing is a statistical method used to make decisions or inferences about population parameters based on sample data. It involves formulating two competing hypotheses and using sample data to determine which hypothesis is more plausible.

1.1 Introduction to Hypothesis Testing

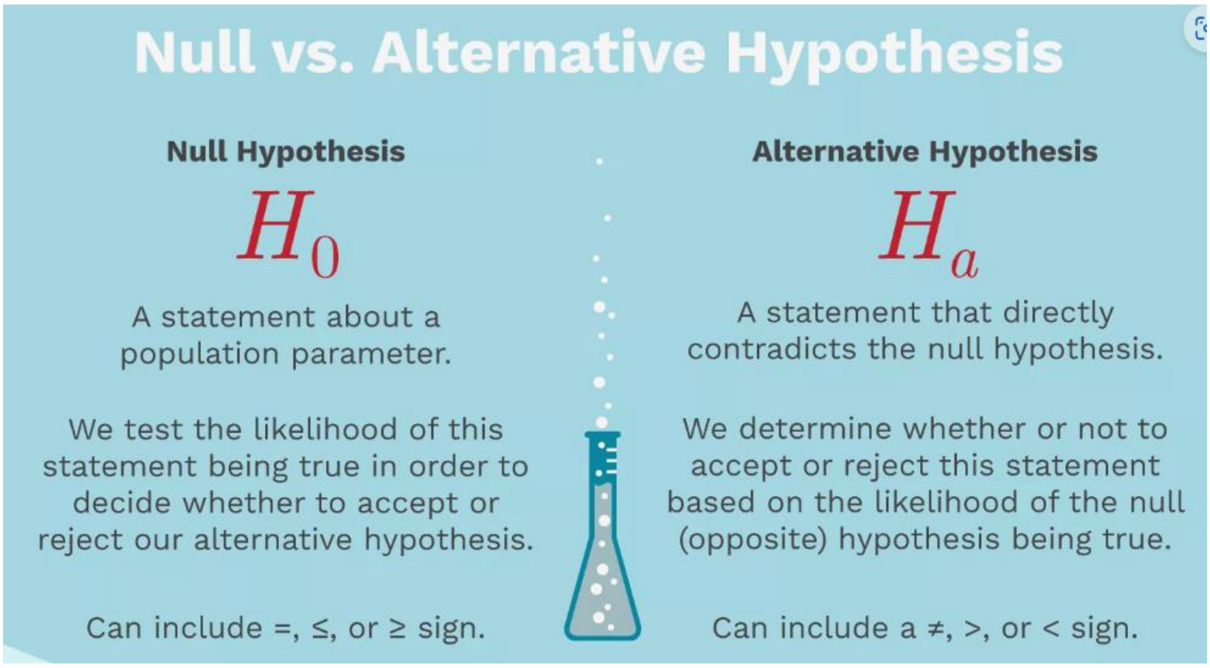
Hypothesis testing involves several key concepts, including the null and alternative hypotheses, Type I and Type II errors, and the significance level. These concepts form the foundation of hypothesis testing and guide the process of making statistical decisions.

1.1.1 Null Hypothesis (H_0)

- **Definition:** The null hypothesis (H_0) is a statement that there is no effect or no difference, and it serves as the starting assumption for the hypothesis test.
- **Purpose:** It provides a baseline to compare against the alternative hypothesis.
- **Example:** In testing a new drug, the null hypothesis might state that the drug has no effect on the disease.

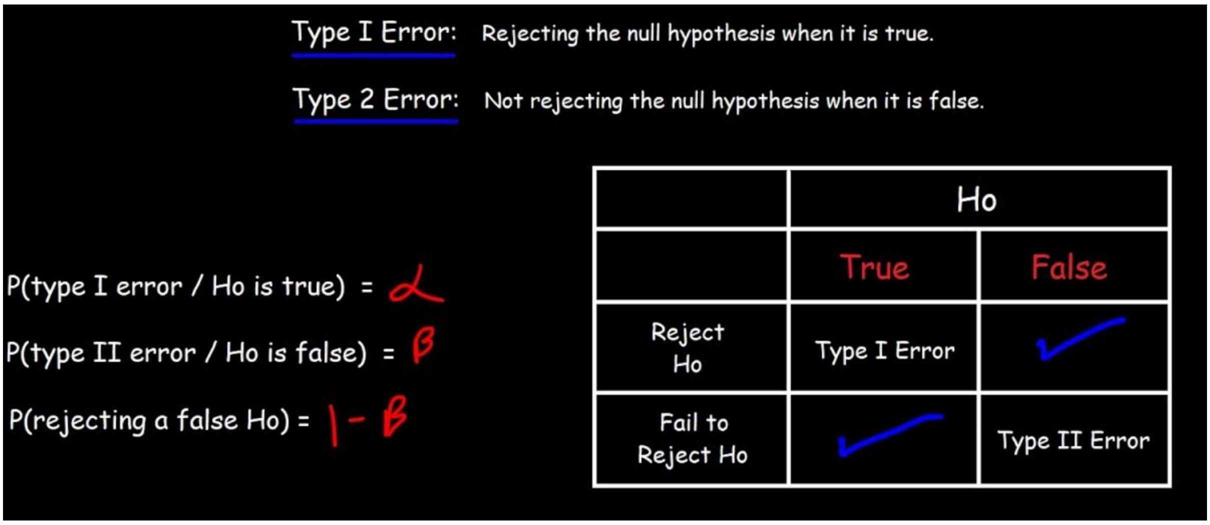
1.1.2 Alternative Hypothesis (H_1)

- **Definition:** The alternative hypothesis (H_1) is a statement that indicates the presence of an effect or a difference.
- **Purpose:** It represents the hypothesis that researchers aim to support with evidence from the sample data.
- **Example:** In the drug testing scenario, the alternative hypothesis might state that the drug has a significant effect on the disease.



1.1.3 Type I and Type II Errors

- **Type I Error (False Positive):** Occurs when the null hypothesis is rejected when it is actually true.
  - **Probability of Type I Error:** Denoted by α (alpha), also known as the significance level.
  - **Consequence:** Incorrectly concluding that there is an effect or difference when there isn't one.
- **Type II Error (False Negative):** Occurs when the null hypothesis is not rejected when it is actually false.
  - **Probability of Type II Error:** Denoted by β (beta).
  - **Consequence:** Failing to detect an effect or difference that actually exists.



	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct outcome! (True positive)
Fail to reject null hypothesis	Correct outcome! (True negative)	Type II Error (False negative)

1.1.4 Significance Level (alpha)

- **Definition:** The significance level (α) is the threshold for rejecting the null hypothesis. It represents the probability of making a Type I error.

- **Common Values:** Typical values for α are 0.05, 0.01, and 0.10.
- **Interpretation:** An α of 0.05 means that there is a 5% risk of rejecting the null hypothesis when it is true.
- **Usage:** If the p-value obtained from the test is less than or equal to α, the null hypothesis is rejected in favour of the alternative hypothesis.

1.2 Steps in Hypothesis Testing

Hypothesis testing is a systematic process used to determine whether there is enough evidence in a sample to infer that a certain condition holds true for the entire population. The following steps outline the typical procedure for conducting a hypothesis test.

1.2.1 Formulating Hypotheses

Formulate the Null and Alternative Hypotheses:

- **Null Hypothesis (H0):** The hypothesis that there is no effect or no difference. It serves as the default assumption.
- **Alternative Hypothesis (H1):** The hypothesis that there is an effect or a difference. It is what you want to provide evidence for.

Types of Alternative Hypotheses:

- **Two-tailed:** Tests for any difference from the null hypothesis value.
- **One-tailed:** Tests for a difference in a specific direction.

1.2.2 Selecting a Significance Level

Choose the Significance Level (α):

- The significance level, denoted by α, is the threshold for deciding whether to reject the null hypothesis.
- Common choices for α\alphaα are 0.05, 0.01, and 0.10.
- Example: α = 0.05 ➔ There is a 5% risk of rejecting the null hypothesis when it is actually true.

1.2.3 Choosing the Appropriate Test

Select the Appropriate Statistical Test:

The choice of test depends on the type of data and the hypotheses.

- **One-sample t-test:** Tests whether the mean of a single sample is different from a known value.
- **Two-sample t-test:** Tests whether the means of two independent samples are different.
- **Paired t-test:** Tests whether the means of two related samples are different.
- **Chi-square test:** Tests for independence between categorical variables.
- **ANOVA:** Tests whether the means of three or more groups are different.
- **Z-test:** Tests whether the mean of a sample is different from a known value, typically used for large sample sizes.

1.2.4 Computing the Test Statistic

Calculate the Test Statistic:

- The test statistic is a standardized value that is calculated from sample data during a hypothesis test.
- It is used to determine whether to reject the null hypothesis.

- Example for a one-sample t-test:
  - Calculate the sample mean ($\bar{x}$).
  - Compute the standard error ($SE = \frac{s}{\sqrt{n}}$), where s is the sample standard deviation and n is the sample size.
  - Calculate the t-statistic: $t = \frac{\bar{x}-\mu}{SE}$.

1.2.5 Making a Decision (Reject or Fail to Reject H0)

Compare the Test Statistic to the Critical Value or P-value:

- **P-value approach:**
  - The p-value is the probability of obtaining a test statistic at least as extreme as the one observed, assuming the null hypothesis is true.
  - If the p-value $\leq \alpha$, reject the null hypothesis.
  - If the p-value $> \alpha$, fail to reject the null hypothesis.
- **Critical value approach:**
  - Determine the critical value(s) from the statistical distribution that corresponds to the test (e.g., t-distribution, z-distribution).
  - Compare the test statistic to the critical value(s).
  - If the test statistic falls in the rejection region, reject the null hypothesis.
  - Otherwise, fail to reject the null hypothesis.

Make a Decision:

- Based on the comparison, decide whether to reject or fail to reject the null hypothesis.
- Example: If the p-value is 0.03 and α is 0.05, reject the null hypothesis.

Example of Hypothesis Testing

1. **Scenario:** A company claims that their new battery lasts longer than the old battery, which has a mean life of 300 hours.
2. **Formulate Hypotheses:**
    - Null Hypothesis (H_0): $\mu \leq 300$ (The mean life of the new battery is less than or equal to 300 hours)
    - Alternative Hypothesis (H_1): $\mu > 300$ (The mean life of the new battery is greater than 300 hours)
3. **Collect Data:** Sample data of the new battery's life spans.
4. **Calculate Test Statistic:** Using appropriate statistical methods (e.g., t-test), calculate the test statistic.
5. **Determine P-value:** Calculate the p-value associated with the test statistic.
6. **Compare P-value with Significance Level (α):**
    - If p-value $\leq \alpha$: Reject H_0
    - If p-value $> \alpha$: Fail to reject H_0.
7. **Make a Decision:** Based on the comparison, conclude whether there is sufficient evidence to support the company's claim.

Case Study:

The null hypothesis (H0) is a statement about the population mean lifespan of the new batteries compared to the known population mean lifespan of the old batteries.

Null Hypothesis (H0):

H0: $\mu=300$.

This means that the mean lifespan of the new batteries is equal to the mean lifespan of the old batteries, which is 300 hours.

Alternative Hypothesis (H1):

H1: $\mu \neq 300$

This would mean that the mean lifespan of the new batteries is different from the mean lifespan of the old batteries (it could be either less than or greater than 300 hours).

Hypothesis Testing:

- **Null Hypothesis (H0):** The mean lifespan of the new batteries is 300 hours.
- **Alternative Hypothesis (H1):** The mean lifespan of the new batteries is not 300 hours

---

2. t-Tests

2.1 One-Sample t-Test

A one-sample t-test is used to determine whether the mean of a single sample is significantly different from a known or hypothesized population mean. This is useful when you have a single sample and you want to compare its mean to a specific value.

2.1.1 Purpose and Application

Purpose:

- To test if the mean of a sample is significantly different from a known or hypothesized population mean.

Application:

- Compare the average test score of a class to the national average.
- Assess whether the average lifespan of a new product differs from an established standard.

2.1.2 Assumptions

1. **Random Sampling:** The sample data should be randomly selected from the population.
2. **Normality:** The distribution of the sample means should be approximately normal. This is especially important for small sample sizes ($n < 30$). For larger sample sizes, the Central Limit Theorem ensures normality.
3. **Scale of Measurement:** The data should be continuous (interval or ratio scale).

2.1.3 Calculation of Test Statistic

The test statistic for a one-sample t-test is calculated using the following formula:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Where:

- $\bar{x}$ is the sample mean.
- μ_0 is the hypothesized population mean
- s is the sample standard deviation.
- n is the sample size.

Steps to calculate the t-statistic:

1. **Calculate the sample mean ($\bar{x}$).**
2. **Calculate the sample standard deviation (s).**
3. **Calculate the standard error of the mean (SE):**
   $SE = \frac{s}{\sqrt{n}}$
4. **Compute the t-statistic:**
   $t = \frac{\bar{x} - \mu_0}{SE}$

2.1.4 Interpreting Results

1. **Calculate the p-value:**
   - The p-value represents the probability of obtaining a test statistic as extreme as the one observed, assuming the null hypothesis is true.
2. **Compare the p-value to the significance level (α\alphaα):**
   - Common significance levels are 0.05, 0.01, or 0.10.
   - If $p \leq \alpha$, reject the null hypothesis.
   - If $p > \alpha$, fail to reject the null hypothesis.
3. **Decision:**
   - **Reject H0:** There is enough evidence to conclude that the sample mean is significantly different from the hypothesized population mean.

- **Fail to reject H0:** There is not enough evidence to conclude that the sample mean is significantly different from the hypothesized population mean.

2.2 Two-Sample t-Test

The two-sample t-test compares the means of two independent or paired groups to determine if there is a significant difference between them.

2.2.1 Independent Samples t-Test

Independent samples t-test is used when comparing the means of two independent groups.

Assumptions and Application

- **Assumptions:**
    1. The samples are independent.
    2. The data in each group are normally distributed.
    3. The variances of the two groups are equal (homogeneity of variance).
- **Application:**
    - Compare the average test scores between two different teaching methods.

Calculation of Test Statistic

The test statistic for an independent samples t-test is calculated using:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where:

- $\bar{X}_1$ and $\bar{X}_2$ are the sample means.
- s_p is the pooled standard deviation.
- n_1 and n_2 are the sample sizes.

2.2.2 Paired Samples t-Test

Paired samples t-test is used when comparing the means of two related groups (e.g., measurements before and after treatment).

Assumptions and Application

- **Assumptions:**
    1. The samples are dependent (paired).
    2. The differences between the paired observations are normally distributed.

- **Application:**
    - Compare the average test scores of students before and after using a new teaching method.

Calculation of Test Statistic

The test statistic for a paired samples t-test is calculated using:

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

Where:

- $\bar{d}$ is the mean of the differences.
- s_d is the standard deviation of the differences.
- n is the number of pairs.

2.3 Assumptions of t-Tests

t-Tests are widely used statistical tests that rely on specific assumptions to provide valid results. Understanding these assumptions is crucial for correctly applying t-tests and interpreting their results.

2.3.1 Normality

Normality refers to the assumption that the data (or the differences between paired data) are approximately normally distributed. This assumption is more critical for small sample sizes. For larger sample sizes ($n > 30$), the Central Limit Theorem suggests that the sampling distribution of the mean will be approximately normal regardless of the shape of the population distribution.

- **How to check for normality:**
    - **Graphical Methods:** Histograms, Q-Q plots (quantile-quantile plots), and box plots can help visualize whether the data follows a normal distribution.
    - **Statistical Tests:** Shapiro-Wilk test, Kolmogorov-Smirnov test, and Anderson-Darling test can be used to formally test for normality.

2.3.2 Homogeneity of Variances

Homogeneity of variances (also known as homoscedasticity) assumes that the variances of the populations being compared are equal. This assumption is crucial for the independent samples t-test but not for the paired samples t-test.

- **How to check for homogeneity of variances:**
    - **Graphical Methods:** Box plots can help visualize the spread of data.
    - **Statistical Tests:** Levene's test and Bartlett's test can be used to formally test for equality of variances.

2.3.3 Independence

Independence assumes that the observations in the data set are independent of each other. This means that the value of one observation does not influence or depend on the value of another observation.

- **How to ensure independence:**
  - Proper experimental design: Random sampling and random assignment can help ensure that the samples are independent.
  - Check the study design: Confirm that the data collection methods do not introduce dependencies.

---

3. Chi-Squared Test

The Chi-Squared test is a non-parametric statistical test used to determine if there is a significant association between categorical variables or if the observed frequencies of events differ from expected frequencies.

3.1 Introduction to Chi-Squared Test

3.1.1 Purpose and Application

The Chi-Squared test is used for:

- **Goodness of Fit Test**: Determines if a sample matches an expected distribution.
- **Test of Independence**: Determines if there is a significant association between two categorical variables.

3.1.2 Types of Chi-Squared Tests

- **Goodness of Fit Test**: Used to compare the observed distribution of a single categorical variable to an expected distribution.
- **Test of Independence**: Used to assess whether two categorical variables are independent of each other.

3.1.3 Assumptions

- The data should be in the form of frequencies or counts of cases.
- The categories are mutually exclusive.
- The expected frequency for each category should be at least 5 to ensure the validity of the test.

3.2 Goodness of Fit Test

The Goodness of Fit test compares the observed frequencies of events to the expected frequencies based on a specific hypothesis.

3.2.1 Formulating Hypotheses

- **Null Hypothesis (H0)**: The observed frequencies match the expected frequencies.
- **Alternative Hypothesis (H1)**: The observed frequencies do not match the expected frequencies.

3.2.2 Calculation of Test Statistic

The test statistic for the Chi-Squared test is calculated as follows:

$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$

where O_i is the observed frequency and E_i is the expected frequency.

3.2.3 Degrees of Freedom

The degrees of freedom for the Goodness of Fit test is calculated as:

$$\text{Degrees of Freedom} = \text{Number of Categories} - 1$$

3.2.4 Interpreting Results

- Compare the calculated Chi-Squared statistic to the critical value from the Chi-Squared distribution table based on the degrees of freedom and the significance level (alpha).
- If the calculated statistic is greater than the critical value, reject the null hypothesis.

3.3 Test of Independence

The Chi-Squared Test of Independence is used to determine whether there is a significant association between two categorical variables.

3.3.1 Formulating Hypotheses

- **Null Hypothesis (H0)**: The two categorical variables are independent (no association).
- **Alternative Hypothesis (H1)**: The two categorical variables are not independent (there is an association).

3.3.2 Contingency Tables

A contingency table (or cross-tabulation) displays the frequency distribution of the variables. Each cell in the table shows the count of occurrences for a combination of the two variables.

3.3.3 Calculation of Test Statistic

The Chi-Squared test statistic for independence is calculated as:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where:

- O_{ij} is the observed frequency in the cell corresponding to the ith row and jth column.
- E_{ij} is the expected frequency, calculated as:
  $$E_{ij} = \frac{(\text{Row Total}_i \times \text{Column Total}_j)}{\text{Grand Total}}$$

3.3.4 Interpreting Results

- Compare the calculated Chi-Squared statistic to the critical value from the Chi-Squared distribution table based on the degrees of freedom and the significance level (alpha).
- **Degrees of Freedom**:

  Calculated as:

  Degrees of Freedom=(Number of Rows−1)×(Number of Columns−1)
- If the calculated statistic is greater than the critical value, reject the null hypothesis.

---

4. Analysis of Variance (ANOVA)

ANOVA (Analysis of Variance) is a statistical method used to compare the means of three or more samples to determine if at least one of the sample means is significantly different from the others. It's particularly useful when comparing more than two groups.

4.1 Introduction to ANOVA

4.1.1 Purpose and Application

Purpose: The main goal of ANOVA is to determine if there are statistically significant differences between the means of three or more independent (unrelated) groups.

Application: ANOVA is commonly used in various fields such as:

- **Psychology**: To compare the effects of different therapies.
- **Medicine**: To compare the efficacy of different drugs.

- **Business**: To compare customer satisfaction across different stores or regions.
- **Agriculture**: To compare crop yields from different treatments or conditions.

Example: Suppose a researcher wants to test the effectiveness of three different diets on weight loss. ANOVA can help determine if the mean weight loss is the same across the three diets or if at least one diet leads to a significantly different weight loss.

4.1.2 Assumptions

For ANOVA to be valid, the following assumptions must be met:

1. **Independence of Observations**: The data in each group should be collected independently of the others.
2. **Normality**: The data in each group should be approximately normally distributed.
3. **Homogeneity of Variances**: The variance among the groups should be approximately equal. This is also known as homoscedasticity.

Testing Assumptions:

- **Normality**: Can be checked using the Shapiro-Wilk test, Q-Q plots, or by looking at skewness and kurtosis.
- **Homogeneity of Variances**: Can be tested using Levene's test or Bartlett's test.

4.2 One-Way ANOVA

One-Way ANOVA is a statistical technique used to compare the means of three or more independent groups to determine if there is a statistically significant difference between them. This method helps in understanding if any of the groups are significantly different from each other in terms of the means of the variable of interest.

4.2.1 Formulating Hypotheses

In a One-Way ANOVA, we set up two hypotheses:

- **Null Hypothesis (H_0)**: All group means are equal.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$

- **Alternative Hypothesis (H_1)**: At least one group mean is different.

$$H_1 : \text{At least one } \mu_i \text{ is different}$$

Where μ_i represents the mean of the i-th group, and k is the total number of groups.

4.2.2 Between-Group and Within-Group Variability

One-Way ANOVA decomposes the total variability in the data into two components:

- **Between-Group Variability (SSB)**: Variability due to the interaction between the different groups. It measures how much the group means differ from the overall mean.

$$\text{SSB} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$$

where n_i is the sample size of the i-th group, $\bar{X}_i$ is the mean of the i-th group, and $\bar{X}$ is the overall mean.

- **Within-Group Variability (SSW)**: Variability due to differences within each group. It measures how much the individual observations differ from their respective group means.

$$\text{SSW} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

where X_{ij} is the j-th observation in the i-th group.

4.2.3 Calculation of F-Statistic

The F-statistic is calculated by taking the ratio of the between-group variability to the within-group variability. This ratio helps in determining if the observed variability between group means is more than would be expected by chance.

1. **Calculate the mean squares**:

   - Mean Square Between (MSB):

$$\text{MSB} = \frac{\text{SSB}}{k-1}$$

   - Mean Square Within (MSW):

$$\text{MSW} = \frac{\text{SSW}}{N-k}$$

   where N is the total number of observations.

2. **Calculate the F-statistic**:

$$F = \frac{\text{MSB}}{\text{MSW}}$$

4.2.4 Interpreting Results

To interpret the results, we compare the calculated F-statistic to the critical value from the F-distribution table, or we use the p-value associated with the F-statistic.

- **If $p \leq \alpha$ (significance level)**: We reject the null hypothesis H0 and conclude that there is a statistically significant difference between the group means.
- **If $p > \alpha$**: We fail to reject the null hypothesis H0, indicating no statistically significant difference between the group means.

4.3 Post-hoc Tests

4.3.1 Purpose of Post-hoc Tests

Post-hoc tests are conducted after an ANOVA test when the null hypothesis is rejected. The purpose of these tests is to identify which specific groups are significantly different from each other. While ANOVA indicates that at least one group mean is different, post-hoc tests help determine exactly which groups differ.

4.3.2 Common Post-hoc Tests

Two commonly used post-hoc tests are Tukey's Honestly Significant Difference (HSD) and the Bonferroni correction.

Tukey's Honestly Significant Difference (HSD):

- Purpose: To compare all possible pairs of means while controlling the overall Type I error rate.
- Method: It calculates a critical value (the HSD) for the difference between means, considering the number of comparisons and the variability within groups.

Bonferroni Correction:

- Purpose: To control the Type I error rate when multiple pairwise comparisons are made.
- Method: It adjusts the significance level by dividing it by the number of comparisons. Each individual test then uses this adjusted significance level to determine if the differences are significant.

4.3.3 Application and Interpretation

Tukey's HSD:

- **Application**: Conducted after a significant ANOVA result.
- **Interpretation**: Provides confidence intervals for the differences between pairs of group means. If the confidence interval for a pair does not include zero, the difference is considered significant.

Bonferroni Correction:

- **Application**: Used when performing multiple pairwise tests.
- **Interpretation**: Adjusted p-values or confidence intervals are used to determine significance. If an adjusted p-value is below the original significance level, the difference is significant.

4.4 Applications in Machine Learning

4.4.1 Comparing Multiple Models

In machine learning, post-hoc tests can be used to compare the performance of multiple models. After running an ANOVA to determine if there is a significant difference in model performances (e.g., accuracy, F1 score), post-hoc tests can identify which models differ significantly from each other.

4.4.2 Experimental Design

Post-hoc tests are also valuable in experimental design for machine learning. When designing experiments to compare different algorithms, parameters, or feature sets, ANOVA followed by post-hoc tests can help in making statistically valid conclusions about which configurations lead to significantly better performance.