MODULE-04

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 analysing 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:

Consider statistical testing in the context of manufacturing quality control. A production line for a light bulb manufacturer needs to ensure that a new way of producing bulbs extends their longevity.

They collect and compare data on the longevity of bulbs generated using the new method to the lifespan of bulbs produced using the old method. They can statistically analyse if the mean lifespan of bulbs generated using the new method is substantially different from those produced using the old method by running a hypothesis test (such as a t-test or ANOVA). If the test reveals a considerable difference in favour of the new procedure, the corporation can confidently implement it, knowing that it results in longer-lasting bulbs.

MODELLING:

statistical modelling involves creating mathematical representations of relationships in data.

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:

Definition: A sampling distribution is a concept used in statistics. It is a probability distribution of a statistic obtained from a larger number of samples drawn from a specific 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.
- This allows entities like governments and businesses to make more well-informed decisions based on the information they gather.
- There are a few methods of sampling distribution used by researchers, including the sampling distribution of a mean.

Example :

Consider a factory that manufactures light bulbs. The firm strives to keep its bulbs at a consistent degree of brightness. All bulbs manufactured have a normal distribution of brightness levels, with a mean () of 800 lumens and a standard deviation () of 20 lumens.

Controlling Quality Through Sampling Distribution:

Quality Control:

The factory conducts frequent quality inspections by sampling samples of bulbs and assessing their brightness.

Sampling Mean Distribution: The factory selects samples of a fixed size (say, 50 bulbs) at random and measures the mean brightness of each sample.

Analysis of Sample Means: The factory can estimate how much the sample means fluctuate around the genuine population mean brightness of 800 lumens by observing the distribution of sample means.

Applications:

Sampling distributions are used in a variety of domains and statistical analyses:

1. **Inference Based on Statistics:**
   - **Hypothesis Testing:** Sampling distributions aid in determining if sample data supports or contradicts a population parameter hypothesis.
   - **Confidence Intervals:** These aid in the construction of intervals that are most likely to contain the genuine population parameter.
2. **Quality Assurance:**
   - **Manufacturing:** Monitoring product quality by sampling items for testing.
   - **Service Industries:** Assessing service quality through periodic customer feedback samples.
3. **Economics and Finance**:
   - **Market Research:** Analysing a subset of a market to infer characteristics about the entire market.

- **Portfolio Management:** Studying the performance of investment portfolios using samples from historical data.

4. Medical and Biological Sciences:

- **Clinical Trials:** Analysing samples to draw conclusions about the effects of a treatment on a population.
- **Genetics:** Studying genetic traits within specific populations by analysing samples.

Public Opinion and Political Polling:

5. **Election Forecasting:** Estimating voting trends based on samples of the population.

1. Policy Making: Assessing public opinion on various policies through sampling surveys.

6. **Education:**

2. **Assessment:** Evaluating student performance by examining a subset of test scores.

Environmental Studies:

7. **Sampling Soil or Water Quality:** Assessing environmental factors through samples from different regions.

Sampling distributions are fundamental in statistical analyses, allowing statisticians, researchers, and professionals in various fields to make informed decisions and draw conclusion.

Types of Sampling Distributions

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. In order to get to this point, the researcher must figure out the mean of each sample group and map out the individual data.
2. **Sampling Distribution of Proportion:** This method involves choosing a sample set from the overall population to get the proportion of the sample.
    - The mean of the proportions ends up becoming the proportions of the larger group.
3. **T-Distribution:** This type of sampling distribution is common in cases of small sample sizes.
    - It may also be used when there is very little information about the entire population.

- T-distributions are used to make estimates about the mean and other statistical points.

Note:

i. probability distribution, the central "balance" point of a sampling distribution is its mean,

ii. The standard deviation of a sampling distribution is referred to as a standard error.

1. DISTRIBUTION FOR A SAMPLE MEAN:

- The sampling distribution of the mean represents the distribution of sample means taken from a population.
- It helps understand how sample means vary and approach the population mean as sample size increases, following the Central Limit Theorem.
- Mathematically, the variability inherent in an estimated sample mean is described as follows: Formally, denote the random variable of interest as X^-. This represents the mean of a sample of n observations from the "raw observation" random variable X, as in x1, x2,. . . , xn.
- Those observations are assumed to have a true finite mean $-\infty < \mu X < \infty$ and a true finite standard deviation $0 < \sigma X < \infty$.
- The conditions for finding the probability distribution of a sample mean vary depending on whether you know the value of the standard deviation.

Situation 1: Standard Deviation Known

- When the true value of the standard deviation σX is known, then the following are true:
- If X itself is normal, the sampling distribution of X^- is a normal distribution, with mean μX and standard error $\sigma X/\sqrt{n}$.
- If X is not normal, the sampling distribution of X^- is still approximately normal, with mean μX and standard error $\sigma X/\sqrt{n}$, and this approximation improves arbitrarily as $n \to \infty$. This is known as the central limit theorem (CLT).

Situation 2: Standard Deviation Unknown

- The standard deviation of the raw measurement distribution that generated your sample data.
- In this eventuality, it's usual to just replace σ_x with S_x which is the standard deviation of the sampled data.

- Standardized values of the sampling distribution of $\bar{X}$ follow a t-distribution with $v = n - 1$ degrees of freedom; standardization is performed using the standard error $sX/\sqrt{n}$.
- In additionally, n is small, then it is necessary to assume the distribution of X is normal for the validity of this t-based sampling distribution of $\bar{X}$.
- The nature of the sampling distribution of $\bar{X}$ therefore depends upon whether the true standard deviation of the observations is known, as well as the sample size n.
- The CLT states that normality occurs even if the raw observation distribution is itself not normal, but this approximation is less reliable if n is small. It's a common rule of thumb to rely on the CLT only if $n \geq 30$.
- If sX, the sample standard deviation, is used to calculate the standard error of $\bar{X}$, then the sampling distribution is the t-distribution(following standardization). Again, this is generally taken to be reliable only if $n \geq 30$.

It is estimated as $\hat{p} = \frac{x}{n}$, where x is the number of successes in a sample of size n. Let the corresponding true proportion of successes (often unknown) simply be denoted with π.

$$\hat{P} \sim \mathrm{N}\left(\hat{p}, \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right),$$

1) Sampling dist In population gathering samples.
- SD provides for Mechanism making Quantitative decession about a process th
- The SD" Effective tool for Research or Researchers. financial analyst and Stock Market Strategies. and RHH Ph.D Resear -ers.

Properties of Sampling Dist
1) Population - Whole data are Complete data
2) Sample - Subset of population or few data. in population
2) Sampling - Getting a Sample from population
4) Random Sampling - Equal chance of Getting every Sample

parameters with Replacement
1 n. - Sample of size
2 N - population of size
3 N_{C_n} - Without replacement (Sd. unknown)
4 N^n - with replacements (Sd, known)
5 (μ, σ) - are represented by population.
Sample mean, Sd

Mean of the sample dist

The Sample taken from population $x_1, x_2 \ldots x_n$.

Mean of the Sample is represented as (μ).

$$\mu = \bar{x} = \frac{x_1 + x_2 \ldots x_n}{n}$$

$$= \frac{\sum x_i}{n}$$

Variance of the Sample dist.

Variance in sampling dist measures how much set of no. of given observation from the mean (Avg)

$$\text{Var}(s)^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

Standard error (or) Standard deviation

Sd of Sampling dist suffered as standard error.

$$SE = \sqrt{\frac{\sum (x_i - \bar{x})}{n}} \quad (\text{or}) \quad \frac{s}{\sqrt{n}}$$

Example

Consider Sample 50 Students marks of 165. SE Sd = 7. To find SE of the 165 marks.

$s = 7$ $n = 50$ Standard Error of 165M $= \frac{s}{\sqrt{n}} = \frac{7}{\sqrt{50}} = 0.99$.

The Sampling dist of a Sample 150 Patients of 198 diseases. Sd = 11.5

To find SE of the 198 diseases.

$s = 11.50$ $n = 150$ $= \frac{11.5}{\sqrt{150}} = \frac{11.5}{12.24} = 0.93$

Example2:

Population parameters

```
population_mean <- 75
population_std <- 10
sample_size <- 25
num_samples <- 1000 # Number of samples to take

# Generating samples and calculating sample means
sample_means <- numeric(num_samples)
for (i in 1:num_samples) {
  sample <- rnorm(sample_size, mean = population_mean, sd = population_std)
  sample_means[i] <- mean(sample)
}

# Calculating the mean of sample means
sampling_distribution_mean <- mean(sample_means)

print(paste("Mean of the sampling distribution of the mean:", sampling_distribution_mean))
```

Output:

```
Mean of the sampling distribution of the mean: 74.9392512095679"
```

Confidence Intervals

- A *confidence interval (CI)* is an interval defined by a lower limit *l* and an upper limit *u*, used to describe possible values of a corresponding true population parameter in light of observed sample data.
- Interpretation of a confidence interval therefore allows you to state a "level of confidence" that a true parameter of interest falls between this upper and lower limit, often expressed as a percentage.
- As such, it is a common and useful tool built directly from the sampling distribution of the statistic of interest.

The following are the important points to note:

- The level of confidence is usually expressed as a percentage, such that you'd construct a $100*(1-\alpha)$ percent confidence interval, where $0 < \alpha < 1$ is an "amount of tail probability."
- The three most common intervals are defined with either $\alpha = 0.1$ (a 90 percent interval), $\alpha = 0.05$ (a 95 percent interval), or $\alpha = 0.01$ (a 99 percent interval).

Confidence intervals may be constructed in different ways, depending on the type of statisti

statistic = critical value × standard error,

A critical value is the value of the test statistic which defines the upper and lower bounds of a confidence interval, or which defines the threshold of statistical significance in a statistical test

Statistic to state the population parameter

standard error, in the sampling distribution standard deviation

* The "Level of confidence" are Expressed in the percentage Such that Construct $100 \times (1-\alpha)$ % confidence where α $\alpha < 1$ is an Range of Amount of test probable

$\alpha = 0.1 \Rightarrow 100 \times (1-\alpha) = 100 \times (1-0.1)$ 90 % Confidence

$\alpha = 0.5 \Rightarrow 100 \times (1-\alpha) = 100 \times (1-0.5)$ 95% confidence rate

$\alpha = 0.01 \Rightarrow 100 \times (1-\alpha) = 100 \times (1-0.01) = 99\%$ Confidence

The formula for a confidence interval for the mean in a sampling distribution, assuming a normal distribution or a sufficiently large sample size (Central Limit Theorem),

Confidence Interval=Sample Mean±(Critical Value×Standard DeviationSample Size)

Confidence Interval=Sample Mean±(Critical Value×Sample Size/Standard Deviation

For example, let's say we want to find a 95% confidence interval for the mean weight of apples sampled from a farm. We collect a sample of 50 apples, measure their weight, and find the sample mean weight to be 150 grams. Let's assume the population standard deviation is 10 grams.

Solution:

Given a 95% confidence level and a normal distribution (z-distribution) with a critical value of 1.96 for a 95% confidence interval:

Confidence Interval=$150 \pm (1.96 \times \frac{10}{\sqrt{50}})$

Confidence Interval=$150 \pm (1.96 \times \frac{10}{\sqrt{50}}) = 150 \pm 2.78$

Therefore, the 95% confidence interval for the mean weight of apples in the population would be approximately from 147.22147.22 grams to 152.78152.78 grams.

```
# Given data
sample_mean <- 2.5
sample_standard_deviation <- 0.8
sample_size <- 500
# Calculate critical value for 95% confidence level
confidence_level <- 0.95
alpha <- 1 - confidence_level
critical_value <- qnorm(1 - alpha / 2)
# Calculate margin of error
margin_of_error <- critical_value * (sample_standard_deviation / sqrt(sample_size))
# Calculate confidence interval
lower_bound <- sample_mean - margin_of_error
upper_bound <- sample_mean + margin_of_error

# Print the confidence interval
cat("The 95% confidence interval for the average time spent on social media is
[", lower_bound, ",", upper_bound, "]\n")
```

Output:

```
The 95% confidence interval for the average time spent on social media is
[ 2.429878 , 2.570122 ]
```

HYPOTHESIS TESTING:

Hypothesis testing is a tool for making statistical inferences about the population data.

- Hypothesis testing can be defined as a statistical tool that is used to identify if the results

of an experiment are meaningful or not.

- It involves setting up a null hypothesis and an alternative hypothesis.
- These two hypotheses will always be mutually exclusive. This means that if the null hypothesis is true then the alternative hypothesis is false and vice versa.

Example of hypothesis testing is setting up a test to check if a new medicine works on a disease in a more efficient manner.

APPICATION OG HYPOTHESIS TESTING

- **Scientific Research:** In biology, psychology, medicine, etc., to test new treatments, study the effects of variables, or validate theories.
- **Quality Control:** In manufacturing to ensure products meet certain standards by testing hypotheses about product quality.
- **Business Decisions:** To analyze market trends, customer behavior, or the effectiveness of strategies.
- **Social Sciences:** To study societal trends, behaviors, and attitudes.
- **Economics:** To analyze the impact of policies, market changes, or economic theories.

TYPES OF HYPOTHESIS

1.Null Hypothesis: The null hypothesis is a concise mathematical statement that is used to indicate that there is no difference between two possibilities.

- In other words, there is no difference between certain characteristics of data.
- This hypothesis assumes that the outcomes of an experiment are based on chance alone.
- It is denoted as H0. Hypothesis testing is used to conclude if the null hypothesis can be rejected or not.

  Example: Suppose an experiment is conducted to check if girls are shorter than boys at the age of 5. The null hypothesis will say that they are the same height.

2.Alternative Hypothesis

- 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.

For the above-mentioned example, the alternative hypothesis would be that girls are shorter than boys at the age of 5.

Hypothesis Testing P Value

- In hypothesis testing, the p value is used to indicate whether the results obtained after conducting a test are statistically significant or not.
- It also indicates the probability of making an error in rejecting or not rejecting the null hypothesis. This value is always a number between 0 and 1.
- The p value is compared to an alpha level, α or significance level.
- The alpha level can be defined as the acceptable risk of incorrectly rejecting the null hypothesis. The alpha level is usually chosen between 1% to 5%.

Hypothesis Testing Critical region

All sets of values that lead to rejecting the null hypothesis lie in the critical region. Furthermore, the value that separates the critical region from the non-critical region is known as the critical value.

Hypothesis Testing Formula:

Depending upon the type of data available and the size, different types of hypothesis testing are used to determine whether the null hypothesis can be rejected or not.

The hypothesis testing formula for some important test statistics are given below:

- $z = \frac{\bar{x} + \mu}{\frac{\sigma}{\sqrt{n}}}$ $\bar{X}$ is the sample mean, μ is the population mean, σ is the population standard deviation and n is the size of the sample.
- $t = \frac{\bar{x} + \mu}{\frac{s}{\sqrt{n}}}$ s is the sample standard deviation.
- $\chi^2 = \sum \frac{(o_i - E_i)^2}{E_i}$ is the observed value and E_i is the expected value.

Hypothesis Testing Z Test

A z test is a way of hypothesis testing that is used for a large sample size ($n \geq 30$).

It is used to determine whether there is a difference between the population mean and the sample mean when the population standard deviation is known.

It can also be used to compare the mean of two samples. It is used to compute the z test statistic.

The formulas are given as follows:

One sample: $z = \frac{\bar{x}+\mu}{\frac{\sigma}{\sqrt{n}}}$

Two samples: $z=\frac{(\bar{X_1}-\bar{X_2})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{\sqrt{n}}+\frac{\sigma_2^2}{\sqrt{n}}}}$

Hypothesis Testing t Test

The t test is another method of hypothesis testing that is used for a small sample size ($n < 30$).

It is also used to compare the sample mean and population mean.

However, the population standard deviation is not known.

Instead, the sample standard deviation is known.

The mean of two samples can also be compared using the t test.

One sample: $t = \frac{\bar{x}+\mu}{\frac{s}{\sqrt{n}}}$

Two samples: $t=\frac{(\bar{X_1}-\bar{X_2})-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{\sqrt{n}}+\frac{s_2^2}{\sqrt{n}}}}$

Hypothesis Testing Chi Square

The Chi square test is a hypothesis testing method that is used to check whether the variables in a population are independent or not. It is used when the test statistic is chi-squared distributed.

One Tailed Hypothesis Testing

- One tailed hypothesis testing is done when the rejection region is only in one direction.
- It can also be known as directional hypothesis testing because the effects can be tested in one direction only.
- This type of testing is further classified into the right tailed test and left tailed test.

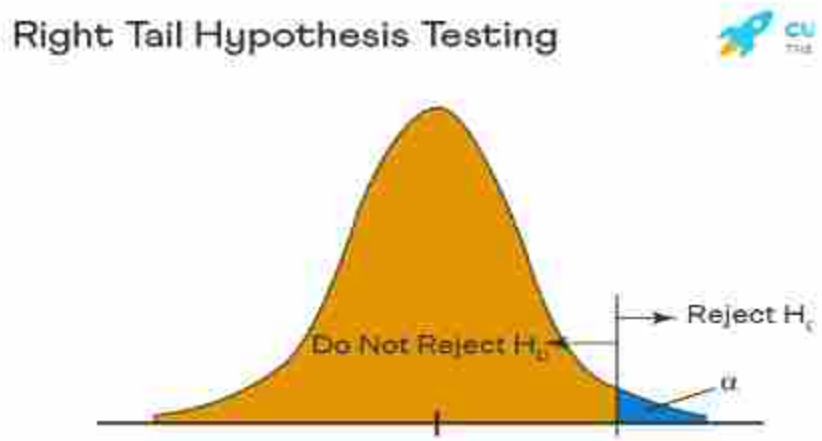
Right Tailed Hypothesis Testing

- The right tail test is also known as the upper tail test. This test is used to check whether the population parameter is greater than some value.
- The null and alternative hypotheses for this test are given as follows:

H0: The population parameter is ≤ some value

H1: The population parameter is > some value.

If the test statistic has a greater value than the critical value then the null hypothesis is rejected



Left Tailed Hypothesis Testing

The left tail test is also known as the lower tail test. It is used to check whether the population parameter is less than some value. The hypotheses for this hypothesis testing can be written as follows:

H0: The population parameter is ≥ some value

H1: The population parameter is < some value.

The null hypothesis is rejected if the test statistic has a value lesser than the critical value.

Two Tailed Hypothesis Testing

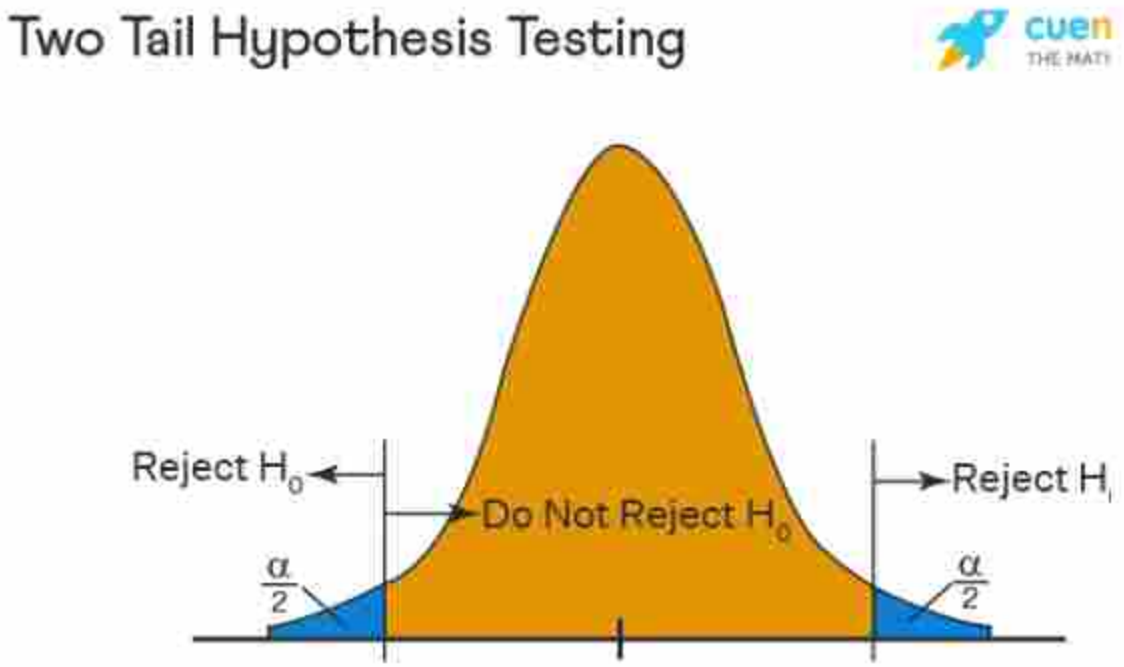
In this hypothesis testing method, the critical region lies on both sides of the sampling distribution. It is also known as a non - directional hypothesis testing method.

The two-tailed test is used when it needs to be determined if the population parameter is assumed to be different than some value. The hypotheses can be set up as follows:

H0: the population parameter = some value

H1: the population parameter ≠ some value

The null hypothesis is rejected if the test statistic has a value that is not equal to the critical value.



Hypothesis Testing Steps

Hypothesis testing can be easily performed in five simple steps. The most important step is to correctly set up the hypotheses and identify the right method for hypothesis testing.

The basic steps to perform hypothesis testing are as follows:

- **Step 1**: Set up the null hypothesis by correctly identifying whether it is the left-tailed, right-tailed, or two-tailed hypothesis testing.
- **Step 2:** Set up the alternative hypothesis.
- **Step 3:** Choose the correct significance level, α, and find the critical value.
- **Step 4:** Calculate the correct test statistic (z, t or χ) and p-value.
- **Step 5:** Compare the test statistic with the critical value or compare the p-value with α to arrive at a conclusion. In other words, decide if the null hypothesis is to be rejected or not.

Hypothesis Testing Example

The best way to solve a problem on hypothesis testing is by applying the 5 steps mentioned in the previous section. Suppose a researcher claims that the mean average weight of men is greater than 100kgs with a standard deviation of 15kgs. 30 men are chosen with an average weight of 112.5 Kgs. Using hypothesis testing, check if there is enough evidence to support the researcher's claim. The confidence interval is given as 95%.

Step 1: **This is an example of a right-tailed test. Set up the null hypothesis as H0: $\mu = 100$.**

Step 2: The alternative hypothesis is given by H1: $\mu > 100$.

Step 3: As this is a one-tailed test, $\alpha = 100\% - 95\% = 5\%$. This can be used to determine the critical value.

$1 - \alpha = 1 - 0.05 = 0.95$

0.95 gives the required area under the curve. Now using a normal distribution table, the area 0.95 is at $z = 1.645$. A similar process can be followed for a t-test. The only additional requirement is to calculate the degrees of freedom given by n - 1.

Step 4: Calculate the z test statistic. This is because the sample size is 30.

Furthermore, the sample and population means are known along with the standard deviation.

$\mu = 100, \bar{x} = 112.5, n = 30, \sigma = 15$

$$z = \frac{\bar{x}+\mu}{\frac{\sigma}{\sqrt{n}}} \quad z = \frac{112.5+100}{\frac{15}{\sqrt{30}}} = 4.56$$

Step 5: Conclusion. As $4.56 > 1.645$ thus, the null hypothesis can be rejected.

Hypothesis Testing and Confidence Intervals

- **Confidence intervals** form an important part of hypothesis testing.

  This is because the alpha level can be determined from a given confidence interval.

- Suppose a confidence interval is given as 95%. Subtract the confidence interval from 100%.

  This gives 100 - 95 = 5% or 0.05.

- This is the alpha value of a one-tailed hypothesis testing.

  To obtain the alpha value for a two-tailed hypothesis testing, divide this value by 2.

  This gives 0.05 / 2 = 0.025

Important Notes on Hypothesis Testing

- Hypothesis testing is a technique that is used to verify whether the results of an experiment are statistically significant.
- It involves the setting up of a null hypothesis and an alternate hypothesis.
- There are three types of tests that can be conducted under hypothesis testing - z test, t test, and chi square test.
- Hypothesis testing can be classified as right tail, left tail, and two tail tests.

Hypothesis testing involves several components:

Null Hypothesis (H0): A statement assumed to be true until evidence suggests otherwise.

Alternative Hypothesis (H1 or Ha): A statement challenging the null hypothesis.

Test Statistic: A numerical summary used to decide between the null and alternative hypotheses.

Significance Level (α): The threshold for accepting or rejecting the null hypothesis.

P-value: The probability of obtaining results as extreme as the observed data, assuming the null hypothesis is true.

Decision Rule: Criteria based on the significance level and the p-value to accept or reject the null hypothesis.

Conclusion: Based on the p-value compared to the significance level, determining whether to reject or fail to reject the null hypothesis.

Hypothesis testing in R

```
# Given values
population_mean <- 100
sample_mean <- 112.5
standard_deviation <- 15
sample_size <- 30
```

```
# Calculating the z-score
z_score <- (sample_mean - population_mean) / (standard_deviation / sqrt(sample_size))

# Specifying the alpha level
alpha <- 0.05

# Finding the critical value for a one-tailed test
critical_value <- qnorm(1 - alpha)
# Making the conclusion
if (z_score > critical_value) {
 cat("Null hypothesis rejected: There is enough evidence to support the claim that the mean average weight of men is greater than 100kgs.")
} else {
 cat("Null hypothesis cannot be rejected: There is not enough evidence to support the claim that the mean average weight of men is greater than 100kgs.")
}
```

OUTPUT:

```
Null hypothesis rejected: There is enough evidence to support the claim t
hat the mean average weight of men is greater than 100kgs.
```

components of hypothesis test

A hypothesis test typically involves several components, and in R programming, different functions and packages are used for each stage of the process. Here are the key components of a hypothesis test in R:

1.**Formulate Hypotheses:**

State the null hypothesis (H0) and alternative hypothesis (H1).

Example:

```
# Null hypothesis: The mean of group1 is equal to the mean of group2
# Alternative hypothesis: The means are not equal
```

2.**Choose Significance Level (Alpha):**

Decide on the significance level (common choices are 0.05 or 0.01).

Example:

```
# Set the significance level
alpha <- 0.05
```

3. **Collect and Analyze Data:** Collect relevant data.

Example:

```
# Collect data into vectors, matrices, or data frames
group1 <- c(25, 30, 35, 40, 45)
group2 <- c(20, 28, 32, 38, 42)
```

4.Choose the Appropriate Test:

Select the appropriate statistical test based on the type of data and research question.

Example:

Two-sample t-test for comparing means.

Chi-squared test for independence.

Wilcoxon rank sum test for non-parametric comparisons

s.Perform the Test:

Use the relevant R function to perform the hypothesis test.

Example:

```
# Two-sample t-test
t_test_result <- t.test(group1, group2)
```

6.Perform the Test:

Use the relevant R function to perform the hypothesis test.

Example:

```
# Two-sample t-test
t_test_result <- t.test(group1, group2)
```

7.Make a Decision:

Based on the p-value and significance level, decide whether to reject the null hypothesis.

Example: # Compare p-value to significance level

```
if (t_test_result$p.value < alpha) {
  cat("Reject the null hypothesis\n")
} else {
  cat("Fail to reject the null hypothesis\n")
}
```

Output: Fail to reject the null hypothesis

8.Draw Conclusions:

Provide a conclusion based on the results of the hypothesis test.

Example:

```
# Conclusion
cat("The means are significantly different\n")
```

Output:

The means are significantly different

TESTING MEANS:

Testing means in sampling distributions typically involves hypothesis testing, where you compare sample statistics to population parameters or compare two sample means to each other. It's a way to infer whether any observed difference between the means is significant or due to random chance.

Types of testing means

TESTING MEANS: testing means in sampling distributions involves various methods to compare means, determine differences or similarities between populations or samples, and infer conclusions about the underlying data.

In testing mean majorly we can use T test or z test Anova,paired test

1Single testing mean or One sample mean

2Two sample testing mean or two sample mean

Single testing mean or One sample mean

Single testing of a mean, also known as a one-sample mean test, is a statistical analysis method used to determine whether the mean of a single sample differs significantly from a known or hypothesized population mean.

1. **Z-Test for One Sample Mean**:

- Utilized when the population variance is known.
- Assesses if a sample mean is significantly different from a known population mean.

One sample $z = \frac{\bar{x}+\mu}{\frac{\sigma}{\sqrt{n}}}$

Example: # Given data

```
sample_data <- c(23, 27, 22, 30, 25, 28, 21, 26, 24, 29)
population_mean <- 26
population_sd <- ... # Assign the known population standard deviation
# Calculate sample statistics
sample_mean <- mean(sample_data)
sample_size <- length(sample_data)
# Calculate Z-score
z_score <- (sample_mean - population_mean) / (population_sd / sqrt(sample_size))
# Specify significance level alpha (e.g., 0.05)
alpha <- 0.05
# Find critical value for a two-tailed test
critical_value <- qnorm(1 - alpha / 2)
# Make a decision
if (abs(z_score) > critical_value) {
```

cat("Reject null hypothesis: The sample mean is significantly different from the population mean.")

} else {

cat("Fail to reject null hypothesis: There is not enough evidence to claim a significant difference.")

}

Output: Fail to reject null hypothesis: There is not enough evidence to claim a significant difference.

2. **T-Test for One Sample Mean**:

- Applied when the population variance is unknown.
- Determines if a sample mean significantly differs from a hypothesized population mean.

One sample: $T = \frac{\bar{x}+\mu}{\frac{s}{\sqrt{n}}}$

Example:

```
# Generate a sample
sample_data <- c(23, 27, 22, 30, 25, 28, 21, 26, 24, 29)
# Perform one-sample t-test
t_test_result <- t.test(sample_data, mu = 26) # mu is the hypothesized mean
# Print the result
print(t_test_result)
```

Output:

```
One Sample t-test
data: sample_data
t = -0.52223, df = 9, p-value = 0.6141
alternative hypothesis: true mean is not equal to 26
```

95 percent confidence interval:

23.33415 27.66585

sample estimates:

mean of x

25.5

2.Two sample testing mean or two sample mean

Two-sample testing for means, also known as a two-sample mean test, involves comparing the means of two independent samples to determine if they come from populations with different means.

This test evaluates whether there's a significant difference between the means of the two groups or populations.

Z-Test for Two Sample Means:

- Compares means of two independent samples.
- Checks if the difference between sample means is statistically significant.

Two samples: $z=\frac{(\overline{X_1}-X_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{\sqrt{n}}+\frac{\sigma_2^2}{\sqrt{n}}}}$

Example: # Generate two samples

```
group1 <- c(23, 27, 22, 30, 25)

group2 <- c(28, 21, 26, 24, 29)


# Perform two-sample t-test

t_test_result <- t.test(group1, group2)

# Print the result
```

```
print(t_test_result)
```

Output: Welch Two Sample t-test

```
data:  group1 and group2
t = -0.098533, df = 8, p-value = 0.9239
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.880677  4.480677
sample estimates:
mean of x mean of y
    25.4      25.6
```

T-Test for Two Sample Means:

- Compares means of two independent samples with unknown population variances.
- Assesses if there's a significant difference between the means of the two groups.

Two samples: $t=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$

Example:

```
# Given data for two samples

sample_data_1 <- c(23, 27, 22, 30, 25)

sample_data_2 <- c(28, 21, 26, 24, 29)



# Calculate sample statistics for each sample

mean_1 <- mean(sample_data_1)

mean_2 <- mean(sample_data_2)

n_1 <- length(sample_data_1)

n_2 <- length(sample_data_2)

var_1 <- var(sample_data_1)

var_2 <- var(sample_data_2)
```

```
# Perform two-sample t-test assuming unequal variances
t_test_result <- t.test(sample_data_1, sample_data_2, var.equal = FALSE)

# Extract p-value and test statistic
p_value <- t_test_result$p.value
test_statistic <- t_test_result$statistic

# Specify significance level alpha (e.g., 0.05)
alpha <- 0.05

# Make a decision based on the p-value or test statistic
if (p_value < alpha) {
 cat("Reject null hypothesis: The means of the two samples are significantly different.")
} else {
 cat("Fail to reject null hypothesis: There is not enough evidence to claim a significant difference
in means.")
}
```

Output: Fail to reject null hypothesis: There is not enough evidence to claim a significant difference in means.

Testing of proportions:

Testing Proportions

Let us consider the parameter p of the population proportion. For instance, we might want toknow the proportion of males within a total population of adults when we conduct a survey. A test of proportion will assess whether or not a sample from a population represents the true proportion of the entire population.

One-Sample Z-Test

In testing for the true value of some proportion of success, let ^p be the sample proportion overn

trials, and let the null value be denoted with 0.

You find the test statistic with the following

$$Z = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

R Function: prop.test

To the **prop.test** function, as used for a single sample test of a proportion, provide the number of successes observed as x, the total number of trials as n, and the null value as p.

Example: prop.test(x=sum(sick),n=length(sick),p=0.2)

Output:

1-sample proportions test without continuity correction data:

sum(sick) out of length(sick), null probability 0.2

X-squared = 1.0431

alternative hypothesis: true p is not equal to 0.2

sample estimates:

0.2758621

Two Proportions

To compare two estimated proportions from independent populations.

In testing for the true difference between two proportions mathematically, π_1 and π_2, let $\hat{p}_1 = x_1/n_1$ be the sample proportion for x_1 successes in n_1 trials corresponding to π_1, and the same quantities as $\hat{p}_2 = x_2/n_2$ for π_2. With a null value of the difference denoted π_0, the test statistic is given by the following:

$$Z = \frac{\hat{p}_2 - \hat{p}_1 - \pi_0}{\sqrt{p^*(1-p^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad (18.9)$$

Example:

prop.test(x=c(x2,x1),n=c(n2,n1),alternative="greater",correct=FALSE)

Output:

1-sample test for equality of proportions without continuity

2- correction data: c(x2, x1) out of c(n2, n1)

X-squared = 9.9395

alternative hypothesis: greater sample estimates:

prop1 prop 2

0.8883249 0.7725322

Testing categorical variables:

- testing a categorical variable typically involves methods to analyse and compare the distribution or proportions of different categories within the variable.
- It often includes techniques like chi-square tests or other hypothesis tests to determine if there are significant differences between groups or categories.

Application / use of testing categorical variable

Market Research: Understanding consumer behaviour by analysing preferences, buying patterns, or demographic characteristics.

Medical Research: Examining the relationship between categorical variables like treatment types and patient outcomes.

Social Sciences: Studying the impact of social factors such as education level, income, or marital status on various aspects of life.

Quality Control: Analysing defects or non-conformities in manufacturing by studying categorical variables related to production lines or product characteristics.

Election Analysis: Investigating voting patterns, demographics, and factors influencing voting

behaviour.

Customer Segmentation: Creating customer profiles based on preferences, demographics, or behaviour to tailor marketing strategies.

Psychology: Understanding behaviour by examining categorical variables like personality types, mental health conditions, or responses to stimuli.

Types of categorical variable

1. Single Categorical Variable
2. Two Categorical Variables

1Single Categorical Variable

The sampling distribution of a single categorical variable, specifically the distribution of sample proportions, can be estimated using the formula for the standard error of a proportion.

Formula for Standard Error of a Proportion:

The standard error (SE) of a sample proportion ($\hat{p}$) is calculated

$$SE(\hat{P} = \sqrt{\frac{\hat{P}*(1-\hat{P})}{n}}$$

Where $\hat{p}$ is the sample proportion.

n is the sample size.

```
population <- c(rep(1, 400), rep(0, 600)) # 1 represents the category of interest
```

Example:

```
# Simulated population data (binary categorical variable)
population <- c(rep(1, 400), rep(0, 600)) # 1 represents the category of interest
# Function to calculate standard error of proportion
se_proportion <- function(sample_prop, sample_size) {
  sqrt((sample_prop * (1 - sample_prop)) / sample_size)
}
# Simulating the sampling distribution by taking multiple samples
set.seed(123) # Setting seed for reproducibility
```

```
num_samples <- 1000
sample_sizes <- c(30, 50, 100) # Different sample sizes to explore
for (size in sample_sizes) {
 sample_props <- replicate(num_samples, mean(sample(population, size = size) == 1))
 se <- se_proportion(sample_props, size)
 cat("Sample size:", size, "\n")
 cat("Mean proportion:", mean(sample_props), "\n")
 cat("Standard Error:", mean(se), "\n\n")
}
```

Output:

```
Sample size: 30
Mean proportion: 0.4003
Standard Error: 0.08780136
Sample size: 50
Mean proportion: 0.40178
Standard Error: 0.06863868
Sample size: 100
Mean proportion: 0.39843
Standard Error: 0.04874045
```

2.Two Categorical Variables

Two categorical variables refer to two types of data that are categorical or qualitative in nature and consist of distinct categories or groups. These variables categorize observations into specific groups or classes without any intrinsic order or numerical value.

To use two categorical variable to use chi square test

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

Example:

```
# Create a contingency table
data <- matrix(c(30, 10, 20, 40), nrow = 2)
# Perform chi-squared test
chi_squared_result <- chisq.test(data)
# Print the result
```

```
print(chi_squared_result)
```

Output:

```
Pearson's Chi-squared test with Yates' continuity correction

data:  data
X-squared = 15.042, df = 1, p-value = 0.0001052
```

ERRORS AND POWER:

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

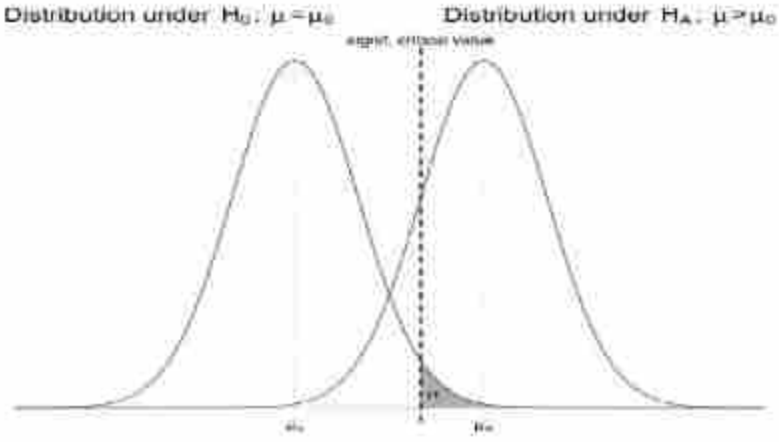
Type I Error (False Positive):

Definition: Type I error occurs when you reject a true null hypothesis. It's the probability of rejecting a null hypothesis when it's actually true.

- 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

Solution: Decrease the significance level (alpha level) for hypothesis testing. Lowering the alpha level decreases the chances of making a Type I error.

A conceptual diagram of the Type I error probability α



Example:

 # sample size

```
n = 10

# standard deviation

sigma = 3

# significance level

alpha = 0.03

typeI.tester(mu0=0,sigma=1,n=40,alpha=0.05)
```

Output:

```
[1] 0.0489
```

Type II Error (False Negative):

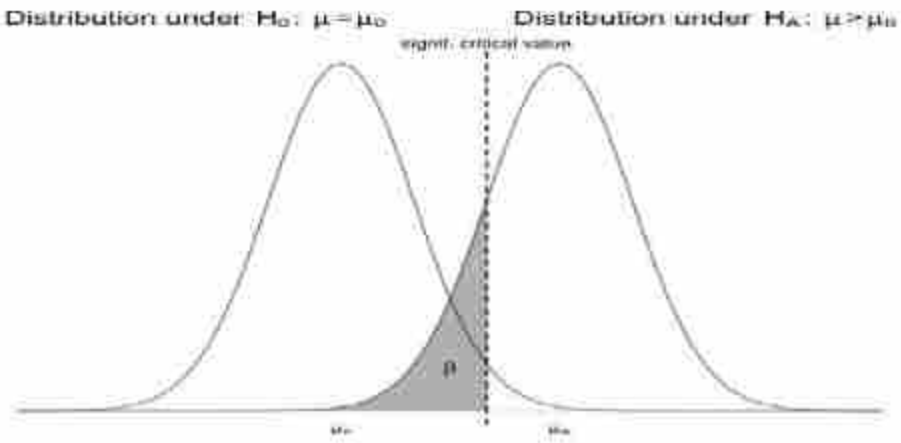
Definition: Type II error occurs when you fail to reject a false null hypothesis. It's the probability of failing to reject a false null hypothesis.

- 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.

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.

A conceptual diagram of the Type II error probability β



Example:

```
# sample size

n = 10

# standard deviation

sigma = 3

# significance level

alpha = 0.03

typeII.test(mu0, TRUEmu, sigma, n, alpha, iterations = 10000)
```

Output:

```
 [1] 3e-04
```

Power or Statistical Power

Power or Statistical power of a hypothesis test is the probability of the test correctly rejects the null hypothesis

★ It is only useful when the null hypothesis is rejected

★ The Statistical power refers to the probability that a Statistical test will correctly reject a false null hypothesis

(★) For any hypothesis test It is useful to consider its potential statistical power.

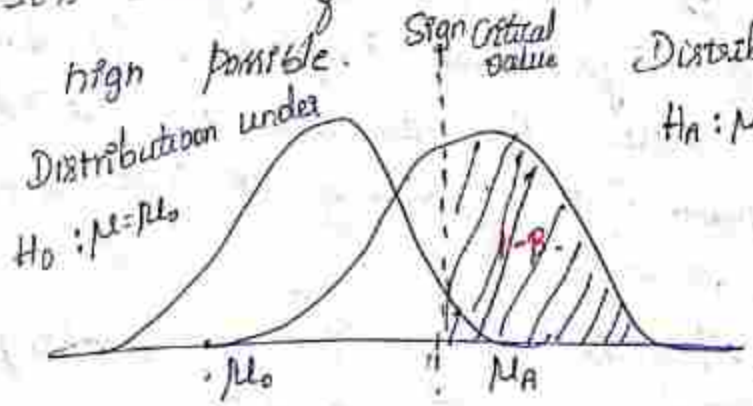
(★) power is the "probability" of correctly rejecting a null hypotheses that is untrue.

Applications / Uses

1. Experimental design :- Determining the sample size needed to detect a specific effect
2. Hypothesis Testing
3. Quality control
4. Decision Making
5. Effect Size Estimation

(★) To test type II Errors, state of β. The Statistical power is found simply. $= 1 - \beta$.

(★) It's desirable for a test to have a power that's as high possible.



A Conceptual diagram of Statistical power $1-\beta$

To Calculate statistical power..

The formula for calculating statistical power depends on various factors including 1. Effect size, 2. level of significance & 3. Sample size Sometimes variability or standard deviation in the data.

(*) The Formula typically involves elements from the specific statistical test being used as Z-test, t-test ANOVA, chi-square test

The General formula for statistical power in the context of hypothesis testing

Statistical power = 1 - β. (2)

Here. Statistical power :- The probability of correctly rejecting a false null hypothesis (i.e ability to detect a true effect)

β (Beta) :- The probability of Type II Error which failing to reject a false null hypothesis (i.e not detecting a true effect

1. A researcher is conducting a study to investigate whether a new teaching method improves student's scores on a standardized test. The researcher is interested in detecting a mean score increase of at least 5 point

To test statistical power

Solutions :- Null hypothesis (H_0) :- The new teaching Method has no effect on test scores ($\mu = 0$)

Alternative hypothesis (H_A) :- The new teaching method increase test scores by at least 5 points ($\mu \geq 5$. a one sided test)

Step 3 :- Effect size (μ_A) : 5 points

Step 4 :- Standard deviation of test scores. Assumed to be 8 points

Step 5 :- Significance level (α) :- 0.05 (3)

Step 6 :- Desired statistical power :- 0.90

Step 7 :- Using the formula for sample size in a one-sided t-test

$$n = \frac{(2_\alpha + 2_\beta)^2 \times \delta^2}{\mu_A^2}$$ [Sample size Calculation]

1) 2_α is the 2-score corresponding to the chosen significance level (0.05)

2) 2_β is the 2-score corresponding to the desired power (0.90)

3) δ is the standard deviation

4) μ_A is the desired effect size.

Let assume $2_\alpha = 1.645$ (From a std. normal dist table for one-sided test $\alpha = 0.05$)

$2_\beta = -1.28$ (From a std normal dist table for 0.90 power)

$$n = \frac{(1.645) - (-1.28)^2 \times 8^2}{5^2}$$

$=$

★ If the achieved statistical power is greater than desired power (0.90) the test is considered statistically powerful.

★ The researcher can conclude that there is high probability 90% that the test correctly identified a significant increase in test scores due to the new teaching Method

'R' pgm

```
library (pwr)
effect size <- 5
sd. score <- 8
alpha <- 0.05
desired power <- 0.90
Sample size <- pwr.t.test (d = effect size / sd. scores,
  sig.level = alpha, power = desired power, type = "one. sample") $n
cat ("Required Sample size :", Round (Sample size))
```

pwr.t.test

One Sample:-

Power, which is the probability of rejecting a false null hypothesis is called as (1-β) (also Expressed) 1-Type II Error probability ;- for a Type II two error 0.15 power

power = 1 - β = 1 - 0.15 = 0.85

```
library (pwr)
effect size <- 0.5
alpha <- 0.05
power <- 0.85
Sample size <- pwr.t.test (d = effect size, sig.level = alpha
  power = power, type = "two. sample") $n
print (Sample size)
```

Analysis of variance [ANOVA]

Analysis of variance (ANOVA):- it's a statistical technique that is used to check if the means of two (or) more groups are significantly different from each other.

It is Developed by Ronald Fisher in 1918.

(A) Compare checks the two/more groups.

(*) Extension model of t-test and z-test (5)

(*) It is also called as Fisher analysis of variance.

T-test - Two mean only but doesn't two/more

(*) ANOVA it's the simple form of compare multiple mean in a test for equivalence. (@) Hypothesis test of Comparing two means

Applications :-

1. Experimental Research 3. Business and Economics
2. Industrial Application 4. Environmental Studies.

Types of Anova

1. One-way Anova
2. Two-way Anova.

One-way Anova

One-way Anova is compared more than two group based "one" factor.

[one independent variable]

A manufacturing company compare productd two/more Employee based and "working hours".

Two-way Anova

Two way Anova means used to Compared more than two groups based "Two" factor.

[Two-Independent variable]

A manufactur company compare Productivity two/more Employee & working hours working condition

One-way Analysis variance.

1. The simplest version of ANOVA is referred as one-way or one-factor Analysis
2. The one-way one is used to test two or more means for Equality

One-way ANOVA

Source of variance	df	Sum of Squares	Mean Square	F-value
Between Groups	K-1	SSB	MSB	F
Within Groups	N-k	SSW	MSW	
✓ Total	N-1	SST		(6)

1. Calculate the Grand mean ($\bar{x}_{Grand}$):

$$\bar{x}_{grand} = \frac{\sum \text{All Scores}}{N}$$

2. Calculate SST :- $\frac{\sum \text{All Scores}}{N}$ $\sum (x - \bar{x}_{Grand})^2$

3. Calculate SSB :- $\sum n_i (\bar{x}_i - \bar{x}_{Grand})^2$

4. Calculate SSW :- $SSW = \sum (x - \bar{x}_i)^2$

5. Degrees of Freedom

Between Groups : $df_{Between} = k - 1$

Within Groups :- $df_{within} = N - k$

6. Calculate Mean Squares

1. $MSB = \frac{SSB}{df_{Between}}$

2. $MSW = \frac{SSW}{df_{within}}$

7. Calculate F-value

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

Two-way ANOVA Table construction

(*) The two-way ANOVA is the Multiple factor ANOVA rather than one way Anova

(*) Two way ANOVA there are independents

Allows a company to compare workers productivity based on two independent on two variable

Source of variance	Sum of Squares (SS)	Degrees of freedom (df)	mean Square (MS)	F departments value gender
Factor A	SSA	df_A	MSA	F_A
Factor B	SSB	df_B	MSB	F_B
Interaction (AXB)	SSAB	df_{AXB}	MSAB	F_{AXB} (7)
Residual/Error	SSE	df_E	MSE	
Total	SST	df_T		✓

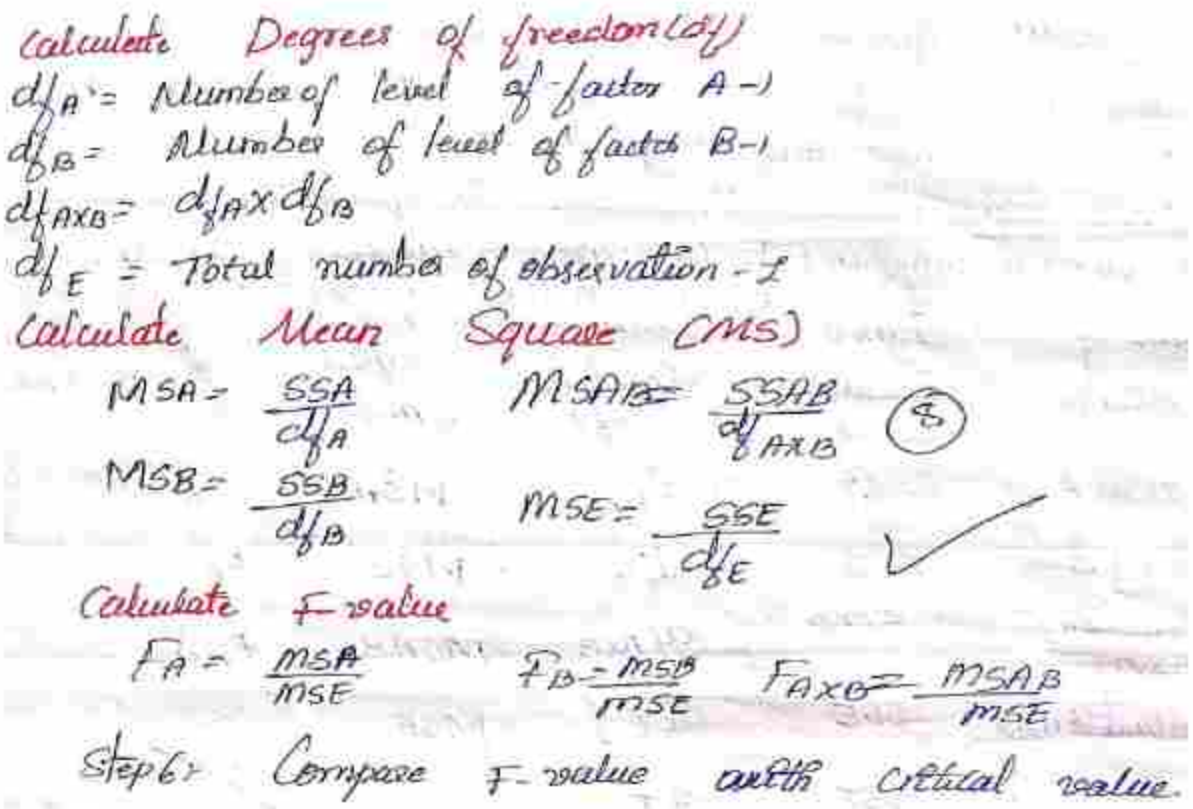
Grand mean = $\frac{\text{Sum of all data points}}{\text{Total number of data points}}$

1) $SSA = \frac{\sum T_i^2}{n_B} - \text{Grand Mean}^2$ (calculate Sum of Squares (SS) for each factors)

2) $SSB = \frac{\sum T_i^2}{n_B} - \text{Grand Mean}^2$

3) $SSE = \sum\sum (X_{ij} - \text{Grand mean})^2$

4) $SSAB = SST - SSA - SSB - SSE$



ONE WAY ANOVA IN IN R PROGRAMMING

A one-way ANOVA, or one-way analysis of variance, is a statistical technique used to compare the means of three or more groups to determine if there are statistically significant differences among these group means.

```
# Generate example data with three groups

group1 <- c(25, 30, 35, 40, 45)

group2 <- c(20, 28, 32, 38, 42)

group3 <- c(22, 29, 34, 39, 44)

# Combine the data into a data frame

data <- data.frame(value = c(group1, group2, group3),

group = rep(c("Group1", "Group2", "Group3"), each = 5))

# Perform one-way ANOVA

anova_result <- aov(value ~ group, data = data)
```

```
# Summarize the ANOVA results

summary(anova_result)
```

OUTPUT:

```
Df Sum Sq Mean Sq F value Pr(>F)
group        2   22.5   11.27   0.161  0.853
Residuals   12  839.2   69.93
```

TWO WAY ANOVA IN IN R PROGRAMMING

A two-way ANOVA, also known as a two-factor ANOVA, is a statistical method used to an alyze the influence of two categorical independent variables (factors) on a continuous depen dent variable. This analysis simultaneously assesses the effects of these two factors, as well as their interaction effect, on the response variable.

```
# Generate example data for a two-way ANOVA
group1 <- c(25, 30, 35, 40, 45)
group2 <- c(20, 28, 32, 38, 42)
group3 <- c(22, 29, 34, 39, 44)

# Creating a second factor
factor2 <- rep(c("FactorA", "FactorB"), each = 15)

# Combine the data into a data frame
data <- data.frame(
  value = c(group1, group2, group3),
  group = rep(c("Group1", "Group2", "Group3"), each = 5),
  factor2 = factor2
)

# Perform two-way ANOVA
anova_result <- aov(value ~ group * factor2, data = data)

# Summarize the ANOVA results
summary(anova_result)
```

OUTPUT:

```
Df Sum Sq Mean Sq F value Pr(>F)
```

group 2 45.1 22.53 0.322 0.728

factor2 1 0.0 0.00 0.000 1.000

group:factor2 2 0.0 0.00 0.000 1.000

Residuals 24 1678.4 69.93

MODULE 4 QUESTIONS:

2 Marks Questions:

1.What is sampling distributions.

2.Define Hypothesis test.

3.Explain errors and power.

4.What are all the components of hypothesis test.

5.Define Analysis if variances(ANOVA).

3 Marks Questions:

1.explain sampling distribution. Distribution for a sample mean.

2.what are all the confidence intervals.

3.Explain testing proportions.

4.Difference between single categorical value and Two categorical values.

5.Explain one way ANOVA

5 Marks Questions:

1.discuss sampling distribution in detail.

2.Explain types testing means testing.

3.Explain type I and type II errors.

4.Explain statistical power and simulating power, power curve.

5.Explain One way ANOVA with example.

10 Marks Questions:

1.Briefly explain components of hypothesis test.

2.Explain Testing proportions with example.

3. write a R program using components of hypothesis test, testing two means.

4.Explainn tesing categorical variables in detail

5.Discuss Two way ANOVA in detail