Unit: IV

Statistical testing and modelling, sampling distributions, hypothesis testing, components of hypothesis test, testing means, testing proportions, testing categorical variables, errors and power, Analysis of variance.

STATISTICAL TESTING AND MODELLING

STATISTICAL TESTING:

- ✓ Statistical testing involves analyzing data to make decision about a population based on a sample.
- ✓ It helps to determine if determine of observed difference or relationships in the data Are statistically significant or due to random chance

Example or Testing
t-Test:

- `t.test()`: Compares means (one-sample, two-sample, paired).

STATISTICAL MODELLING:

- ✓ statistical modelling involves creating mathematical representations of relationships in data.
- ✓ These models can help in making predictions, testing hypotheses, and understanding complex data patterns.
- ✓ Statistical models are widely used across various fields like economics, biology, social sciences, engineering, and more

Example: For instance, predicting house prices based on factors like square footage, location, and number of bedrooms using regression analysis is a statistical modeling task. The model helps estimate how these factors influence the price and make predictions for new houses.

SAMPLING DISTRIBUTIONS: -

- ✓ A sampling distribution is a probability distribution of a statistic (such as the mean, median, or mode) calculated from multiple samples drawn from a population.
- ✓ It's essentially a distribution of the possible values that a statistic could take if we were to repeatedly sample from the same population.
- ✓ The sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for a statistic of a population.

- ✓ In statistics, a population is an entire pool from which a statistical sample is drawn. A population may refer to an entire group of people, objects, events, hospital visits, or measurements.

A population can thus be said to be an aggregate observation of subjects grouped together by a common feature.

- ✓ A sampling distribution is a statistic that is arrived out through repeated sampling from a larger population.
- ✓ It describes a range of possible outcomes that of a statistic, such as the mean or mode of some variable, as it truly exists a population.
- ✓ The majority of data analyzed by researchers are actually drawn from samples, and not populations.

Characteristics of Sampling Distributions

Mean: The mean of the sampling distribution of the sample mean is equal to the population mean μ.

Standard Error: The standard error of the sample mean is $\frac{\sigma}{\sqrt{n}}$, and it decreases with larger sample sizes.

Shape: The sampling distribution of the sample mean approaches normality as the sample size increases (Central Limit Theorem).

Variance: The variance of the sampling distribution of the sample mean is $\frac{\sigma^2}{n}$, and it decreases as sample size increases.

Bias: The sample mean is an **unbiased estimator** of the population mean.

Consistency: The sample mean is a **consistent estimator** of the population mean as the sample size increases.

POPULATION refers to the entire group of individuals or objects that you're interested in studying. Here are some examples:

1. Human populations:

- ✓ All adult residents of a particular city
- ✓ All students enrolled in a university

2. Animal populations:

- ✓ All dogs in a city

3. **Object populations:**
   - ✓ All cars produced by a specific manufacturer

4. Data populations:

   - ✓ All responses **to a surv**ey
   - ✓ All transactions recorded by a company

1. **Sampling Distribution of the Mean:** This method shows a normal distribution where the middle is the mean of the sampling distribution. As such, it represents the mean of the overall population.
2. **Sampling Distribution of the Proportion:**
   - ✓ The sampling distribution of the sample proportion (p^) refers to the probability distribution of the proportion of a certain characteristic in a sample drawn from a population
   - ✓ **Applicability:** It is used when dealing with categorical data, where observations are classified into categories.

The standard deviation (also known as the standard error) of the sampling distribution is given by:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

3. **Sampling Distribution of the Variance**:
   - ✓ The sampling distribution of the sample variance (S2) describes the distribution of the variance calculated from multiple samples drawn from the same population.
   - ✓ **Used in:** Analysis of variance (ANOVA) and other situations where variance is a key parameter.
4. **Sampling Distribution of the Difference between Two Means:**
   - ✓ Examines the distribution of the difference between two sample means.
   - ✓ Used in: Two-sample hypothesis tests and confidence intervals.
5. **Sampling Distribution of the Difference between Two Proportions:**
   - ✓ Deals with the distribution of the difference between two sample proportions.
   - ✓ Applicability: Commonly used when comparing proportions from two independent groups.

Hypothesis

- ✓ A hypothesis is a statement or idea that is suggested as a possible explanation or solution to a problem, but has not yet been proven. It's an educated guess or assumption that is tested through research, experiments or data analysis to determine its validity.

Key characteristics of a hypothesis:

- ✓ Specific and clear statement
- ✓ Testable and falsifiable
- ✓ Based on limited data or observations
- ✓ May be supported or rejected through experimentation or further research

Example: "Eating less sugar reduces weight gain" is a hypothesis.

Types of Hypothesis:

- ✓ **Null Hypothesis**
- ✓ **Alternative Hypothesis**

Null Hypothesis:

- ✓ The null hypothesis is a statement that indicates no effect, no difference, or no relationship between variables in a study.
- ✓ it is a basic assumption or made based on the problem knowledge
- ✓ It is denoted as H0

Example

- ✓ There is no difference in the average test scores between students who study in groups and those who study alone.
- ✓ The new drug has no effect on blood pressure compared to a placebo.

Alternative hypothesis (H1):

- ✓ It suggests that there is an effect, a difference, or a relationship between variables, contrary to the null hypothesis (H0).
- ✓ The alternative hypothesis is an alternative to the null hypothesis.
- ✓ It is used to show that the observations of an experiment are due to some real effect.

- ✓ It indicates that there is a statistical significance between two possible outcomes and can be denoted as H1 or Ha.

  Example

  - There is a difference in the average test scores between students who study in groups and those who study alone.
  - The new drug has effect on blood pressure compared to a placebo.

Example for Null Hypothesis and Alternative Hypothesis

- *Null Hypothesis:* H_0: There is no difference in the salary of factory workers based on gender. *Alternative Hypothesis*: H_a: Male factory workers have a higher salary than female factory workers.

- *Null Hypothesis*: H_0: There is no relationship between height and shoe size. *Alternative Hypothesis*: H_a: There is a positive relationship between height and shoe size.

Hypothesis testing:

- Hypothesis testing is a statistical process used to determine whether a hypothesis about a population is true or false based on a sample of data.

***Steps in Hypothesis Testing: ***

- ✓ Formulate a null hypothesis (H0) and alternative hypothesis (H1).
- ✓ Choose a significance level (α).
- ✓ Collect a random sample of data.
- ✓ Calculate a test statistic.
- ✓ Determine the p-value.
- ✓ Compare the p-value to α.
- ✓ Make a decision: reject H0 or fail to reject H0.

Key Steps in Hypothesis Testing

1. **State the Hypotheses:**
   - **Null Hypothesis (H_0)**: This is a statement of no effect or no difference, serving as the default assumption. For example, $H_0 : \mu = \mu_0$ (the population mean equals a specified value).
   - **Alternative Hypothesis (H_a or H_1)**: This statement reflects what you aim to prove, indicating an effect or difference. It can be one-tailed (e.g., $H_a : \mu > \mu_0$ or $H_a : \mu < \mu_0$) or two-tailed (e.g., $H_a : \mu \neq \mu_0$).
2. **Select a Significance Level (α):**
   - Commonly set at 0.05, 0.01, or 0.10, this level defines the threshold for rejecting the null hypothesis. It represents the probability of making a Type I error (rejecting H_0 when it is true).
3. **Choose the Appropriate Test:**
   - Depending on the data type and sample size, choose a suitable statistical test (e.g., t-test, z-test, chi-square test).
4. **Collect Data:**
   - Gather your sample data that will be used to conduct the test.
5. **Calculate the Test Statistic:**
   - Compute the statistic (like t or z) based on the sample data. This statistic helps determine how far your sample statistic is from the null hypothesis value.
6. **Determine the Critical Value or p-value:**
   - **Critical Value Approach**: Compare the test statistic to critical values from the statistical distribution relevant to your test.

7. **Make a Decision:**
   - If using the critical value approach, reject H_0 if the test statistic falls into the critical region.
   - If using the p-value approach, reject H_0 if the p-value is less than or equal to α.
8. **Draw Conclusions:**
   - Report the results, including whether the null hypothesis was rejected or not, and provide context for the findings.

COMPONENTS OF HYPOTHESIS TEST

Hypothesis testing is a statistical method used to determine whether or not to reject a null hypothesis. It involves several key components:

- **Research Question:** This is the question that the study aims to answer.
- **Null Hypothesis (H_0)** This is the statement of "no effect" or "no difference." It's the default assumption that there's no significant relationship between the variables.
- **Alternative Hypothesis (H_1):** This is the statement of "effect" or "difference." It's the opposite of the null hypothesis and suggests that there is a significant relationship between the variables.
- **Test Statistic:** This is a numerical value calculated from the sample data that is used to evaluate the evidence against the null hypothesis. Common test statistics include the t-test, z-test, chi-square test
- **Significance Level (α):** This is the probability of rejecting the null hypothesis when it is actually true. It's often set at 0.05 (5%), but it can be adjusted depending on the desired level of confidence.
- **P-Value**: This is the probability of obtaining a test statistic as extreme or more extreme than the observed one, assuming the null hypothesis is true. A smaller p-value indicates stronger evidence against the null hypothesis.
- **Decision Rule:** This is a set of criteria used to decide whether to reject or fail to reject the null hypothesis. It typically involves comparing the p-value to the significance level.
- **Conclusion:** This is the final statement about whether or not the null hypothesis is rejected based on the evidence from the data.

Testing means

Testing means typically involves comparing the mean of a sample to a known or hypothesized population mean. This is commonly done using hypothesis testing.

Single testing means or One sample mean

- ✓ In statistics, a single mean is a method used to compare the mean of a sample to a hypothesized mean of a population.

OR

- ✓ refers to testing a hypothesis about a **single population** based on a sample drawn from that population.

Types of Single Mean Test:

- ✓ One-Sample z-test: Compares sample mean to known population mean.
- ✓ One-Sample t-test: Compares sample mean to known population mean (used when population standard deviation is unknown).

One-Sample t-test:

- ✓ A one-sample t-test is a statistical test used to determine if the mean of a sample is significantly different from a known or hypothesized population mean. (used when population standard
- ✓ deviation is unknown).
- ✓ It's commonly applied when you have a small sample size (typically $n<30$) and the population standard deviation is unknown.

When to Use:

- ✓ When you have a small sample size.
- ✓ When the population standard deviation is unknown.
- ✓ When you want to compare the sample mean to a specific value.

Conduct a One-Sample T-Test:

Formulate Hypotheses:

- **Null Hypothesis (H_0)**: The population mean is equal to a specified value (e.g., $H_0 : \mu = \mu_0$).
- **Alternative Hypothesis (H_1)**: The population mean is not equal to the specified value (e.g., $H_1 : \mu \neq \mu_0$). This can also be one-tailed.

Calculate the Test Statistic (t):

- Use the formula:

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

Where:

- $\bar{X}$ = sample mean
- μ_0 = hypothesized population mean
- s = sample standard deviation
- n = sample size

One-Sample z-test:

- ✓ A one-sample z-test is used to determine whether the mean of a single sample differs significantly from a known population mean
- ✓ when the population standard deviation is known.
- ✓ This test is typically applied when the sample size is large (n≥30).

Calculate the Test Statistic (z):

- Use the formula:

$$z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Where:

- $\bar{X}$ = sample mean
- μ_0 = hypothesized population mean
- σ = population standard deviation
- n = sample size

When to Use a One-Sample Z-Test

- ✓ You have a sample of data.
- ✓ You want to compare the sample mean to a known population mean.
- ✓ The population standard deviation is known, or the sample size is large

Steps to Conduct a One-Sample Z-Test

1. **Formulate Hypotheses**:
   - Null Hypothesis (H_0): The sample mean is equal to the population mean (e.g., $\mu = \mu_0$).
   - Alternative Hypothesis (H_a): The sample mean is not equal to the population mean (e.g., $\mu \neq \mu_0$).
2. **Collect Data**: Gather your sample data.
3. **Calculate the Sample Mean** ($\bar{x}$) and identify the known population mean (μ_0) and population standard deviation (σ).
4. **Determine the Sample Size** (n).
5. **Calculate the Z-Statistic**:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

6. **Find the Critical Z-Value**: Use a z-table or software to find the critical z-value for your desired significance level (e.g., $\alpha = 0.05$).
7. **Make a Decision**:
   - If the absolute value of the z-statistic is greater than the critical value, reject the null hypothesis.
   - If it is less, do not reject the null hypothesis.

Two sample testing mean or two sample mean:

- ✓ A two-sample t-test is a statistical test used to compare the means of two independent samples.
- ✓ It is commonly used to determine if there is a significant difference between the means of two groups.

Types of Two Mean Tests

1. **Independent Samples T-Test (Unpaired t-Test)**
2. **Paired Samples T-Test (Dependent Samples T-Test)**
   - ✓ is used to determine whether there is a significant difference between the means of two related groups.
   - ✓ when the sample sizes are large ($n<30$)
3. **Independent Samples Z-Test**
   - ✓ Similar to the independent samples t-test but used when the sample sizes are large ($n\geq30$) and the population standard deviations are known.

Independent Samples T-Test(Unpaired t-Test) with pooled variance or unpooled variance

- ✓ An independent samples t-test, also known as a two-sample t-test, is used to determine whether there is a significant difference between the means of two independent groups.
- ✓ This test is appropriate when you want to compare two different groups or treatments.
- ✓ when the sample sizes are large ($n<30$)
- ✓ assuming that the **population variances are equal or unequal**

When to Use

- You have two independent groups (e.g., treatment vs. control).
- You want to test if their means are significantly different.
- **The variances of the two groups can either be equal (pooled) or unequal**

Hypotheses:

- **Null Hypothesis (H0)**: The means of the two groups are equal ($\mu_1 = \mu_2$).
- **Alternative Hypothesis (H1)**: The means are not equal ($\mu_1 \neq \mu_2$).

Types of T-Tests

1. **Pooled T-Test**: Assumes equal variances.
2. **Welch's T-Test**: Does not assume equal variances and is more robust in cases of unequal variances.

Formulas

Pooled T-Test

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

Where:

- $\bar{x}_1, \bar{x}_2$ = sample means
- s_p = pooled standard deviation
- n_1, n_2 = sample sizes

Pooled Standard Deviation:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Welch's T-Test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

- s_1, s_2 = sample standard deviations

Degrees of Freedom

- Pooled: $df = n_1 + n_2 - 2$
- Welch's:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Paired Samples T-Test (Dependent Samples T-Test)

- ✓ The paired samples t-test, also known as the dependent samples t-test or the matched pairs t-test, is used to compare the means of two related groups.
- ✓ This test is particularly useful when the same subjects are measured under two different conditions or at two different times.

When to Use:

- ✓ Two related groups (e.g., measurements taken at two different times, or under two different conditions).
- ✓ The differences between the paired observations should be normally distributed.

Hypotheses

- **Null Hypothesis** (H_0): The mean difference between the paired observations is zero (e.g., $H_0 : \mu_d = 0$).
- **Alternative Hypothesis** (H_1): The mean difference is not equal to zero (e.g., $H_1 : \mu_d \neq 0$). This can also be one-tailed, depending on the research question.

Formula

The test statistic is calculated using the following formula:

$$t = \frac{\bar{D}}{s_D/\sqrt{n}}$$

Where:

- $\bar{D}$ = mean of the differences between paired observations
- s_D = standard deviation of the differences
- n = number of paired observations

Independent Samples Z-Test

- ✓ The independent samples z-test is a statistical method used to compare the means of two independent groups when the population standard deviations are known.
- ✓ It is most appropriate for large sample sizes, typically when both groups have at least 30 observations.

When to Use

- ✓ You have two independent samples.
- ✓ The population standard deviations are known.
- ✓ The sample sizes are large (n1≥30 and n2≥30)

Hypotheses:

- **Null Hypothesis (H0)**: The means of the two groups are equal ($\mu_1 = \mu_2$).
- **Alternative Hypothesis (H1)**: The means are not equal ($\mu_1 \neq \mu_2$).

Formula

The z-statistic is calculated using the following formula:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Where:

- $\bar{x}_1, \bar{x}_2$ = sample means
- σ_1^2, σ_2^2 = population variances
- n_1, n_2 = sample sizes

Steps to Conduct the Test

1. **Collect Data**: Obtain data from two independent groups.
2. **Set Hypotheses**:
   - H0: $\mu_1 = \mu_2$
   - H1: $\mu_1 \neq \mu_2$
3. **Calculate Sample Statistics**: Compute means and population variances for each group.
4. **Calculate the z-Statistic**: Use the formula above.
5. **Determine the z-Critical Value**: Based on the significance level (α) and whether the test is one-tailed or two-tailed.
6. **Find the p-value**: Use the standard normal distribution to find the p-value corresponding to the calculated z-statistic.
7. **Make a Conclusion**: Compare the p-value to your significance level to accept or reject the null hypothesis.

Testing proportions

- ✓ Testing proportions involves comparing two or more proportions to determine if there's a statistically significant difference between them or from a hypothesized proportion

Types of Testing Proportions

1. Single Proportion test
2. Two Proportions test

Single Proportion test

- ✓ A single proportion test is used to determine if a sample proportion is significantly different from a known population proportion
- ✓ is used to test whether the proportion of successes in a single sample is equal to a hypothesized proportion. This test is useful when you're dealing with binary data, such as "yes/no" responses, or "success/failure" outcomes.

- **Use**: To test if the sample proportion differs from a known or hypothesized population proportion.
- **Hypotheses**:
  - Null Hypothesis (H_0): $p = p_0$ (the sample proportion is equal to the hypothesized proportion).
  - Alternative Hypothesis (H_1: $p \neq p_0$ (the sample proportion is not equal to the hypothesized proportion).
- **Test Statistic**:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Where:

- $\hat{p}$ = sample proportion
- p_0 = hypothesized population proportion
- n = sample size

Two Proportions test

✓ The two-proportion test is a statistical method used to compare the proportions of a specific outcome between two independent groups. It helps determine if there is a significant difference between the two proportions.

When to Use

- You have two independent groups and want to compare the proportions of a certain characteristic or outcome (e.g., success rates, yes/no responses).
- The sample sizes should be sufficiently large (generally at least 5 successes and 5 failures in each group).

2. **Two-Sample Proportion Test**

- **Use**: To compare the proportions of two independent groups.
- **Hypotheses**:
  - Null Hypothesis (H_0): $p_1 = p_2$ (the proportions are equal).
  - Alternative Hypothesis (H_1: $p_1 \neq p_2$ (the proportions are not equal).
- **Test Statistic**:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

.

Where:

- $\hat{p}_1$ = sample proportion for group 1
- $\hat{p}_2$ = sample proportion for group 2
- $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$ (pooled proportion)
- x_1 = number of successes in group 1
- x_2 = number of successes in group 2
- n_1 and n_2 = sample sizes of the two groups

testing categorical variables

✓ Testing categorical variables typically involves determining if there are significant associations or differences between categories in one or more groups.

- ✓ The most common statistical tests for categorical variables include the Chi-Square Test of Independence

Chi-Square Test of Independence

- ✓ **Purpose**: To determine if there is a significant association between two categorical variables.

When to Use:

- ✓ You have two categorical variables.
- ✓ You want to assess whether the distribution of one variable is independent of the other.

Hypotheses:

- **Null Hypothesis** (H_0): The two categorical variables are independent (no association).
- **Alternative Hypothesis** (H_1): The two categorical variables are not independent (there is an association).

Test Statistic: The test statistic is calculated as:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Where:

- O = observed frequency
- E = expected frequency under the null hypothesis

ERRORS AND POWER

ERRORS

- ✓ An error refers to the discrepancy or deviation between an observed or calculated value and the true value or the value that one would expect under ideal conditions

Types of Errors in R

- ✓ Type1 Error (false positive)
- ✓ Type 2 Error (False Negative)

Type I Error (false positive):

- ✓ A Type I error occurs when the null hypothesis is rejected when it is actually true.
- ✓ **Example**: The test indicates that the patient has the disease (positive result), but in reality, the patient does not have it.

- ✓ **Consequence**: The patient might undergo unnecessary treatments or experience anxiety due to a false diagnosis.
- ✓ Type I Error (also known as alpha error): Type I error occurs when we reject the Null hypothesis but the Null hypothesis is correct. This case is also known as a false positive.

Type 2 Error (False Negative)

- ✓ A Type II error occurs when the null hypothesis is not rejected when it is actually false.
- ✓ **Example**: The test indicates that the patient does not have the disease (negative result), but in reality, the patient does have it.
- ✓ **Consequence**: The patient may not receive necessary treatment, leading to potential worsening of their condition.
- ✓ Type II Error (also known as beta error): Type II error occurs when we fail to remove the Null Hypothesis when the Null hypothesis is incorrect/the alternative hypothesis is correct. This case is also known as a false negative.
- ✓ Denoted by β (beta), which represents the probability of making a Type II error.
- ✓ Solution: Increase the sample size or adjust the significance level. Increasing the sample size reduces the chance of a Type II error by providing more information for analysis. Alternatively, adjusting the significance level can also affect the likelihood of a Type II error.

Power

- ✓ In statistics, **power** refers to the probability that a statistical test will correctly reject a false null hypothesis. Essentially, it measures a test's ability to detect an effect or difference when one truly exists.
- ✓ The power of a test is denoted by $1-\beta$, where β is the probability of making a Type II error (failing to reject a false null hypothesis).
- ✓ The probability of correctly rejecting a false null hypothesis (i.e., the probability of avoiding a Type II error).

Formula: Power=1−β

Factors Influencing Power:

- **Sample Size:** Increasing the sample size generally increases power.
- **Effect Size:** A larger effect size increases power.
- **Significance Level (α):** Lowering α increases power, but it also increases the risk of Type I error.

- **Variability:** Lower variability in the data increases power.

Balancing Type I and Type II Errors:

- ✓ There is often a trade-off between Type I and Type II errors. Lowering the significance level (α) to reduce Type I error increases the risk of Type II error, and vice versa.
- ✓ Researchers typically choose a significance level based on the importance of avoiding Type I or Type II errors in a particular context.

Analysis of variance.

- ✓ ANOVA (Analysis of Variance) is a statistical technique used to compare the means of multiple groups. It's particularly useful when you have more than two groups and want to determine if there are significant differences between them.

Types of ANOVA:

1. One-Way ANOVA: Compares the means of multiple groups from a single factor.
2. Two-Way ANOVA: Compares the means of multiple groups from two factors, allowing for the analysis of both main effects and interactions between the factors.
3. Repeated Measures ANOVA: Used when the same individuals are measured multiple times under different conditions.

One-Way ANOVA:

- ✓ One-Way ANOVA is a statistical technique used to compare the means of multiple groups (or levels) of a single independent variable (factor). It's a common tool in various fields, including psychology, biology, and social sciences.

Steps to Conduct One-Way ANOVA

1. **Define the Hypotheses**:
   - ✓ **Null Hypothesis (H0)**: All group means are equal (no effect).
   - ✓ **Alternative Hypothesis (H1)**: At least one group mean is different.
2. **Collect Data**:
   - ✓ Gather samples from three or more independent groups.

3. **Check Assumptions**:
   - ✓ **Independence**: Samples must be independent.
   - ✓ **Normality**: Data in each group should be approximately normally distributed.
   - ✓ **Homogeneity of Variance**: Variances among the groups should be roughly equal.

4. **Calculate Group Means and Overall Mean**:
   - Compute the mean for each group and the overall mean across all groups.
5. **Calculate the Sums of Squares**:
   - **Between-Group Sum of Squares (SSB)**:

$$SSB = n \sum (\text{Group Mean} - \text{Overall Mean})^2$$

   - **Within-Group Sum of Squares (SSW)**:

$$SSW = \sum (\text{Individual Value} - \text{Group Mean})^2$$

6. **Calculate Degrees of Freedom**:
   - **Between-Groups**: $df_{between} = k - 1$ (where k is the number of groups).
   - **Within-Groups**: $df_{within} = N - k$ (where N is the total number of observations).

level, e.g., $\alpha = 0.05$), reject the null hypothesis.

10. **Post-hoc Tests (if needed)**:
    - If you reject the null hypothesis, conduct post-hoc tests (e.g., Tukey's HSD) to determine which specific group means are different.

7. **Calculate Mean Squares**:
   - **Mean Square Between (MSB)**:

$$MSB = SSB/df_{between}$$

   - **Mean Square Within (MSW)**:

$$MSW = SSW/df_{within}$$

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

$$F = MSB/MSW$$

9. **Determine Significance**:
   - Compare the calculated F-value to the critical F-value from the F-distribution table, or use a p-value.
   - If the calculated F is greater than the critical F (or if the p-value is less than the significance

Two-way ANOVA

✓ Two-way ANOVA is a statistical method used to assess the impact of two independent categorical variables (factors) on a continuous dependent variable. It allows researchers to examine not only the main effects of each factor but also any interaction effects between them.

1. **Factors**: Two independent variables. For example:
   - Factor A: Diet Type (A, B, C)
   - Factor B: Exercise Level (Low, High)
2. **Null Hypotheses**:
   - H_{0A}: No difference in means due to Factor A.
   - H_{0B}: No difference in means due to Factor B.
   - H_{0AB}: No interaction effect between Factors A and B.
3. **Interaction**: Examines whether the effect of one factor depends on the level of the other factor.

Steps to Conduct Two-Way ANOVA

1. **Define the Hypotheses**:
   - Set up null and alternative hypotheses for both factors and their interaction.
2. **Collect Data**:
   - Gather samples across all combinations of the factors.
3. **Check Assumptions**:
   - Independence of observations.
   - Normality within each group.
   - Homogeneity of variance across groups.
4. **Calculate Group Means**:
   - Compute means for each combination of factor levels.

5. **Calculate Sums of Squares**:
   - **Total Sum of Squares (SST)**:

$$SST = \sum(X_{ij} - \bar{X})^2$$

   - **Between-group Sum of Squares for Factor A (SSA)**:

$$SSA = n_B \sum(\bar{X}_{A_i} - \bar{X})^2$$

   - **Between-group Sum of Squares for Factor B (SSB)**:

$$SSB = n_A \sum(\bar{X}_{B_j} - \bar{X})^2$$

   - **Interaction Sum of Squares (SSAB)**:

$$SSAB = \sum(\bar{X}_{A_iB_j} - \bar{X}_{A_i} - \bar{X}_{B_j} + \bar{X})^2$$

- **Within-group Sum of Squares (SSW)**:

$$SSW = \sum(X_{ij} - \bar{X}_{A_iB_j})^2$$

. **Calculate Degrees of Freedom**:

- Total: $df_{total} = N - 1$
- For Factor A: $df_A = a - 1$ (where a is the number of levels of Factor A).
- For Factor B: $df_B = b - 1$ (where b is the number of levels of Factor B).
- For Interaction: $df_{AB} = (a - 1)(b - 1)$
- Within-groups: $df_W = N - ab$

7. **Calculate Mean Squares**:
   - Mean Square for Factor A:

$$MSA = SSA/df_A$$

   - Mean Square for Factor B:

$$MSB = SSB/df_B$$

   - Mean Square for Interaction:

$$MSAB = SSAB/df_{AB}$$

   - Mean Square Within:

$$MSW = SSW/df_W$$

8. **Calculate F-statistics**:
   - For Factor A:

$$F_A = MSA/MSW$$

- For Factor B:

$$F_B = MSB/MSW$$

- For Interaction:

$$F_{AB} = MSAB/MSW$$

9. **Determine Significance**:
   - Compare each F-statistic to the critical F-value from the F-distribution table or use p-values.
   - Reject the null hypothesis for a factor if the F-statistic is greater than the critical value (or if the p-value is less than the significance level, e.g., $\alpha = 0.05$).
10. **Post-hoc Tests (if needed)**:
    - If significant effects are found, conduct post-hoc tests to identify specific group differences.