



# Probability and Statistics



# Pre-requisites

Hope you have gone through the self-learning content for this session on the PRISM portal.



# Recap



# Exploring Hypothesis Testing in Data Science



# By the End of this Session:

## Core Concepts

- Formulate and test hypotheses
- Interpret the results of statistical tests
- Identify and avoid errors
- Communicate the results of statistical tests

## Analytical Skills Development

- **Data analysis:** Students learn how to collect, organize, and analyze data.
- **Critical thinking:** Students learn how to think critically about the data and to draw sound conclusions.
- **Communication:** Students learn how to communicate the results of their analysis in a clear and concise way.

# Poll Time

Q. A company claims that their new energy drink increases focus and concentration. They want to test this claim by giving the drink to some of their employees and seeing if they perform better on a cognitive test than employees who don't drink the drink. What is the purpose of the test?

- a. To prove that the new energy drink works for everyone in the company.
- b. To count the number of employees who experienced improved focus and concentration after consuming the new energy drink.
- c. To compare the taste preferences of the employees for the new energy drink versus the placebo.
- d. To determine if there is a significant difference in the performance of the employees who drank the new energy drink and those who drank the placebo on the cognitive test.



# Poll Time

Q. A company claims that their new energy drink increases focus and concentration. They want to test this claim by giving the drink to some of their employees and seeing if they perform better on a cognitive test than employees who don't drink the drink. What is the purpose of the test?

- a. To prove that the new energy drink works for everyone in the company.
- b. To count the number of employees who experienced improved focus and concentration after consuming the new energy drink.
- c. To compare the taste preferences of the employees for the new energy drink versus the placebo.
- d. To determine if there is a significant difference in the performance of the employees who drank the new energy drink and those who drank the placebo on the cognitive test.**



# Hypothesis Testing

---

Hypothesis testing is a statistical method for making inferences about populations based on samples.

It is like solving a captivating mystery, where we use data to uncover insights about populations.

There are two rival statements in hypothesis testing: The null hypothesis ( $H_0$ ) and the alternative hypothesis ( $H_1$ ).

The level of significance ( $\alpha$ ) sets the threshold for accepting or rejecting the null hypothesis.

Confidence intervals provide a range of plausible values for the population parameter, and the margin of error measures the uncertainty in our estimate.





# Foundation of Hypothesis Testing

---

	Null Hypothesis	Alternate Hypothesis
Definition	The null hypothesis is a statement that there is no difference or effect in the population and that any observed difference or effect is due to chance.	The alternative hypothesis is a statement that contradicts or challenges the null hypothesis.
Denoted as	$H_0$	$H_a$
Purpose	It represents the default or baseline assumption. This is typically what researchers attempt to disprove or reject.	This statement is often called the "research hypothesis" because it is the hypothesis that the researcher is trying to prove.

# Pop Quiz

Q. Which of the following statements accurately describes the null and alternative hypotheses in hypothesis testing?

- a. The null hypothesis assumes a significant difference, while the alternative hypothesis assumes no difference
- b. The null hypothesis assumes no significant difference, while the alternative hypothesis assumes a difference
- c. The null hypothesis and alternative hypothesis both assume a significant difference
- d. The null hypothesis and alternative hypothesis both assume no difference



# Pop Quiz

Q. Which of the following statements accurately describes the null and alternative hypotheses in hypothesis testing?

- a. The null hypothesis assumes a significant difference, while the alternative hypothesis assumes no difference
- b. The null hypothesis assumes no significant difference, while the alternative hypothesis assumes a difference**
- c. The null hypothesis and alternative hypothesis both assume a significant difference
- d. The null hypothesis and alternative hypothesis both assume no difference



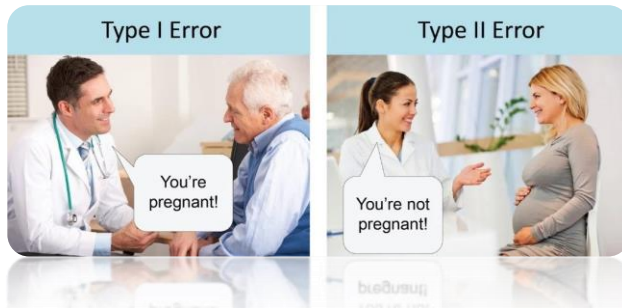
# The Risk of Hypothesis Testing

---

Hypothesis testing is a powerful tool for making inferences about populations, but it is important to be aware of the risks involved.

**There are two main types of errors that can occur in hypothesis testing:**

- Type 1 errors and Type 2 errors.
- A Type 1 error occurs when the null hypothesis is rejected when it is actually true. This is also known as a false positive.
- A Type 2 error occurs when the null hypothesis is not rejected when it is actually false. This is also known as a false negative.
- The probability of making a Type 1 error is denoted by alpha ( $\alpha$ ). The probability of making a Type 2 error is denoted by beta ( $\beta$ ).



# Making Informed Decision in Hypothesis Testing

---

In hypothesis testing, there are two important concepts:

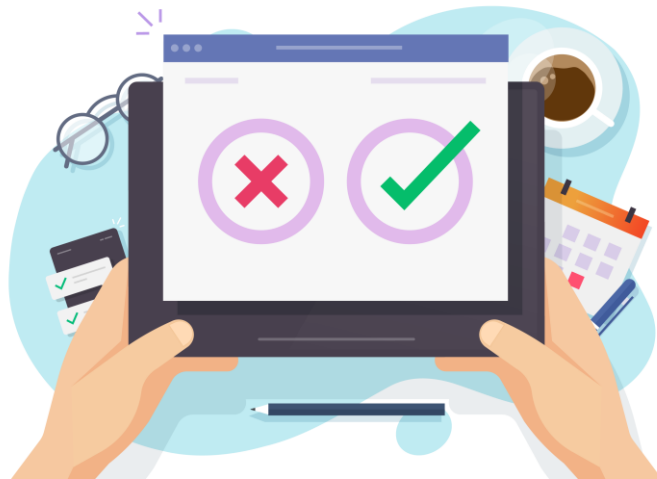
1. The level of significance
2. The level of confidence

Level of Significance	Level of Confidence
The probability of rejecting the null hypothesis when it is actually true	The probability of correctly accepting the null hypothesis when it is actually true.
Denoted by the Greek letter alpha ( $\alpha$ ).	Denoted by the Greek letter beta ( $\beta$ ).
The most common value is 0.05, which means that there is a 5% chance of making a Type 1 error.	The most common value is 0.95, which means that there is a 95% chance of correctly accepting the null hypothesis when it is actually true.
A lower level of significance means that there is a lower chance of making a <b>Type 1</b> error, but it also means that there is a lower chance of rejecting the null hypothesis when it is actually false.	A higher level of confidence means that there is a higher chance of correctly accepting the null hypothesis, but it also means that there is a higher chance of making a <b>Type 2</b> error.

# Poll Time

Q. In a criminal trial, which of the following scenarios represents a Type I error?

- a. Convicting an innocent person and sending them to prison
- b. Acquitting a guilty person and allowing them to go free
- c. Convicting a guilty person and ensuring justice is served
- d. Acquitting an innocent person and protecting their rights



# Poll Time

Q. In a criminal trial, which of the following scenarios represents a Type I error?

- a. **Convicting an innocent person and sending them to prison**
- b. Acquitting a guilty person and allowing them to go free
- c. Convicting a guilty person and ensuring justice is served
- d. Acquitting an innocent person and protecting their rights



# How Accurate Are Your Results

---

When conducting research, it is important to be able to quantify the accuracy of your results.

This is where the concept of margin of error comes in.

**The Margin of error** is a statistical term that refers to the amount of uncertainty associated with a sample statistic. It is used to determine the range within which the true population parameter is likely to fall.

The smaller the margin of error, the more accurate the results are likely to be.

$$\text{MOE}_{\gamma} = z_{\gamma} \times \sqrt{\frac{\sigma^2}{n}}$$

MOE = margin of error

$\gamma$  = confidence level

$z_{\gamma}$  = quantile

$\sigma$  = standard deviation

$n$  = sample size



# How Accurate Are Your Results

---

## Need for Margin of Error:

- In hypothesis testing, the margin of error is used to calculate the confidence interval.
- The confidence interval is a range of values likely to contain the true population parameter.
- The confidence interval is calculated by adding and subtracting the margin of error from the sample statistic.

$$\text{MOE}_{\gamma} = z_{\gamma} \times \sqrt{\frac{\sigma^2}{n}}$$

MOE = margin of error

$\gamma$  = confidence level

$z_{\gamma}$  = quantile

$\sigma$  = standard deviation

$n$  = sample size

# Example

---

## Scenario:

Imagine that you are a political pollster and you want to know how many people in your city support a particular candidate.

You randomly survey 100 people and find that 55% of them support the candidate.

You know that the true percentage of people in the city who support the candidate may not be exactly 55%.

There is always some uncertainty in survey results.

The margin of error is a way of quantifying this uncertainty.

In this example, the margin of error for the survey is 3% at a 95% confidence level.

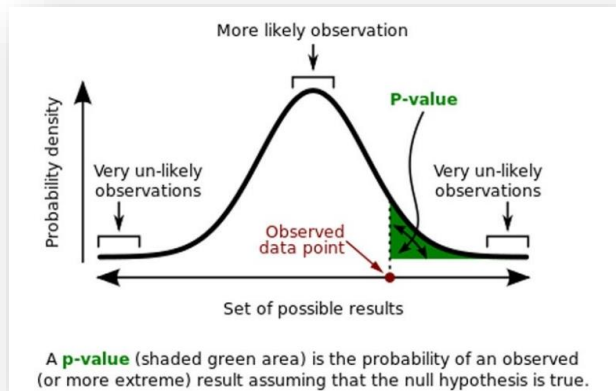
This means that the true population percentage of people who support the candidate is likely to fall between 52% and 58%.

In other words, you can be 95% confident that the true population percentage of people who support the candidate is within this range.



# The P-value: A Statistical Measure of Evidence

- P-value is a measure of how likely the results you observed are if the null hypothesis is true.
- The null hypothesis is a statement about the population mean or proportion assumed to be true before the data is collected.
- If the P-value is small, it means that the observed results are unlikely to have occurred if the null hypothesis is true.
- In this case, the researcher would reject the null hypothesis and conclude that there is evidence to support the alternative hypothesis.



# Example

---

## Scenario:

You are a marketing manager for a company that sells widgets. You want to know if a new advertising campaign has increased sales.

You collect data on sales for the past 6 months, both before and after the advertising campaign started.

1. The P-value is 0.01, which means that there is a 1% chance of obtaining the results that you observed if the advertising campaign did not have a significant effect on sales.

Since the P-value is less than the significance level of 0.05, you would reject the null hypothesis and conclude that the advertising campaign did have a significant effect on sales.



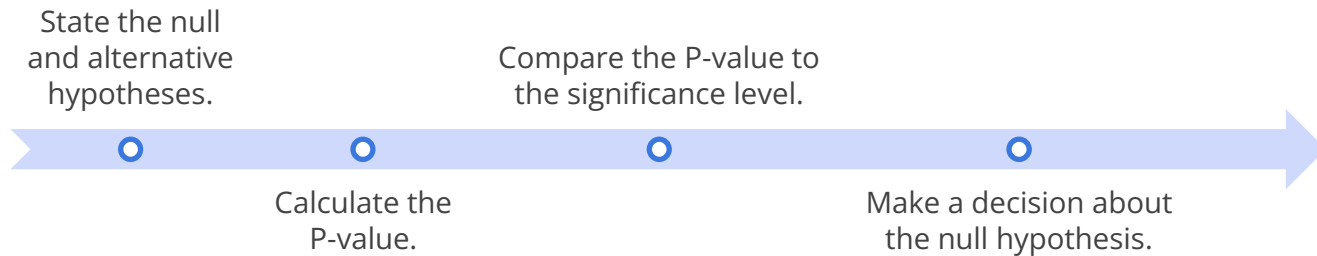
# Inference From Hypothesis Testing

---

## Case Study:

A company wants to know if a new advertising campaign has increased sales. They collect data on sales for the past 6 months, both before and after the advertising campaign started.

## Roadmap for solution:



# Inference From Hypothesis Testing

---

## **Solution:**

- **H<sub>0</sub>:** The advertising campaign did not have a significant effect on sales.
- **H<sub>A</sub>:** The advertising campaign did have a significant effect on sales.
- P-value: 0.01
- Significance level: 0.05
- Decision: Reject the null hypothesis.

**Inference:** The results of the study are unlikely to have occurred by chance. Therefore, we can conclude that the new advertising campaign did increase sales.

# Poll Time

Q. A company wants to know if a new advertising campaign has increased sales. They collect data on sales for the past 6 months, both before and after the advertising campaign started. The P-value is 0.01. Should the null hypothesis be rejected? Consider 5% Level of significance.

- a. Yes
- b. No



# Poll Time

Q. A company wants to know if a new advertising campaign has increased sales. They collect data on sales for the past 6 months, both before and after the advertising campaign started. The P-value is 0.01. Should the null hypothesis be rejected? Consider 5% Level of significance.

- a. **Yes**
- b. No







# T-test

---

- The t-test is a statistical test used to compare the means of two groups.
- It is a parametric test, which means that it assumes that the data is normally distributed.

Type of t-test	Description
One-sample t-test	Compares the mean of a single group to a known value.
Independent two-sample t-test	Compares the means of two independent groups.
Paired t-test	Compares the means of two related groups.

# One Sample T-test

---

The one-sample t-test is a test that helps you see if the average of a group of data points is significantly different from a known value.

The known value is often called the hypothesized mean, and it is denoted by  $\mu$ .

## Assumptions of One Sample T-test:

- Data is independent.
- Data is collected randomly. For example, with simple random sampling.
- The data is approximately normally distributed.

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

- The sample mean ( $\bar{x}$ )
- The population mean ( $\mu$ )
- The sample standard deviation ( $s$ )
- Number of observations ( $n$ )

# One Sample T-test

---

## **Example:**

A school administrator wants to know if the average height of the students in her school is significantly different from the national average height of 5 feet 10 inches. She collects data on the heights of a sample of 10 students, and the mean height of the sample is 5 feet 11 inches. The sample standard deviation is 1 inch.

## **Solution Steps:**

1. State the null hypothesis and the alternative hypothesis.
2. Calculate the t-statistic.
3. Calculate the P-value.
4. Compare the P-value to the significance level.
5. Make a decision about the null hypothesis.

# One Sample T-test

---

## **Solution:**

### **Null Hypothesis:**

The average height of the students in the school is equal to 5 feet 10 inches.

$H_0: \mu = 5.5$  feet

### **Alternative Hypothesis:**

The average height of the students in the school is significantly different from 5 feet 10 inches.

$H_a: \mu \neq 5.5$  feet

### **Calculation of the t-statistic:**

$$t = (\bar{x} - \mu) / s \sqrt{(1/n)} = (5.75 - 5.5) / 1 \sqrt{(1/10)} = 2.25$$

### **Calculation of the P-value:**

P-value = 0.03

### **Decision about the null hypothesis:**

The P-value is less than the significance level of 0.05, so we reject the null hypothesis. This means that we can conclude that the average height of the students in the school is significantly different from the national average height of 5 feet 10 inches.

# Independent T-test

---

The independent t-test is a statistical test used to compare the means of two independent groups.

Independent groups mean that the data points in one group are not related to the data points in the other group.

Assumptions of Independent T-test:

- **Assumption of Independence:** You need two independent, categorical groups that represent your independent variable. In the above example of test scores “males” or “females” would be your independent variable.
- **Assumption of normality:** The dependent variable should be approximately normally distributed. The dependent variable should also be measured on a continuous scale. In the above example on average test scores, the “test score” would be the dependent variable.
- **Assumption of Homogeneity of Variance:** The variances of the dependent variable should be equal.

# Independent T-test - formula

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

**$\bar{x}_1$**  is the mean of the first sample

**$\bar{x}_2$**  is the mean of the second sample

**$\mu_1$**  is the mean of the first population

**$\mu_2$**  is the mean of the second population

**$s_1$**  is the standard deviation of the first sample

**$s_2$**  is the standard deviation of the second sample

**$n_1$**  is the size of the first sample

**$n_2$**  is the size of the second sample

# Independent T-test - formula

---

## Example:

A researcher is interested in knowing whether there is a difference in the average IQ scores of children raised in single-parent households and two-parent households.

## Data:

Group	Mean IQ Score	Standard Deviation	Sample Size
Single Parent	100	15	50
Two Parent	110	10	50



# Independent T-test - Formula

---

## **Solution:**

## **Hypothesis:**

$H_0$ : MIQ(single-parent) = MIQ(two-parent)

$H_a$ : MIQ(single-parent) < MIQ(two-parent)

MIQ stands for mean IQ score

## **Calculate the t-statistic:**

The t-statistic is calculated as follows:

$$t = (100 - 110) / 12.25 \sqrt{1/50 + 1/50} = -2.236$$

## **Calculate the P-value:**

The critical value depends on the significance level, which is typically set to 0.05.

In this case, the significance level is 0.05, so the critical value is 1.96.

$$P\text{-value} = P(t < -2.236) = 0.025$$

## **Make a decision:**

The decision about whether to reject or fail to reject the null hypothesis is based on the P-value.

# Pop Quiz

Q. Independent t-test is used to compare the means of two independent groups. Which of the following is NOT a requirement for independent t-test?

- a. The two groups must be randomly sampled.
- b. The two groups must have equal variances.
- c. The two groups must be normally distributed.
- d. The two groups must be independent of each other.



# Pop Quiz

Q. Independent t-test is used to compare the means of two independent groups. Which of the following is NOT a requirement for independent t-test?

- a. The two groups must be randomly sampled.
- b. The two groups must have equal variances.**
- c. The two groups must be normally distributed.
- d. The two groups must be independent of each other.



# Dependent T test

---

The dependent t-test is a statistical test used to compare the means of two related groups.

Related groups means that the data points in one group are related to the data points in the other group.

## **Assumptions of Paired T test:**

- The dependent variable must be continuous (interval/ratio).
- The observations are independent of one another.
- The dependent variable should be approximately normally distributed.
- The dependent variable should not contain any outliers.

# Dependent T-test

---

## Example:

Let's say you have a group of 10 participants who are all taking a new medication for high blood pressure. You measure their blood pressure before they start taking the medication and then again after they have been taking it for 6 weeks.

## Data:

Participant	Blood Pressure (Before)	Blood Pressure (After)
1	140	120
2	150	130
3	160	140
4	170	150
5	180	160
6	190	170
7	200	180
8	210	190
9	220	200
10	230	210

# Dependent T-test

---

## **Solution:**

### **Hypothesis:**

H0: MBP(before) = MBP (after)

Ha: MBP(after) < MBP (before)

Where:

MBP stands for Mean Blood Pressure

(before) and (after) refer to the time point when the blood pressure was measured

### **Calculate the t-statistic:**

The t-statistic is calculated as follows:

$$t = (\bar{x}_1 - \bar{x}_2) / s \sqrt{(1/n)}$$

$$t = (130 - 140) / 10 \sqrt{(1/10)} = -2$$

### **Calculate the P-value:**

The P-value is calculated by comparing the t-statistic to a critical value from a t-distribution. The critical value depends on the significance level, which is typically set to 0.05.

$$P\text{-value} = P(t < -2) = 0.025$$

### **Make a decision:**

The decision about whether to reject or fail to reject the null hypothesis is based on the P-value.

# Pop Quiz

Q. Which of the following is true about a dependent t-test?

- a. It is used to compare the means of two independent groups.
- b. It is used to compare the means of two dependent groups.
- c. It is used to compare the means of two groups, but the groups must be randomly selected.
- d. It is used to compare the means of two groups, but the groups must be matched.



# Pop Quiz

Q. Which of the following is true about a dependent t-test?

- a. It is used to compare the means of two independent groups.
- b. It is used to compare the means of two dependent groups.**
- c. It is used to compare the means of two groups, but the groups must be randomly selected.
- d. It is used to compare the means of two groups, but the groups must be matched.

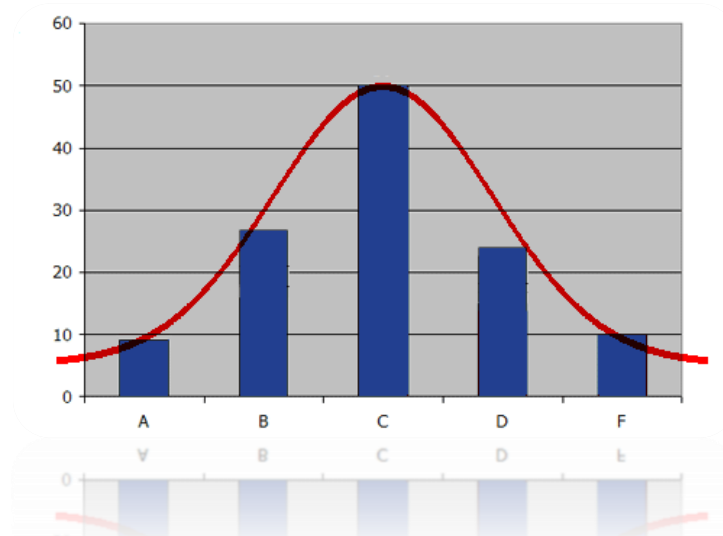




# Z-test

---

- A **Z-test** is a type of hypothesis test, which is a way for you to figure out if results from a test are valid or repeatable.
- For example, if someone said they had found a new drug that cures cancer, you would want to be sure it was probably true.
- A Z test is used when your data is approximately normally distributed i.e., the data has the shape of a bell curve when you graph it).



## When to Run Z-test

---

The sample size is greater than 30. Otherwise, use a t-test.

Data points should be independent of each other.

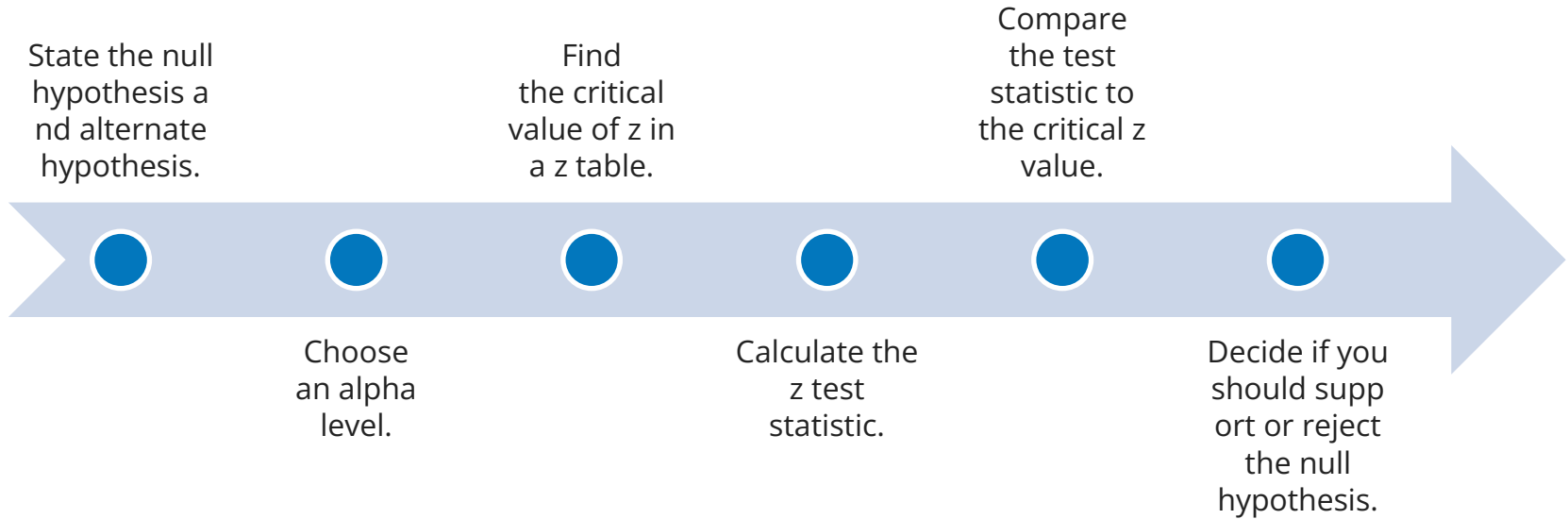
Data should be normally distributed. However, for large sample sizes, (over 30) this doesn't always matter.

Data should be randomly selected from a population, where each item has an equal chance of being selected.

Sample sizes should be equal if at all possible.

# How To Run Z-test?

---



## Formula for Z-test

---

$$Z = \frac{(\bar{X} - \mu_0)}{s}$$

$Z$  = Z-test

$\bar{X}$  = sample average

$\mu_0$  = mean

$s$  = standard deviation

## Example

---

A researcher wants to know if the average height of men in the US is different from 6 feet. They collect a sample of 100 men and find the average height to be 6'1". The population standard deviation is known to be 2 inches.

**Solution:**

$$z = (\bar{x} - \mu) / \sigma$$

z is the z-score

$\bar{x}$  is the sample mean (6'1")

$\mu$  is the population mean (6 feet)

$\sigma$  is the population standard deviation (2 inches)

$$z = (6'1" - 6') / 2" = 0.5$$

The z-score is calculated as 0.5, which is less than the critical value of 1.96.

Therefore, the null hypothesis is not rejected, meaning the average height of men in the US is not significantly different from 6 feet.

# Summary

---

- ✓ Hypothesis testing is a statistical method that is used to determine whether there is a significant difference between two or more populations.
- ✓ T-tests and z-tests are two types of hypothesis tests used to compare the means of two populations.
- ✓ T-tests are used when the population standard deviation is unknown, while z-tests are used when the population standard deviation is known.
- ✓ The main difference between t-tests and z-tests is that z-tests require a larger sample size than t-tests.

## Activity

---

A pharmaceutical company claims that its new drug is effective in reducing blood pressure. A sample of 100 patients is randomly selected and given the drug. The mean blood pressure of the sample is 120 mmHg, with a standard deviation of 10 mmHg. The company claims that the mean blood pressure of the population is 130 mmHg.

### **Solution Hint:**

1. Calculate the z-score.
2. Determine the critical value.
3. Make a decision about the null hypothesis.





## Next Session:

Hypothesis Testing Case Study

# THANK YOU!

Please complete your assessments and review the self-learning content for this session on the **PRISM** portal.





# Hypothesis Testing Case Study



# Pre-requisites

Hope you have gone through the self-learning content for this session on the PRISM portal.



By the end of  
this Session,  
you will:

- Identify the difference between a null hypothesis and an alternative hypothesis.
- Set the level of significance for a hypothesis test.
- Calculate the test statistic for a T-test or Z-test.
- Interpret the p-value of a hypothesis test.
- Make decisions about the null hypothesis based on the p-value.

# What Have You Learned So Far?

---

The different types of hypothesis tests and when to use them.

How to calculate the test statistic and p-value for a hypothesis test.

How to interpret the p-value and make decisions about the null hypothesis.

The importance of choosing the appropriate level of significance.

The role of margin of error in statistical estimation.

## Poll Time

Q. A telecom company wants to test whether the average customer satisfaction score is different for customers who live in urban areas versus customers who live in rural areas. Which of the following hypothesis tests should the company use?

- a. Independent T test
- b. Dependent T test
- c. Z test



## Poll Time

Q. A telecom company wants to test whether the average customer satisfaction score is different for customers who live in urban areas versus customers who live in rural areas. Which of the following hypothesis tests should the company use?

- a. **Independent T test**
- b. Dependent T test
- c. Z test



# Case Study – Problem Statement



# Problem Statement

---

## **Problem:**

Ankur wants to understand which attributes of a car insurance dataset impact insurance claims.

## **Approach:**

Ankur will use statistical techniques and visualization tools to analyze the dataset.

## **Goal:**

Ankur will use hypothesis testing, correlation analysis, and regression analysis to identify the attributes that contribute to insurance claims.

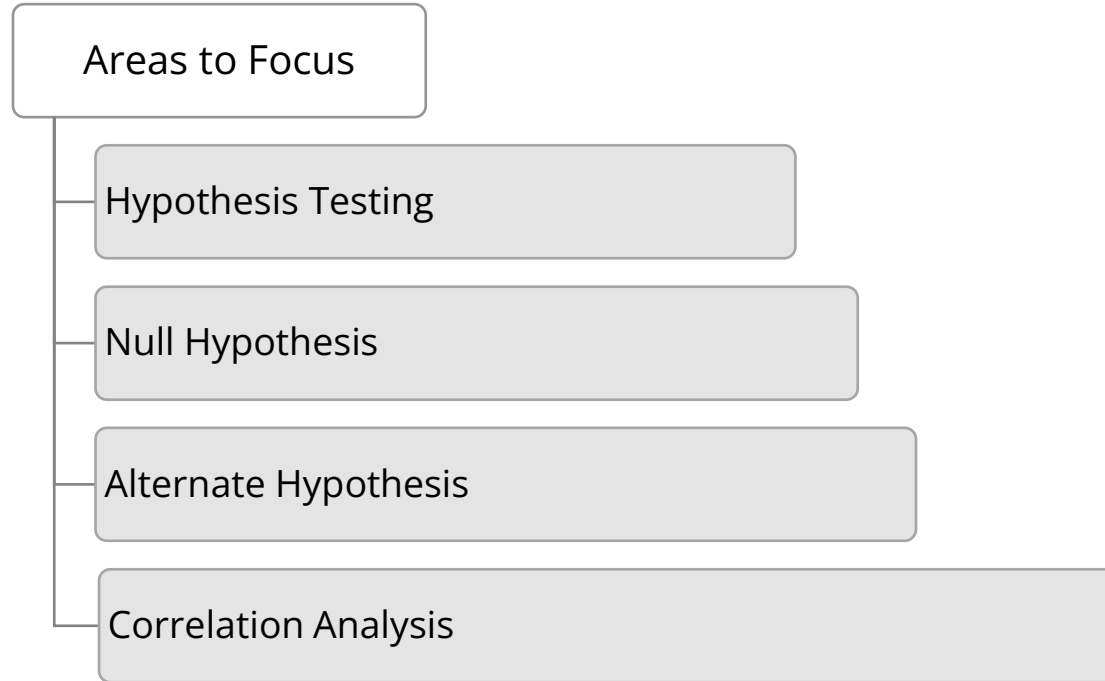
## **Expected Outcome:**

Ankur expects to be able to identify the attributes that are most strongly correlated with insurance claims.



## Areas to Focus

---



## Poll Time

Q. Which of the following is NOT an area of focus for Ankur in his analysis of the data?

- a) Identifying the attributes that are most likely to be correlated with insurance claims.
- b) Performing hypothesis testing to determine whether there is a statistically significant relationship between attributes and insurance claims.
- c) Using correlation analysis to measure the strength of the relationship between each attribute and insurance claims.
- d) Using visualization tools to help understand the data and the results of the analysis.
- e) Building a Predictive Model



# Poll Time

Q. Which of the following is NOT an area of focus for Ankur in his analysis of the data?

- a) Identifying the attributes that are most likely to be correlated with insurance claims.
- b) Performing hypothesis testing to determine whether there is a statistically significant relationship between attributes and insurance claims.
- c) Using correlation analysis to measure the strength of the relationship between each attribute and insurance claims.
- d) Using visualization tools to help understand the data and the results of the analysis.

**e) Building a Predictive Model**





# Understanding the Data

# Sneak Peak into the Data

---

## Policy holder Information

Attribute	Description
policy_id	Unique identifier of the policy holder
policy_tenure	Time period of the policy
age_of_policyholder	Normalised age of policyholder in years
area_cluster	Area cluster of the policyholder
population_density	Population density of the city (Policyholder City)

## Vehicle Information

Attribute	Description
make	Encoded Manufacturer/Company of the car
segment	Segment of the car(A/ B1/B2/ C1/ C2)
model	Encoded name of the car
fuel_type	Type of fuel used by the car
max_torque	Maximum torque generated by Car(Nm@rpm)
max_power	Maximum power generated by Car(Nm@rpm)
engine_type	Type of engine used in the car
ncap_rating	Safety rating given by NCAP (out of 5)

# Sneak Peak into the Data

---

## Safety Features

Attribute	Description
airbags	Number of airbags
is_esc	ESC present or not
is_tpms	TPMS present or not
is_parking_sensors	Parking sensors present or not
is_parking_camera	Parking camera present or not

## Comfort Features

Attribute	Description
is_front_fog_lights	Boolean flag indicating whether front fog lights are available in the car or not.
is_rear_window_wiper	Boolean flag indicating whether the rear window wiper is available in the car or not.
is_rear_window_washer	Boolean flag indicating whether the rear window washer is available in the car or not.
is_rear_window_defogger	Boolean flag indicating whether rear window defogger is available in the car or not.
is_brake_assist	Boolean flag indicating whether the brake assistance feature is available in the car or not.

# Sneak Peak into the Data

---


## Driving Convenience Features

Attribute	Description
is_power_door_locks	Boolean flag indicating whether power door locks are present in the car or not.
is_central_locking	Boolean flag indicating whether central locking feature is available in the car or not.
is_power_steering	Boolean flag indicating whether power steering is available in the car or not.
is_driver_seat_height_adjustable	Boolean flag indicating whether the height of the driver seat is adjustable or not.



# What Are You Going to Build?

---



We perform Hypothesis Testing that can be used to form decisions about whether one hypothesis is true or false based on different attributes.

We will use the dataset to test different hypotheses about the factors that contribute to policyholders filing claims.

We will identify the attributes that have the strongest relationship with policyholders filing claims.

# Poll Time

Q. Ankur and his friend want to perform a hypothesis testing on the dataset to make decisions about if one hypothesis is true or false based on different attributes.

Hypothesis: Older cars are more likely to file a claim than newer cars.  
Which of the following attributes would be used to test this hypothesis?

- a. Policy tenure
- b. Age of the policyholder
- c. Population density of the city
- d. Make and model of the car
- e. Age of the car

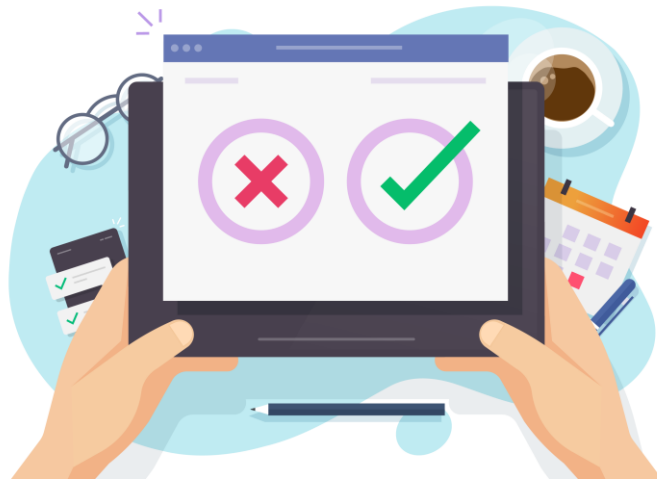


# Poll Time

Q. Ankur and his friend want to perform a hypothesis testing on the dataset to make decisions about if one hypothesis is true or false based on different attributes.

Hypothesis: Older cars are more likely to file a claim than newer cars.  
Which of the following attributes would be used to test this hypothesis?

- a. Policy tenure
- b. Age of the policyholder
- c. Population density of the city
- d. Make and model of the car
- e. Age of the car**





# Hands-on: Case Study Questions

## Poll Time

Q. Which attribute is most important to consider when testing the hypothesis that policy holders who live in cities are more likely to file a claim than policy holders who live in rural areas?

- a. Policy tenure
- b. Age of the policy holder
- c. Age of the car
- d. Population density of the city
- e. Make and model of the car



## Poll Time

Q. Which attribute is most important to consider when testing the hypothesis that policy holders who live in cities are more likely to file a claim than policy holders who live in rural areas?

- a. Policy tenure
- b. Age of the policy holder
- c. Age of the car
- d. Population density of the city**
- e. Make and model of the car







# Activity 1

---

## **Pre-requisites:**

- Hypothesis Testing
- T test

## **Scenario:**

1. Load the breast cancer dataset from the sklearn library.
2. Split the dataset into a training set and a test set.
3. Fit a t-test to the training set.
4. Test the hypothesis that the mean tumor size for malignant tumors is different from the mean tumor size for benign tumors.
5. Report the results of the t-test.

## **Hint:**

The hypothesis that we are testing is that the mean tumor size for malignant tumors is different from the mean tumor size for benign tumors.

## Summary

---

- Hypothesis testing is a statistical method used to determine whether there is a significant difference between two groups or sets of data.
- T-test is a parametric test, which means that it makes certain assumptions about the data, such as the normality of the distributions and the equality of the variances. If these assumptions are not met, then the results of the t-test may be unreliable.
- Z-test is a non-parametric test, which means that it does not make any assumptions about the data. This makes it a more robust test than the t-test, but it also means that the results of the z-test may be less precise.

# Session Feedback



## Next Session:

Excel Case Study - II

# THANK YOU

Please complete your assessments and review the self-learning content for this session on the **PRISM** portal.

