



Hypothesis Testing, Regressions and ANOVA



Pre-requisites

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

What Have You Learned So Far?

- You can use hypothesis testing to determine whether there is a significant difference between the means of two groups.
- You can use hypothesis testing to determine whether there is a significant difference between the proportions of two groups.
- You can use hypothesis testing to determine whether there is a significant difference between the variances of two groups.



By the End of this Session, You Will be Able To:

- Measure the strength and direction of the relationship between two variables.
- Compare the means of two or more groups.
- Compare the frequencies of categorical variables.
- Develop and test hypotheses about the relationship between variables.
- Interpret the results of statistical tests and draw conclusions.
- Communicate the results of statistical tests to others in a clear and concise way.

Pop Quiz

Q. Which statistical test should you use to compare the average weight of two groups of adults if you only have data for a small sample of each group? (less than 30)

- a. T-test
- b. Z-test



Pop Quiz

Q. Which statistical test should you use to compare the average weight of two groups of adults if you only have data for a small sample of each group? (less than 30)

a. **T-test**

b. Z-test





Hypothesis Testing, Regressions and ANOVA (analysis of variance)

Covariance

Covariance is a measure of the relationship between two random variables.

Covariance is a statistical measure that indicates the extent to which two variables are related to each other.

It is calculated as the average of the product of the deviations of two variables from their means.

A positive covariance indicates that the two variables tend to move in the same direction, while a negative covariance indicates that they tend to move in opposite directions.

Covariance

Formula

$$\text{cov}_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

$\text{cov}_{x,y}$ = covariance between variable x and y

x_i = data value of x

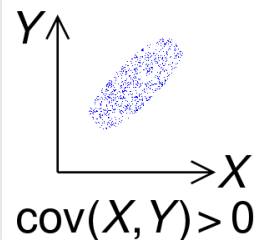
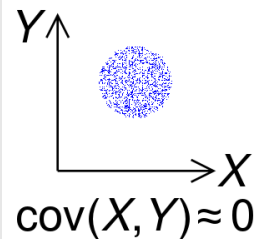
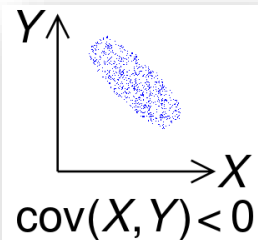
y_i = data value of y

\bar{x} = mean of x

\bar{y} = mean of y

N = number of data values

What does it means by Covariance?



Correlation

Correlation is a statistical concept that measures the strength and direction of the linear relationship between two variables.

A correlation of 0 indicates that there is no linear relationship between two variables.

A correlation of -1 indicates that there is a perfect negative relationship between two variables.

A correlation of 1 indicates that there is a perfect positive relationship between two variables.

Correlation

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

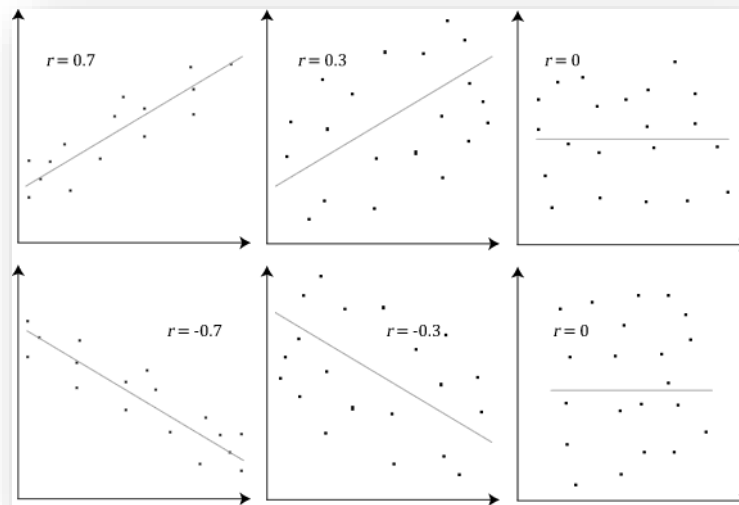
x_i = values of the x-variable in a sample

\bar{x} = mean of the values of the x-variable

y_i = values of the y-variable in a sample

\bar{y} = mean of the values of the y-variable

What does it mean by Correlation?



scatterplots of various datasets with various correlation coefficients.

Causation

Causation is the relationship between two events or variables where one event or variable (the cause) brings about the other event or variable (the effect).

Factors which can help us infer Causation:

Temporal precedence: The cause must come before the effect.
For example, you cannot get lung cancer before you start smoking.

The cause and effect must be linked.
For example, it would not make sense to say that smoking causes lung cancer if there is no evidence that smoking and lung cancer are related.

The cause and effect must be seen over time.
For example, the correlation between smoking and lung cancer.

Removal of the cause: If you remove the cause, the effect should also disappear.
For example, if you stop smoking, your risk of developing lung cancer should decrease.

Difference

Concept	Definition	Relationship
Correlation	Measures the strength and direction of the linear relationship between two variables.	A correlation coefficient of +1 indicates a perfect positive relationship, a correlation coefficient of -1 indicates a perfect negative relationship and a correlation coefficient of 0 indicates no relationship.
Covariance	Measures the extent to which two variables vary together.	A positive covariance indicates that when one variable increases, the other variable also increases. A negative covariance indicates that when one variable increases, the other variable decreases.
Causation	Is the relationship between two variables where one variable (the cause) brings about the other variable (the effect).	The cause is said to be responsible for the effect.

Practical Application

Scenario:

Researchers conducted a study among a group of students to collect data on their study time and exam scores.

Data:

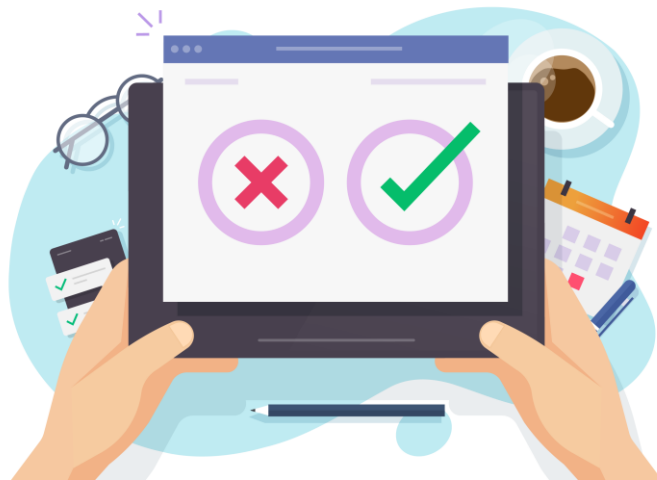
Student	Study Time (hours)	Exam Performance (score)
1	2	70
2	4	80
3	3	75
4	6	90
5	5	85

- The data showed a strong positive correlation between study time and exam scores, meaning that as study time increased, exam scores tended to increase as well.
- The correlation coefficient between study time and exam scores was approximately **0.98**.
- The covariance between study time and exam scores was approximately **6.5**.
- To establish causation, researchers would need to conduct controlled experiments where they manipulate study time while keeping other factors constant.

Poll Time

Q. Which of the following statements is true regarding covariance, correlation, and causation?

- a. Covariance and correlation can be used interchangeably to measure the relationship between variables
- b. Causation can be established solely based on high correlation or covariance between variables
- c. Covariance measures the strength and direction of the linear relationship between variables, while correlation provides a standardized measure of the strength of the relationship
- d. Establishing causation requires meeting specific criteria, including temporal order, association, non-spuriousness, and mechanism



Poll Time


Q. Which of the following statements is true regarding covariance, correlation, and causation?

- a. Covariance and correlation can be used interchangeably to measure the relationship between variables
- b. Causation can be established solely based on high correlation or covariance between variables
- c. Covariance measures the strength and direction of the linear relationship between variables, while correlation provides a standardized measure of the strength of the relationship
- d. Establishing causation requires meeting specific criteria, including temporal order, association, non-spuriousness, and mechanism**





ANOVA



Analysis of Variance (ANOVA) is a statistical technique used to compare means across two or more groups or treatments.

ANOVA can be used to determine whether there is a significant difference between the means of the groups.

Assumptions of ANOVA

The data must be normally distributed.

The variances of the groups must be equal.

The data must be independent.

Types of ANOVA

1. **One-way ANOVA**

- One-way ANOVA is a statistical test that is used to compare the means of two or more groups.
- It is called one-way because there is only one independent variable (the factor).

2. **Two-way ANOVA**

- It is a statistical test that is used to compare the means of two or more groups while also taking into account the effects of another variable.

Two-way ANOVA partitions the total variation in the data into two components:

- Between-group variation: This is the variation that is due to the differences between the means of the groups.
- Within-group variation: This is the variation that is due to other factors, such as individual differences.

One-way ANOVA

Scenario:

Experiment to study the effect of fertilizer on plant growth:

- Three levels of fertilizer (a_1 , a_2 , a_3) were tested.
- Six observations were taken for each level.

The data can be written in a table like this:

a_1	a_2	a_3
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

Perform one-way ANOVA to test the hypothesis that the three levels of fertilizer have the same effect on plant growth.

One-way ANOVA

Solution:

H_0 : All three levels of the factor produce the same response, on average

H_a : One of the three levels of the factor produces the same response, on average

Step 1: Calculate the mean within each group.

Group	Mean
a1	5 (Y1)
a2	9 (Y2)
a3	10 (Y3)

Step 2: Calculate the overall mean.

Overall mean (\bar{Y}) = $(5+9+10)/3 = 8$

Step 3: Calculate the "between-group" sum of squared differences:

Formulae: $n \{(Y1 - \bar{Y})^2 + (Y2 - \bar{Y})^2 + (Y3 - \bar{Y})^2\}$

Where n is the number of data values per group

SB = $6(5-8)^2 + 6(9-8)^2 + 6(10-8)^2 = 84$

One-way ANOVA

The between-group degrees of freedom is one less than the number of groups

$$\mathbf{Fb} = 3 - 1 = 2$$

Thus, between group mean square is

$$\mathbf{MSb} = 84/2 = 42$$

Step 4: Calculate the "within-group" sum of squares. Begin by centering the data in each group.

a_1	a_2	a_3
$6-5=1$	$8-9=-1$	$13-10=3$
$8-5=3$	$12-9=3$	$9-10=-1$
$4-5=-1$	$9-9=0$	$11-10=1$
$5-5=0$	$11-9=2$	$8-10=-2$
$3-5=-2$	$6-9=-3$	$7-10=-3$
$4-5=-1$	$8-9=-1$	$12-10=2$

Sw = 68 (Adding square of all the resultant from above table)

One-way ANOVA

The within-group degrees of freedom is

$$\mathbf{Fw} = a(n-1) = 3(6-1) = 15$$

Thus, the within-group mean square value is

$$\mathbf{MSw} = \mathbf{Sw} / \mathbf{fw} = 68/15 = 4.5$$

Step 5: The *F*-ratio is

$$F = \mathbf{MSb} / \mathbf{MSw} = 42/4.5 = 9.3$$

$$F_{\text{crit}}(2, 15) = 3.68 \text{ at } \alpha = 0.05$$

$F=9.3 > 3.68$, the results are significant at the 5% significance level. One would not accept the null hypothesis, concluding that there is strong evidence that the expected values in the three groups differ.

Poll Time

Q. Which of the following is NOT an assumption of one-way ANOVA?

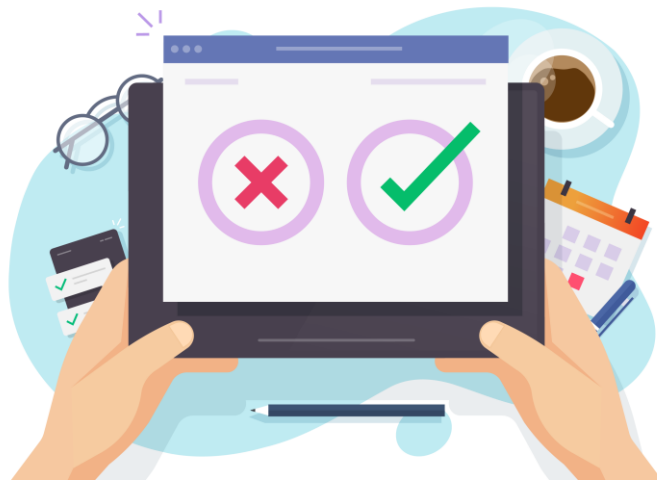
- a. The data is normally distributed
- b. The samples are independent
- c. The population variances are equal
- d. The dependent variable is continuous
- e. The independent variable is categorical



Poll Time

Q. Which of the following is NOT an assumption of one-way ANOVA?

- a. The data is normally distributed
- b. The samples are independent
- c. The population variances are equal
- d. The dependent variable is continuous
- e. The independent variable is categorical**



Two-way ANOVA

- Two-way ANOVA is an extension of one-way ANOVA.
- It is used to test the effect of two independent variables on a dependent variable.
- The independent variables are called factors.

Use Cases:

1. To test the effect of different teaching methods and different levels of teacher experience on student test scores.
2. To test the effect of different marketing channels and different product promotions on website traffic.
3. To test the effect of different customer demographics and different product features on product return rates.
4. Two-Way ANOVA can be utilized to assess the main effects of demographic segment and purchase behavior on customer lifetime value.

Chi-square

The chi-square statistic is a measure of how far the observed values deviate from the expected values. A large chi-square statistic indicates that there is a significant difference between the observed and expected values.

Chi-square test is a statistical test that compares the distribution of observed values with the distribution of expected values.

Formula

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

χ^2 = chi squared

O_i = observed value

E_i = expected value

Example

Scenario:

A researcher is interested in whether there is a relationship between the gender of a student and their preference for a particular type of music. The researcher surveys 100 students and asks them to identify their gender and their favorite type of music. The results of the survey are shown in the following table:

Data:

Gender	Music preference	Observed value	Expected value
Male	Rock	30	25
Female	Rock	20	25
Male	Pop	20	25
Female	Pop	30	25

The researcher wants to test the hypothesis that there is no relationship between gender and music preference. The alternative hypothesis is that there is a relationship between gender and music preference.

Example

Solution:

- Null hypothesis: Gender and music preference are independent.
- Alternative hypothesis: Gender and music preference are not independent.

Expected value = (Total number of observations) * (Proportion of observations in cell) = $100 * (25 / 100) = 25$

chi-square statistic:

$$\chi^2 = (30 - 25)^2 / 25 + (20 - 25)^2 / 25 + (20 - 25)^2 / 25 + (30 - 25)^2 / 25 = 1 + 1 + 1 + 1 = 4$$

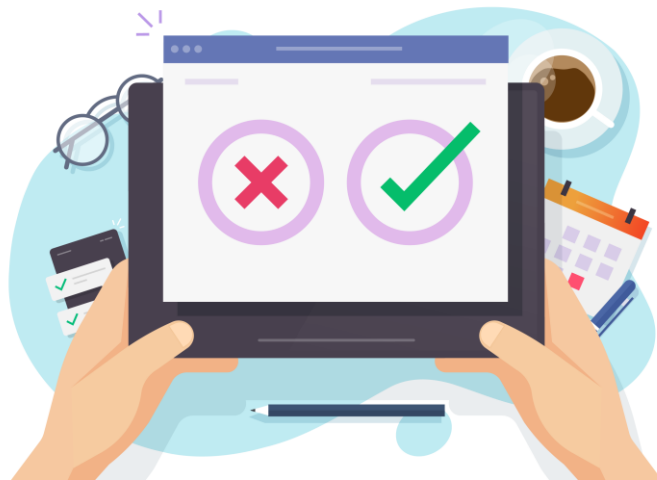
- The p-value for this test can be calculated using a chi-square table.
- The degrees of freedom for this test is $(2 - 1)(2 - 1) = 1$.
- The p-value for a chi-square statistic of 4 with 1 degree of freedom is 0.5.

Since the p-value is greater than the significance level ($\alpha = 0.05$), you cannot reject the null hypothesis. This means that there is not enough evidence to conclude that there is a relationship between gender and music preference.

Poll Time

Q. A researcher conducts a chi-square test of independence and finds a p-value of 0.05. What can the researcher conclude?

- a. There is a significant relationship between the two variables
- b. There is no significant relationship between the two variables
- c. The sample size is too small to draw any conclusions
- d. The data is not normally distributed



Poll Time

Q. A researcher conducts a chi-square test of independence and finds a p-value of 0.05. What can the researcher conclude?

- a. There is a significant relationship between the two variables
- b. There is no significant relationship between the two variables**
- c. The sample size is too small to draw any conclusions
- d. The data is not normally distributed





Summary

- ✓ **Covariance:** Measured how much two variables vary together.
- ✓ **Correlation:** Measured the strength of the relationship between two variables.
- ✓ **Causation:** Explored how one variable causes the other to change.
- ✓ **ANOVA:** Compared means of two or more groups.
- ✓ **Chi-square:** Compared observed and expected frequencies of categorical variables.

Activity

A study was conducted to investigate the relationship between smoking habits (smoker or non-smoker) and the incidence of lung cancer in a sample of 500 individuals. The following data was collected:

Among 300 smokers, 80 individuals had lung cancer.

Among 200 non-smokers, 30 individuals had lung cancer.

Using the Chi-Square test, determine if there is a significant association between smoking habits and the incidence of lung cancer in this sample.

Activity

Solution Steps:

Step 1: Set up hypotheses.

Step 2: Set the significance level.

Step 3: Calculate the Chi-Square test statistic.

Step 4: Calculate the degrees of freedom.

Step 5: Find the critical value.

Step 6: Calculate the Chi-Square test statistic.

Step 7: Compare the test statistic with the critical value.

Step 8: Interpret the results.

THANK YOU!

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





Hypothesis Testing Case Study – Part II



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 be Able To:

- Define null and an alternative hypotheses for statistical tests.
- Set the level of significance for a hypothesis test.
- Calculate the test statistic for an ANOVA and Chi-square 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?

Measuring the
relationship among
variables.

Cause and Effect
analysis.

Comparing means of
multiple continuous
variables.

Analyzing the
frequencies of
categorical variables to
find the relation
among them.

Identifying the
strength of
relationships among
variables.

Poll Time

Q. Which of the following is the formula for the F-statistic in ANOVA?

- a. $(\text{Sum of squares between groups}) / (\text{Sum of squares within groups})$
- b. $(\text{Mean of group 1} - \text{Mean of group 2}) / (\text{Standard deviation of group 1})$
- c. $(\text{Mean of group 1} + \text{Mean of group 2}) / 2$
- d. $(\text{Total sum of squares}) / (\text{Number of groups})$



Poll Time

Q. Which of the following is the formula for the F-statistic in ANOVA?

- a. **(Sum of squares between groups)/(Sum of squares within groups)**
- b. (Mean of group 1 - Mean of group 2)/(Standard deviation of group 1)
- c. (Mean of group 1 + Mean of group 2)/2
- d. (Total sum of squares)/(Number of groups)



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.

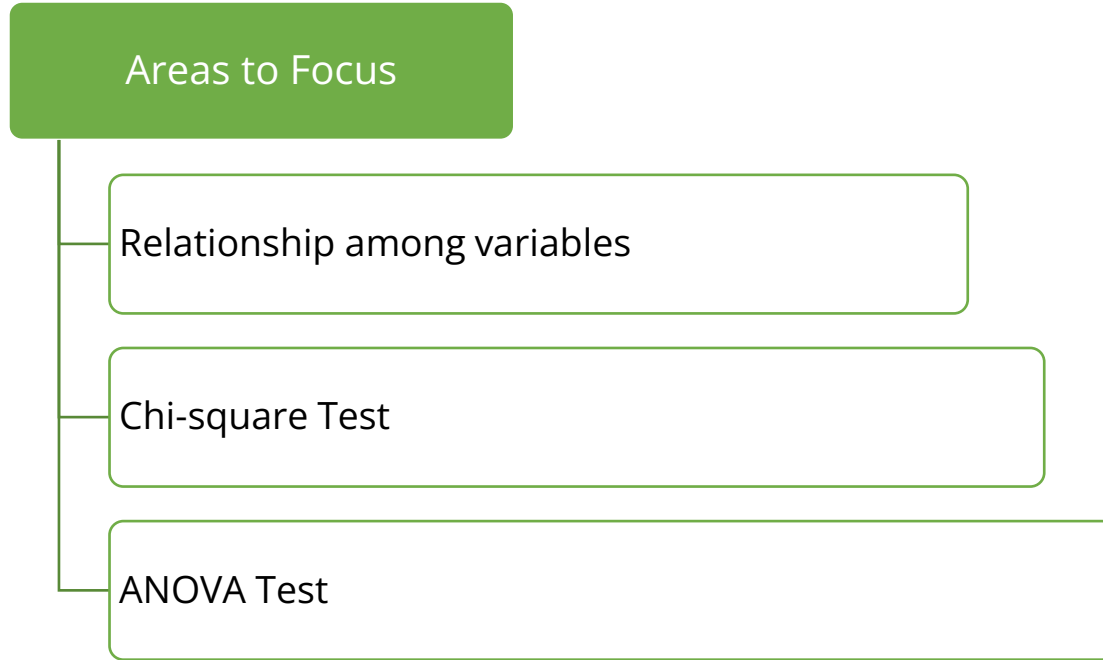
Goal:

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. Ankur wants to test whether the length and width of a vehicle are independent of each other. Which of the following statistical tests should he use?

- a. Chi square test of independence
- b. One-way ANOVA
- c. Pearson correlation coefficient
- d. T-test



Poll Time

Q. Ankur wants to test whether the length and width of a vehicle are independent of each other. Which of the following statistical tests should he use?

- a. **Chi-square test of independence**
- b. One-way ANOVA
- c. Pearson correlation coefficient
- d. T-test





Understanding the Data

Sneak Peak into the Data

Policyholder Information

Attribute	Description
policy_id	Unique identifier of the policyholder
policy_tenure	Time period of the policy
age_of_policyholder	Normalized 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

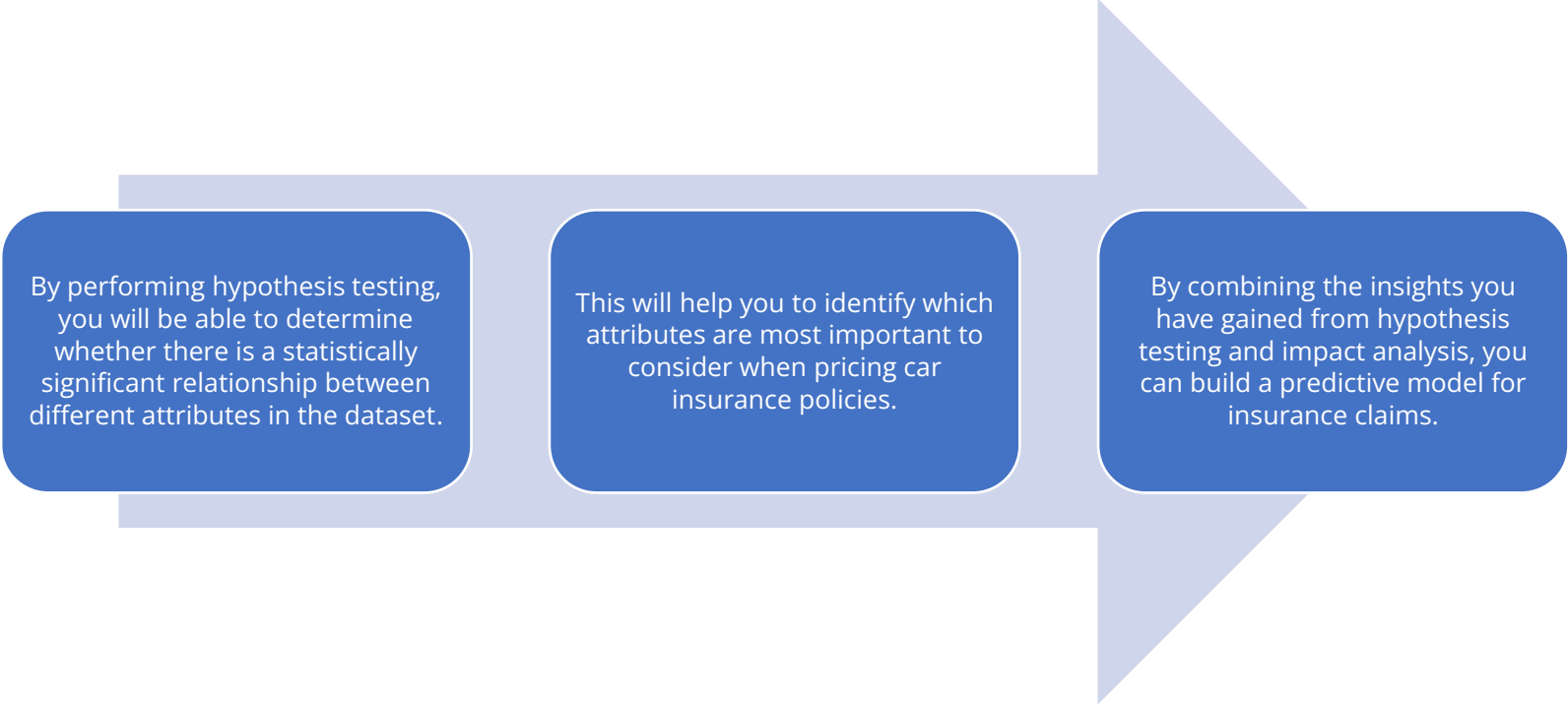
Attribute	Description
is_front_fog_lights	Boolean flag indicating whether front fog lights are available in the car or not.
is_rear_window_wiper	Boolean flag indicating whether the rear window wiper is available in the car or not.
is_rear_window_washer	Boolean flag indicating whether the rear window 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?



By performing hypothesis testing, you will be able to determine whether there is a statistically significant relationship between different attributes in the dataset.

This will help you to identify which attributes are most important to consider when pricing car insurance policies.

By combining the insights you have gained from hypothesis testing and impact analysis, you can build a predictive model for insurance claims.

Poll Time

Q. Which of the following statements is true about the relationship between the type of fuel and the likelihood of an insurance claim?

- a. There is a positive correlation between the type of fuel and the likelihood of an insurance claim
- b. There is a negative correlation between the type of fuel and the likelihood of an insurance claim
- c. There is no correlation between the type of fuel and the likelihood of an insurance claim
- d. It is impossible to say whether there is a correlation between the type of fuel and the likelihood of an insurance claim without further analysis



Poll Time

Q. Which of the following statements is true about the relationship between the type of fuel and the likelihood of an insurance claim?

- a. There is a positive correlation between the type of fuel and the likelihood of an insurance claim
- b. There is a negative correlation between the type of fuel and the likelihood of an insurance claim
- c. There is no correlation between the type of fuel and the likelihood of an insurance claim
- d. It is impossible to say whether there is a correlation between the type of fuel and the likelihood of an insurance claim without further analysis**





Hands-on: Case Study Questions

Poll Time

Q. Is there a correlation between the maximum torque and maximum power of a vehicle?

- a. Yes, there is a positive correlation between the maximum torque and maximum power of a vehicle
- b. Yes, there is a negative correlation between the maximum torque and maximum power of a vehicle
- c. No, there is no correlation between the maximum torque and maximum power of a vehicle



Poll Time

Q. Is there a correlation between the maximum torque and maximum power of a vehicle?

- a. **Yes, there is a positive correlation between the maximum torque and maximum power of a vehicle**
- b. Yes, there is a negative correlation between the maximum torque and maximum power of a vehicle
- c. No, there is no correlation between the maximum torque and maximum power of a vehicle





Summary

- ✓ Understood that hypothesis testing is a statistical method used to determine whether there is a significant difference between two groups or sets of data.
- ✓ Dived deep into ANOVA, a parametric test, which means that it makes certain assumptions about the data.
- ✓ Performed a nonparametric test (Chi-square), which means that it does not make any assumptions about the data. This makes chi-square a more robust test than ANOVA, but it also means that the results of chi-square may not be as precise.

Activity 1

Pre-requisites:

- Hypothesis Testing
- Chi Square Test

Scenario:

A smartphone manufacturing company, XYZ Technologies, wants to understand whether there is a relationship between smartphone brand preference and age groups among consumers.

They conducted a survey with a sample of 500 smartphone users, asking them to indicate their preferred smartphone brand and their age group.

The age groups considered were 18-25, 26-35, 36-45, and 46 and above. The goal is to determine if there is a significant association between smartphone brand preference and age groups.

Data:

	Apple	Samsung	Google	Other
18-25	60	80	20	40
26-35	50	70	30	40
36-45	40	60	30	20
46 and above	30	50	10	10

Activity 2

Pre-requisites:

- Correlation and Covariance

Scenario:

A company is interested in understanding whether there is a relationship between the amount of time a person spends on social media and their GPA. They have collected data on the amount of time spent on social media and GPA for a sample of 100 students. Help the company by testing the given hypothesis by calculating the correlation and covariance coefficients.

Hypothesis:

The company hypothesizes that people who spend more time on social media will have lower GPAs.

Data:

Amount of Time Spent on Social Media (hours/week)	GPA
0-5	3.5
5-10	3.0
10-15	2.5
15-20	2.0

Session Feedback



Next Session: Introduction to NumPy

THANK YOU

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

