

CHAPTER-7

HYPOTHESIS TESTING

INTRODUCTION:

In this chapter, students will learn the concept of hypothesis testing and its role in data analysis. They will understand what a hypothesis is and how research questions are converted into testable hypotheses. The chapter introduces different types of hypotheses, including the null hypothesis (H_0), and explains how to formulate correct hypotheses.

7.1 Introduction to Hypothesis Testing

Ex.O.1. What is hypothesis? Explain the concept of hypothesis testing and its role in data analysis.

Hypothesis

A hypothesis is an **assumption or statement** in research that is made before testing begins. The statement can be tested using data. It predicts an expected result. A good hypothesis should be clear and measurable. Data is collected to check whether the hypothesis is correct or not. A correct hypothesis helps to guide the analysis and helps to focus the study.

A hypothesis is a possible explanation or prediction about a problem that can be tested through experiments or observations. It is an educated guess made by researchers before conducting a study. A hypothesis helps to guide the research process and find the relationship between variables.

The hypothesis function is written as: $h_{\theta}(x) = y$

Where:

- $h_{\theta}(x)$: The hypothesis function with parameters (θ) applied to input (x).
- x : The input features or data points.
- y : The predicted output or target value.
- θ : The parameters or weights of the model.

Example:

"If the amount of sunlight increases, then the growth of plants will increase." This is a hypothesis because it can be tested through an experiment.

Hypothesis Testing

Hypothesis testing is a method used to make decisions by using data. It checks whether an assumption is true or false. Hypothesis testing is widely used in research and data analysis. It follows a step-by-step process to test assumptions.

Data is collected and analyzed carefully. The results help to accept or reject an assumption. This method reduces guesswork because decisions are based on evidence instead of personal opinions.

Short Question

What is Hypothesis testing?

Research Questions and Hypotheses

A research question is the **main question** that a researcher wants to answer. It helps to identify the problem and guides the research study.

Research questions tell us what we want to study. Research questions help to decide the purpose of the study.

A hypothesis is created from the research question. Research questions guide data collection. Hypotheses give possible answers. Clear research questions lead to clear hypotheses. Good research begins with good questions. A hypothesis keeps the analysis focused. They improve study direction.

Research Question: Does daily exercise improve students' health?

Hypothesis: Students who exercise daily have better health than students who do not exercise.

Role Of Hypothesis Testing in Data Analysis

- It helps analyze data logically to find out whether the results are meaningful.
- It checks whether the collected data supports the hypothesis.
- It helps avoid random conclusions.
- It compares the results with the expected outcome.
- It improves the accuracy of experiments and surveys.
- It helps make decisions based on evidence instead of guessing.
- It increases confidence in the final results.

Short Question

Describe role of Hypothesis Testing in Data Analysis?



TIDBIT

- Hypothesis testing checks assumptions using data. It helps decide whether a claim is true. This method is used in research and experiments. It reduces guesswork. Results depend on proper testing.

7.2 Types of Hypotheses

Ex.Q.2. Describe different types of hypotheses and explain how to formulate correct hypotheses.

Hypotheses are divided into different types. Each type has a different purpose.

The two most common types are the Null Hypothesis and the Alternative Hypothesis. These two hypotheses are used together in hypothesis testing. Correct hypotheses improve the quality of analysis by reducing confusion. They help in making better decisions. Understanding these types is important because they are the foundation of hypothesis testing.

Types of Hypotheses

1. Null Hypothesis (H_0)

The Null Hypothesis (H_0) states that there is **no effect** or no difference. It is the default assumption in hypothesis testing. The collected data is tested against this hypothesis. If the evidence is strong, the Null Hypothesis is rejected. It provides a base for comparison. It also helps reduce bias. Most hypothesis tests begin with the Null Hypothesis. It plays an important role in decision-making. It is represented by H_0 .

Example

A school wants to check whether a new teaching method improves students' marks.

H_0 : The new teaching method does not improve students' marks.

2. Alternative Hypothesis (H_1 or H_a)

The Alternative Hypothesis (H_1 or H_a) states that there is an **effect** or a difference. It is the opposite of the Null Hypothesis. It shows that a change or effect exists. The purpose of hypothesis testing is usually to find evidence to support the Alternative Hypothesis. Strong evidence is needed before accepting it. It helps explain the expected result and gives direction to the research. It is represented by H_1 or H_a .

Short Question

What is Null Hypothesis?

Example:

Sample Research Questions and Hypotheses:

Research Question:

Does the amount of study time affect students' exam scores?

Hypothesis:

H_0 (Null Hypothesis): There is no significant relationship between study time and exam scores.

H_1 (Alternative Hypothesis): The amount of study time significantly affects students' exam scores.

Formulating Correct Hypotheses

Correct hypotheses should be clear, specific, and testable. A hypothesis should match the research question. It should be tested by using data. Variables should be clearly defined. Vague statements should be avoided because they reduce accuracy. Simple language makes the hypothesis easier to understand. A well-written hypothesis saves time and produces more reliable results. It also improves the quality of data analysis.

Example

Research Question: Does drinking enough water improve students' concentration?

Correct Hypothesis: Students who drink enough water have better concentration than students who do not drink enough water.



CLASS ACTIVITY

Students write a null and alternative hypothesis for a marks or percentage situation. Discuss correctness.

Situation:

A teacher wants to check whether the average marks of Class 11 students are 70.

Students should write the hypotheses:

- **Null Hypothesis (H_0):** The average marks of the students are 70.
- **Alternative Hypothesis (H_1):** The average marks of the students are not 70.

Discussion:

- Ask students to identify which statement is the **Null Hypothesis** and which is the **Alternative Hypothesis**.
- Check whether the null hypothesis includes the original claim and the alternative hypothesis shows a difference.
- Discuss why both hypotheses are needed before analyzing the data.



Additional Multiple Choice Questions

1. An assumption or statement in research that is made before testing begins is called:
(a) Analysis (b) Programming (c) Data storage (d) hypothesis
2. A good hypothesis should be:
(a) Clear and measurable (b) Long and complex (c) Difficult to test (d) Based on guessing
3. Hypothesis testing is a method used to make decisions by using:
(a) Opinions (b) Guesses (c) Data (d) Assumptions
4. A research question is:
(a) The final result (b) The predicted output
(c) The main question that a researcher wants to answer (d) The collected data
5. The Null Hypothesis (H_0) states that there is:
(a) An effect (b) A prediction (c) A relationship (d) No effect or no difference
6. The Alternative Hypothesis (H_1 or H_a) states that there is:
(a) No effect (b) No relationship (c) An effect or a difference (d) No data
7. Which hypothesis is the default assumption in hypothesis testing?
(a) Alternative Hypothesis (b) Research Hypothesis
(c) Working Hypothesis (d) Null Hypothesis
8. Why is hypothesis testing used in data analysis?
(a) It helps analyze data logically to find out whether the results are meaningful.
(b) It stores data permanently.
(c) It creates databases automatically.
(d) It increases internet speed.

9. Correct hypotheses should be:
 (a) Long and detailed (b) Clear, specific and testable
 (c) Difficult to understand (d) Based on opinions
10. Why should vague statements be avoided while formulating a hypothesis?
 (a) They reduce accuracy. (b) They improve testing. (c) They increase data. (d) They simplify programming.

Answers:

1	D	2	A	3	C	4	C	5	D	6	C	7	D	8	A	9	B	10	A
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	----	---

7.3 Concepts in Hypothesis Testing

Ex.Q.3. Explain key concepts in hypothesis testing, including test statistics, critical region, and p-value.

Concepts Used in Hypothesis Testing

Hypothesis testing uses several important concepts to make correct decisions. Each concept has a specific role in hypothesis testing. These concepts help avoid confusion and guide the acceptance or rejection of hypotheses. They work together to improve the accuracy of data analysis. Learning these concepts is important for understanding hypothesis testing.

Test Statistics

Test statistics are numerical values calculated from sample data. They summarize the behavior of the data. Test statistics help compare the results with the expected values. They show how far the results are from the expected outcome. Different statistical tests use different test statistics. Common examples are the t-value and the chi-square value.

t-value Test

Where:

- $\bar{X}$: Sample mean
- μ_0 : Population mean under the null hypothesis
- s : Sample standard deviation
- n : Sample size

Example:

Suppose we have a sample of 25 students with an average score of 80, and we want to test if the mean score is significantly different from the population mean of 75. The sample standard deviation is 10.

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{80 - 75}{\frac{10}{\sqrt{25}}} = \frac{5}{\frac{10}{5}} = \frac{5}{2} = 2.5$$

Chi-Square Test:

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

Where:

- O_i : Observed frequency in each category
- E_i : Expected frequency in each category

Example:

Suppose you roll a fair die 60 times and observe the following frequencies:

$$\begin{aligned} \chi^2 &= \frac{(10 - 10)^2}{10} + \frac{(12 - 10)^2}{10} + \frac{(8 - 10)^2}{10} + \frac{(14 - 10)^2}{10} + \frac{(9 - 10)^2}{10} + \frac{(7 - 10)^2}{10} \\ &= 0 + \frac{4}{10} + \frac{4}{10} + \frac{16}{10} + \frac{1}{10} + \frac{9}{10} \\ &= 0 + 0.4 + 0.4 + 1.6 + 0.1 + 0.9 = 3.4 \end{aligned}$$

Now, the calculated χ^2 value is 3.4, and we would compare this value against the critical value from the Chi-Square distribution table at a certain significance level (e.g., $\alpha = 0.05$) and degrees of freedom to make a decision.

Face of Die	Observed Frequency (O_i)	Expected Frequency (E_i)
1	10	10
2	10	10
3	10	10
4	10	10
5	9	10
6	7	10

Critical Region

The critical region is a **range of values** where the Null Hypothesis (H_0) is rejected. This region is decided before testing begins. It depends on the significance level. If the test statistic falls inside the critical region, the Null Hypothesis is rejected. The critical region represents extreme results. It helps reduce uncertainty and makes decision-making easier. It is often shown in statistical graphs.

Significance Level

The significance level is the accepted risk of making an error in hypothesis testing. It is represented by **alpha (α)**. A common significance level is **0.05**. It is selected before data analysis starts. A smaller significance level means stricter testing. The significance level helps control false results and improves the reliability of decisions.

Short Question

What is significance level?

P-value and its Importance

The p-value shows the **strength of evidence** against the Null Hypothesis (H_0). It measures the probability of getting the observed result, or a more extreme result, if the Null Hypothesis is true. A small p-value gives stronger evidence against the Null Hypothesis. The p-value is compared with the significance level such as 0.05, to make a decision.

- If the p-value is less than the significance level (p-value < 0.05), the Null Hypothesis is **rejected**.
- If the p-value is greater than or equal to the significance level (p-value $\geq$ 0.05), the Null Hypothesis is **not rejected**.

Example 1:

A teacher wants to check whether a new teaching method improves students' marks.

H_0 : The new teaching method has no effect on students' marks.

After analyzing the test scores, the p-value is 0.03. Since 0.03 is less than 0.05, the Null Hypothesis is rejected. This shows that there is statistical evidence that the new teaching method may have improved students' performance.

Example 2:

Researchers test whether a new medicine works better than an old one. If the p-value is 0.01, it gives strong evidence that the improvement is unlikely to have happened by chance.

Q. What is t-Test and t-value? Give examples

t-Test

A t-test is used to compare the means of two groups. It is used when the **sample size is small** or the population standard deviation is unknown. The t-test helps determine whether the difference between the two means is statistically significant or occurred by chance.

Example:

A company wants to check whether the average daily working hours of employees are 8 hours. It collects data from 12 employees and uses a t-test to compare the sample mean with 8 hours.

t-value

The t-value is the calculated value in a t-test. It is calculated from the sample data. It shows how much the sample mean differs from the expected population mean. The t-value helps decide whether to accept or reject the Null Hypothesis. It is compared with the critical value or used to calculate the p-value.

Example:

After performing a t-test on the employees' working hours, the calculated t-value is 2.45. This t-value is compared with the critical value or used to find the p-value to decide whether the average working hours are significantly different from 8 hours.

Short Question

What is t-Test?

Short Question

What is t-value?

7.4 Performing Hypothesis Tests

Ex.O.4. Discuss the steps involved in performing hypothesis testing with a suitable example.

Performing Hypothesis Testing

Hypothesis testing follows a clear step-by-step process to make correct decisions.

The process starts by defining the hypotheses. Then data is collected carefully. Test statistics are calculated from the data. The results are compared with the selected criteria. Finally, a decision is made. Following these steps reduces errors and makes testing fair. It also helps produce reliable conclusions.

Steps in Hypothesis Testing

1. State the Hypotheses

Define the Null Hypothesis (H_0) and the Alternative Hypothesis (H_1 or H_a).

2. Select the Significance Level: Choose the significance level (α), such as 0.05, to decide the acceptance criteria.

3. Collect the Data: Gather data from a suitable sample.

4. Calculate the Test Statistic: Use the collected data to calculate the appropriate test statistic.

5. Find the p-value: Determine the p-value from the test statistic.

6. Compare the p-value with the Significance Level: Compare the p-value with the selected significance level (α).

Make the Decision:

If the p-value is less than α , reject the Null Hypothesis. Otherwise, fail to reject the Null Hypothesis.

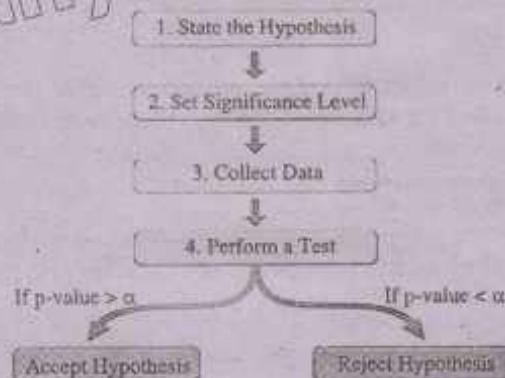


Figure 7.1: Hypothesis Testing Process

Numerical Example: Hypothesis Testing

A school wants to check whether a new teaching method has improved students' marks. The previous average marks were 70. A sample of 10 students taught with the new method obtained the following marks:

72, 75, 78, 74, 77, 80, 76, 79, 81, 78

Step 1: State the Hypotheses

Null Hypothesis (H_0): The new method does not improve marks ($\mu = 70$)

Alternative Hypothesis (H_1): The new method improves marks ($\mu > 70$)

Step 2: Set Significance Level

$$\alpha = 0.05$$

Step 3: Calculate Sample Mean

$$\bar{x} = \frac{\text{Sum of all values}}{n} = \frac{770}{10} = 77$$

Step 4: Calculate Standard Deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = \sqrt{\frac{70}{9}} = 2.79$$

Step 5: Calculate t-value

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{77 - 70}{\frac{2.79}{\sqrt{10}}} = \frac{7}{0.88} = 7.93$$

Step 6: Decision Rule

$$df = n - 1 = 10 - 1 = 9$$

Using the t-distribution table or statistical software, the p-value for $t = 7.93$ and $df = 9$ is approximately: $p = 0.00002$. For ($df = 9$), the p-value is very small ($p < 0.05$).

Step 7: Conclusion

Since: $p\text{-value} < \alpha$

We reject the null hypothesis.

Final Conclusion: The new teaching method significantly improves students' marks.

Q. Describe the F-Test and Chi-Square Test. Also explain the F-Test with a suitable example.

F-Test and Chi-Square Test.

The F-Test and Chi-Square Test are advanced hypothesis testing methods.

F-Test

The F-Test compares the variances of different groups. It is commonly used in analysis of variance (ANOVA). It helps determine whether the variation between groups is significant.

Chi-Square Test

The Chi-Square Test checks the relationship between categorical data. It is used to find whether two categories are related. This test follows fixed statistical rules and is useful for data analysis.

F-test Example:

$$F = \frac{S_1^2}{S_2^2}$$

Where:

- S_1^2 = Variance of sample 1
- S_2^2 = Variance of sample 2

Example:

Suppose we have two groups of students:

Sample 1 variance = 16

Sample 2 variance = 9

$$F = 1.78$$

The calculated F-value (1.78) is compared with a critical value from the F-table to decide whether the variances are significantly different.

Short Question

What are F-Tests?

Short Question

What are Chi-Square Tests?



TIDBIT

- Testing follows clear steps. Data is collected and analyzed. Results are compared with criteria. Conclusions are drawn carefully. Correct tests give reliable outcomes.



Additional Multiple Choice Questions

- Test statistics are numerical values calculated from:
 - Population data
 - Sample data
 - Random values
 - Survey forms
- A common example of a test statistic is the:
 - F-value
 - Mean
 - Median
 - t-value
- The accepted risk of making an error in hypothesis testing is called:
 - Critical region
 - p-value
 - Significance level
 - Sample size
- The range of values where the Null Hypothesis (H_0) is rejected is called:
 - Sample space
 - Test statistic
 - Confidence interval
 - Critical region
- The p-value is compared with the:
 - Population mean
 - Standard deviation
 - Sample size
 - Significance level
- A t-test is used to compare means when the sample size is small or the population standard deviation is:
 - Known
 - Very large
 - Unknown
 - Equal
- The Chi-Square Test is used to find whether two _____ are related.
 - Variables
 - Categories
 - Samples
 - Means

8. The F-Test compares the variances of:
 A) Different groups B) One sample C) Two means only D) Categorical data
9. If the p-value is less than the significance level (0.05), then:
 A) Reject the Null Hypothesis B) Accept the Null Hypothesis
 C) Collect more data D) Calculate the mean again
10. If the p-value is greater than or equal to the significance level (0.05), then:
 A) Reject the Null Hypothesis B) Do not reject the Null Hypothesis
 C) Collect more data D) Calculate the mean again

Answers:

1 B 2 D 3 C 4 D 5 D 6 C 7 B 8 A 9 A 10 B

7.5 Data Visualization For Hypothesis Testing

Ex.Q.5. What is data visualization? Explain the importance of data visualization in hypothesis testing and how graphs help present results.

Data visualization

Data visualization is the graphical presentation of data using charts, graphs, maps, or diagrams. It makes data easier to understand, analyze, and interpret.

Importance of Data Visualization.

- Data visualization changes raw data into meaningful visual forms.
- It helps researchers understand large amounts of data quickly.
- It helps to identify patterns, trends, relationships, and outliers in data.
- Charts and graphs make complex data easier to understand and compare.
- It presents results clearly to both technical and non-technical people.
- It provides visual evidence for better decision-making.
- It improves the clarity and accuracy of hypothesis testing.

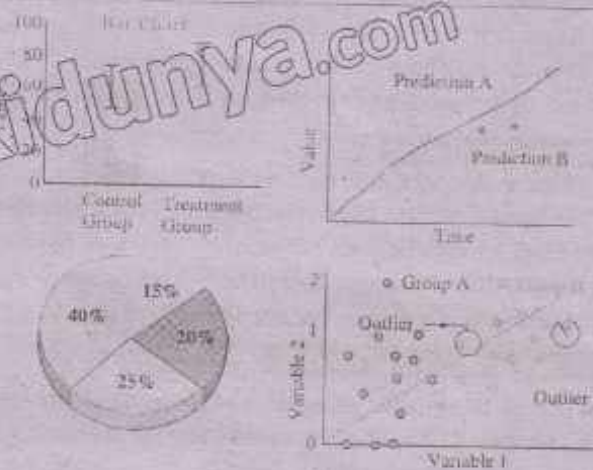


Figure 7.2: Data Visualization

Graphs and charts for presenting results.

Graphs and charts are used to present the results of hypothesis testing in a clear way. Different graphs and charts are used for different purposes.

i. Bar Chart

A bar chart is used to compare different categories or groups.

ii. Line Graph

A line graph shows trends and changes over time.

iii. Pie Chart

A pie chart shows proportions and percentages.

iv. Scatter Plot

A scatter plot shows the relationship between two variables.

Linking data visuals with hypotheses.

Data visuals help connect the results with the proposed hypothesis. Researchers use charts and graphs to check whether the data support or reject the hypothesis. A comparison graph can show the difference between control and experimental groups. A line graph can show whether the expected trend is present. A scatter plot can show whether two variables are related. Visual evidence makes the results easier to understand and supports logical decision-making.

7.6 Ethical Issues in Data Science and Analysis

Ex.Q.6. Discuss ethical issues in data science, including bias, ethical use of data, and responsible communication.

Ethical Issues in Data Science

Ethical issues in Data Science are the moral principles and challenges involved in collecting, using, storing, and sharing data. They ensure that data is used fairly, responsibly, and without harming individuals or society.

Ethical issues are important in data science and data analysis.

Data is collected and analyzed on a large scale, so it must be handled responsibly. Ethical practices ensure fairness, accuracy, and trust in data-driven decisions. Poor ethical practices can lead to the misuse of data and may harm individuals or groups. Data scientists should follow ethical guidelines while collecting, processing, and interpreting data. Transparency and accountability are important at every stage of data analysis. Ethical awareness helps maintain public confidence in data science applications.

Short Question

What are Ethical issues in Data Science?

Bias in data collection and analysis.

Bias in data collection is an **error** that makes the collected data unfair or unbalanced. It occurs when the data does not fairly represent all groups. It can happen because of limited data sources or unfair sampling methods. Biased data produces inaccurate and misleading results. During data analysis, personal assumptions can also introduce bias. Bias can affect decision-making and lead to unfair outcomes.

Therefore, data should be collected carefully and analyzed objectively to ensure fairness and reliable results.

Common Examples of Bias:

Survey Bias: A survey about student performance is conducted only in top-performing schools.

Sampling Bias: A health study includes only young adults, so the results do not represent older people.

Gender Bias: A hiring dataset contains mostly male candidates, so the model favours male applicants.

Geographical Bias: Data is collected only from urban areas and used to make decisions for rural populations.

Confirmation Bias: A researcher considers only the data that supports a personal belief and ignores other results.

Ethical use of data and models.

Ethical use of data means using data responsibly while respecting people's rights. Personal data should be collected with proper consent. It should be used only for its intended purpose. Privacy and confidentiality should always be protected. Machine learning models should avoid discrimination and unfair decisions. Sensitive data should be protected from unauthorized access. Transparency in model design builds trust among users. Responsible use of data and models creates a positive social impact.

Responsible communication of findings.

Responsible communication of findings means presenting results honestly and clearly. The results should not be exaggerated or manipulated. The limitations of the data and analysis should be explained clearly. Misleading graphs, charts, or selective reporting should be avoided. Clear explanations help people understand the results correctly. Ethical communication prevents misunderstanding and misuse of findings. It also improves trust and credibility in data analysis.

Short Question

What do you mean by responsible communication of findings?



TIDBIT

- Biased data can produce unfair results. Ethical data science protects privacy and avoids discrimination. Responsible data use builds trust in technology.

7.7 Communicating Results and Conclusions

Ex.Q.7. Explain how to interpret hypothesis testing results and connect conclusions to hypotheses.

Communicating results and conclusions is an important step in data analysis and research. It helps readers understand what the data shows and what decisions can be made. Clear communication prevents misunderstanding and misuse of results. Researchers should explain the findings in a simple and logical way. Conclusions should be based only on the analyzed data. Proper communication improves transparency and credibility. It also supports future research and decision-making.

Interpreting test results

Interpreting test results means understanding the meaning of statistical results. It includes studying values such as the p-value, confidence level, and test statistics. These results help decide whether the hypothesis is accepted or rejected. Researchers should also consider the context of the data while interpreting the results. Wrong interpretation can lead to incorrect conclusions. Clear interpretation makes the findings accurate, reliable, and meaningful.

Presenting findings clearly

Presenting findings clearly means explaining research results in an easy-to-understand way. The results should be arranged in a logical and organized manner. Tables, charts, and graphs can be used to support the explanation. Simple language should be used instead of difficult technical words. Important points should be highlighted so that readers can understand the results easily. Clear presentation reduces confusion and improves communication. Well-presented findings make the research more useful.

Connecting conclusions to hypotheses

Connecting conclusions to hypotheses means linking the final decision with the original hypothesis. Researchers should clearly state whether the hypothesis is supported or not supported by the data. The conclusion should be based on the analyzed data and test results. Personal opinions should not affect the conclusion. This connection makes the research logical and valid. It also completes the hypothesis testing process.

Ex.Q.8. Describe how findings should be communicated clearly in data analysis and why it is important.

Communication of Findings

After data analysis, the findings should be communicated in a clear and organized way. Clear communication helps readers understand the results correctly.

1. Use Simple and Clear Language

The findings should be written in simple and easy language. Difficult words and confusing sentences should be avoided.

2. Present Data Visually

Tables, charts, graphs, and diagrams should be used to present the results. Visual presentation makes the findings easier to understand and compare.

3. Explain the Main Results

The main findings should be explained clearly. The conclusion should be based only on the analyzed data and evidence.

4. Avoid Personal Opinions

The findings should be objective. Personal opinions, assumptions, and exaggerated claims should not be included.

5. Report All Results Honestly

Researchers should report both positive and negative results. No data should be hidden, changed, or removed without a valid reason.

6. Explain the Limitations

The limitations of the study should be mentioned honestly. This helps readers understand the scope and reliability of the research.

Importance of Clear Communication

- It helps readers understand the findings correctly.
- It prevents misunderstanding and misleading conclusions.
- It improves the accuracy and reliability of the research.
- It increases the credibility and trustworthiness of the findings.
- It helps others make correct decisions based on the research.
- It allows other researchers to review and repeat the study.



Additional Multiple Choice Questions

1. Data visualization is the graphical presentation of data using:
A) Charts and graphs B) Computer programs C) Database tables D) Text documents
2. Data visualization helps researchers to understand:
A) Small files only B) Large amounts of data quickly
C) Programming languages D) Network devices
3. A line graph shows:
A) Proportions and percentages B) Relationship between variables
C) Trends and changes over time D) Different categories
4. Bias in data collection is an error that makes the collected data:
A) Accurate B) Fair C) Unfair or unbalanced D) Complete
5. Ethical use of data means using data responsibly while:
A) Increasing profits B) Respecting people's rights
C) Reducing storage D) Creating software
6. Responsible communication of findings means presenting results:
A) Honestly and clearly B) With personal opinions
C) By hiding limitations D) Through selective reporting
7. Which graph is most suitable to show the relationship between two variables?
A) Pie chart B) Bar chart C) Line graph D) Scatter plot
8. Researchers should clearly state whether the hypothesis is:
A) Supported or not supported by the data C) Based on personal opinions
B) Written in difficult language D) Ignored after testing
9. Which of the following is an example of sampling bias?
A) A researcher ignores unwanted results. C) A health study includes only young adults.
B) Data is collected only from urban areas. D) A hiring dataset contains mostly male candidates.
10. Why should misleading graphs, charts, or selective reporting be avoided?
A) They increase processing speed. C) They improve data collection.
B) They reduce storage space. D) They can cause misunderstanding and misuse of findings.

Answers:

- | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|----|---|
| 1 | A | 2 | B | 3 | B | 4 | C | 5 | B | 6 | A | 7 | D | 8 | A | 9 | C | 10 | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|----|---|

TEXTBOOK EXERCISE (SOLVED)



MULTIPLE CHOICE QUESTIONS

- A hypothesis is:**
(a) A final result (b) A test value (c) A proposed explanation (d) A dataset
- The null hypothesis represents:**
(a) An alternative idea (b) No effect or no difference (c) A final conclusion (d) A prediction model
- The symbol used for the null hypothesis is:**
(a) H_1 (b) H_2 (c) H_0 (d) H_x
- A p-value is used for:**
(a) Measuring accuracy (b) Testing a hypothesis (c) Measuring variance (d) Data storage
- If ($p\text{-value} < \alpha$), the decision is to:**
(a) Accept H_0 (b) Reject H_0 (c) Ignore data (d) Recalculate the mean
- The critical region is the:**
(a) Safe area (b) Rejection area (c) Data range (d) Sample space
- An example of a test statistic is:**
(a) Mean (b) t-value (c) Mode (d) Median
- Data visualisation is used to:**
(a) Store data (b) Present results clearly (c) Delete data (d) Calculate the mean
- A chart useful for comparison is:**
(a) Pie chart (b) Bar chart (c) Table (d) Text
- Bias in data means:**
(a) Correct data (b) Balanced data (c) Unfair representation (d) Random data

Answers:

1	C	2	B	3	C	4	B	5	B	6	B	7	B	8	B	9	B	10	C
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	----	---



SHORT QUESTIONS

- What is a hypothesis?**

Ans: A hypothesis is an assumption or statement in research that is made before testing begins. The statement can be tested using data. It predicts an expected result. A good hypothesis should be clear and measurable. Data is collected to check whether the hypothesis is correct or not. A correct hypothesis helps to guide the analysis and helps to focus the study.

- What is a research question?**

Ans: A research question is the main question that a researcher wants to answer. It helps to identify the problem and guides the research study. Research questions tell us what we want to study. Research questions help to decide the purpose of the study. A hypothesis is created from the research question.

- Define null hypothesis (H_0).**

Ans: The Null Hypothesis (H_0) states that there is no effect or no difference. It is the default assumption in hypothesis testing. The collected data is tested against this hypothesis. If the evidence is strong, the Null Hypothesis is rejected. It provides a base for comparison.

4. What is a test statistic?

Ans: Test statistics are numerical values calculated from sample data. They summarize the behavior of the data. Test statistics help to compare the results with the expected values. They show how far the results are from the expected outcome. Different statistical tests use different test statistics. Common examples are the t-value and the chi-square value.

5. What is meant by p-value?

Ans: The p-value shows the strength of evidence against the Null Hypothesis (H_0). It measures the probability of getting the observed result, or a more extreme result, if the Null Hypothesis is true. A small p-value gives stronger evidence against the Null Hypothesis.

6. What is a critical region?

Ans: The critical region is a range of values where the Null Hypothesis (H_0) is rejected. This region is decided before testing begins. It depends on the significance level. If the test statistic falls inside the critical region, the Null Hypothesis is rejected. The critical region represents extreme results.

7. What are the basic steps in hypothesis testing?

Ans:

- State the Hypothesis
- Select the significance level
- Collect data.
- Perform a test
- Accept Hypothesis or Reject Hypothesis

8. Why is data visualization important?

Ans: Data visualization presents data in a clear and easy-to-understand form. It helps researchers see patterns, trends, and relationships in data before applying statistical tests. Charts and graphs make complex data easier to understand. They help compare differences between groups.

9. What is bias in data collection?

Ans: Bias in data collection is an error that makes the collected data unfair or unbalanced. It occurs when the data does not fairly represent all groups. It can happen because of limited data sources or unfair sampling methods. Biased data produces inaccurate or misleading results.

10. Why is ethical use of data important?

Ans: Ethical use of data protects people's privacy and rights. It ensures that data is collected, used, and shared fairly and responsibly. It also builds trust and prevents the misuse of data.



LONG QUESTIONS

All Qs are answered in chapter

1. Explain the concept of hypothesis testing and its role in data analysis.
2. Describe different types of hypotheses and explain how to formulate correct hypotheses.
3. Explain key concepts in hypothesis testing, including test statistics, critical region, and p-value.
4. Discuss the steps involved in performing hypothesis testing with a suitable example.
5. Explain the importance of data visualization in hypothesis testing and how graphs help present results.
6. Discuss ethical issues in data science, including bias, ethical use of data, and responsible communication.
7. Explain how to interpret hypothesis testing results and connect conclusions to hypotheses.
8. Describe how findings should be communicated clearly in data analysis and why it is important.

ADDITIONAL SHORT QUESTIONS

Hypothesis Testing

1. What is hypothesis testing and why is it used?

Ans: Hypothesis testing is a method used to make decisions by using data. It checks whether an assumption is true or false. Data is collected and analyzed carefully before making a decision. This method reduces guesswork because decisions are based on evidence.

2. What is the relationship between a research question and a hypothesis?

Ans: A research question is the main question that a researcher wants to answer. A hypothesis is created from the research question. Research questions guide data collection, while hypotheses give possible answers. Good research begins with clear research questions.

Types of Hypotheses

3. What is the main difference between the Null Hypothesis (H_0) and the Alternative Hypothesis (H_1)?

Ans:

Null Hypothesis (H_0)	Alternative Hypothesis (H_1)
The Null Hypothesis (H_0) states that there is no effect or no difference.	The Alternative Hypothesis (H_1) states that there is an effect or a difference.
It is the initial assumption in hypothesis testing.	It is the opposite of the Null Hypothesis.
It is accepted if there is not enough evidence to reject it.	It is accepted when the Null Hypothesis is rejected.
It is written as H_0 .	It is written as H_1 .

4. Explain the importance of the Null Hypothesis (H_0) in hypothesis testing.

Ans: The Null Hypothesis is the default assumption in hypothesis testing. The collected data is tested against this hypothesis. If strong evidence is found, the Null Hypothesis is rejected. It provides a base for comparison and helps reduce bias.

5. How can a correct hypothesis be formulated?

Ans: A correct hypothesis should be clear, specific, and testable. It should match the research question and be tested using data. The variables should be clearly defined. A well-written hypothesis improves the quality of data analysis.

6. Differentiate between a Research Question and a Hypothesis.

Ans:

Research Question	Hypothesis
A research question is the main question that a researcher wants to answer.	A hypothesis is a possible explanation or prediction that can be tested.
It helps identify the problem and guides the research study.	It gives a possible answer to the research question and guides the analysis.

Concepts Used in Hypothesis Testing

7. What is the significance level and why is it important in hypothesis testing?

Ans: The significance level is the accepted risk of making an error in hypothesis testing. It is represented by alpha (α), and a common value is 0.05. It is selected before data analysis starts. It helps control false results and improves the reliability of decisions.

8. Why is the p-value important in hypothesis testing?

Ans: The p-value shows the strength of evidence against the Null Hypothesis (H_0). It is compared with the significance level to make a decision. If the p-value is less than 0.05, the Null Hypothesis is rejected. If it is greater than or equal to 0.05, the Null Hypothesis is not rejected.

9. Why must the critical region be selected before looking at the test results?

Ans: The critical region should be selected before testing to keep the process fair. It prevents changing the decision after seeing the results. This reduces bias in research. It also improves the reliability of the conclusion.

Performing Hypothesis Testing

10. Discuss the steps involved in performing hypothesis testing.

Ans: Hypothesis testing starts by stating the Null Hypothesis and the Alternative Hypothesis. Then the significance level is selected and data is collected. The test statistic and p-value are calculated from the data. Finally, the p-value is compared with the significance level to make a decision.

Advanced Tests

11. What is the F-Test and where is it used?

Ans: The F-Test compares the variances of different groups. It is commonly used in Analysis of Variance (ANOVA). It helps determine whether the variation between groups is significant. It is an advanced hypothesis testing method.

12. What is the Chi-Square Test and why is it used?

Ans: The Chi-Square Test checks the relationship between categorical data. It is used to find whether two categories are related. This test follows fixed statistical rules and is useful for data analysis. It is also an advanced hypothesis testing method.

13. Differentiate between the F-Test and the Chi-Square Test.

Ans:

F-Test	Chi-Square Test
The F-Test compares the variances of different groups.	The Chi-Square Test checks the relationship between categorical data.
It is commonly used in Analysis of Variance (ANOVA).	It is used to find whether two categories are related.

14. Why is collecting accurate data important in hypothesis testing?

Ans: Accurate data produces reliable results. Poor-quality data can lead to wrong conclusions. Good data improves decision-making. It also increases confidence in the findings.

15. What happens if the test statistic falls outside the critical region?

Ans: The Null Hypothesis is not rejected. The evidence is not strong enough to support the Alternative Hypothesis. The observed result may have occurred by chance. The researcher keeps the original assumption.

16. Why can't a researcher make a decision without calculating a test statistic?

Ans: The test statistic summarizes the sample data. It is needed to compare the observed results with the expected results. Without it, the researcher cannot make an objective decision. It is an important part of hypothesis testing.

Data Visualization

17. What is data visualization?

Ans: Data visualization is the graphical presentation of data using charts, graphs, maps, or diagrams. It makes data easier to understand, analyze, and interpret. It converts raw data into meaningful visual forms. This helps people understand information quickly.

18. Explain the purpose of a bar chart and a line graph.

Ans: A bar chart is used to compare different categories or groups. It helps show differences between values clearly. A line graph shows trends and changes over time. It helps researchers observe how data changes from one point to another.

19. How do data visuals help connect results with hypotheses?

Ans: Data visuals help researchers check whether the data support or reject the hypothesis. A comparison graph can show differences between groups. A line graph can show expected trends, while a scatter plot can show relationships between variables. Visual evidence makes conclusions easier to understand.

Ethical Issues in Data Science

20. List bias in data collection?

Ans: Survey bias, sampling bias, gender bias, geographical bias and confirmation bias

21. Describe any two types of bias in data collection.

Ans: **Survey Bias:** A survey about student performance is conducted only in top-performing schools.
Sampling Bias: A health study includes only young adults, so the results do not represent older people.

22. Why is responsible communication of findings important?

Ans: Responsible communication means presenting results honestly and clearly. Results should not be exaggerated or manipulated. The limitations of the data and analysis should be explained clearly. This prevents misunderstanding and improves trust in data analysis.

23. Why should conclusions always be linked to the original hypothesis?

Ans: The conclusion should answer the original hypothesis. It shows whether the evidence supports or rejects the hypothesis. This completes the research process. It also makes the study logical.

CONCEPTUAL SHORT QUESTIONS

1. If you reject the Null Hypothesis at a 5% significance level ($\alpha = 0.05$), what is the percentage chance that you might be making a mistake and rejecting a true hypothesis?

Ans: At a significance level of 0.05, there is a 5% chance of making a mistake by rejecting a true Null Hypothesis. This type of mistake is called a Type I error.

2. Draft a clear Null Hypothesis (H_0) and Alternative Hypothesis (H_1) for an experiment testing whether rolling a standard 6-sided die is fair or biased.

Ans: H_0 : The 6-sided die is fair, and each number has an equal chance of appearing.
 H_1 : The 6-sided die is biased, and the numbers do not have equal chances of appearing.

3. Why is it considered an ethical issue to completely delete outliers from your dataset just to make your hypothesis look correct?

Ans: Deleting outliers only to support a hypothesis is unethical because it changes the original data. This can produce false or misleading conclusions. Researchers should report data honestly and remove outliers only with a valid scientific reason. Honest reporting improves the credibility of research.

4. A researcher says, "Students who sleep 8 hours score higher in exams." Write a suitable Null Hypothesis (H_0) and Alternative Hypothesis (H_1).

Ans: H_0 : Sleeping 8 hours does not improve students' exam scores.
 H_1 : Sleeping 8 hours improves students' exam scores.
These hypotheses can be tested by collecting students' sleep and exam data. The results will help decide whether H_0 should be rejected.

5. A student writes, "Social media is bad." Why is this not a good hypothesis?

Ans: This statement is too vague and cannot be measured easily. A good hypothesis should be clear, specific, and testable. It should include measurable variables. A clear hypothesis gives more reliable research results.

6. A researcher wants to study whether online learning improves exam results. Write a suitable research question.

Ans: A suitable research question is: "Does online learning improve students' exam results?" This question clearly identifies the problem to be studied. It helps the researcher collect relevant data and develop a testable hypothesis.

7. A researcher chooses $\alpha = 0.01$ instead of $\alpha = 0.05$. Is the test stricter or less strict? Why?

Ans: The test becomes stricter because only a 1% chance of error is accepted. Stronger evidence is needed to reject the Null Hypothesis. This reduces the chance of making a wrong decision. It increases the reliability of the results.

8. A hypothesis test gives a p-value of 0.02 and $\alpha = 0.05$. What decision should be made?

Ans: Since the p-value (0.02) is less than the significance level (0.05), the Null Hypothesis is rejected. There is enough evidence to support the Alternative Hypothesis. The observed difference is considered statistically significant.

9. Which graph would you use to show students' marks over six months? Why?
Ans: A line graph should be used because it shows changes over time. It helps identify trends and patterns easily. Students' marks can be compared from one month to another. This makes the results easier to understand.
10. Which graph is most suitable for comparing the marks of four classes? Why?
Ans: A bar chart is the most suitable graph because it compares different groups. It clearly shows the differences between the four classes. The graph is easy to read and understand. It presents the data clearly.
11. Which graph would you use to show the percentage of students choosing different subjects? Why?
Ans: A pie chart should be used because it shows proportions and percentages. Each section represents part of the whole. It makes comparisons simple and clear. It is the best choice for percentage data.
12. Which graph is best for showing the relationship between study time and exam marks?
Ans: A scatter plot is the best graph because it shows the relationship between two variables. It helps identify patterns and trends in the data. Researchers can easily see whether the variables are related. It supports data analysis.
13. A researcher changes some results to support the hypothesis. Why is this unethical?
Ans: Changing research results is unethical because it gives false conclusions. It reduces the accuracy and reliability of the research. Researchers should report all results honestly. Honest reporting builds trust in research.
14. A survey is conducted only in private schools, but the results are used for all students. What type of problem is this?
Ans: This is an example of sampling bias. The sample does not represent all students fairly. Therefore, the conclusions may not be accurate for the whole population. Data should be collected from different groups.
15. A hiring system selects mostly male candidates because of past data. What ethical issue does this show?
Ans: This shows gender bias in the data. The system does not treat all candidates fairly. It may lead to unfair hiring decisions. Researchers should remove bias to ensure fairness.
16. A graph clearly shows that students using a new teaching method scored higher than the control group. How does this graph help the researcher?
Ans: The graph presents the results in a clear visual form. It makes it easy to compare the two groups. It provides evidence to support the analysis. It also helps explain the research findings more effectively.
17. A school wants to know whether a new timetable improves attendance. What should the school do before making a conclusion?
Ans: The school should first collect attendance data. The data should be analyzed using hypothesis testing. The results should be compared with the selected criteria. The final conclusion should be based on evidence.
18. A researcher studies internet usage but collects data only from university students. What problem may occur?
Ans: This may cause sampling bias because the sample does not represent all age groups. The results may not be accurate for the whole population. A fair sample should include people from different groups. This improves the reliability of the study.
19. A company uses customers' personal information without their permission for research. Why is this unethical?
Ans: Using personal information without permission violates people's privacy. Data should be collected only after obtaining proper consent. Researchers should respect participants' rights. Ethical data collection builds trust and protects privacy.
20. If you reject the Null Hypothesis at a 5% significance level ($\alpha = 0.05$), what is the percentage chance that you might be making a mistake?
Ans: At a significance level of 0.05, there is a 5% chance of rejecting a true Null Hypothesis. This mistake is called a Type I error. It means there is a small risk of making an incorrect decision during hypothesis testing.
21. A researcher wants to compare the average heights of boys and girls in a school. Which statistical test should be used?
Ans: A t-test should be used because it compares the means of two groups. It helps determine whether the difference between the average heights is statistically significant. The test uses sample data to support the final decision.

TEST YOUR SKILLS

Total Marks: 40

Time: 1:30 Hour

Q.1 MULTIPLE CHOICE QUESTIONS

8 x 1 = 8

1. The Alternative Hypothesis (H_1 or H_a) states that there is:
(a) No effect (b) No relationship (c) An effect or a difference (d) No data
2. The symbol used for the null hypothesis is:
(a) H_1 (b) H_2 (c) H_0 (d) H_x
3. A p-value is used for:
(a) Measuring accuracy (b) Testing a hypothesis
(c) Measuring variance (d) Data storage
4. The Chi-Square Test is used to find whether two _____ are related.
A) Variables B) Categories C) Samples D) Means
5. If (p-value < α), the decision is to:
(a) Accept H_0 (b) Reject H_0
(c) Ignore data (d) Recalculate the mean
6. A line graph shows:
A) Proportions and percentages B) Relationship between variables
C) Trends and changes over time D) Different categories
7. Bias in data means:
(a) Correct data (b) Balanced data
(c) Unfair representation (d) Random data
8. Why should misleading graphs, charts, or selective reporting be avoided?
A) They increase processing speed. C) They improve data collection.
B) They reduce storage space. D) They can cause misunderstanding and misuse of findings.

Q.2 SHORT QUESTIONS:

Note: Attempt any 8 short questions

8 x 2 = 16

- i. Describe role of Hypothesis Testing in Data Analysis
- ii. What is significance level?
- iii. What is the main difference between the Null Hypothesis (H_0) and the Alternative Hypothesis (H_1)?
- iv. What is a critical region?
- v. What is the F-Test and where is it used?
- vi. What happens if the test statistic falls outside the critical region?
- vii. List bias in data collection?
- viii. Why is ethical use of data important?
- ix. Why is data visualization important?
- x. If you reject the Null Hypothesis at a 5% significance level ($\alpha = 0.05$), what is the percentage chance that you might be making a mistake?

LONG QUESTIONS:

Note: Attempt any 2 long questions

2 x 8 = 16

3. Explain key concepts in hypothesis testing, including test statistics, critical region, and p-value.
4. Discuss the steps involved in performing hypothesis testing with a suitable example.
5. Discuss ethical issues in data science, including bias, ethical use of data, and responsible communication.