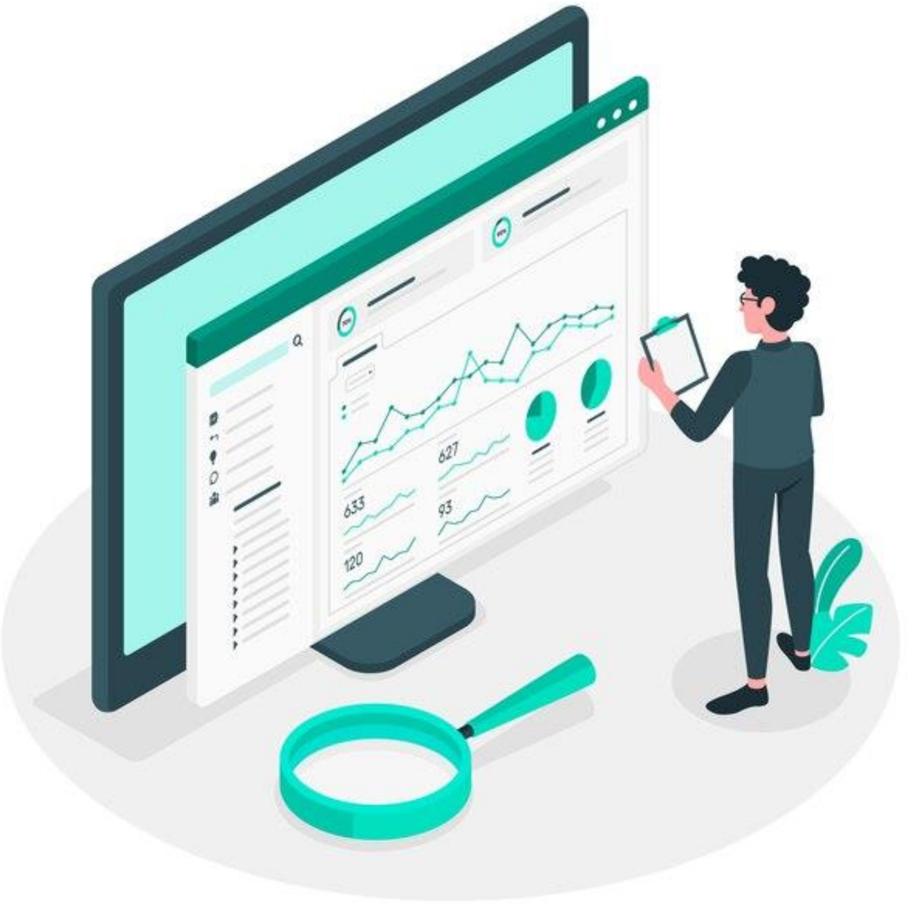
Data Analytics with R

Hypothesis Testing

Business Scenario

- A store is planning to buy the EAS systems of a company. The company claims that no more than 5% of all the consumers would say that they would never shop in a store again if the store subjected them to a false alarm.
- The store hires Archie’s company to study the claim and check the validity of this claim.

Approach: The claim made by the EAS system selling company is an assumption that needs to be tested. The claim can be tested by constructing a hypothesis test.
Archie needs to get a complete understanding of how to construct a hypothesis test and should provide the final conclusion.



Learning Objectives

By the end of this lesson, you will be able to:

- Define hypothesis and develop null and alternate hypothesis
- Describe type 1 and type 2 errors in hypothesis testing
- Perform one-sample and two-sample hypothesis tests
- Perform ANOVA test
- Perform chi-square test of independence

Null and Alternate Hypothesis

Hypothesis Testing

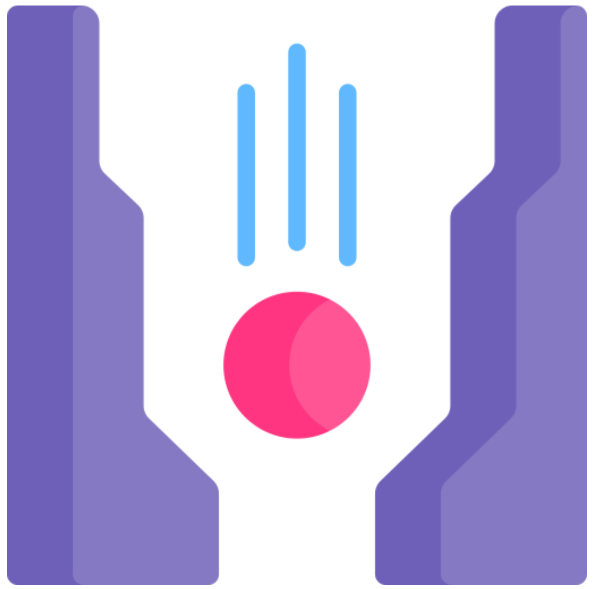
A hypothesis is a potential explanation for something that happens or is observed and thought to be true. Generally, business hypotheses are educated assumptions.

Hypothesis testing is a method of statistical inference to assess whether the statement (called a hypothesis) made about the population is consistent with the observed data (sample).

Hypothesis Testing: Scientific Approach

In Science, we cannot prove something to be TRUE.

For example, we cannot prove that things fall when dropped. However, we can prove that a ball will NOT “NOT FALL” when dropped.

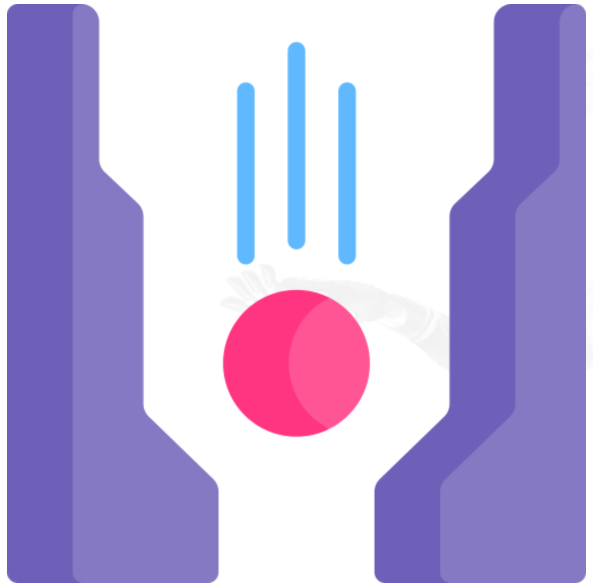


Science provides evidence that something is not true.

Hypothesis Testing: Intuition

It cannot be proved that a ball will always fall down.

However, the claim that a ball will not fall and will stay at the same place if it is dropped can be disproved by actually dropping the ball to provide evidence for this. This event can be repeated to provide even more evidence.



Forming Hypothesis: Null Hypothesis

The first step in a hypothesis test is to formalize it by specifying the null hypothesis.

It is a statement about the value of population parameter that holds true unless there is sufficient sample evidence to conclude otherwise.

It is represented by H_0. It is formed by expressing a condition on population parameters that would lead to the status quo.

Alternate Hypothesis or Research Hypothesis

It is denoted by H_a and is the negation of the null hypothesis.

It is formed by expressing a condition on population parameters that would lead to change.

Note:

H_0 and H_a are mutually exclusive and collectively exhaustive statements.

Example: Setting Up H_0 and H_a

A manufacturer of golf balls claims that the variance in the weights of the balls is controlled to be within 0.0028 oz^2

How to set up the H_0 and H_a to test this claim?

H_0: variance <= .0028 oz^2

H_a: variance > .0028 oz^2

Type 1 and Type 2 Errors

Error in Hypothesis Testing

Court of law always assumes the accused to be INNOCENT until PROVEN GUILTY.

H_0: Accused = Innocent

H_a: Accused = Guilty

Prosecution must provide enough evidence to reject innocence and conclude guilty.

		VERDICT	
		NOT GUILTY	GUILTY
TRUTH	INNOCENT	Correct	ERROR
	GUILTY	ERROR	Correct

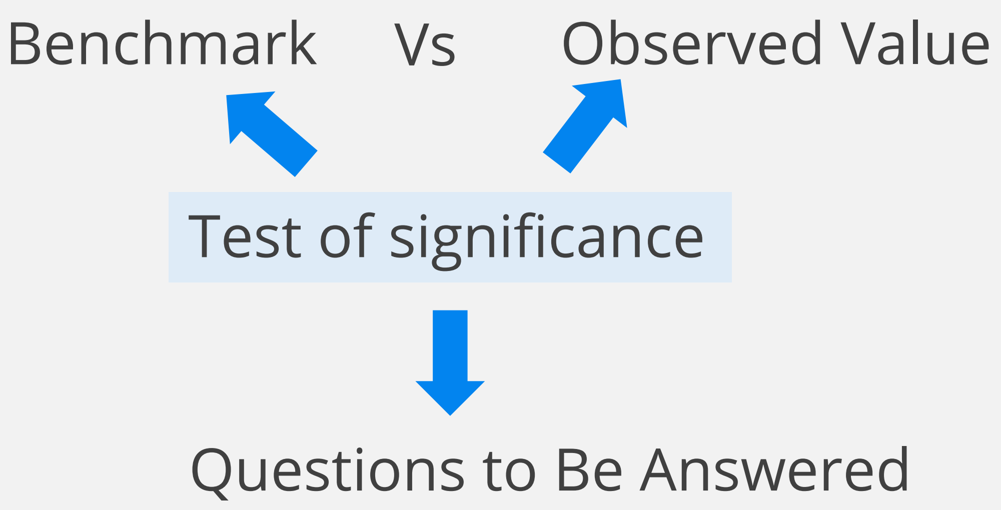
Statistical Significance

Significance level (α) is the probability of rejecting the null hypothesis when it is true.

1. It is the probability of Type 1 error. The standard value of α are 10%, 5%, and 1%. It is the risk of false positive results.
2. Probability of Type 2 error is denoted by β. It is the risk of false negative results.
3. 1- β is called as the power of test.

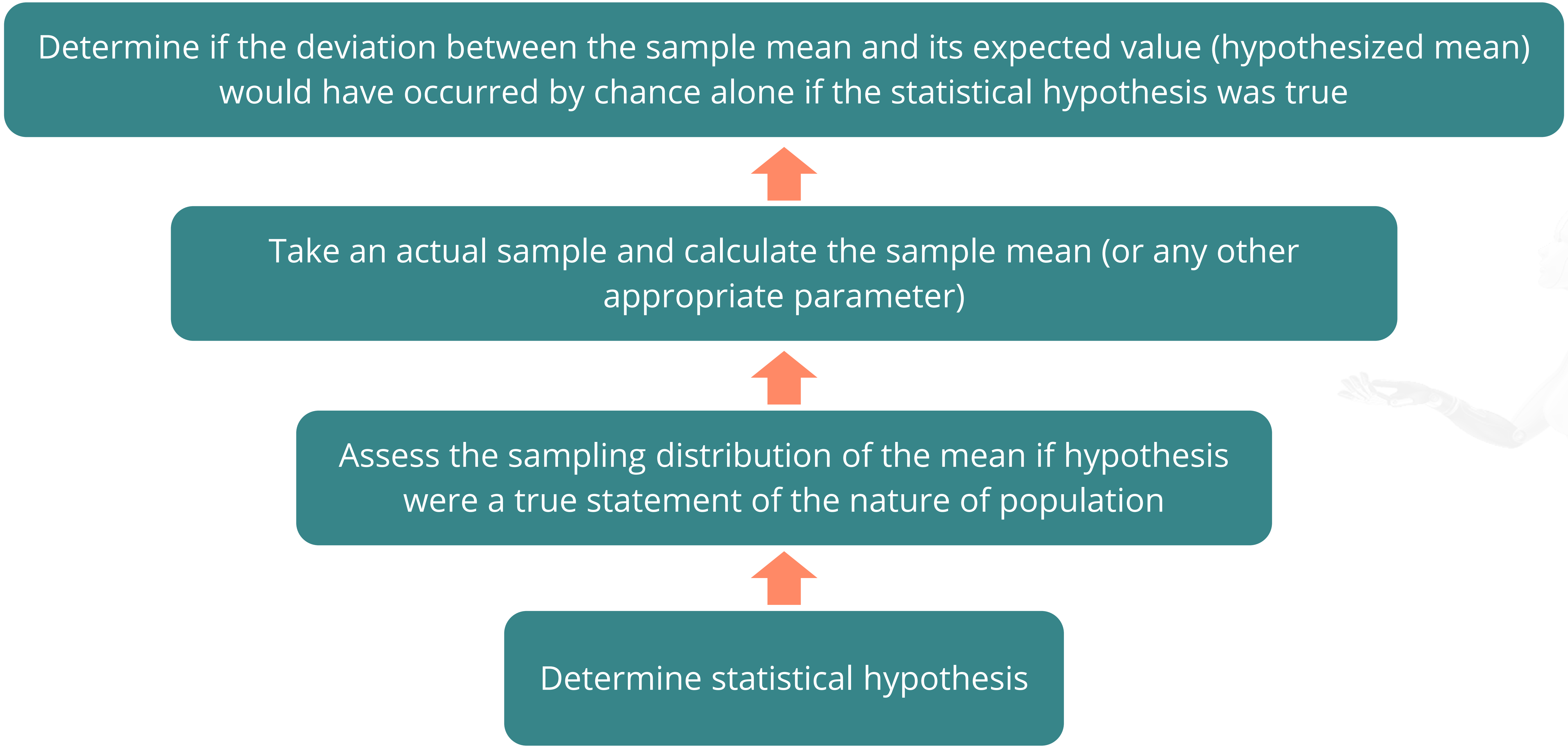
		DECISION	
		Don’t reject null	Reject null
STATE OF NATURE	H_o True	Correct Decision	TYPE 1 ERROR
	H_o False	TYPE 2 ERROR	Correct Decision

Typical Hypothesis Test

Situation	Evidence of phenomenon or behavior
A hardware franchise has almost 1,000 retail outlets. Monthly sales of a particular tool averaged 612 units for each outlet. The sales manager believed that a competitor's new price on a similar item may have an impact on sales. Based on a random sample of 64 outlets, the sales manager wishes to test if the observed sample mean differs from the benchmark figure.	 Benchmark Vs Observed Value Test of significance Questions to Be Answered 1. Is the difference between the benchmark and the observed values statistically significant? 2. Does the magnitude of the increase (or decrease) in the phenomenon justify a change in business strategy?

Hypothesis Test Procedure

Hypothesis Testing Procedure



Determine statistical hypothesis

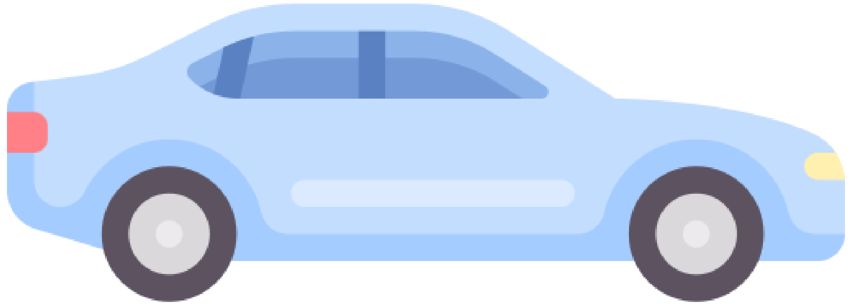
Assess the sampling distribution of the mean if hypothesis were a true statement of the nature of population

Take an actual sample and calculate the sample mean (or any other appropriate parameter)

Determine if the deviation between the sample mean and its expected value (hypothesized mean) would have occurred by chance alone if the statistical hypothesis was true

The Audi R-18 e-Tron Quattro Case

Audi R-18 e-Tron Quattro is equipped with a top-class brake system with astonishing braking distance.



Braking distance is the distance required to bring the vehicle to a complete stop from a speed of 60mph.

(Imagine) One of the competitors has advertised to achieve an average braking distance of 20 m. In the new ad, Audi would like to claim that Audi R-18 e-Tron Quattro achieves a better braking distance.

The Audi R-18 e-Tron Quattro Case

According to the protocols of the broadcasting company, this advertisement can be broadcasted only if Audi convinces them that its average braking distance is less than 20 m.

If a random sample of 70 Audi R-18s has an average stopping distance of 19.5 m, will National Motors advertise the claim?

Population standard deviation is 1.5 m.

We can perform a hypothesis test to test this claim.

Solution Steps: Decide H_0, H_a, and α

Step 1: Set up H_0 and H_a

H_0: μ (Average braking distance) >= 20
H_a: μ (Average braking distance) < 20

Step 2: Ascertain significance level (α)

- Use standard value or establish it based on business requirements.
- Take 5%

Calculate Test Statistic

Step 3: Utilize test statistic

$$\text{test statistic} = \frac{\bar{X}-\mu}{\sigma_{\bar{x}}}$$

$z = {}^{(19.5-20)}/_{1.5/\sqrt{70}} = -2.78$

- The test statistic measures the distance between sample mean ($\bar{X}$) and hypothesized mean (μ).
- Division by $\sigma_{\bar{X}}$ says that this distance is measured in the units of the standard deviation of all possible means.

Note:

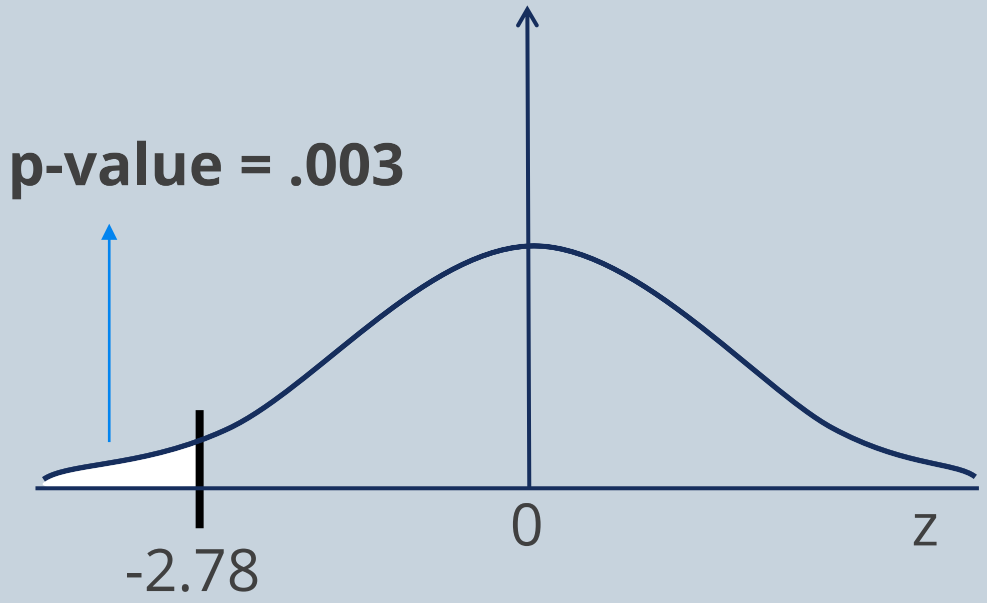
- If population standard deviation (σ) is given, test statistic = z and $\sigma_{\bar{X}} = \sigma/\sqrt{n}$.
- If population standard deviation is not given, test statistic = t and $\sigma_{\bar{X}}$ = s/$\sqrt{n}$, where s is the standard deviation of the sample used to test the hypothesis assuming normal population distribution.

p-value in a Hypothesis Test

Find p-value

The farther the value of the test statistic is below 0 (or the farther $\bar{x}$ is below 20), the stronger is the evidence to support the rejection of H_0 in favour of H_a.

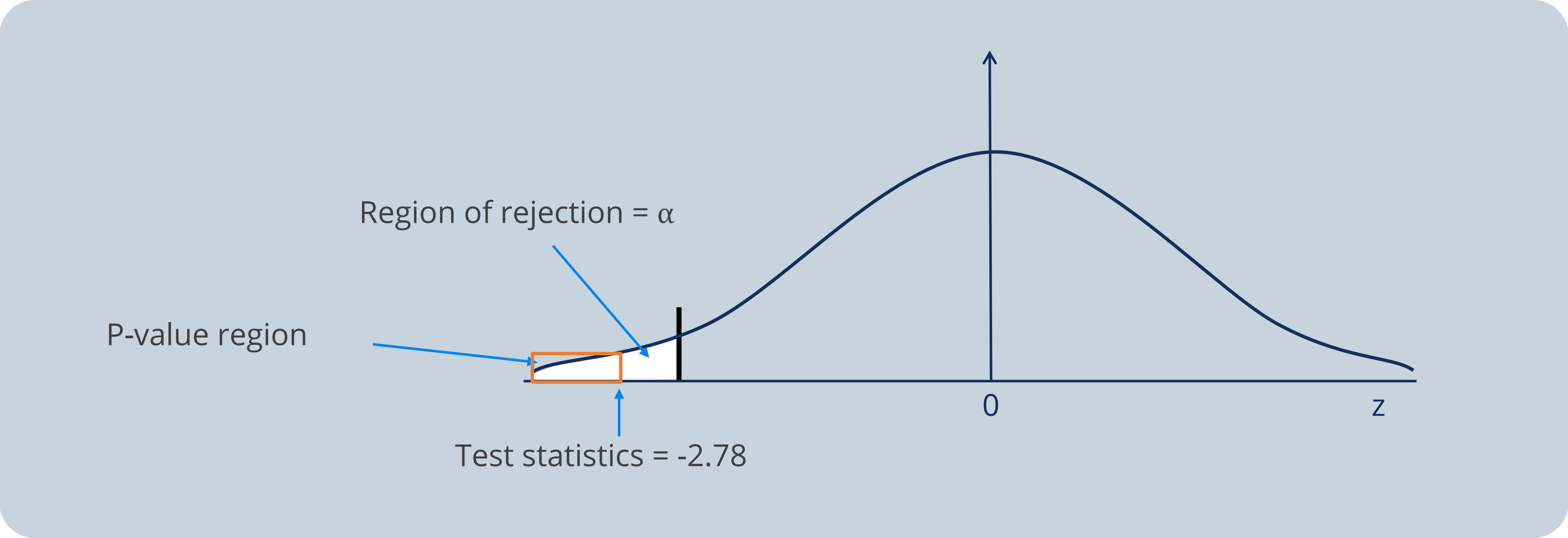
To see how small z must be in order to reject $H_{0,}$ we use the p-value approach. It is calculated using test statistic.



p-value is a probability that provides a measure of the evidence against the null hypothesis provided by the sample.

Comparing p-value with Significance Level

Since test statistic z falls into rejection region or p-value < α, reject the null hypothesis.



Note:

This is an example of one sample left tail hypothesis test.

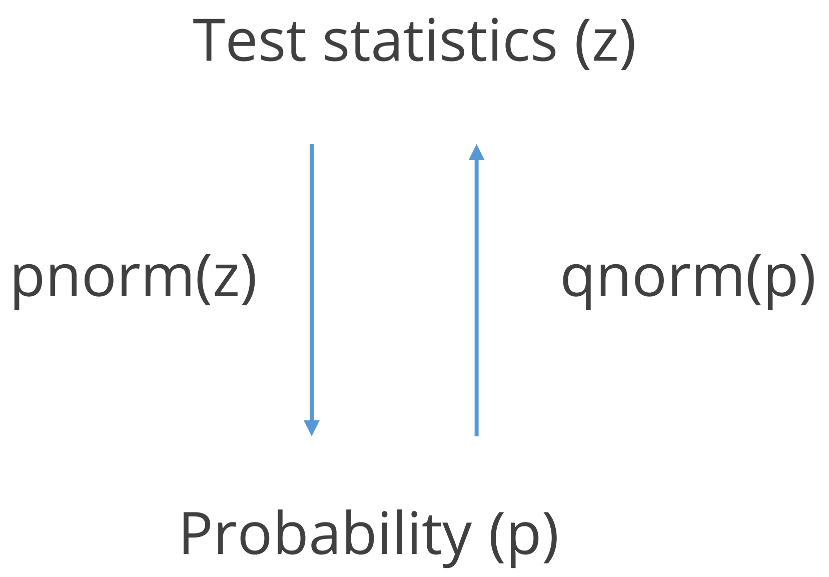
p-value Computation in R

For estimation of probability (area) value corresponding to a z test statistics, pnorm() function is used.

For estimation of z statistics corresponding to a probability (area) value, qnorm() function is used.

Syntax

```
pnorm(q, lower.tail = TRUE)
```



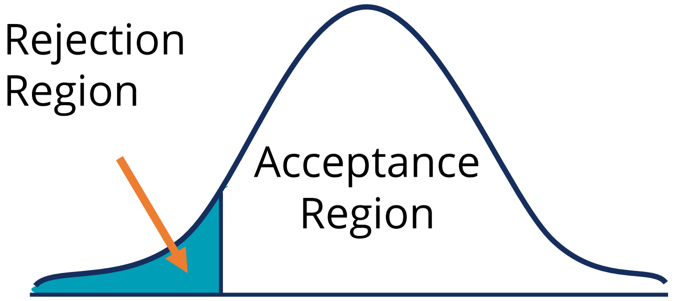
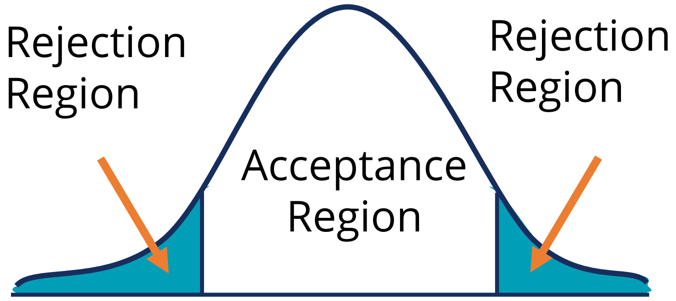
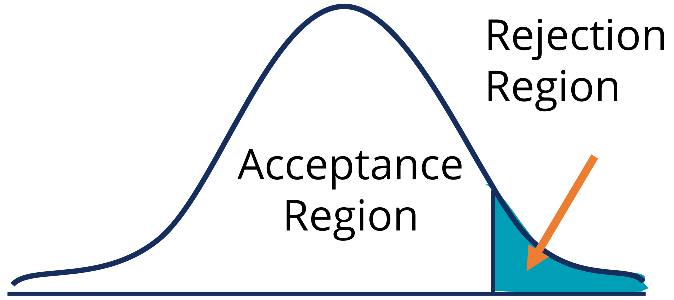
Syntax

```
qnorm(p, lower.tail = TRUE)
```

One-Sample Hypothesis Test

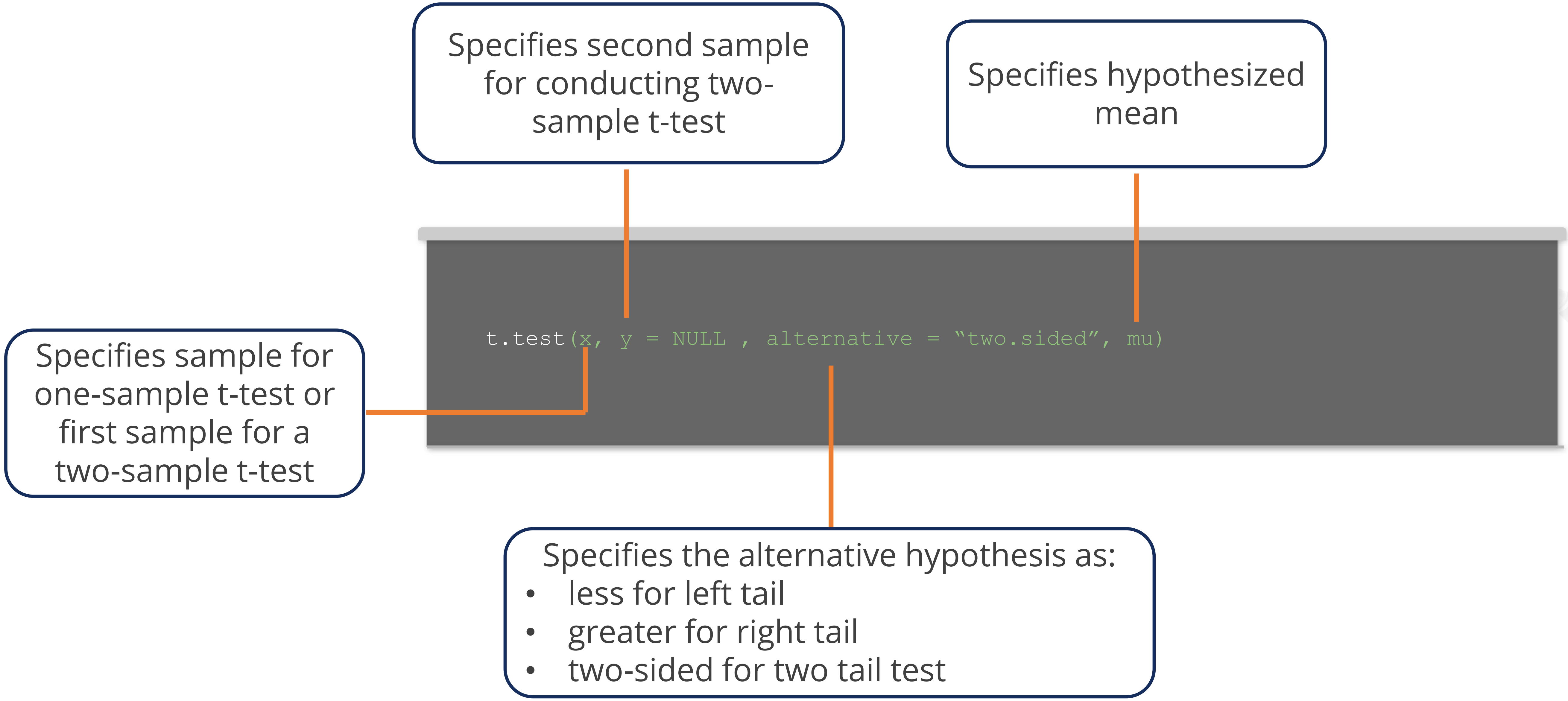
One Sample Hypothesis Testing

A statistical hypothesis test is conducted to determine whether an unknown population mean is different from a specific value.

One-Tailed Test (Left Tail)	Two-Tailed Test	One-Tailed Test (Right Tail)
$H_0 : \mu_x = \mu_0$ $H_1 : \mu_x < \mu_0$	$H_0 : \mu_x = \mu_0$ $H_1 : \mu_x \neq \mu_0$	$H_0 : \mu_x = \mu_0$ $H_1 : \mu_x > \mu_0$
		

T-Test in R

To perform t-test in R, t.test() function is used.



```
t.test(x, y = NULL , alternative = “two.sided”, mu)
```

Specifies second sample for conducting two-sample t-test

Specifies hypothesized mean

Specifies sample for one-sample t-test or first sample for a two-sample t-test

Specifies the alternative hypothesis as:

- less for left tail
- greater for right tail
- two-sided for two tail test

One Sample Hypothesis Testing

Duration: 3 minutes

Problem Scenario: For a particular location, it is advertised that residential property is available at an average cost of $250,000 or less per lot.

Dave is willing to invest into a property in the locality and hence wants to test the said claim. He gathers a sample data of 40 houses for validating the claim made in the advertisement.

Perform a statistical significance test at α = .05 to help Dave validate the claim.

Note: Please download the data set and the solution document from the **Course Resources** section and follow the steps given in the document

ASSISTED PRACTICE

Two-Sample Hypothesis Test

Comparing Two Populations

Independent samples t-test	Paired samples t-test
Independent random samples are collected	Samples collected are paired in some way
Number of datapoints may be different in the two samples	Number of datapoints are same
Examples: • Comparing performance of girls and boys • Comparing per capita income of two regions	Examples: • Testing efficacy of a medicine • Assessing the effectiveness of a marketing campaign by launching a survey before and after the marketing campaign

Assumptions

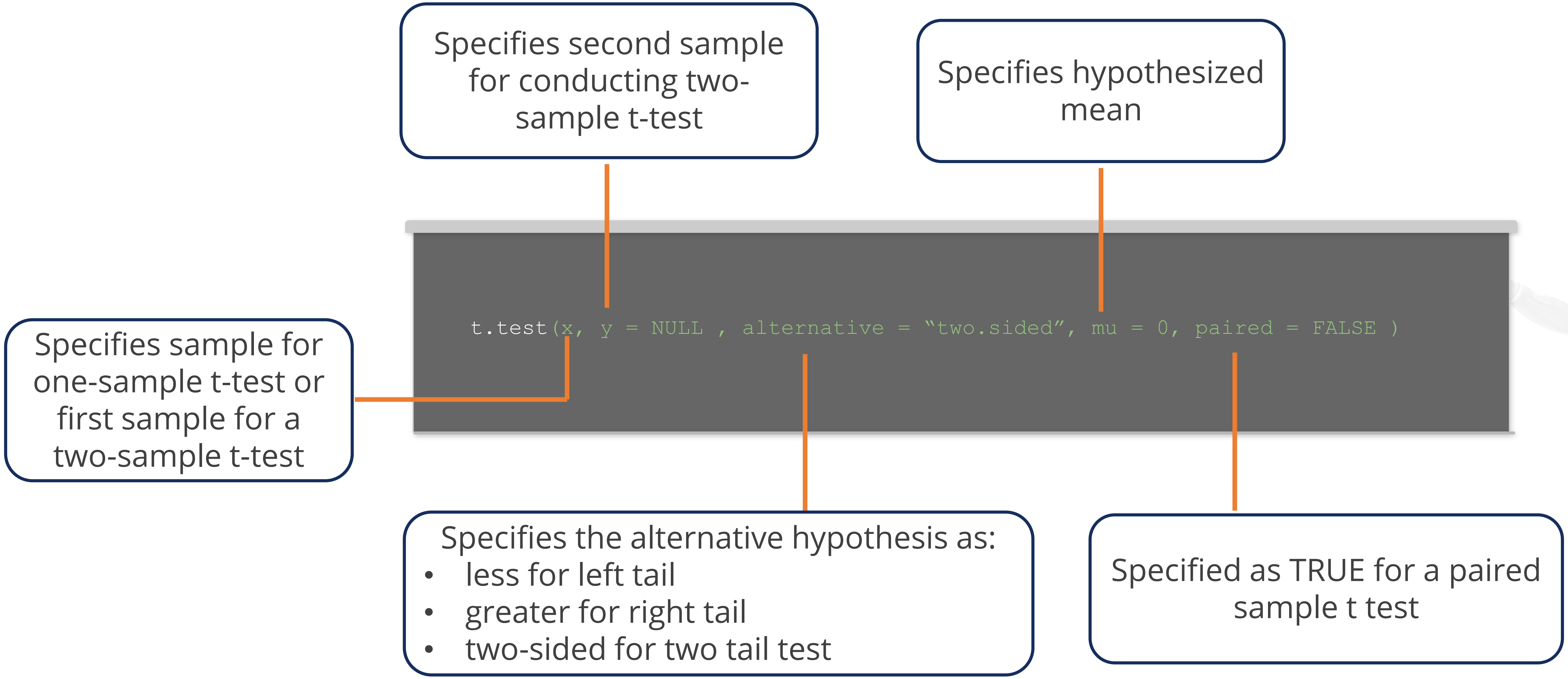
1. The data is continuous (not discrete).

2. Datapoints from the two groups follow a normal distribution.

3. The variances of the two populations are equal. If not, the Welch Unequal-Variance test is used.

T-Test in R

To perform t-test in R, t.test() function is used.



Specifies second sample for conducting two-sample t-test

Specifies hypothesized mean

```
t.test(x, y = NULL , alternative = “two.sided”, mu = 0, paired = FALSE )
```

Specifies sample for one-sample t-test or first sample for a two-sample t-test

Specifies the alternative hypothesis as:

- less for left tail
- greater for right tail
- two-sided for two tail test

Specified as TRUE for a paired sample t test

Comparing Two Populations

Duration: 3 minutes

Problem Scenario: A study claimed that boys scored better or the same as girls in the Mathematics section of competitive tests. A researcher wanted to check the validity of this claim.

The researcher collected a random sample of 80 female test takers and 120 male test takers to disprove the claim made in the study.

Construct a hypothesis test to help the researcher validate the hypothesis at a significance level of 0.05.

Note: Please download the data set and the solution document from the **Course Resources** section and follow the steps given in the document

Analysis of Variance (ANOVA)

Comparing Three or More Populations

An oil company wishes to develop a reasonably priced gasoline that will deliver improved mileage. The company prepares three different formulations of gasoline for use.

Analysis of variance is used to compare the effects of the three types of gasoline based on mileage in order to find the gasoline that delivers the highest average mileage.

ANOVA: Analysis of Variance

Hypothesis test of analysis of variance:

H_0: $\mu 1 = \mu 2 = \mu 3 = \ldots \ldots$

H_a: Not all μ are equal

Assumptions:

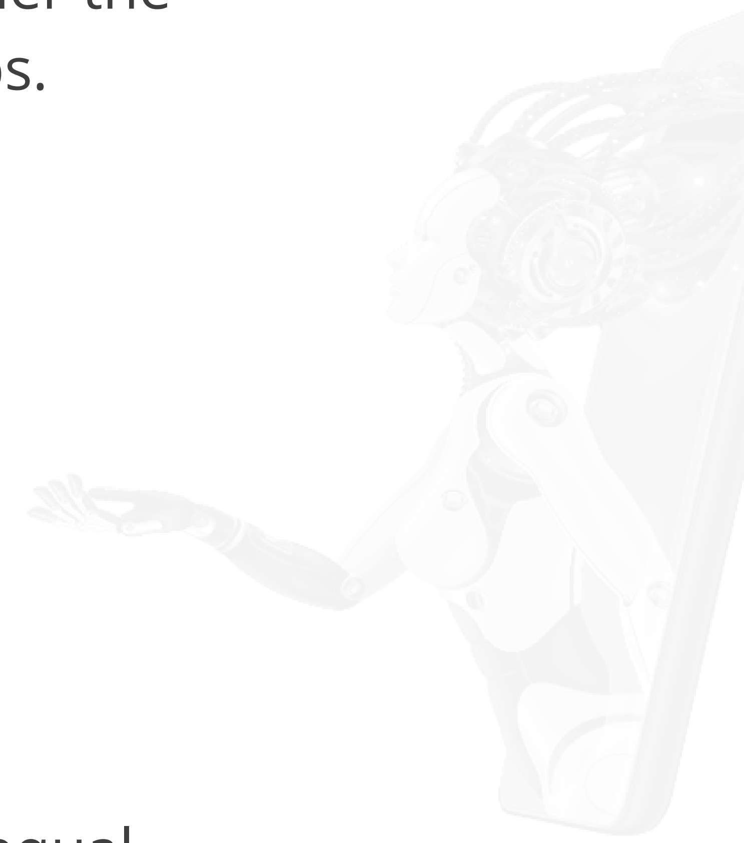
1. Groups have homogeneity of variance among the groups.
2. Populations maintain independent random sampling.
3. Population has normal distribution.

F-Statistic for ANOVA

F-statistic is a ratio of two variances.

In ANOVA, F-statistic is used to test statistics. F-statistic helps to determine whether the variance between the group means is larger than the variance within the groups.

$$F = \frac{Variance\ between\ sample\ means}{Variance\ within\ the\ samples}$$



If the F-statistic is sufficiently large, it means that not all group means are equal.

Steps for One-way ANOVA

1. State null hypothesis (H_0) and alternate hypothesis (H_a)

2. State significance level (α)

3. Calculate the Grand Mean: $\bar{\bar{x}} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{n_T}$, $n_T = n_1 + n_2 + n_3 + n_4 \ldots + n_k$

   k is the number of groups and n represents number of data points

4. Calculate Group Means: $\bar{x_j} = \sum_{i=1}^{n_j} x_i$

   n_j is number of samples in group j

simplilearn

Steps for One-way ANOVA

5 Calculate $SS_{between}$: Sum of squares of deviations between groups: $SS_{between} = \sum_{j=1}^{k} n_j\left(\bar{x}_j - \bar{\bar{x}}\right)^2$

6 Calculate SS_{within}: Sum of squares of deviations within groups: $SS_{within} = \sum_{j=1}^{k}(n_j - 1){s_j}^2$

7 Calculate degrees of freedom for:

a. Between Group: k - 1

b. Within Group: n - k

8 Calculate

a. Mean sum of squares of deviation between: $MS_{Between} = \frac{SS_{between}}{df_{between}}$

b. Mean sum of squares of deviation within: $MS_{Within} = \frac{SS_{within}}{df_{within}}$

Steps for One-way ANOVA

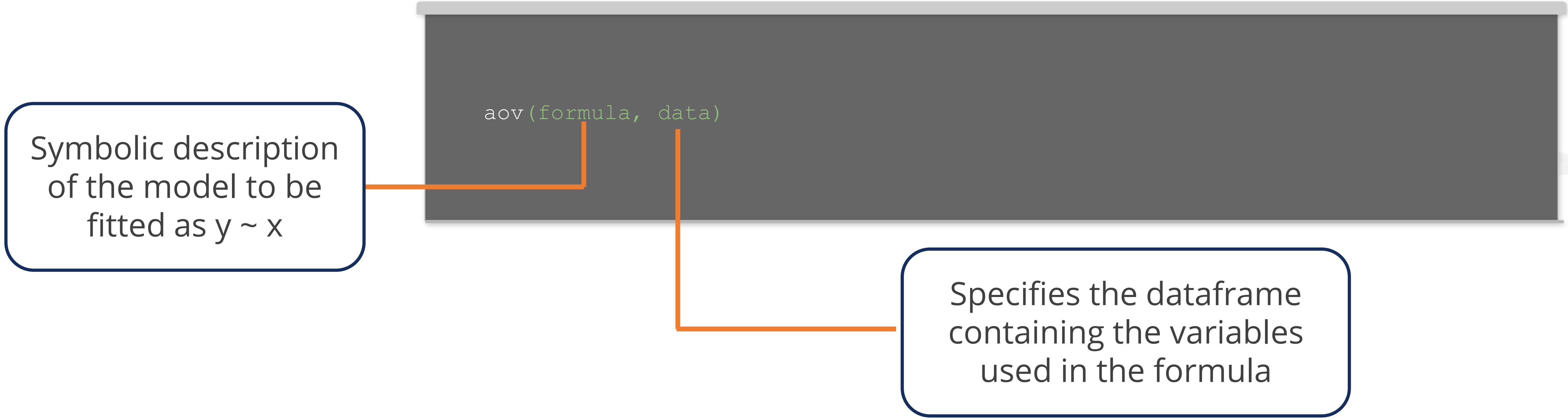
9 Calculate F-statistics: $F = \frac{MS_{between}}{MS_{within}}$

10 Get p-value using degrees of freedom k-1 and n-k

11 Compare p-value with significance level to conclude

ANOVA in R

To perform ANOVA in R, aov() function is used.



```
aov(formula, data)
```

Symbolic description of the model to be fitted as y ~ x

Specifies the dataframe containing the variables used in the formula

ANOVA

Duration: 3 minutes

ASSISTED PRACTICE

Problem Scenario: Three different assembly methods have been proposed for a new product. The plant operations head wants to find out which of these methods produces the most units per hour.

The manager selected 150 workers and assigned them to test the three methods.

Construct a test of significance at a significance level of 0.05.

Note: Please download the data set and the solution document from the **Course Resources** section and follow the steps given in the document

Non-Parametric Tests

Parametric- vs. Non-parametric Tests

Parametric	Nonparametric
Assumptions are made about the population distribution	Does not require any assumption about specific population distribution
Uses a mean value for central tendency	Does not deal with specific parameters like mean, variance, or standard deviation; uses median for central tendency
Deals with interval and ratio type of data	Deals with enumerative data (frequency counts)

Nonparametric Test and Their Parametric Equivalent

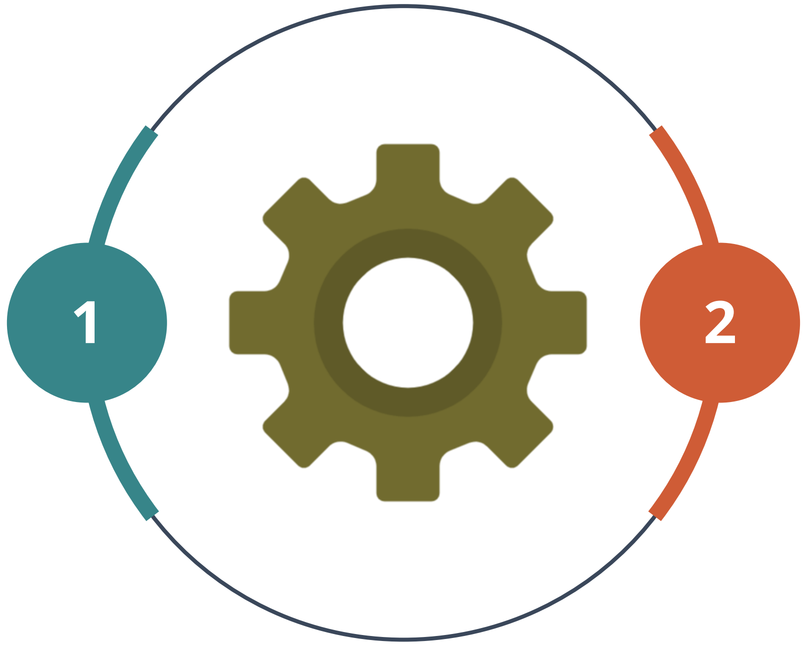
Nonparametric test	Parametric equivalent
Mann-Whitney U test	Independent t-test
Wilcoxon Signed-Rank test	Paired t-test
Kruskal-Wallis H test	One-way ANOVA
Spearman rank correlation coefficient	Pearson correlation coefficient

Note:

The chi-square test is the most used nonparametric test.

Chi-Square Tests

There are two types of Chi-square tests:



1. Chi-square test for goodness of fit

2. Chi-square test for independence of two variables

Chi-square test statistics:

$$\chi^2 = \sum \frac{(Observed\ Frequency\ - Expected\ Frequency)^2}{Expected\ Frequency}$$

Chi-Square Test for Independence

- A city has a newly opened nuclear plant, and there are families staying close to the plant. A health and safety officer wants to take this case up to provide relocation for the families that live in the surrounding area.
- To make a strong case, they want to prove with data that an exposure to radiation levels is leading to an increase in the number of people who live around the plant getting ill. He formulates a contingency table of exposure and disease. This table contains observed frequencies.
- Does the data suggest an association between the disease and exposure?

Contingency table	**Disease**		**Total**
Exposure	Yes	No	
Yes	37	13	50
No	17	53	70
Total	54	66	120

Steps For Chi-Square Test for Independence

1 State the null and alternative hypotheses

H_0: The two categorical variables are independent
H_a: The two categorical variables are not independent

2 Select a random sample and record the observed frequencies for each cell of the contingency table

3 Compute the expected frequency for each cell

Expected cell frequency = $\frac{(row\ total) * (column\ total)}{(grand\ total\ of\ all\ cells)}$

Steps For Chi-Square Test for Independence

4 Compute the value of the test statistic

$$\chi^2 = \sum \frac{(observed\ frequency - expected\ frequency)^2}{expected\ frequency}$$

5 Obtain p-value for degrees of freedom (# of rows – 1) * (# of columns – 1)

6 Reject H_0 if p-value is less than α

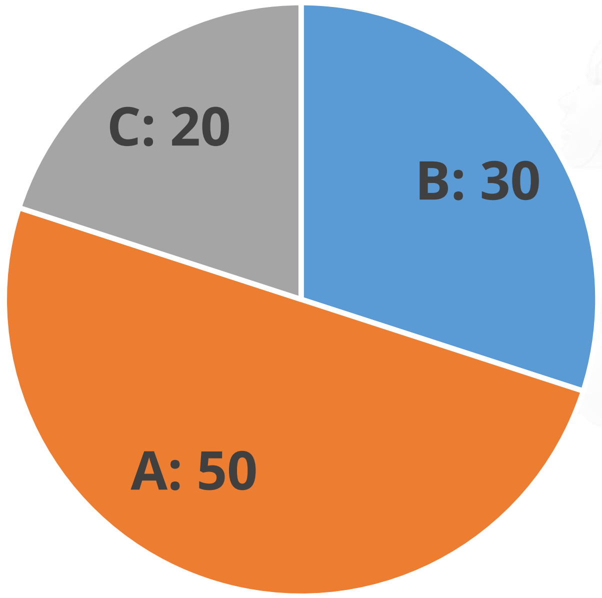
Chi-Square Test for Goodness of Fit

A reach firm published the results of a study on the market share of various companies who all operate in the same industry. The study showed that the market share stabilized at 50% for company A, 30% for company B, and 20% for company C.

To expand its market share, company C developed an improved version of their product and conducted a study to see if that changes their market shares.

The population of interest here would be a multinomial population consisting of customers of each of the three company’s products. The survey would show the proportion of population preferring each company’s product.

To check if the introduction of the new product has brought a change in market share, a chi-square test for goodness of fit has to be done.



Steps for Chi-square Test for Goodness of Fit

1

State the null and alternative hypotheses.
H_0: The population follows a multinomial distribution with specified probabilities for each of the k categories.
H_a: The population does not follow a multinomial distribution with the specified probabilities for each of the k categories.

2

Select a random sample and record the observed frequencies (f_i) for each category.

3

Assume the null hypothesis is true and determine the expected frequency (e_i) in each category by multiplying the category probability by the sample size.

Steps for Chi-square Test for Goodness of Fit

4 Compute the value of test statistic.

$$\chi 2 = \sum \frac{(fi - ei)2}{ei}$$

5 Calculate p-value and compare with α to conclude.

Chi-Square Test of Independence

Duration: 3 minutes

Problem Scenario: A study about the subscribers of a magazine collected data on the subscribers' employment status (shown in the table on screen).

Using alpha = 0.05, test if the employment status is independent of region.

Employment Status	**Region**	
	North Edition	South Edition
Full time	1105	574
Part time	31	15
Self employed	229	186
Not employed	485	344

Note: Please download data set and the solution document from the **Course Resources** section and follow the steps given in the document

Key Takeaways

- Hypothesis testing is a method of statistical inference to assess whether the statement (hypothesis) made about the population is consistent with the observed data (sample).

- Type 1 error is committed when a true null hypothesis is rejected by the statistical test.

- Type 2 error is committed when the test fails to reject a false null hypothesis.

- P-value is known as the credibility rating for null hypothesis. The lower the p-value, the lesser the evidence against the null hypothesis.

Key Takeaways

- A test null hypothesis will be rejected if p-value is less than the significance level.

- ANOVA is performed to compare means of more than two populations.

- Chi-square test is a nonparametric test performed to test independence of variables.

Knowledge Check

Knowledge Check

1

Which of the following are assumptions for ANOVA?

A. Homogeneity of variance among the groups

B. Independent random sampling from the populations

C. Normally distributed population

D. All of the above

Knowledge Check

1

Which of the following are assumptions for ANOVA?

A. Homogeneity of variance among the groups

B. Independent random sampling from the populations

C. Normally distributed population

D. All of the above

The correct answer is **D**

All of these are assumptions to perform ANOVA test.

Knowledge Check

2

Which of the following statements is false?

A. Null hypothesis describes the status quo.

B. We reject the null hypothesis if p-value is less than significance level.

C. T-test is used to compare the population mean of not more than two populations.

D. ANOVA is a test of variance.

Knowledge Check 2

Which of the following statements is false?

A. Null hypothesis describes the status quo.

B. We reject the null hypothesis if p-value is less than significance level.

C. T-test is used to compare the population mean of not more than two populations.

D. ANOVA is a test of variance.

The correct answer is **D**

ANOVA is a test of means, that is, the hypothesis is constructed to compare the means of the populations.

Knowledge Check

3

Which of these is not an example of a parametric test ?

A. Z-test

B. T-test

C. ANOVA

D. Chi-square test

Knowledge Check

3

Which of these is not an example of a parametric test ?

A. Z-test

B. T-test

C. ANOVA

D. Chi-square test

The correct answer is **D**

Chi-square test is a non-parametric test.